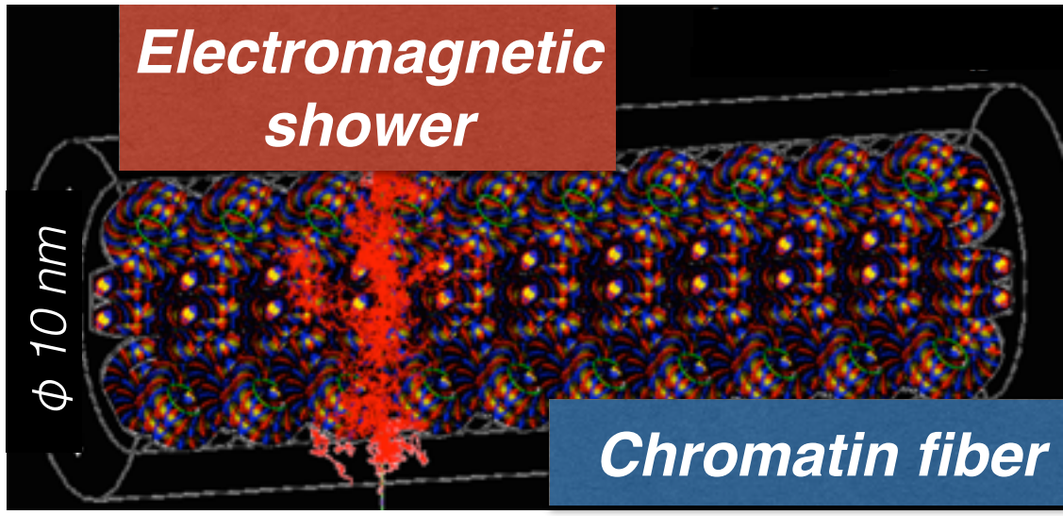
GPU Acceleration of Physics Models in Geant4-DNA

11th Geant4 Space Users' Workshop
@ Hiroshima
August 28, 2015

S. Okada, K. Murakami, K. Amako, T. Sasaki; *KEK, Japan*
S. Incerti, M. Karamitros; *CENBG, France*
N. Henderson, M. Gerritsen; *Stanford University, ICME, US*
M. Asai, A. Dotti; *SLAC, US*

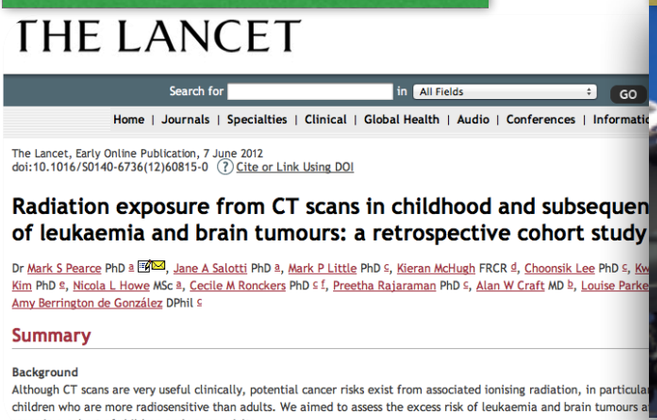
Geant4-DNA

- Extension of Geant4 physics models for radiation simulation at DNA scale.
  - EM physics in low energy scale (*down to a few eV*)
  - Chemical Processes (*e.g. Brownian motion*)
  - Outputs:
    - Local energy deposition
    - Quantification of DNA damage



- Quantitative estimation of biological effects of radiation:
  - Estimation of effects on human health for chronic radiation exposure
    - Medical diagnosis, Airline crew, Space mission, …
  - Radiation therapy simulation

