CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Performance Evaluation of Serverless Applications and Infrastructures

PhD Defense

Opponent: Petr Tůma
Committee:
Evgenia Smirni
Vittorio Cortellessa
Alessandro Papadopoulos

Joel Scheuner
scheuner@chalmers.se
joe4dev
joelscheuner.com

Supported by WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

2022-09-08

Goal of the PhD

To enable reproducible performance evaluation of serverless applications and their underlying cloud infrastructure.

Progression of Deployment Options

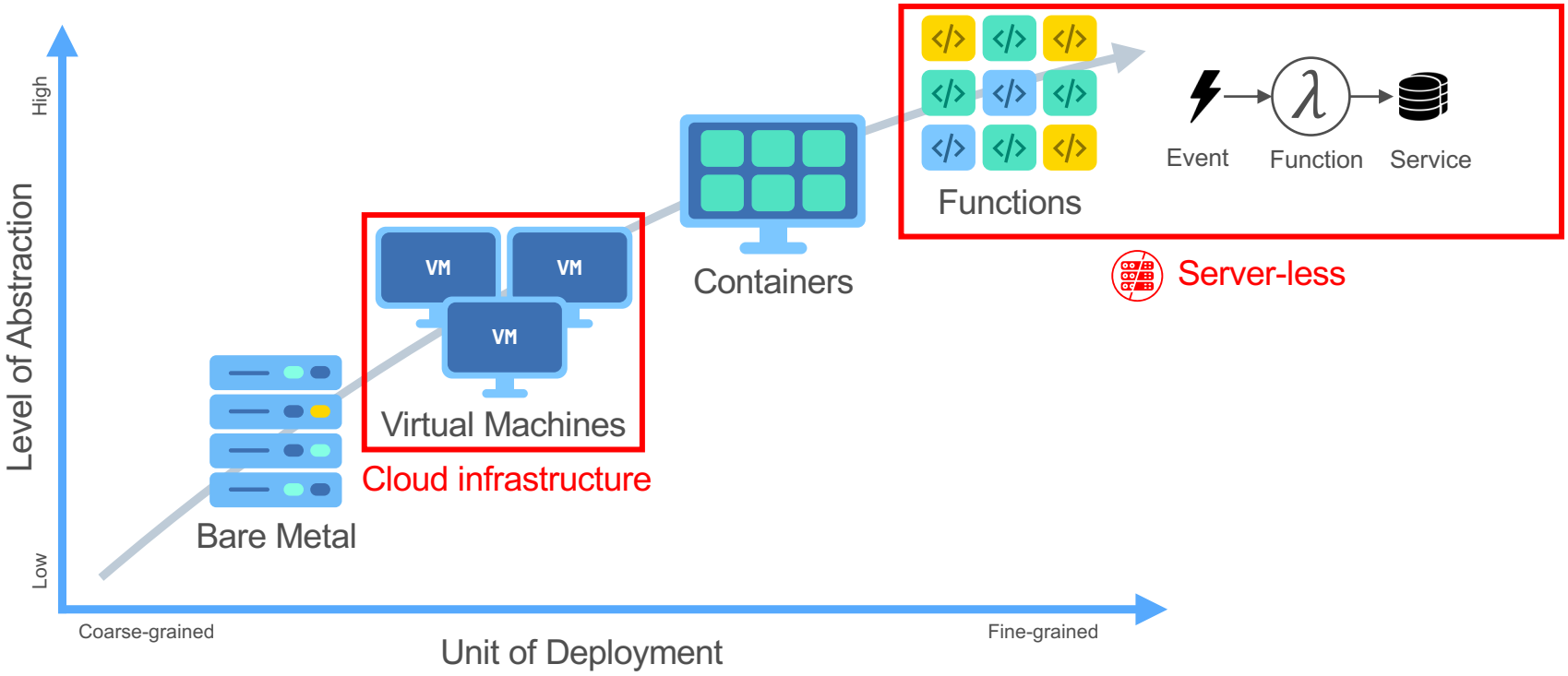
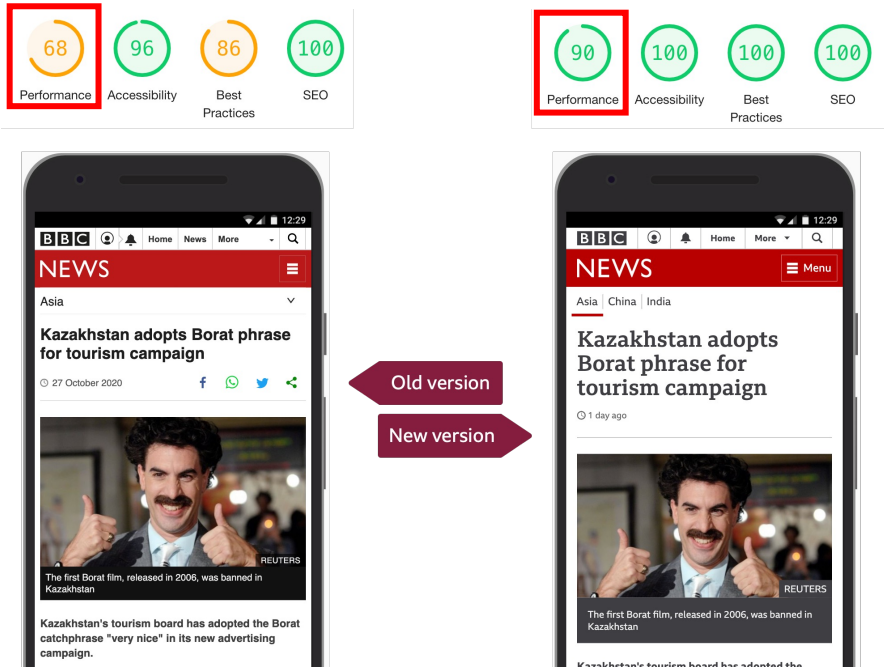


