Optimizing Concurrent Inferences with Fine-grained GPU Sharing

Zixi CHEN, Junqiao QIU
City University Of Hong Kong

Introduction

Modern data centers increasingly co-locate multiple ML inference workloads on a single GPU device for well utilizing the computational resources. However, existing temporal and spatial multiplexing techniques often fall short in accurately allocating resources and meeting performance expectations. We identify that this inefficiency stems from neglecting the importance of latency-critical kernels during scheduling.

To address this, we propose ***Fico***, a fine-grained GPU sharing framework that adjusts resource allocation based on kernel-level priorities. ***Fico*** can dynamically update GPU quotas based on critical kernel detection and schedule execution accordingly to maximize efficiency and fairness. It integrates lightweight profiling and runtime control to enforce tenant quotas while reducing end-to-end latency.

Evaluation across diverse DNN workloads show that ***Fico*** lowers inference latency by 37.3% on average compared to prior methods, while faithfully preserving quota guarantees.

Objectives

- Enable Low-Latency Inference in Multi-Tenant GPU Systems
- Enforce flexible and lightweight GPU Quota Guarantees
- Prioritize Dominant ML Application Kernels to Improve Scheduling Efficiency
- Demonstrate Practical Effectiveness Through Real-World Evaluation

Framework

Our proposed GPU sharing system Fico consists of **two major components**:

- an interference-aware offline profiling which collects the important and necessary information about kernels;
- critical kernel prioritized runtime scheduling which determine how to fine-grained share the GPU resources when each launched kernel is assigned.

It is developed to be transparent to end users and requires no API changes. The current prototype is implemented as a dynamically linked library that controls GPU operations submitted by an application framework

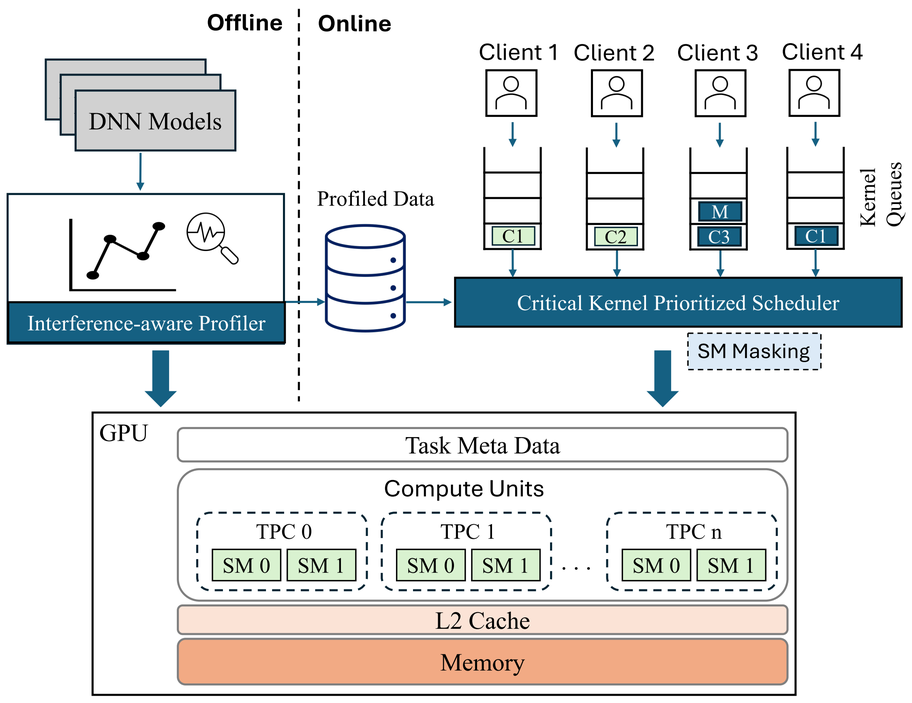

*Figure 1: Overview of **Fico***

ML Apps	Dominant Kernels	Neutral Kernels	In-sensitive Kernels
Resnet152	46.20%	11.84%	41.97%
Resnet101	42.73%	14.79%	42.48%
Densenet201	29.08%	29.05%	41.87%
VGG	76.01%	0	23.99%
MobilenetV2	22.01%	39.44%	38.55%

Table 1: Offline profiling about the kernels in the ML applications

Evaluation

Experimental Setup

- Benchmarks: Several representative DNN models
- Platform: NVIDIA A6000 and PyTorch, and 5 different workload distributions
- Baselines: Include comparison with MPS and state-of-the-art systems (REEF[1], KRISP[2], ORION[3]).