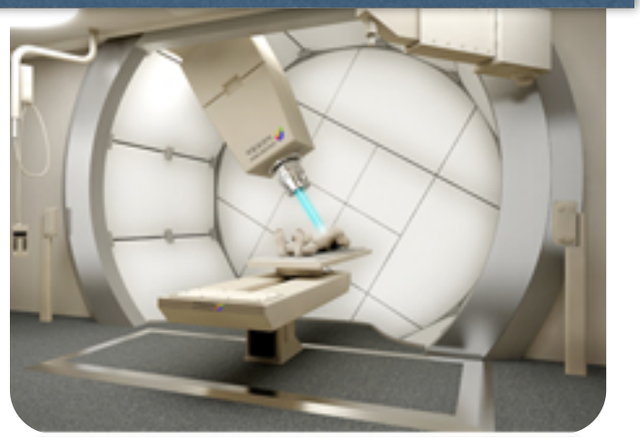
Diagnosis



Airline crew

Space mission

Radiation therapy



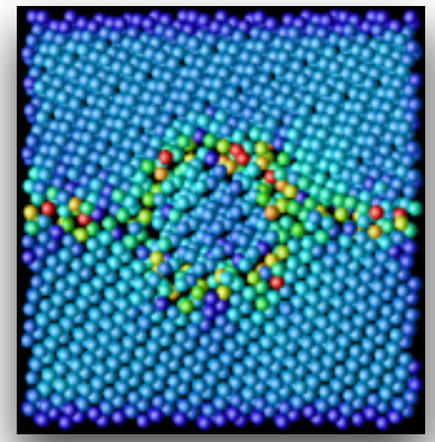
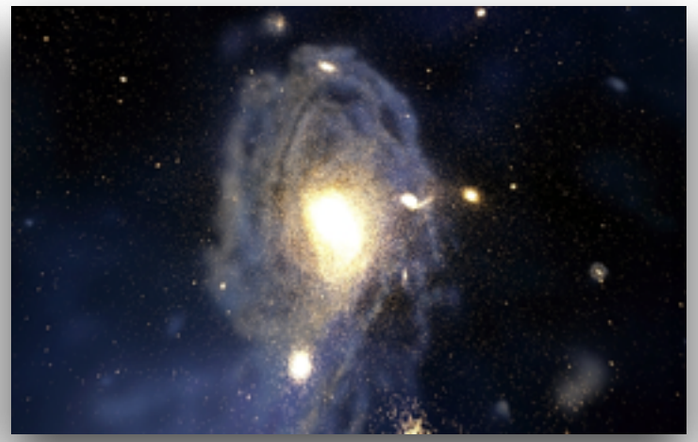
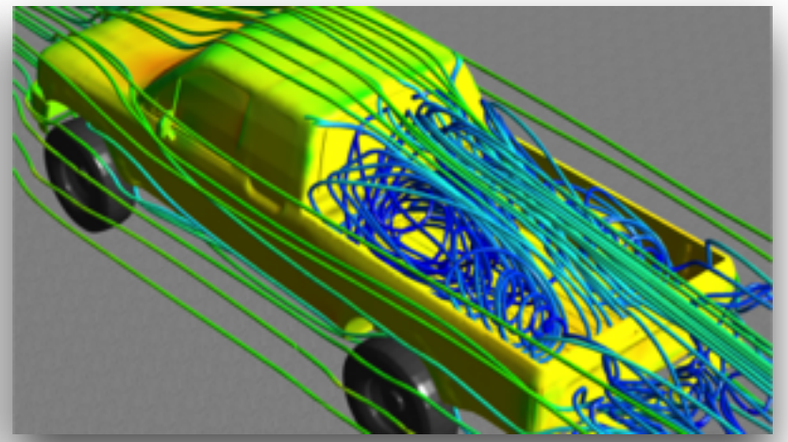
Matter of Geant4-DNA

- ***Geant4-DNA*** has matter on its simulation speed.
  - *Days-Weeks* simulation on CPU cluster.
    - Due to large amount of low energy secondary particles generated.

- With focusing on Geant4-DNA physics processes/models, we are attacking to improve its computational performance using ***GPGPU***.
  - Expect scientific breakthrough of DNA-level particle simulation.

GPGPU: General Purpose on GPU

- Using many GPU cores, computational performance of various types of scientific calculation have been improved against CPU.
    - Molecular Dynamics, Astronomy, Thermal & Fluid Dynamics, …

    - NVIDIA Tesla K20c GPU (Kepler Architecture)
        - 2,496 CUDA cores, 706 MHz
        - 5 GB Memory (GDDR5)

- Some domains of Geant4 are hoping improvement of its computational performance.
    - GPU can track many particles in parallel, and gain calculation speed drastically.
    - Achieve up to *250 times* speedup in the Standard EM Physics simulations.
      (-> see later slides)