Figure adapted from S. Fink. Serverless – Where Have We Come? Where Are We Going? Keynote at WoSC@CLOUD'18.

Serverless in the Wild



Performance in Serverless

High latency is a problem
[Leitner et al., JSS'19.]

Top 2 non-functional challenge
[Wen et al., ESEC/FSE'21.]

Most popular topic within serverless
[Yussupov et al., UCC'19.]

→ No consolidated view

Moving BBC Online to the cloud: https://medium.com/bbc-design-engineering/moving-bbc-online-to-the-cloud-afdfb7c072ff

Research Questions

RQ1 What is the current state of serverless applications and their performance?

RQ2 What are the performance challenges of serverless applications?

RQ3 How can limitations of benchmarking cloud infrastructure be addressed?

Contributions Overview

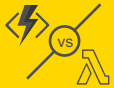
RQ1: Current state of serverless

Paper α (JSS'19)
Performance evaluation
literature review

Paper β (TSE'21)
Application characteristics
sample study

RQ2: Serverless application performance

Paper γ (journal submission)
ServiTrace application
benchmarking suite



Paper δ (conference submission)
CrossFit: Cross-provider
application benchmarking

Paper ε (IC2E'22)
TriggerBench: Function
trigger benchmark

RQ3: Limitations of cloud benchmarking

Paper ζ (QUDOS'18)
Integrated micro and
application benchmark suite

Paper η (CLOUD'18)
Application performance
estimation

Paper θ (EMSE'19)
Reliable cloud
benchmarking

RQ1: Current State of Serverless

Literature review [α]

112 serverless performance studies

2016 – 2019

2x studies since

Studies and their design
→ Secondary research

Sample study [β]

89 serverless applications

22 characteristics

Triangulate with 10 related sources

Documentation and code
→ Primary research

α Function-as-a-Service Performance Evaluation: A Multivocal Literature Review. JSS'20.

