Profiling & Monitoring Deep Learning Training Tasks

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

GPU Underutilization for ML Workloads

- An analysis of 100,000 jobs run by 100s of users for ~2 months on a real-world cluster shows ~52% GPU utilization on average*
  - Energy-inefficient & waste of hardware resources
- Compute/memory requirements of models don't match with the giant GPUs
  - e.g., transfer learning, small models

Thus, understanding the profilers and monitoring tools for GPUs is necessary.

Profilers

PyTorch profiler

Nsight Compute

Nsight Systems

Monitoring tools

NVIDIA System Management Interface (nvidia-smi)

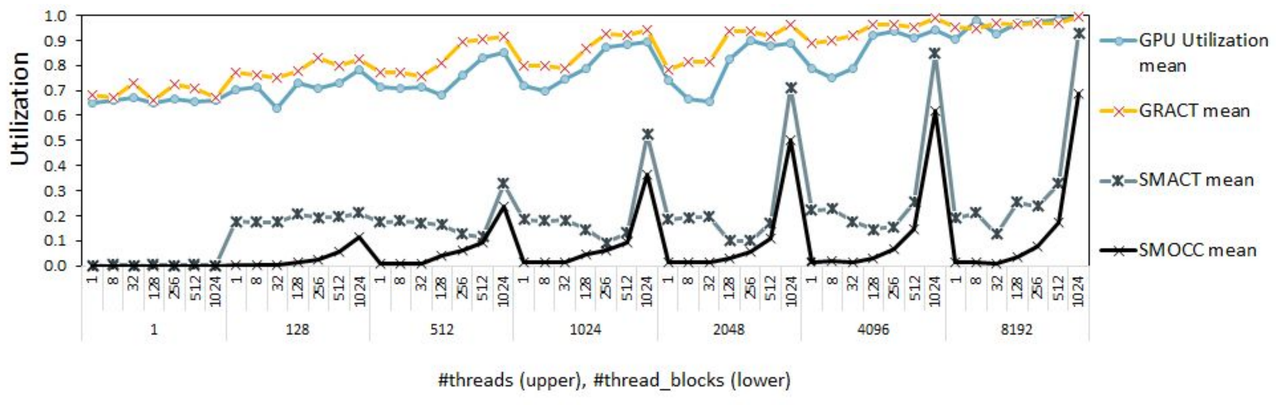
NVIDIA Data Center GPU Manager (dcgm)

Which metrics for utilization?

Coarse-grained:
GPU Utilization: % time one or more kernels execute on the GPU
GRACT: % time the graphics or compute engines were active

Fine-grained:
SMACT: the fraction of active time on an SM, averaged over all SMs
SMOCC: degree of parallelism / max supported parallelism on SM

A100 40GB **A CUDA micro-benchmark**



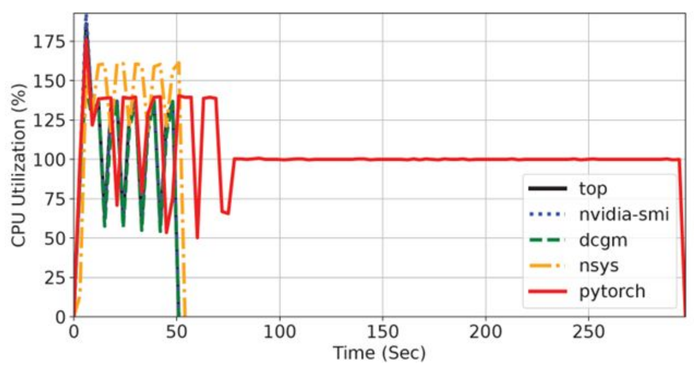
Coarser-grained utilization metrics can be misleading.

Tool overhead

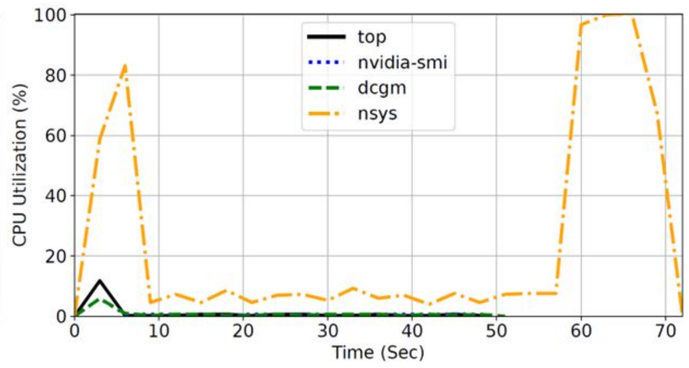
Simple CNN on MNIST (light) **DGX A100 stationA100 40GB**

Tools	simple CNN		ResNet50	
	Time	Space	Time	Space
no tool	9.61sec	NA	37.06min	NA
top	9.66sec	~20KB	37.11min	~2MB
nvidia-smi	9.61sec	~20KB	37.04min	~2MB
dcgm	9.68sec	~85KB	37.19min	~8MB
nsys	9.88sec	~40MB	39.13min	~5GB
pytorch	13.65sec	~1.4GB	NA	

Training processes



Tools' processes



⇒ Monitoring tools have negligible time and space overhead.
⇒ Profiling tools' CPU and memory overheads are higher than monitoring tools.
⇒ GPU overhead of tools is negligible.

Insights

- For model level optimization purposes
  - Use framework-specific profilers
- For digging deeper into OS and system
  - Use Nsight Systems
- For kernel-level optimization
  - Use Nsight Compute
- Profile only the needed amount
  - e.g., an iteration of the model may be enough
- For online decision-making purposes
  - Use monitoring tools with representative fine-grained metrics

INDEPENDENT RESEARCH FUND DENMARK
dff.dk

IT UNIVERSITY OF COPENHAGEN
www.itu.dk

DASYA
www.dasya.dk
@dasyaITU

RAD
https://rad.itu.dk