Status of the GPU Simulation prototype

Soon Yung Jun
Philippe Canal

Connection To Vector Prototype

- Broker than can schedule the processing of tracks with maximum flexibility:
  - Transforms the CPU tracks in GPU tracks.
  - Stages the tracks coming from one or more until there is either
    - Enough track to efficiently process
    - No more interesting tracks
  - Reschedules the *uninteresting* tracks to the CPU
  - Continues to bundle and upload tracks while a kernel is executing.
    - By overlapping upload/download/kernel using streams.

Connection To Vector Prototype

- Geometry
  - Connection between the two geometry with (set of) indices for both physical and logical volume

- Next:
  - Investigate how to download a variable size array from the GPU in a efficient manner
  - Rebuild the transportation history/location information on CPU (based on the particle path on the GPU).

EM Physics for GPU

- EM physics for electrons and photons were ported for CUDA kernels based on Geant4.
- List of EM physics processes/models and associated secondaries

Primary	Process	Model	Secondaries	Survivor
e^-	Bremsstrahlung	SeltzerBerger	γ	e^-
	Ionization	MollerBhabhaModel	e^-	e^-
	Multiple Scattering	UrbanMscModel95	–	e^-
γ	Compton Scattering	KleinNishinaCompton	e^-	γ
	Photo Electric Effect	PEEffectFluoModel	e^-	–
	Gamma Conversion	BetheHeitlerModel	$e^- e^+$	–

Device Memory Accesses

- EM processes/models require frequent data accesses from/to global memory
- Input
  - material information
  - physics tables (lambda, dedx, range and others)
- Output
  - secondary particles (electrons, positrons, photons)
  - hits (energy, position)

Strategies for Secondaries/Hits

- N-secondaries per step (N=0,1,2)
  - pre-allocated memories (a fixed size stack)
  - dynamic memory allocations per-thread-basis (local) or per-block-basis (shared), and reallocate them into a single stack (global) – need 2-kernels
- One hit for each step (only in sensitive detector)
  - Need a temporary hit container on the global memory for multiple stepping
  - make hits in CPU if tracks are transferred to CPU after one step on GPU

Baseline Performance

- Test configuration
  - thread organization: 32 blocks x128 threads
  - one step for 100K primary tracks
  - maximum number of secondaries per track = 2
- Memory transaction (atomic add)

	GPU [ms]	CPU [ms]
Fixed memory	1.5	35
Dynamic per thread	130	60
Dynamic per block	60	60

Preliminary Performance Evaluation

- One stepping for 20K electrons/photons
  - 32 blocks x128 threads
  - a single material ($PbWO_4$)
  - momenta range ~ 10-100 MeV
  - use tables (lambda, dEdx, range, bream_SB tables)
  - write secondaries to a pre-allocated fixed size stack on the global memory
- Compile flags for nvcc
  - -arch=sm_20 –use_fast_math --optimize 2

Preliminary Performance Evaluation

- Photon processes

Accumulative	CPU[ms]	Kernel/GPU	CPU/Kernel	CPU/GPU
Compton	16.8	0.93/2.35	18.3	7.2
Conversion	43.6	1.52/2.95	28.7	14.8
PhotoElectric	13.9	3.91/5.40	12.0	4.1

- Electron processes

Accumulative	CPU[ms]	Kernel/GPU	CPU/Kernel	CPU/GPU
Bremsstrahlung	635	14.5/15.9	43.7	40.0
Ionisation	103	2.04/3.46	50.7	29.8
MultipleScattering	38	1.61/3.02	23.9	12.7

Managers

- Process managers (build a physics list)
- Stepping manager
  - physical interaction length
  - AlongStepDoIt and transportation
  - PostStepDoIt (secondary sampling) and hit making
- Tracking manager
  - the main driver for processing one track
  - start/end tracking
  - do N-stepping while a track is alive (N is set-able)

Preliminary Performance Evaluation

- One step with a simple calorimeter (CMS Ecal)
  - tracking particles with all EM processes including transportation with a realistic (CMS) magnetic field
  - separate kernels for photons and electrons
  - host (CPU) code is optimized with -O2 (factor 1/2)
- Performance

	CPU [ms]	Kernel/GPU	CPU/Kernel	CPU/GPU
Photon Kernel	109	9.4/10.8	14.2	11.9
Electron Kernel	285	21.2/22.7	13.4	12.6

Overview of the GPU Prototype

- Primary CUDA components
  - particle transportation in a magnetic field
  - geometry and material
  - EM physics processes/models (electrons/photons)
  - managers (process, tracking, stepping)
  - random numbers generators
- Validation framework
  - identical host codes (executed on CPU)
  - Geant4 application for each/combined process(es)

Current Status and Future Plan

- Consolidating transportation, geometry and EM physics into managers and perform stress-tests
- Building and testing a framework for physics validations as well as performance evaluations
- Build a realistic example of an experiment and (detector, magnetic field, input tracks) and test the GPU prototype
- Optimize, optimize and optimize (simplify, reorganize and develop strategies)