IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, November 2025
Impact Factor: 7.67

A Review of Dynamic Resource Allocation Algorithms for Machine Learning Workloads

Sanjeev Kumar Shukla[1] and Dr. Sanmati Kumar Jain[2]
[1]Research Scholar, Department of Computer Science & Engineering
[2]Research Guide, Department of Computer Science & Engineering
Vikrant University, Gwalior (M.P.)

Abstract: *Dynamic resource allocation is central to efficient training and serving of machine learning workloads across modern cloud, on-premise, and edge infrastructures. This review surveys algorithmic strategies used to allocate CPU/GPU/accelerator, memory, and network resources for ML tasks. We present a taxonomy covering heuristic and rule-based methods, optimization-based approaches, elastic and autoscaling systems, straggler-mitigation and coding techniques, predictive and workload-forecasting algorithms, and machine learning / deep reinforcement learning schedulers. Strengths, limitations, implementation considerations, and open research directions are discussed to guide both researchers and practitioners.*

Keywords: ML Workloads, Scheduling Algorithms, GPU/CPU Provisioning

DOI: 10.48175/568