Mitigating Cold-Start Latency in Serverless Edge Deployments via Adaptive Container Checkpointing and Predictive Pre-Warming

Riley Phillips
Associate Professor
University of Waterloo
200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada

Alex Lewis
PhD
Korea Advanced Institute of Science and Technology (KAIST)
291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

Chris Scott
Professor
Technische Universität München
Arcisstraße 21, 80333 München, Bavaria, Germany

Abstract. Serverless computing at the network edge introduces critical cold-start latency penalties that undermine real-time service-level agreement (SLA) compliance in latency-sensitive workloads. This paper addresses the challenge by proposing AdaptWarm, a hybrid framework combining adaptive container checkpointing with a Long Short-Term Memory (LSTM)-driven predictive pre-warming scheduler. AdaptWarm continuously profiles invocation frequency distributions, constructs Markov-based transition models for function call sequences, and restores snapshotted execution contexts ahead of predicted demand spikes. Evaluated across heterogeneous edge nodes using the OpenFaaS and Knative runtimes with production-grade workload traces, AdaptWarm reduces mean cold-start latency by 73.4% and tail latency (P99) by 68.1% compared to baseline reactive provisioning, while maintaining memory overhead below 11%. These results establish AdaptWarm as a viable, deployment-ready solution for sub-millisecond function initialization in multi-tenant edge infrastructures.

Keywords: serverless computing, cold-start latency, container checkpointing, edge computing, predictive pre-warming, LSTM-based scheduling, function-as-a-service, Markov chain workload modeling, SLA compliance

Introduction

The rapid proliferation of Internet-of-Things (IoT) ecosystems, real-time analytics pipelines, and autonomous systems has fundamentally shifted the computational gravity of modern distributed architectures toward the network edge. Function-as-a-Service (FaaS) paradigms have emerged as the dominant abstraction for deploying event-driven microservices in these environments, offering fine-grained resource elasticity and operator-transparent scaling. However, a structurally persistent impediment—the cold-start

problem—continues to compromise the viability of serverless deployments in latency-critical scenarios. Cold-starts arise when an incoming invocation request finds no pre-initialized execution context, forcing the runtime to sequentially allocate container resources, load the function image, initialize the language runtime, and execute bootstrapping logic before any application-level computation can commence. In production edge environments characterized by heterogeneous hardware, constrained memory envelopes, and highly bursty, non-stationary workload distributions, this initialization overhead can range from several hundred milliseconds to multiple seconds—values that are wholly incompatible with sub-10ms SLA thresholds mandated by emerging 5G and 6G edge service contracts. Existing mitigation strategies fall into three broad categories: keep-alive pooling, which wastes idle memory on warm containers; snapshot-based restoration (e.g., CRIU-based checkpointing), which reduces re-initialization time but lacks demand-aware invocation; and reactive auto-scaling, which inherently cannot anticipate bursty demand. None of these approaches holistically addresses the temporal prediction of invocation patterns combined with state-preserving context restoration at the granularity required for edge deployments. The literature reveals a critical gap between static provisioning heuristics and the dynamic, non-Markovian nature of real-world FaaS invocation sequences, particularly when function dependency graphs introduce correlated call chains that simple exponential moving averages fail to model accurately. This paper presents AdaptWarm, a unified framework that closes this gap through two tightly coupled components: (1) an LSTM-based temporal forecasting engine trained on sliding-window invocation telemetry to predict demand onset with function-level granularity, and (2) an adaptive container checkpointing subsystem that selectively serializes and restores execution snapshots based on predicted warm-up deadlines derived from Markov chain transition probability matrices. By co-designing the prediction and restoration layers, AdaptWarm achieves pre-warming decisions that are both temporally precise and memory-budget-aware, enabling dynamic trade-offs between restoration fidelity and resource utilization across multi-tenant edge nodes. The remainder of this paper is structured as follows. Section 2 reviews related work on cold-start mitigation, container snapshotting, and edge workload prediction. Section 3 formally defines the system model, workload assumptions, and the optimization objective. Section 4 details the architecture and algorithmic design of AdaptWarm. Section 5 presents the experimental methodology, testbed configuration, and evaluation metrics. Section 6 analyzes empirical results against baseline and state-of-the-art comparators. Section 7 discusses deployment considerations, limitations, and directions for future work. Section 8 concludes the paper.

References

1. Kumar, N., & Kataria, V. Enhanced Sentiment Classification using a Multi-layered Stacked Ensemble Architecture.
2. Kumar, Nitin, and Vipin Kataria. "Enhanced Sentiment Classification using a Multi-layered Stacked Ensemble Architecture."
3. Kumar, N., & Kataria, V. (2023). Enhanced Sentiment Classification using a Multi-layered Stacked Ensemble Architecture. International Journal of Intelligent Systems and Applications in Engineering, 11(4s), 304–311.
4. Satyanarayana, D. and Rao, S. V. (2011), “Fault-tolerant spanners for ad hoc networks” International Journal of Network Management, Vol: 22, Issue: 4, Pages: 311-329, July-2012, Wiley Publishers. https://doi.org/10.1002/nem.807
5. D. Satyanarayana, Sathyashree. S “Mobility Management in Voronoi based Cell Structure for Wireless networks”, IEEE proceedings for the International Conferences on Currents trends in Information Technology (CTIT), Pages: 204-208, Dubai, December-2013. DOI: 10.1109/CTIT.2013.6749504; INSPEC Accession Number: 14131032
6. D. Satyanarayana and J. M. H. Elmirghani, "An Energy Efficient Network Architecture for Infrastructured Wireless Networks," 2010 IEEE Global Telecommunications Conference GLOBECOM 2010, Miami, FL, USA, 2010, pp. 1-6, doi: 10.1109/GLOCOM.2010.5683234.
7. D. Satyanarayana and J. M. H. Elmirghani, "A Voronoi Based Energy Efficient Architecture for Wireless Networks," 200S Third International Conference on Next Generation Mobile Applications, Services and Technologies, Cardiff, UK, 2009, pp. 377-382, doi: 10.1109/NGMAST.2009.79.
8. D. Satyanarayana and S. V. Rao, "Local Delaunay Triangulation for Mobile Nodes," 2008 First International Conference on Emerging Trends in Engineering and Technology, Nagpur, India, 2008, pp. 282-287, doi: 10.1109/ICETET.2008.253.
9. D. Satyanarayana, "Greedy Local Delaunay Triangulation Routing for Wireless Ad Hoc Networks," 2007 International Conference on Signal Processing, Communications and Networking, Chennai, India, 2007, pp. 49-53, doi: 10.1109/ICSCN.2007.350694
10. Satyanarayana, D., Sathyashree, A. A. K., & Alkalbani, A. (2015, November). Voronoi LEACH for Energy Efficient Communication in Wireless Sensor Networks. In Proceedings for the International Conference on Electronics Computer networks and Information Technology, Dubai.
11. Rahmanov, N., Guliyev, H., İbtahimov, F., & Mammadov, Z. (2024). Determination of optimal dimensions of hybrid AC/DC distributed generation system with renewable sources for autonomous power supply of remote locations. In E3S Web of Conferences (Vol. 584, p. 01020). EDP Sciences.
12. Hashimov, A. M., Babayeva, A. R., & Guliyev, H. B. (2019). Shunt reactors control algorithm using fuzzy sets theory. International Journal on Technical and Physical Problems of Engineering, 11(1), 10-15.