GPU Efficiency through Intelligent Collocation

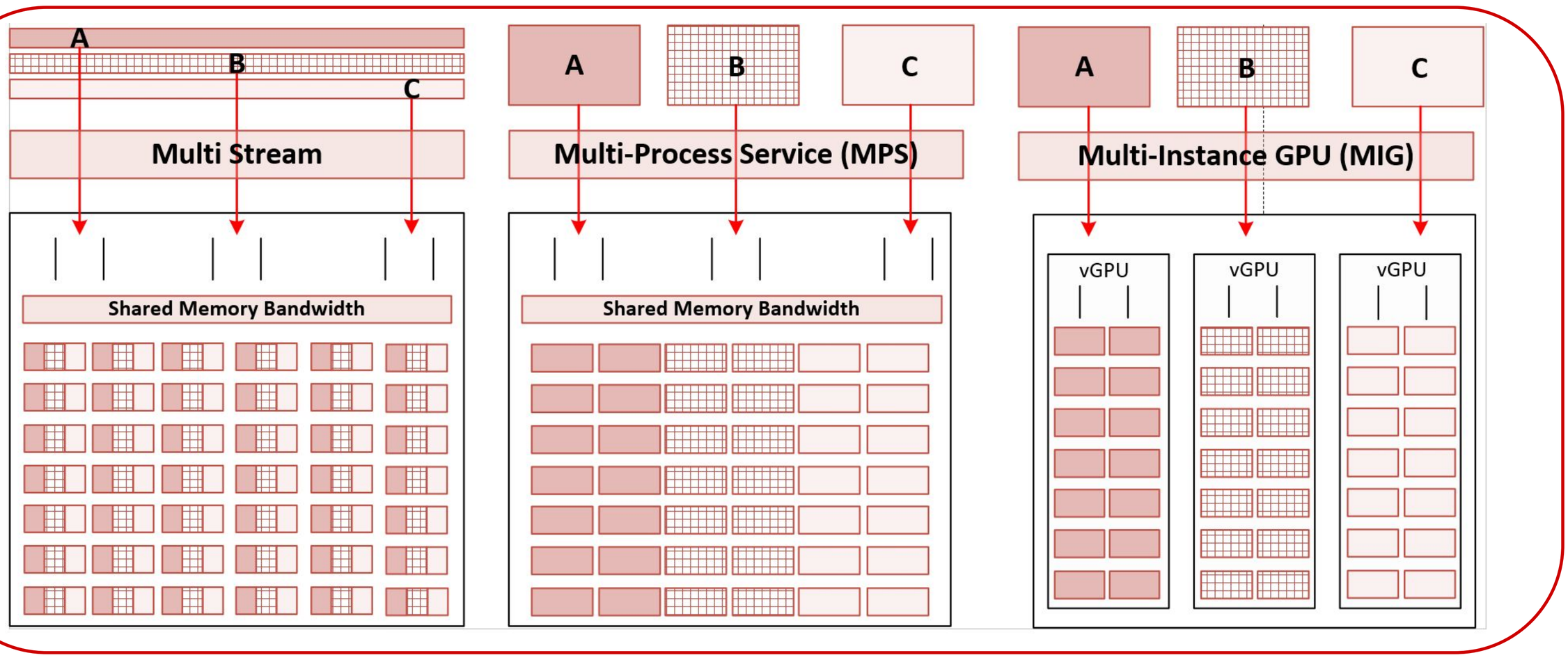
Ehsan Yousefzadeh-Asl-Miandoab (ehyo@itu.dk) Ties Robroek (titr@itu.dk) Pınar Tözün (pito@itu.dk)

GPU Underutilization and Causes

1- The lack of fine-grain sharing mechanism of GPUs
2- Lack of advanced virtual memory for GPUs
3- Adopting big-data fit schedulers (black box scheduling)
4- Gang scheduling nature of distributed learning tasks.