β The State of Serverless Applications: Collection, Characterization, and Community Consensus. TSE'21.

RQ1: Current State of Serverless

Benchmark Type [α]

75% Micro-benchmarks

44% Application-benchmarks

External Service

12% Database *used* in performance studies [α]

48% Database *found* in real applications [β]

Reproducibility [α]

40% Sufficient experimental setup description

22% Open access artifact

Workload Burstiness [β]

84% Bursty vs. non-bursty

α Function-as-a-Service Performance Evaluation: A Multivocal Literature Review. JSS'20.

β The State of Serverless Applications: Collection, Characterization, and Community Consensus. TSE'21.

Contributions Overview

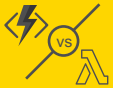
RQ1: Current state of serverless

Paper α (JSS'19)
Performance evaluation
literature review

Paper β (TSE'21)
Application characteristics
sample study

RQ2: Serverless application performance

Paper γ (journal submission)
ServiTrace application
benchmarking suite



Paper δ (conference submission)
CrossFit: Cross-provider
application benchmarking

Paper ε (IC2E'22)
TriggerBench: Function
trigger benchmark

RQ3: Limitations of cloud benchmarking

Paper ζ (QUDOS'18)
Integrated micro and
application benchmark suite

Paper η (CLOUD'18)
Application performance
estimation

Paper θ (EMSE'19)
Reliable cloud
benchmarking

RQ2: Serverless Application Performance

RQ3: Limitations of Cloud Benchmarking

Engineering research

Field experiment

① Benchmark design

② Benchmark execution

③ Data pre-processing

④ Data analysis

Benchmarks < 13 apps
50 micro

Languages: python, JavaScript, Java, Ruby, GO, php, BASH, …

Cloud providers: amazon web services, Google Cloud, Microsoft Azure, IBM Cloud

Many dimensions (e.g., workloads, repetitions, instrumentation)

5 public datasets

Data volume: >200 GB

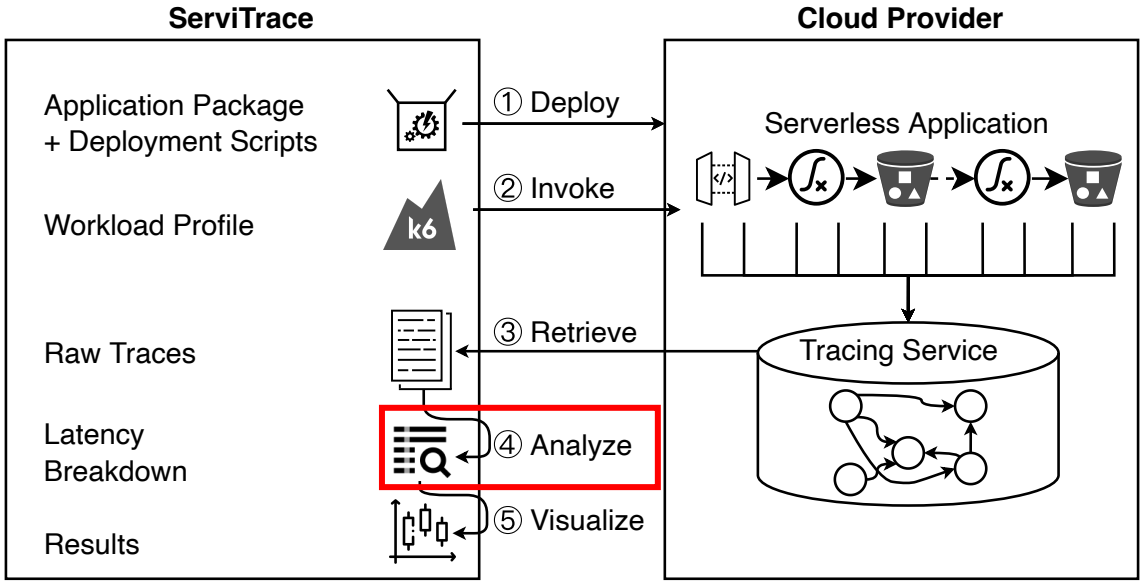
Data types: execution times, timestamps, traces