Key words on GPU Programming

SIMT: Single Instruction, Multiple Threads

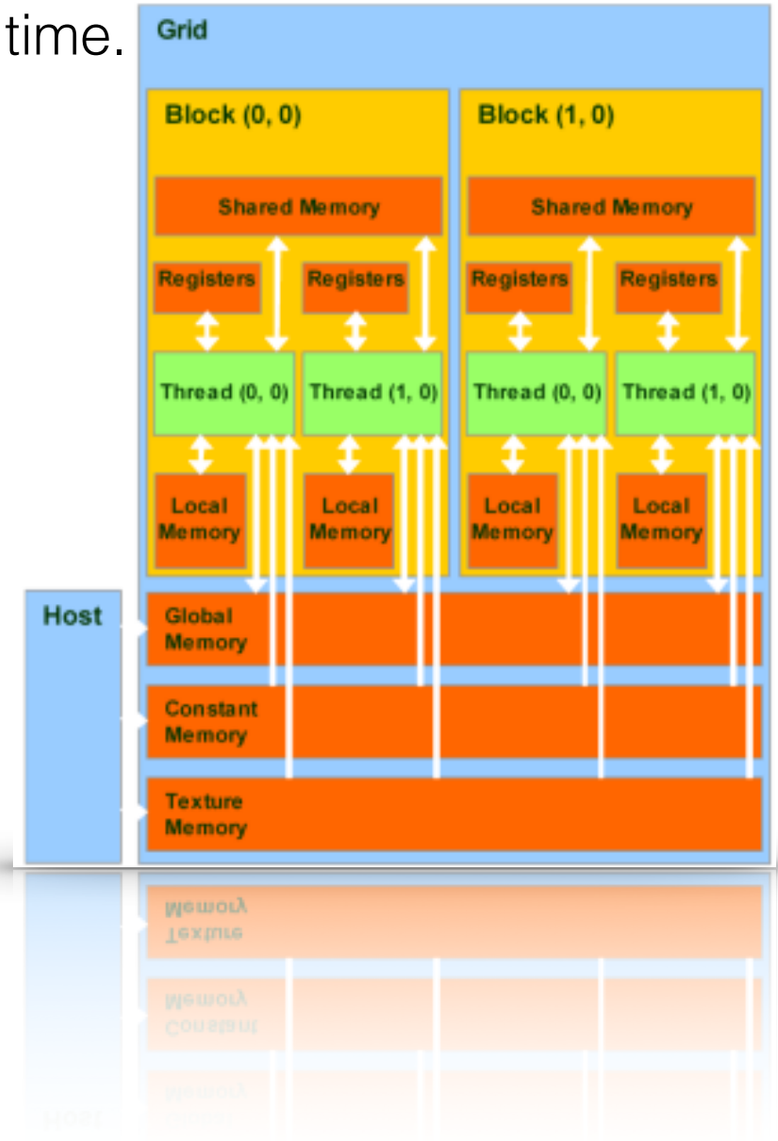
- Each instruction is issued to all threads in a “warp” at the same time.

Thread divergence

- If the code contains branches, some threads in a warp may have to stall until the branches converge.
- Minimizing branching is an important aspect in achieving good performance.

Memory hierarchy

- NVIDA GPUs have a large bank of “global” memory and a set of “on-chip” cashes.
- G4CU uses only global memory.



Coalesced memory access

- The hardware can bundle the requests by contiguous threads to contiguous memory locations.
- Crucial component in achieving high performance.

Race conditions

- Could arise when multiple CUDA threads attempt to read/write to same location in global memory.
- G4CU uses atomicAdd operations to avoid race conditions in dose accumulation.

Our GPU simulator: G4CU

Motivation to developing of GPU simulator

Application to Medical science by Geant4

- Geant4 is used in QA of treatment planning for *"radiation therapy"*.
  - Dose calculation with high accuracy (a few %)
    (Energy range: *1 keV ~ 100 MeV)*
  - One day simulation on CPU cluster.

- ***G4CU***, GPU-powered Geant4 EM physics
  - Up to 250 times speed-up simulation against CPU single core
    (see later slides)
  - Treatment planning using MC with higher accuracy

