Graduate School of Science and Technology Master's Thesis Abstract

Laboratory name (Supervisor)	Software Design and Analysis (Hajimu Iida (Professor))		
Student ID	2311408	Submission date	2025 / 7 / 22
Name	CHOONHAKLAI PAPON		
Thesis title	Efficient GPU Sharing for Machine Learning in Kubernetes: A Comparative Study of Time-Slicing and Metric-Based MPS Scheduling		

Abstract

Efficient GPU utilization remains a critical challenge in Kubernetes-based machine learning (ML) pipelines, especially when deploying diverse workloads across heterogeneous environments. This thesis presents a comparative study of two GPU sharing approaches—time-slicing with KubeRay and metric-driven Multi-Process Service (MPS) scheduling—for optimizing resource usage and performance in ML workloads. The first approach utilizes KubeRay with NVIDIA time-slicing to dynamically assign GPU resources to distributed training jobs. The second approach introduces a custom Kubernetes operator that leverages real-time GPU metrics collected via the Data Center GPU Manager (DCGM) exporter to perform fine-grained, dynamic scheduling of inference workloads through MPS, eliminating the need for static resource definitions. Experimental evaluations were conducted using a variety of training and inference workloads. In addition to comparing the two proposed approaches, this study also benchmarks them against existing GPU sharing techniques—including NVIDIA's native time-slicing, NVIDIA MPS, and the standalone NOS framework. The results show that while time-slicing with KubeRay improves memory efficiency in distributed training scenarios, the proposed metric-driven MPS approach achieves higher throughput, lower latency, and better GPU utilization in inference workloads. This thesis provides practical insights into GPU sharing strategies for both training and inference, highlighting trade-offs and best practices for deploying scalable ML workloads on Kubernetes.