Performance Profiling Update of madgraph4gpu with Nvidia A100 GPU

Carl Vuosalo
Harshit Sharma
University of Wisconsin-Madison

Profiling Overview

- centos:7 Docker container:
  - NVIDIA-Linux-x86_64-525.116.04
  - cuda_12.1.1_530.30.02_linux
  - HEAD of https://github.com/madgraph5/madgraph4gpu.git
  - epochX/cudacpp/gg_ttgg.sa/SubProcesses/P1_Sigma_sm_gg_ttxgg
  - AVX=avx2 FPTYPE=d HELINL=0
  - Nvidia A100-SXM4 with 40GB global GPU memory
- Tests run via HTCondor jobs
- Trying to use Nsight Compute tool from Nvidia to perform profiling

A100 Performance

- Find optimum max register value
- Double precision
- 256 threads/block
- Number of iterations = 30 x 5

- Best performance with 255 max registers and 216 blocks/grid
- Achieves 8.82e5 ME/s

108 blocks/grid	120	140	160	180	200	220	240	255	260
ME/s (e5)	4.86±.01	5.58±.01	6.38±.01	7.18±.03	7.68±.05	8.05±.05	8.41±.03	8.77	8.76±.06

216 blocks/grid	120	140	160	180	200	220	240	255	260
ME/s (e5)	4.00±.01	5.63±.01	6.44±.005	7.23±.01	7.75±.01	8.14±.06	8.62±.02	8.82±.03	8.81±.02

Summary

- Best performance with A100 GPU is 8.82e5 ME/s using:
  - 255 max registers
  - 216 blocks/grid
- Will explore profiling of A100 next

Backup

Profiling Overview--Previous

- Performance profiling to find parameters for highest performance
- Platform at T2_US_Wisconsin:
  - CentOS 7.9.2009 (Core)
  - 125.74 GiB RAM
  - 1x AMD EPYC 7402P 24-Core Processor, 48 threads
  - GPU 2x TU104GL [Tesla T4] (only 1 GPU used)
- Physics process to profile: g g → t t g g
- Using Nsight Compute tool from Nvidia to perform profiling
- Majority of time (~90%) spent on SigmaKin kernel that does matrix element (ME) calculation (not counting I/O)

Tesla T4 Performance -- Old

- Check gg_ttgg GPU performance with varied max registers
- Double precision
- Number of blocks = 65536
- Number of iterations = 12

- Best performance with 160 max registers
- Memory usage:
  - 369 MB for 32 threads/block
  - 873 MB for 128 threads/block

32 threads per block	128	140	160	180	216	240	256
Events/s (ME)	1.13e4	1.12e4	2.30e4	1.11e4	1.11e4	1.11e4	1.11e4

128 threads per block	128	140	160	180	216	240	256
Events/s (ME)	1.91e4	1.90e4	2.31e4	1.89e4	1.90e4	1.89e4	1.90e4