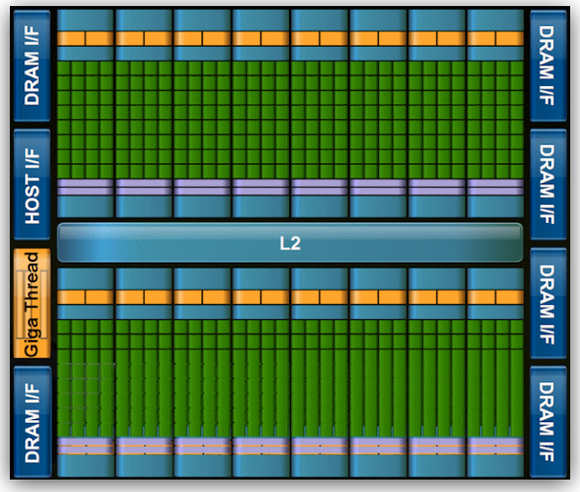
G4CU Overview

Main Objective

- Speed up dose calculation for “radiation therapy” using GPU computing

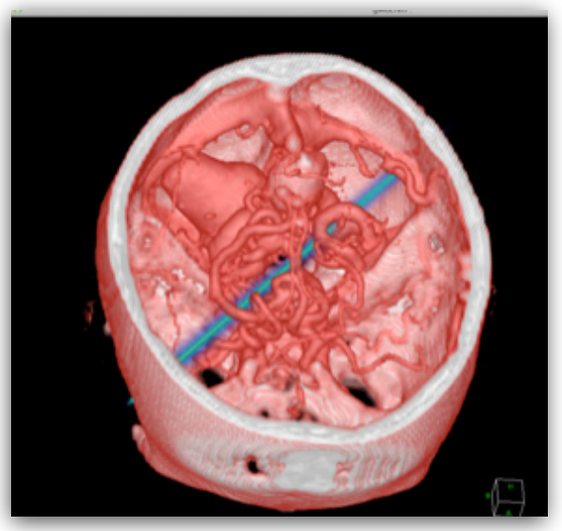
GPU-powered

- Implementing of EM physics of Geant4 in **CUDA**
- Tracking many particles in parallel



Current Functionality

- Transportation in voxelized geometry
- Water with variable densities (Water equivalent model)
- Standard EM processes for e+/e- and γ
- Scoring dose in each voxel.



Collaboration:

