

A Review of Scheduling Algorithms for the Prevention of GPU Fragmentation

Irakli Koberidze^a

e-mail: irakli.koberidze776@ens.tsu.edu.ge

^aComputer Science Department, Faculty of Exact and Natural Sciences, Ivane Javakhishvili Tbilisi State University.

Abstract

Modern AI infrastructure faces a critical challenge: the severe underutilization of expensive GPU resources. While GPU sharing enables fine-grained allocation, it introduces a fragmentation problem, where free resource slices become unusable due to the hardware's contiguity constraint. This report analyzes the evolution of scheduling heuristics designed to mitigate this issue. We trace the path from pragmatic, cluster-specific policies and refined multi-dimensional bin-packing algorithms to advanced theoretical models. We demonstrate that while these approaches offer incremental improvements, they are fundamentally limited by treating fragmentation as a static, geometric packing problem. The analysis culminates with Fragmentation Gradient Descent (FGD), a state-of-the-art algorithm that represents a paradigm shift. FGD introduces a statistical, workload-aware measure of fragmentation and schedules tasks to minimize future resource unusability. This approach directly addresses the core limitation of prior methods, offering a principled solution to transform GPU clusters into efficiently shared utilities, which is essential for sustainable and scalable AI progress.