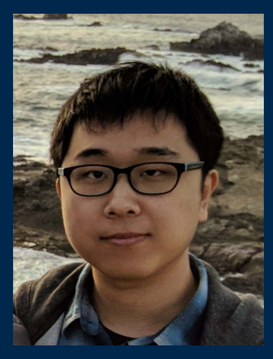
Salus: Fine-Grained GPU Sharing Among CNN Applications

COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

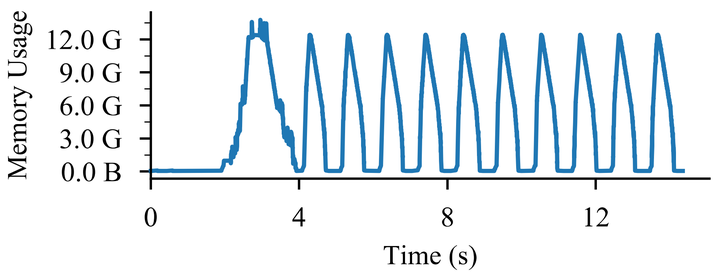
Peifeng Yu, Mosharaf Chowdhury



INTRODUCTION

A deep learning (DL) job can have multiple GPUs, but each GPU belongs to exactly one job regardless of its utilization level.

➢High peak-to-average usage ratio creates chronic under-utilization.

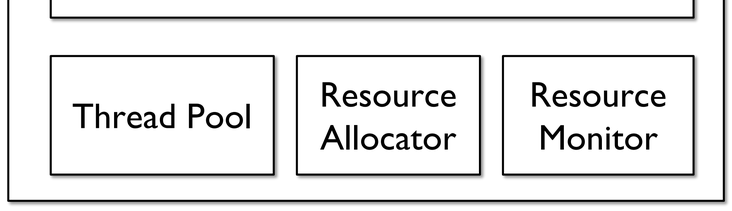


Memory usage when training ResNet152.

➢Prior GPU sharing work is not suitable for DL.

Approach	CUDA API	Backward Compatible	Fine-grained Sharing	Sharing Policies	DL Support
Cluster managers	Yes	Yes	No	Yes	Yes
Library Interception	Partial	Yes	Yes	No	No
New set of API	No	No	Yes	Yes	No
Salus	Yes	Yes	Yes	Yes	Yes

DESIGN

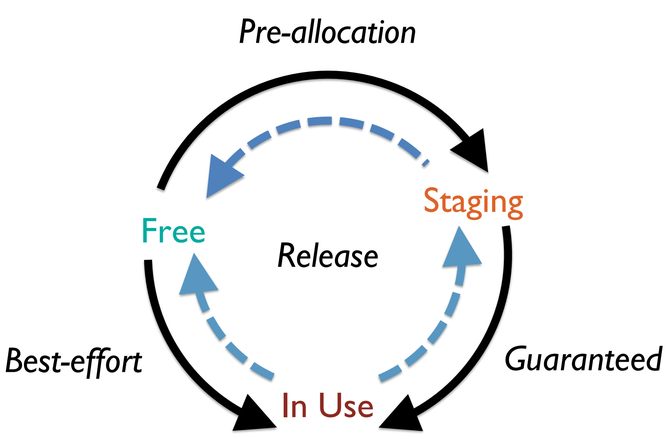


Salus *enables fine-grained sharing of individual GPUs among multiple CNN applications, without modifying any user scripts or operating systems.*

➢***Salus Execution Service*** consolidates all GPU accesses and provides primitives like fair sharing, preemption and admission control.

➢***Salus Adaptor*** collects information inside each framework and submits it to execution service.

MEMORY



Three stage memory life cycle

➢Task- and Job-level memory management

➢Deadlock avoidance: safety condition and paging

➢Job admission control

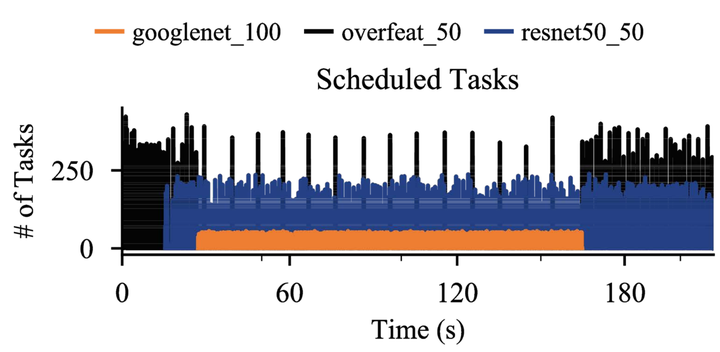
SCHEDULING

Job: one session of training of a computation graph

Task: a closure that executes a certain op.

Three simple policies to show possibilities of a huge design space to replace FIFO.

1. ***Packing*** packs tasks together for higher utilization.
2. ***Preempt*** enables prioritization based on arbitrary rules.
3. ***Fairness*** equalizes the resource usage of active jobs.



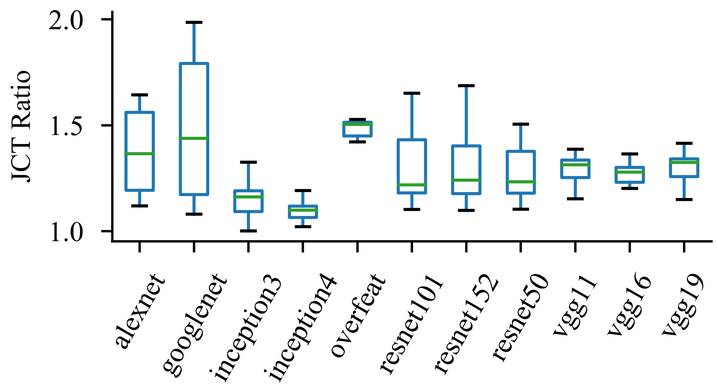
Three training jobs running with preemption

EVALUATION

➢Salus can share GPU among multiple jobs using various policies.

➢Salus can handle 20 jobs simultaneously and equalize their memory usage.

➢Salus still has some overhead in terms of JCT when comparing raw speed.



JCT ratios of Salus w.r.t. baseline