INTRODUCTION

GPU: Lack of flexibility
A deep learning (DL) job can have many GPUs, but each GPU belongs to exactly one application.

• Hinders the scheduling ability of GPU cluster managers
• Underutilization
  • Hyper-parameter tuning (AutoML)
  • Model serving (Inference)

MEMORY USAGE

3 types of memory:
• Model
• Ephemeral
• Framework-internal

GPU LANE

• Continuous physical memory + GPU stream
• Time-slicing within lane, parallel across lanes
• Dynamic re-partitioning
• Avoid in-lane fragmentation

EVALUATION

2 sets of hyper-parameter exploration

100 jobs from a production trace: SRTF vs FIFO: 3.19x improvement in Avg. JCT

PACK
packs tasks together for higher utilization.

SRTF
prioritize based on shortest remaining time.

FAIR
equalizes the resource usage of active jobs.

Still a Huge design space to explore