Upscaling Messaging and Stateful Computation

Josef Spillner
josef.spillner@zhaw.ch
Zurich University of Applied Sciences
Winterthur, Switzerland

ABSTRACT

Large-scale, production-grade cloud applications are no longer black boxes for academic researchers. They are observable subjects under test in an increasing number of projects, with the aim to quantify and improve their runtime characteristics, including performance. With more meaningful measurements available, data-driven approaches have matured and advanced the knowledge in particular around conventional stateless workloads such as functions and containers. A few less explored areas still exist. They are fueled by the increasing number of atypical function deployments for instance in message brokers, in intelligent switches and in blockchains. This talk summarises reference architectures for large-scale applications, sometimes resulting in nation-scale deployments, discusses performance numbers in this context, and elaborates on whether more focus on performance is needed.

ACM Reference Format:
Josef Spillner. 2024. Upscaling Messaging and Stateful Computation. In *Companion of the 15th ACM/SPEC International Conference on Performance Engineering (ICPE '24 Companion), May 7–11, 2024, London, United Kingdom.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3629527.3652885

1 INTRODUCTION

Industrially relevant systems engineering has led to high computational complexities, both in pure software systems (e.g. business applications and middleware) and in an increasing variety of systems that depend on combinations of specific hardware and sensors. The complexity is partly driven by the sheer scale, with even small and medium engineering companies often facing the demand to produce, deploy and operate large-scale and even nation-scale applications. With concurrent invocations exceeding the limits of commercial off-the-shelf offerings (e.g. FaaS), engineers are then faced with the tough choice to revert back to self-managed VMs or containers, or to become creative especially with atypical function deployments. This keynote talk encourages to try the latter and to dare looking forward to software and system architectures where the simplicity is maintained and yet the posssible scale is increased significantly. Can entry-level engineers or students be tasked with such a task now? Probably not, but within a few years that may change.

ICPE '24 Companion, May 7–11, 2024, London, United Kingdom

ACM ISBN 979-8-4007-0445-1/24/05
https://doi.org/10.1145/3629527.3652885

2 EXPLORATION

To explore the topic area systematically, a few ideas around the notion of functions must be combined and critically examined. Raw concurrency? Can be achieved with big data processing frameworks, all of which fit into modern cloud hosting by now, but require understanding user-defined functions especially for event-driven (non-batch) stream processing [3], as well as prediction of computation time bases on input sources. Ease of deployment and portability? Second-generation serverless frameworks, running pre-containerised functions, help but come with fixed isolation levels and severe limitations per deployment and per region. In-situ processing of logic within message brokers [4], or even within switches and network interfaces [2]? Brings computation closer to the communication path but often hits limits of devices designed for the latter but not for the former. Blockchains running smart contracts with function semantics? Have recently brought an interesting perspective to stateful functions [1] but, due to the nature of decentralisation, may not be suitable for raw throughput and low latency. Liquid functions across edge-cloud continuums? Prototypes exist...

Given such a large number of choices, systematic evaluations and practical experience are both required to come up with useful advice to engineers in the form of checklists, patterns, metaprogramming and other simplifications at the textbook level. The challenges are not only in the infrastructure with hardware constraints and arbitrarily set limits, but deeply routed in the messaging protocols (e.g. privacy considerations) and compute units. With reference to currently ongoing industry-funded research and innovation projects, the talk therefore goes beyond the encouragement and demonstrates the practical need to find convincing system designs, under the hypothesis that atypical function deployments (beyond mainstream FaaS) plays a role in this search process. It gives examples from several cyber-physical application domains such as monitoring people flows and herd movement, as well as digital pandemic and academic credentials management.

REFERENCES

[1] Maksym Arutyunyan, Andriy Berestovskyy, Adam Bratschi-Kaye, Ulan Degenbaev, Manu Drijvers, Islam El-Ashi, Stefan Kaestle, Roman Kashitsyn, Maciej Kot, Yvonne-Anne Pignolet, Rostislav Rumenov, Dimitris Sarlis, Alin Sinpalean, Alexandru Uta, Bogdan Warinschi, and Alexandra Zapuc. 2023. Decentralized and Stateful Serverless Computing on the Internet Computer Blockchain. In *2023 USENIX Annual Technical Conference, USENIX ATC 2023, Boston, MA, USA, July 10-12, 2023*, Julia Lawall and Dan Williams (Eds.). USENIX Association, 329–343.

[2] Nilanjan Daw, Umesh Bellur, and Purushottam Kulkarni. 2021. Speedo: Fast dispatch and orchestration of serverless workflows. In *SoCC '21: ACM Symposium on Cloud Computing, Seattle, WA, USA, November 1 - 4, 2021*, Carlo Curino, Georgia Koutrika, and Ravi Netravali (Eds.). ACM, 585–599. https://doi.org/10.1145/3472883.3486982

[3] Yannis Foufoulas and Alkis Simitsis. 2023. Efficient Execution of User-Defined Functions in SQL Queries. *Proc. VLDB Endow.* 16, 12 (aug 2023), 3874–3877. https://doi.org/10.14778/3611540.3611574

[4] K. Sundar Rajan, A. Vishal, and Chitra Babu. 2021. A Scalable Data Pipeline for Realtime Geofencing Using Apache Pulsar. In *Comp. Intellig. in Data Science - 4th IFIP TC 12 Intl. Conf., ICCIDS 2021, Chennai, India, March 18-20, 2021 (IFIP Advances in Information and Communication Technology, Vol. 611)*, Vallidevi Krishnamurthy, Suresh Jaganathan, Kanchana Rajaram, and Saraswathi Shunmuganathan (Eds.). Springer, 3–14. https://doi.org/10.1007/978-3-030-92600-7_1