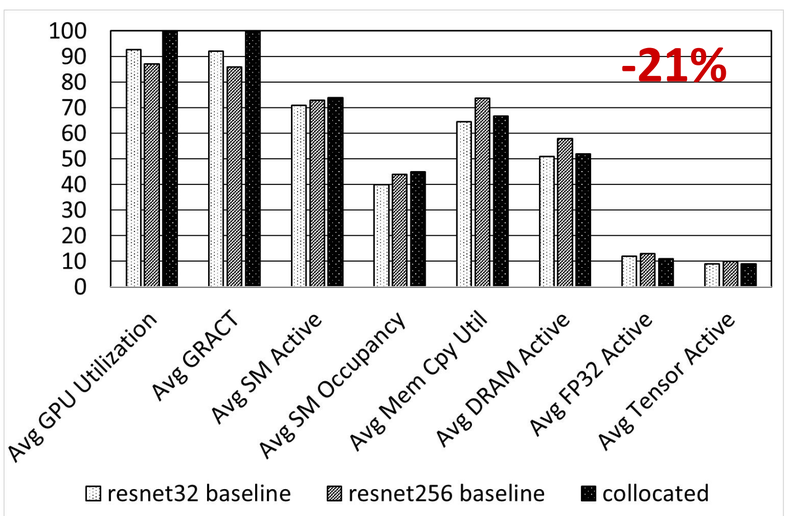
Collocating is Challenging!

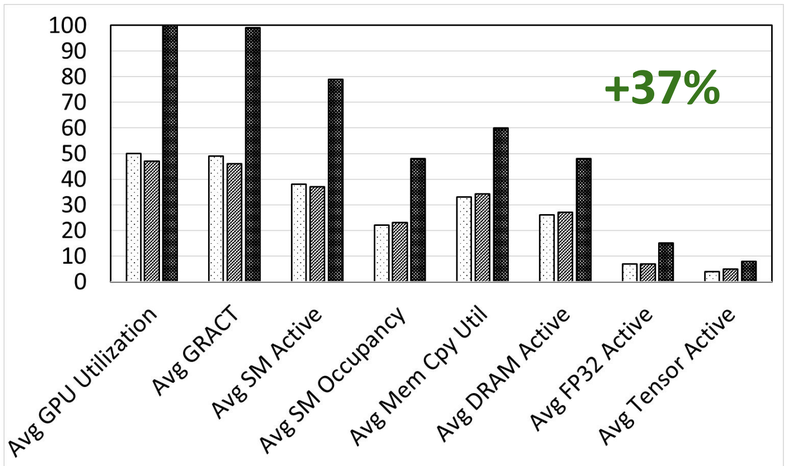
1- Crashing due to lack of memory (batch size and TTA tradeoff)
2- Memory Fragmentation
3- Model Signature



