Salus: Fine-Grained GPU Sharing Primitives for Deep Learning Applications

COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

Peifeng Yu, Mosharaf Chowdhury

Open sourced at: https://github.com/SymbioticLab/Salus

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)

DESIGN

Salus is a consolidated execution service enabling sharing primitives:

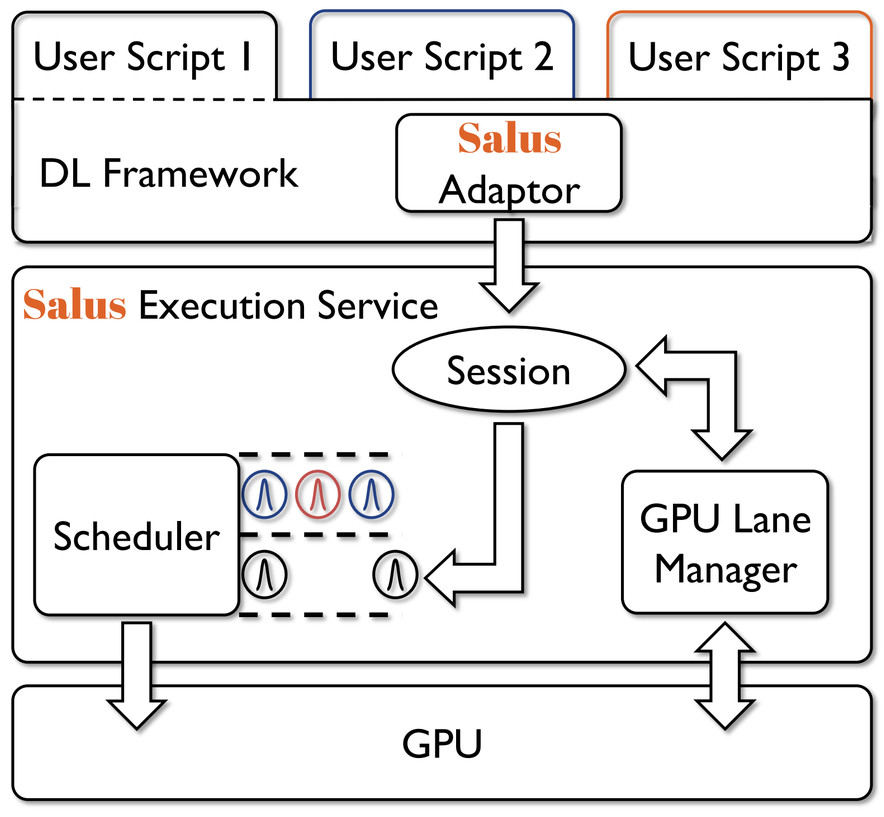
- Fast job switching,
- Memory sharing

Without modifying any

- User scripts
- Operating systems, or
- Hardware

With the goal to

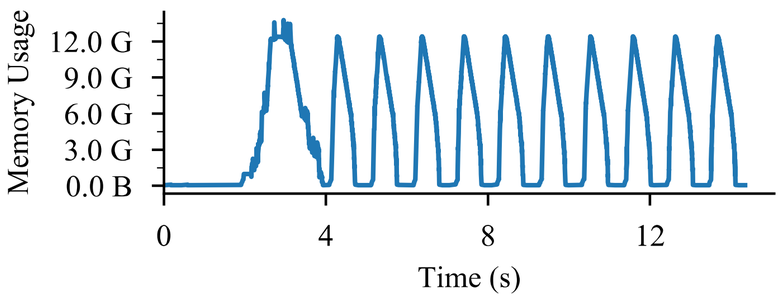
- Support new GPU schedulers,
- Improve GPU utilization



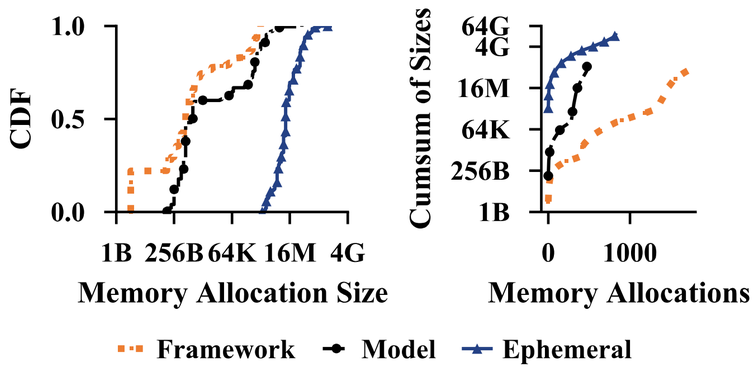
MEMORY USAGE

3 types of memory:

