International Journal of Computer Science and Engineering (IJCSE)
ISSN (P): 2278–9960; ISSN (E): 2278–9979
Vol. 10, Issue 2, Jul – Dec 2021; 193–232

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

OPTIMIZING SERVERLESS ARCHITECTURES: STRATEGIES FOR REDUCING COLD STARTS AND IMPROVING RESPONSE TIMES

Akash Balaji Mali[1], Ashvini Byri[2], Sivaprasad Nadukuru[3], Om Goel[4], Niharika Singh[5] & Prof.(Dr.) Arpit Jain[6]

[1]State University of New York at Binghamton, Binghamton NY, US

[2]Scholar, University of Southern California, Parel, Mumbai, India

[3]Andhra University, Muniswara Layout, Attur, Yelahanka, Bangalore, India

[4]ABES Engineering College Ghaziabad, India

[5]ABES Engineering College Ghaziabad, India

[6]KL University, Vijaywada, Andhra Pradesh, India

ABSTRACT

Serverless architectures have become an essential component of modern cloud computing, enabling scalable, event-driven services without the burden of server management. However, one of the significant challenges faced in serverless environments is the occurrence of cold starts—delays caused when inactive serverless functions are initialized—resulting in higher response times. This abstract explores various strategies for optimizing serverless architectures to mitigate cold starts and improve overall performance. The discussion focuses on pre-warming techniques, effective resource allocation, and the integration of predictive scaling models. Additionally, it covers innovative approaches like leveraging container-based function environments and caching mechanisms to enhance responsiveness. The paper aims to provide a comprehensive overview of best practices for developers and enterprises striving to optimize serverless workloads, ensuring seamless user experience and efficient resource consumption. These strategies are increasingly critical as serverless adoption grows across industries, demanding solutions that balance performance, cost, and scalability.

KEYWORDS*: Serverless Architectures, Cold Starts, Response Time Optimization, Pre-Warming Techniques, Resource Allocation, Predictive Scaling, Container-Based Functions, Caching Mechanisms, Cloud Computing, Performance Enhancement, Scalability*

Article History

Received: *24 Jul 2021* | ***Revised:*** *27 Jul 2021* | ***Accepted:*** *29 Jul 2021*