Naive Collocation

Degradations - Interference

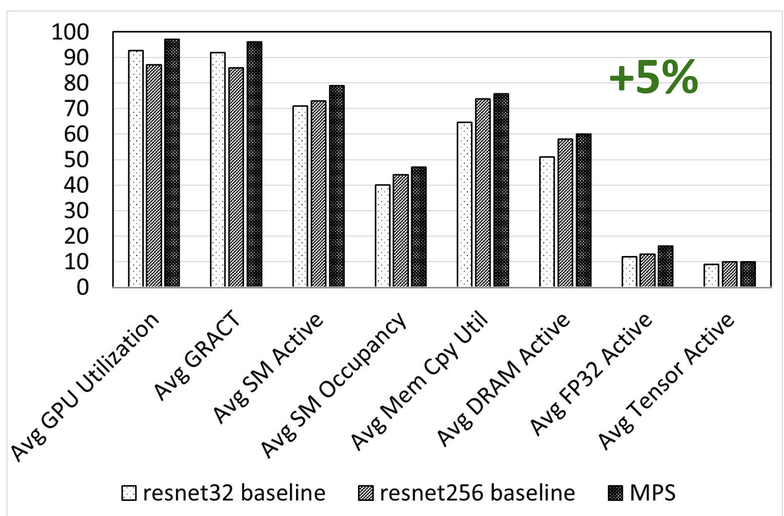


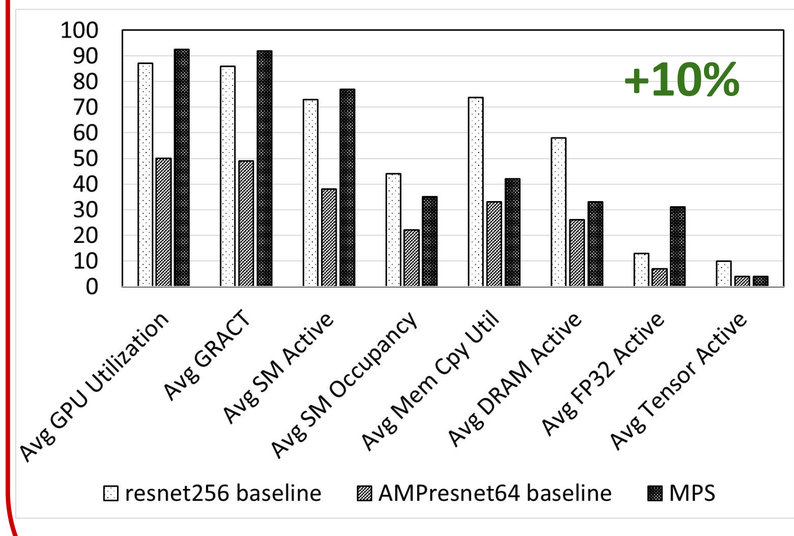


Efficient when models are light!

MPS

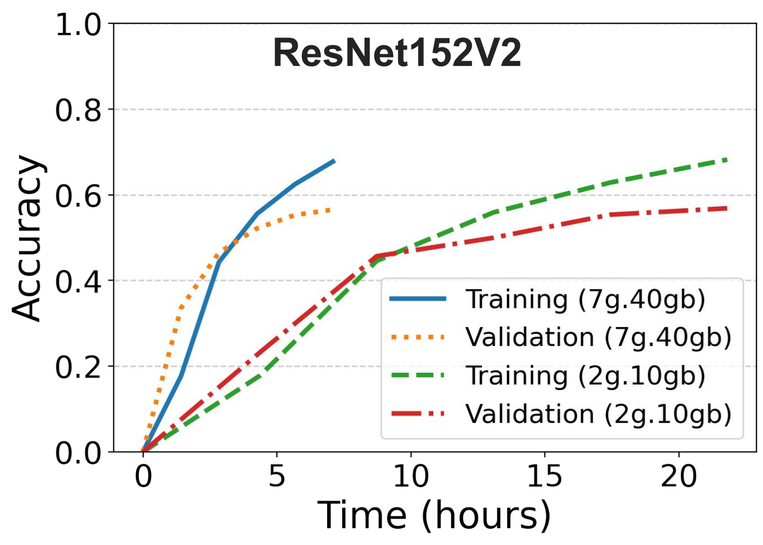
Effective when kernels are light

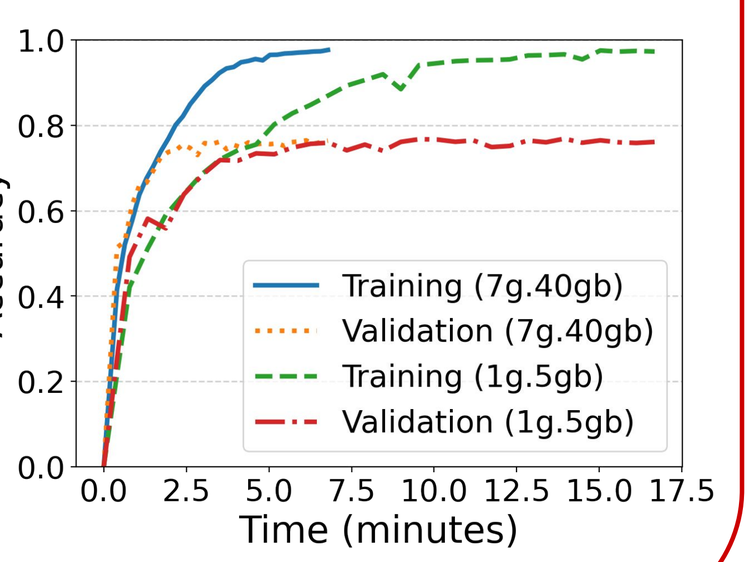




MIG

Isolation/ Non-Linear Improvement





ResNet26V2

Schedulers should become more intelligent, accelerators must get more specialized!

1- Features of models
2- Features of general programs
3- Machine Learning based estimation
4- Reinforcement Learning
5- Scheduling Granularity
6- GPGPU Utilization!

IT UNIVERSITY OF COPENHAGEN
itu.dk

RAD
rad.itu.dk

DASYA
Data-Intensive Systems and Applications
dasya.itu.dk