Chapter 13
Revolutionizing Load Balancing in Cloud Computing With Genetic Algorithms

Abhipsha Das
https://orcid.org/0009-0004-4800-7671
School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar, India

Swayam Yadav
https://orcid.org/0009-0004-4339-6209
School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar, India

Neetu Dey
https://orcid.org/0009-0009-5806-0886
School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar, India

Aayushma Gautam
https://orcid.org/0009-0000-1576-0860
School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar, India

Hitesh Mohapatra
https://orcid.org/0000-0001-8100-4860
School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar, India

DOI: 10.4018/979-8-3693-9984-2.ch013

ABSTRACT

Load balancing is vital in cloud computing for efficiently distributing workloads and preventing resource bottlenecks. This review explores using Genetic Algorithms (GAs), known for their optimization strength, to improve load balancing by minimizing task execution times and boosting resource utilization. The GA-based approach adapts to changing tasks and conditions, evolving solutions through iterative natural selection processes. This paper examines the key principles and real-world applications of this method, showing its potential to transform traditional load balancing and enhance cloud system performance and scalability.

1. INTRODUCTION

Cloud computing is a fresh way to handle large-scale distributed computing. It works by using on-demand services from cloud providers to deliver key IT resources like storage, applications, infrastructure, communication tools, and teamwork platforms. A core part of making this work is strong cloud management, which offers secure data backups, 24/7 availability, load balancing, and flexible resource scheduling. As the size of the cloud grows, providers must handle huge numbers of user requests efficiently. One of the big benefits of cloud computing is its flexibility and ability to deliver high-performance, pay-as-you-go, and on-demand services (Das et al., 2024). But when sudden spikes in demand happen, the challenge is to keep performance steady or even improve it. Cloud load balancing plays a key role here. It's the process of distributing workloads smartly across the cloud's computing resources and balancing the network traffic hitting those resources. In distributed environments like cloud or grid computing, many connected nodes work together to complete tasks. To avoid situations where some nodes are overloaded while others sit idle, load balancing ensures the local workloads are spread evenly across all nodes (Dasgupta et al., 2013).

To handle this efficiently, solutions often rely on optimization techniques, where possible load-balancing setups are represented as individuals in a population, and over time, these configurations are refined to find the best or nearly best fit. Adding to this, modern systems also explore combining load balancing with energy-efficient practices and fault tolerance to not just improve performance but also reduce power use and increase system reliability — areas that are becoming more important as cloud systems scale even bigger.