Quantitative methods (e.g., violin or ECDF plots)

RQ2: Serverless Application Performance

ServiTrace [γ]

- Automates full benchmarking lifycycle
- 10 diverse applications (based on RQ1)
- Well-tested (unit, integration, 7.5 million traces)

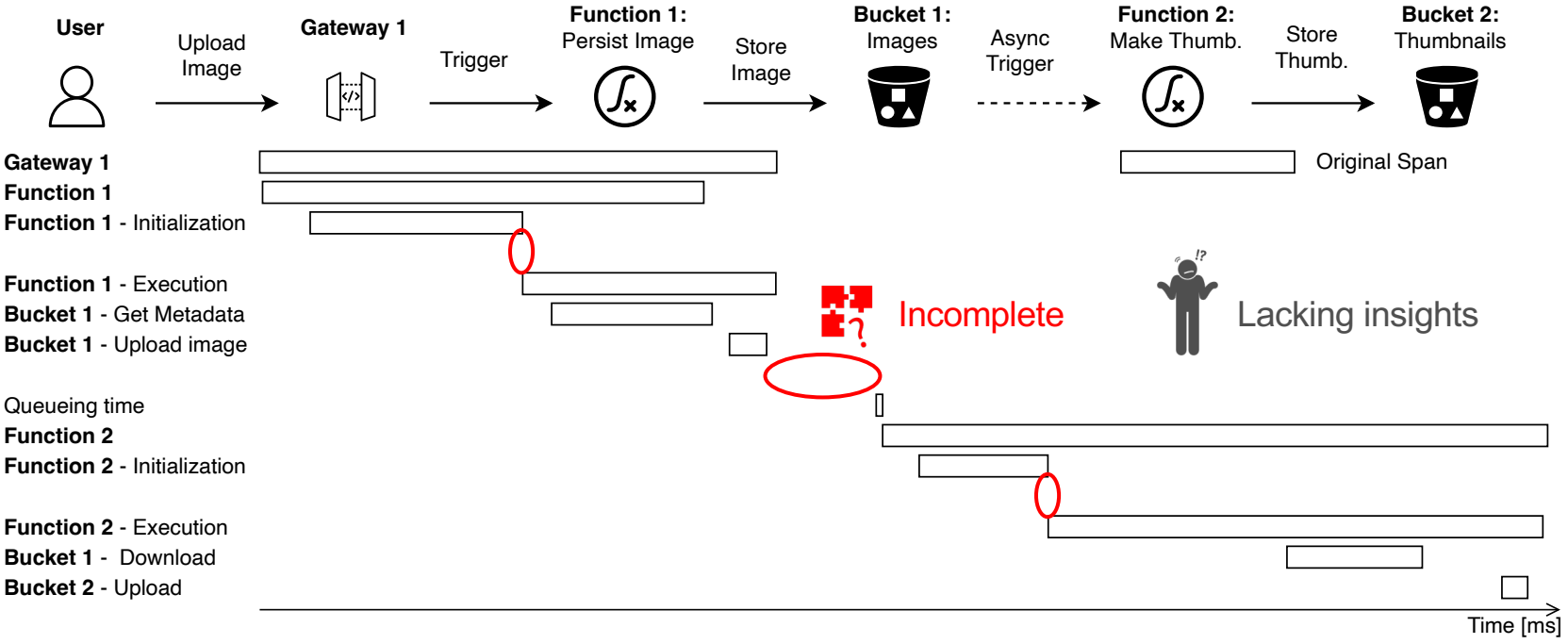


γ Let's Trace It: Fine-Grained Serverless Benchmarking for Synchronous and Asynchronous Applications. Under submission to a journal.