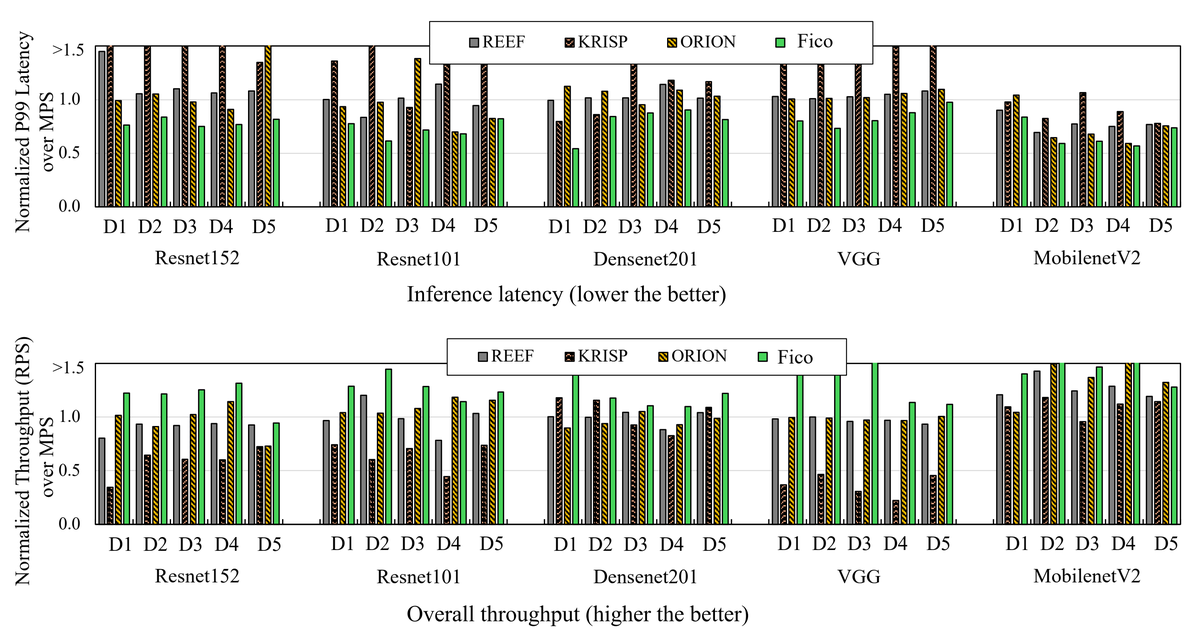

Figure 2: Performance comparison when applying two homo-models

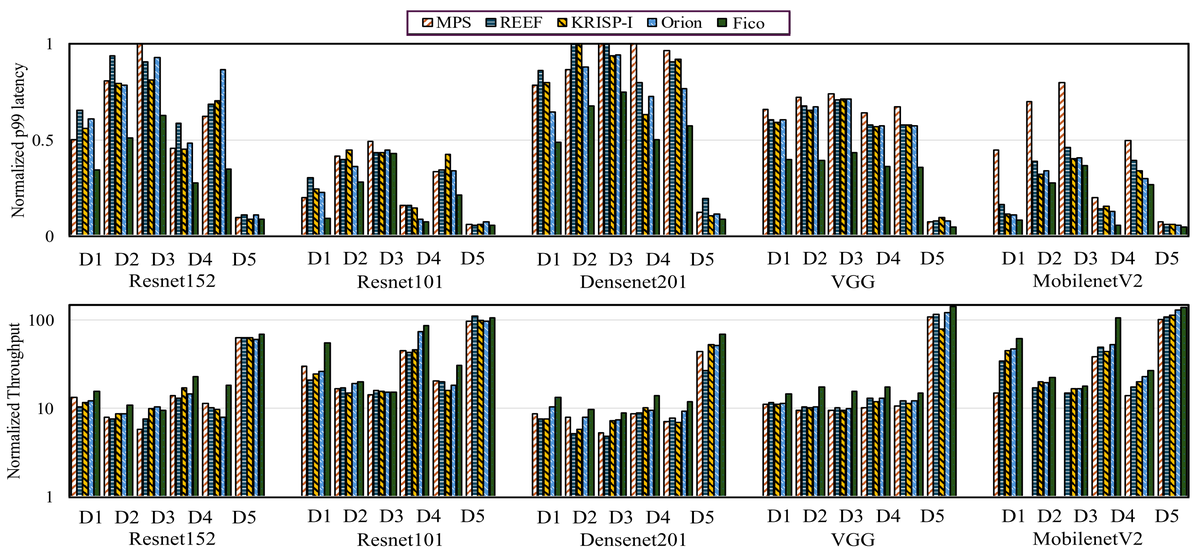

Figure 3: Performance comparison when applying four homo-models

Conclusion

This poster presents *Fico*, a fine-grained GPU sharing system designed to improve inference latency and resource fairness in multi-tenant environments.

By identifying and prioritizing dominant ML kernels during scheduling, Fico mitigates the inefficiencies of existing quota-based approaches that treat all workloads uniformly. It integrates lightweight runtime profiling and quota-aware scheduling to provide strict resource guarantees while enhancing responsiveness.

Extensive evaluations across diverse deep learning workloads demonstrate that Fico achieves an average latency reduction of 37.3% compared to state-of-the-art systems, without violating tenant quotas.

References

[1] Mingcong Han, Hanze Zhang, Rong Chen, and Haibo Chen. 2022. Microsecond-scale Preemption for Concurrent GPU-accelerated DNN Inferences. In OSDI 2022, 539– 558.

[2] Marcus Chow, Ali Jahanshahi, and Daniel Wong. 2023. Krisp: Enabling kernel-wise right-sizing for spatial partitioned gpu inference servers. In HPCA 2023, 624–637

[3] Foteini Strati, Xianzhe Ma, and Ana Klimovic. 2024. Orion: Interference-aware, fine-grained GPU sharing for ML applications. In EuroSys 2024, 1075–1092

香港城市大學
City University of Hong Kong