GPU Working Group Outbrief

Scalable Tools Workshop July 2018

GPU Performance and Correctness

Issues of interest

- Attribution of performance to computational kernels
- Control of WARP scheduler for reproducibility
  - without control, hard to do correctness analysis
- Why unpredictable slowdowns?
  - hidden synchronization
  - CUPTI slow
  - driver+library change causes slowdowns
- Performance counters and power management
  - ACPI allows host to sleep a GPU
  - cap power used by nvidia GPU using nvidia-smi
- Missing information in nvprof

Challenges

- Understanding performance with PC sampling
  - very high sampling rates
  - delivers histograms of PC samples to the host
  - unknown how to tune performance based on PC sample information
- Capturing missing synchronization
  - NVIDIA omits mention of “activities”, e.g., synchronization
    - without NVIDIA’s help, can only catch these with binary instrumentation
  - capture GPU memory access behavior and synchronizations
    - PTX binary analysis - Hollingsworth
- Understanding needs for FP precision
  - Are FP precision analysis tools of use for ML?
    - FPTuner (Utah) github.com/soarlab/fptuner
    - Verificarlo

Other Topics

- Interest in performance data related to
  - threads, thread blocks,
  - See cudaOccupancyCalculator
- What's next from AMD - APU for exascale design?
  - http://www.computermachines.org/joe/publications/pdfs/hpca2017_exascale_apu.pdf
- Are new ML HW designs of interest for HPC?

Action Items

Identify issues of broader interest to NVIDIA to engage them

- Performance analysis of “altcoin” mining
  - Levinthal, Mellor-Crummey, Welton
- Work with machine learning community to identify performance infrastructure needs
  - David to suggest ML benchmark that we should measure
    - analysis of LSTM language translator
      - Pytorch/tensorflow calls nvidia libCUdnn
    - David to provide URL, instructions, etc.