RQ2: Serverless Application Performance

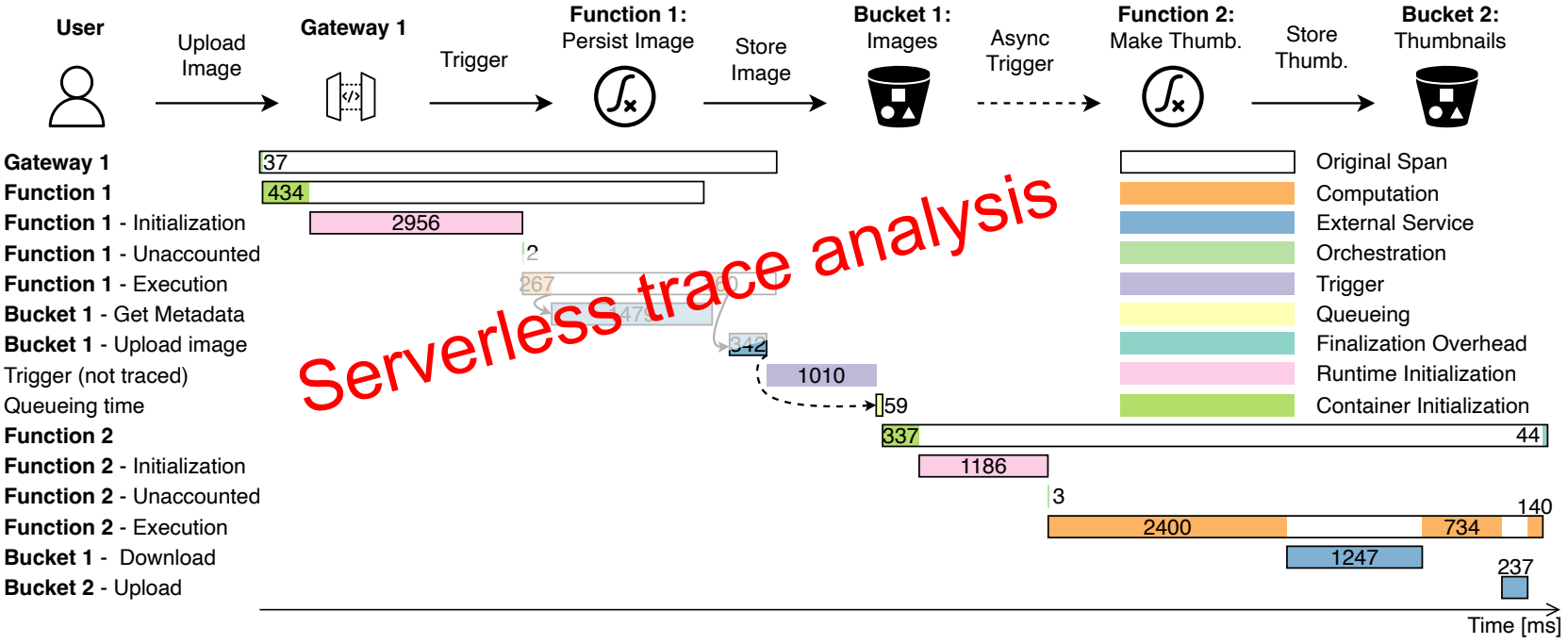
γ Let's Trace It: Fine-Grained Serverless Benchmarking for Synchronous and Asynchronous Applications. Under submission to a journal.

