An Adaptive PPO-Based Approach for Real-Time Autoscaling in Serverless Computing

Jasmine Kaur *
Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala, Punjab, India

Anju Bala
Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala, Punjab, India

Inderveer Chana
Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala, Punjab, India

Divyanshu Garg
Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala, Punjab, India

Abstract

Serverless computing has revolutionized cloud computing by allowing developers to build and deploy applications without managing the underlying infrastructure. However, efficiently allocating resources to handle dynamic workloads remains a significant challenge. This paper presents an approach for auto-scaling in serverless environments using Proximal Policy Optimization (PPO), a reinforcement learning technique that optimizes resource allocation in real-time. Unlike previous methods that relied on Deep Q-Learning (DQL) or Q-Learning (QL), PPO enhances stability and scalability by directly learning optimal policies. A synthetic workload dataset is used to simulate realistic traffic patterns for model training. Experimental results on AWS Lambda demonstrate that PPO reduces average response time by 35% compared to QL and 20% compared to DQL, ensuring faster job execution. Energy consumption is lowered by 25% and 15%, respectively, improving efficiency. Additionally, throughput increases by 18% over QL and 10% over DQL,while success rate improves by 12% and 8%, ensuring more reliable task execution. These findings highlight PPO's superior effectiveness in reinforcement learning based resource management, making it a promising solution for autoscaling in serverless computing.

Keywords

Serverless Computing, Proximal Policy Optimization (PPO), Auto-Scaling, AWS Lambda.