FaaSPR: Latency-oriented Placement and Routing Optimization for Serverless Workflow Processing

Yunshan Jia, Chao Jin, Qing Li, Xuanzhe Liu, *Senior Member, IEEE,* and Xin Jin, *Senior Member, IEEE*

***Abstract*—**

***Index Terms*—Serverless, scheduling, placement and routing.**

I. Introduction

Serverless and pod resource management in Serverless
overhead: resource utilization and cold start
current solution and problem: pallete and faascache (other sota)
challenge:
technique: mention inplace scaling in k8s
performance
contribution

II. Background and Motivation

A. resource management in Serverless

B. Job scheduling in Serverless

add demo experiment here, sota -¿ overhead detail

III. FaaSPR Overview

Instance placement based on heuristic algorithm

job scheduling based on dynamic instance resizing

IV. FaaSPR Deisgn

A. Instance placement based on heuristic algorithm

B. Cold-start-free hybrid placement updating

C. job scheduling(todo: name)

V. Implementation

VI. Evaluation

A. Evaluation Setup

Benchmarks.

Workload generation.

Cluster configuration.

Baseline.

Yunshan Jia, Chao Jin, Xuanzhe Liu, and Xin Jin are with the School of Computer Science, Peking University, Beijing, China. E-mail: {jiayunshan;chaojin;liuxuanzhe;xinjinpku}@pku.edu.cn.

Qing Li is with the School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China. E-mail: qingli@bupt.edu.cn.

This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB4500700, in part by the National Natural Science Foundation of China under Grant 62172008 and 62325201, and in part by the National Natural Science Fund for the Excellent Young Scientists Fund Program (Overseas). (Corresponding author: Xin Jin.)

B. Benefits of FaaSPR

Comparison under different benchmark number.

Comparison under different resource distributions.

Performance analysis.

C. Scalability

Larger-scale cluster experiment.

System overhead.

D. Effectiveness of FaaSPR

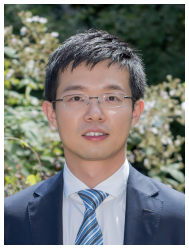
Technique effectiveness.

Parameter effectiveness.

VII. Related Work

VIII. Conclusion

Xuanzhe Liu (Senior Member, IEEE) is a Full Professor in the School of Computer Science at Peking University, Beijing, China. His research interests mainly fall in service-based software engineering and systems. Most of his recent efforts have been published at prestigious conferences including WWW, ICSE, FSE, SOSP, SIGCOMM, NSDI, MobiCom, MobiSys, and in journals including ACM TOSEM/TOIS and IEEE TSE/TMC/TSC. He is a distinguished member of the ACM and the CCF. Web page: http://www.liuxuanzhe.com/.



Xin Jin (Senior Member, IEEE) received the PhD degree from Princeton University, in 2016. He is currently an associate professor (with Tenure) with the School of Computer Science, Peking University. His research interests include computer systems, networking, and cloud computing.