RQ2: Serverless Application Performance



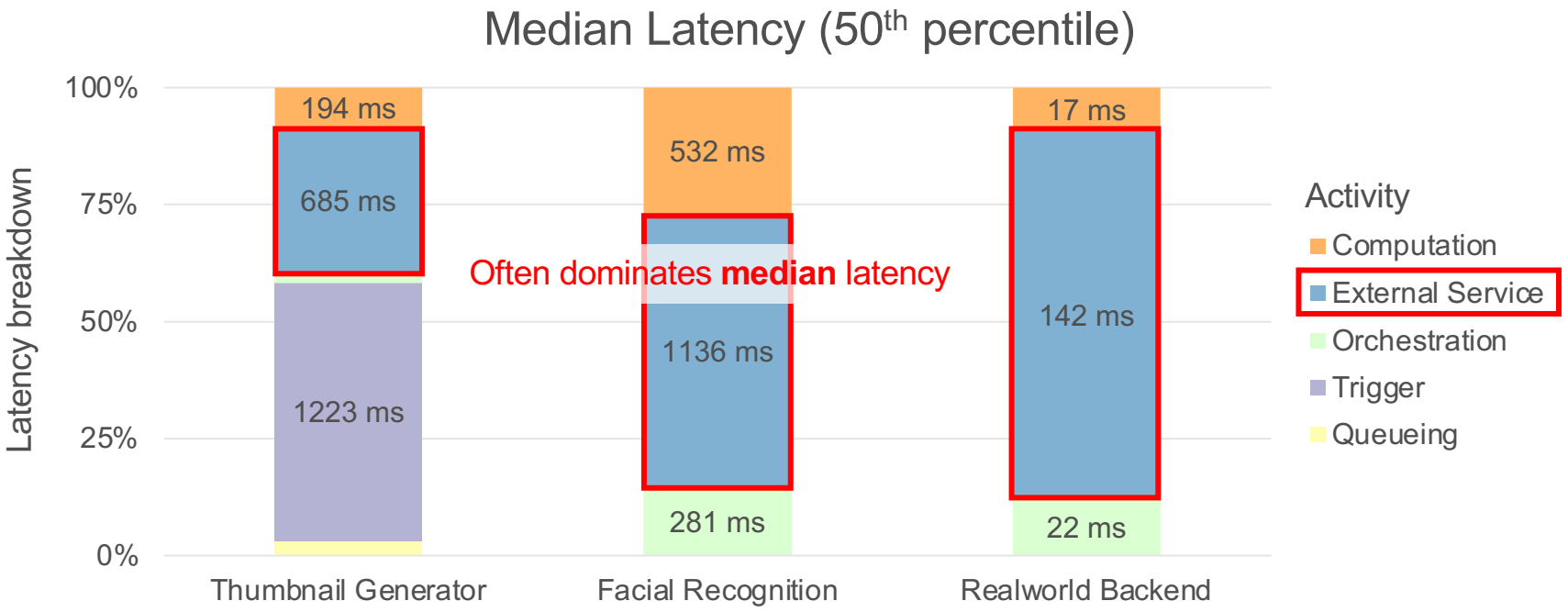
γ Let's Trace It: Fine-Grained Serverless Benchmarking for Synchronous and Asynchronous Applications. Under submission to a journal.