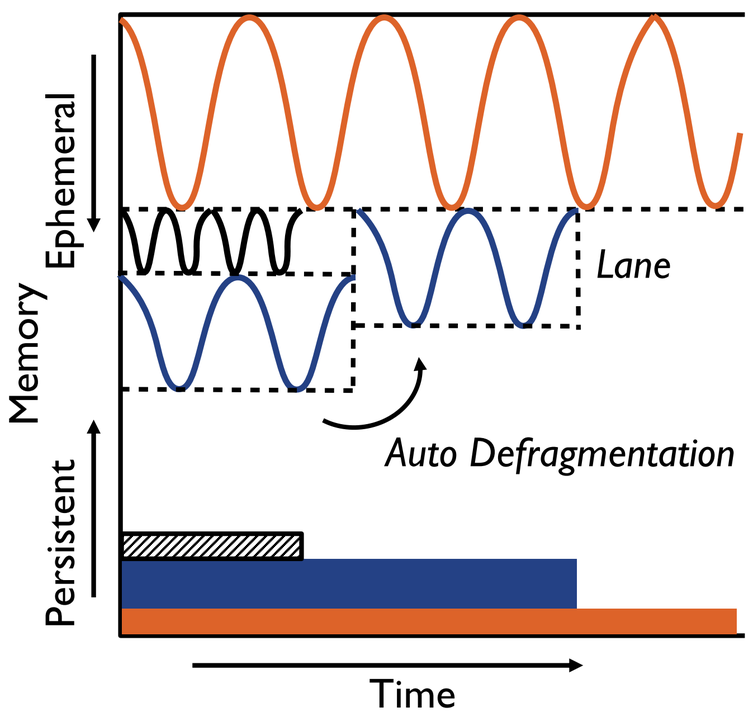
- Model
- Ephemeral
- Framework-internal



Memory usage when training ResNet152



GPU LANE



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

SCHEDULING

PACK

packs tasks together for higher utilization.

SRTF

prioritize based on shortest remaining time.

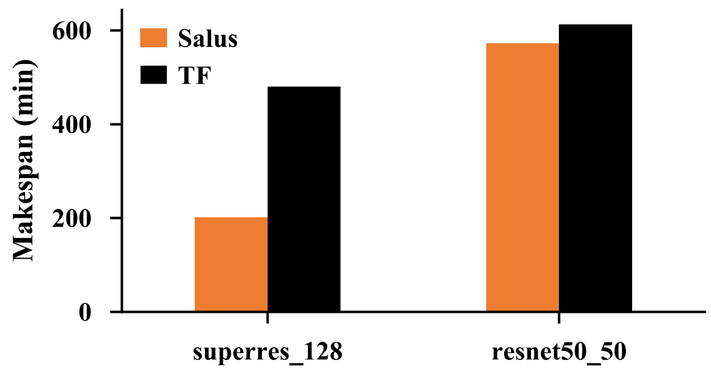
FAIR

equalizes the resource usage of active jobs.

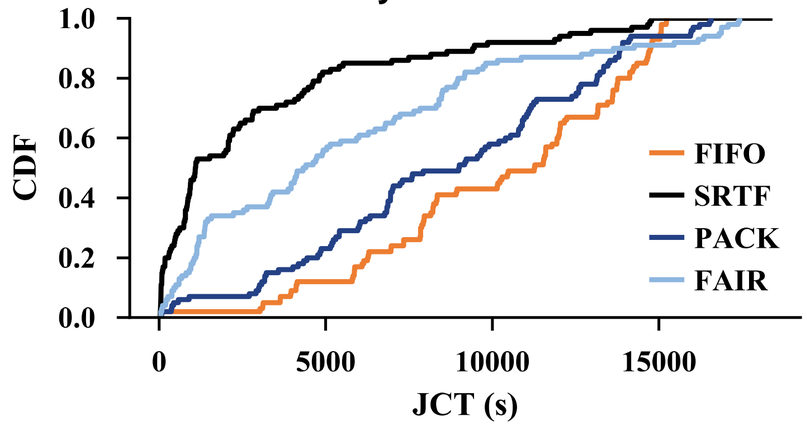
Still a Huge design space to explore

EVALUATION

2 sets of hyper-parameter exploration



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



42 DL inference applications in 1 GPU