RQ2: Serverless Application Performance



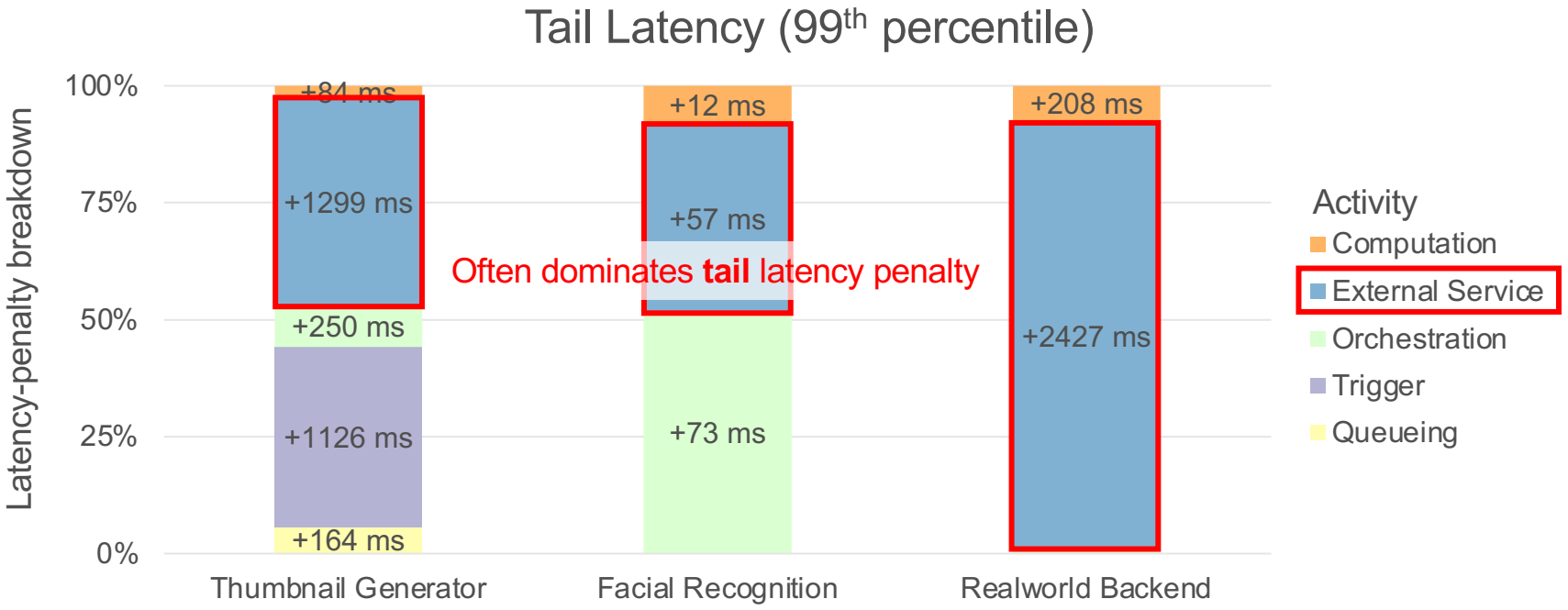
γ Let's Trace It: Fine-Grained Serverless Benchmarking for Synchronous and Asynchronous Applications. Under submission to a journal.

RQ2: Serverless Application Performance



γ Let's Trace It: Fine-Grained Serverless Benchmarking for Synchronous and Asynchronous Applications. Under submission to a journal.

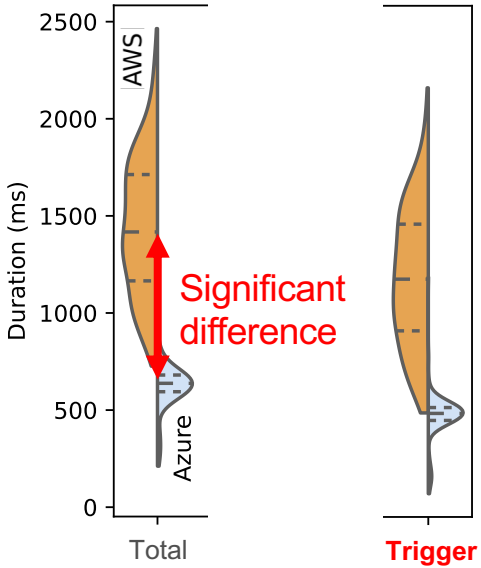
RQ2: Serverless Application Performance



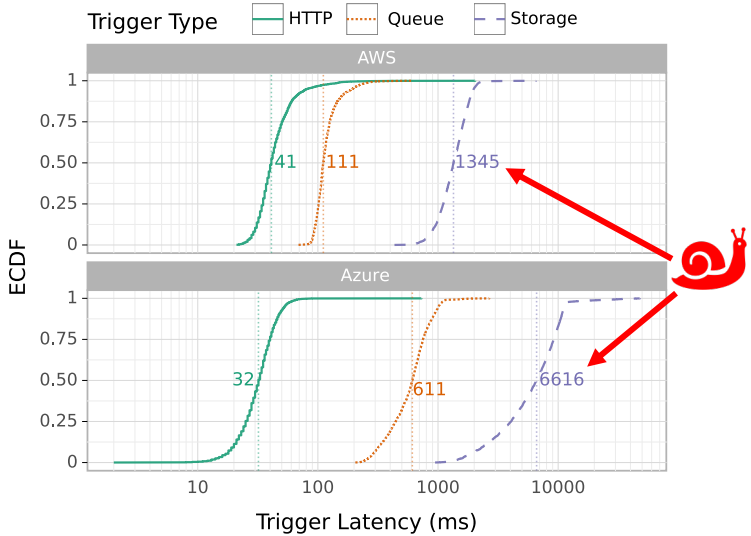
γ Let's Trace It: Fine-Grained Serverless Benchmarking for Synchronous and Asynchronous Applications. Under submission to a journal.

RQ2: Serverless Application Performance

CrossFit [δ]: Cross-provider application insights



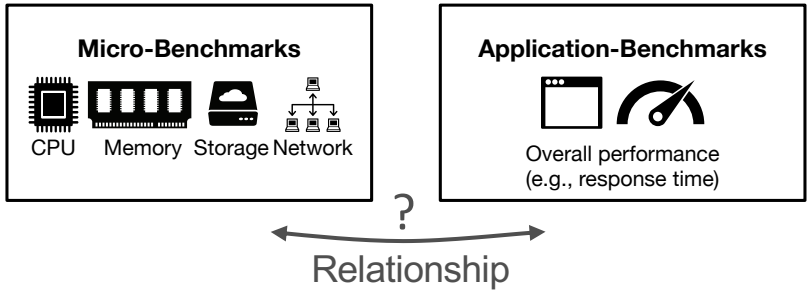
TriggerBench [ε]: Latency of trigger types

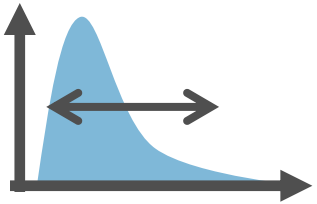


δ CrossFit: Fine-grained Benchmarking of Serverless Application Performance across Cloud Providers. Under submission to a conference.

ε TriggerBench: A Performance Benchmark for Serverless Function Triggers. IC2E'22. To appear as short paper.

RQ3: Limitations of Cloud Benchmarking





Depends on benchmark and environment

Often possible with repetitions within and across virtual machines

●○ **Selected** micro-benchmarks are better application performance predictors than static baselines.

η Estimating Cloud Application Performance Based on Micro-Benchmark Profiling. CLOUD'18.

θ Software Microbenchmarking in the Cloud. How Bad is it Really? EMSE'19.

Results Summary

RQ1: Current state of serverless

Synthetic micro-benchmarks have been studied extensively but we need more realistic application-benchmarks that integrate with external services.

RQ2: Serverless application performance

External service calls and trigger-based function coordination are often slow and suffer from long tail latency.

RQ3: Limitations of cloud benchmarking

Only selected micro-benchmarks are suitable for application performance estimation and repetitions at different levels should be used for reliable performance testing.

Take-Away

Enables reproducible performance evaluation of serverless applications and their underlying cloud infrastructure.

Conclusions

Design better cloud performance studies


Improve the performance of serverless applications


All artefacts are available



Joel Scheuner

scheuner@chalmers.se

joe4dev

joelscheuner.com

Credits

Icons created by:

- Freepik, Becris – Flaticon
- Puckung graphic design factory – Iconfinder
- Documentation by Eucalyp, Graph by Wuppdidu, serverless by Juicy Fish from Noun Project