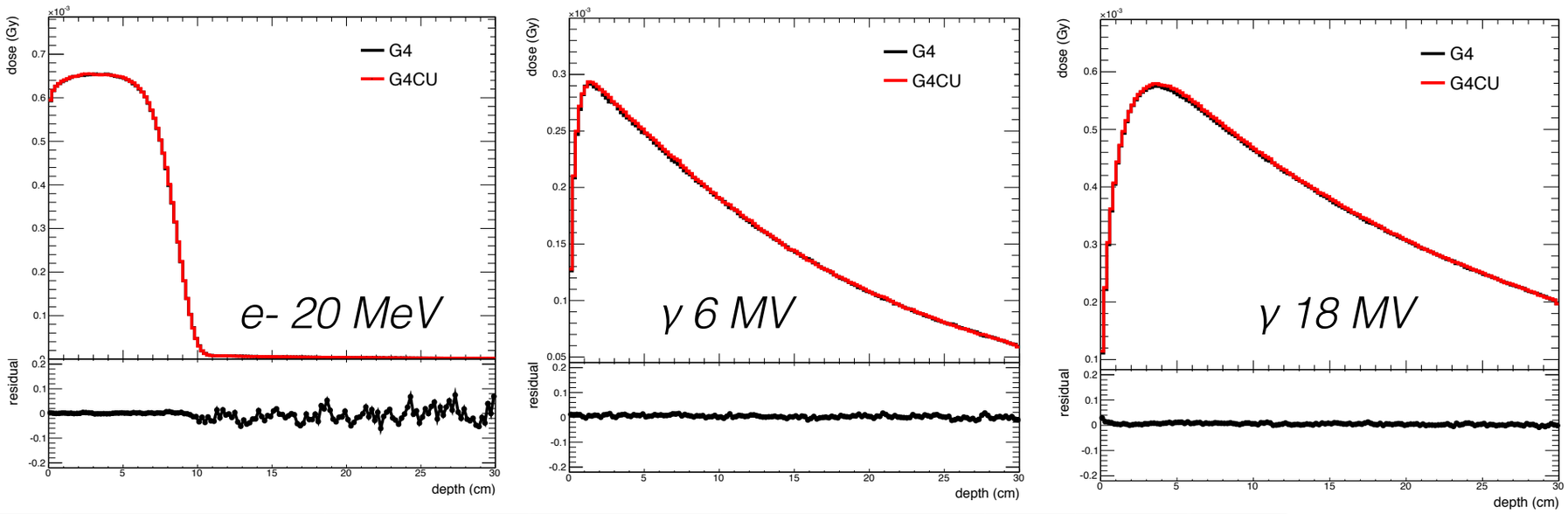
Geant4 @ SLAC

ICME INSTITUTE for COMPUTATIONAL & MATHEMATICAL ENGINEERING at STANFORD UNIVERSITY

Geant4 @

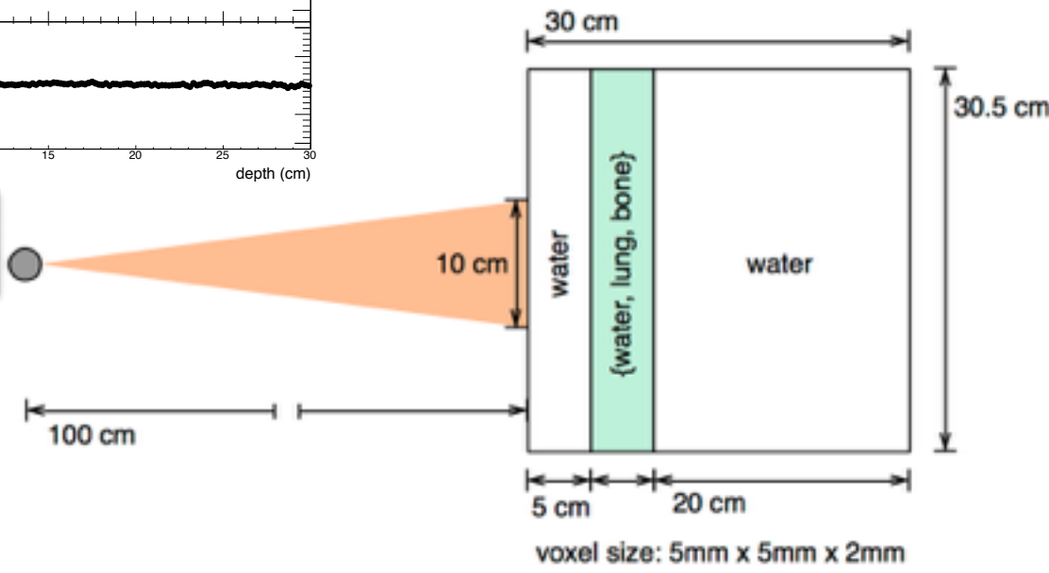
G4CU Performance

Depth dose comparisons



— ***Geant4 (v9.6.4)***
— ***G4CU***

- e- 20 MeV
- γ 6, 18MV Linac



Perfect agreements in physics performance

Computation time

water phantom	e- 20MeV	γ 6MV	γ 18MV
Geant4 [msec/particle]	1.84	0.78	0.803
G4CU [msec/particle]	0.00881	0.00336	0.00433
× speedup factor (= G4 / G4CU)	***208***	***232***	***185***

Currently, *up to 250 times* faster than CPU single core

- GPU (NVIDIA, Tesla K20c, 2496cores, 706MHz)
  - *4096 blocks x 128 threads / block*
- CPU single core(Intel, Xeon E5-2643 v2, 3.50 GHz)

Ref.) Nick Henderson, et. al, “Dealing with Thread Divergence in a GPU Monte Carlo Radiation Therapy Simulator” @ GPU Technology Conference 2015

G4CU-DNA:

Implementing Geant4-DNA physics models in CUDA as an extension of physics processes of G4CU.

G4CU-DNA Overview

GPGPU benefits to Geant4-DNA physics

- expect breakthrough in computational time

Implementing physics processes / models of Geant4-DNA in CUDA

- Based on Geant4 10.2 beta 01
- Particles: *electrons, protons, hydrogen atoms, and helium atoms with charged states (He^{++}, He^{+}, and He^{0}).*

Data tables

- All cross-section tables, reformatted with uniform energy bin intervals.
  -> much faster bin-searching for a given energy than bisection method.
  (*reduce memory access*)

Material and Geometry

- Transportation in voxelized geometry
- Water with variables densities (Water equivalent model)

supported by

Geant4 @

Physics Processes / Models in G4CU-DNA

	Particles	Electron	Proton	Hydrogen	Helium (He^{++}, He^{+}, He^{0})
Physics Processes	**Elastic scattering**	9 eV - 10 keV Uehara 10 keV - 1 MeV Champion	100 eV - 1 MeV Hoang		100 eV - 10 MeV Hoang
	Excitation	10 eV - 10 keV Emfietzoglou 10 keV - 1 MeV Born	10 eV - 500 keV Miller Green 500 keV - 100 MeV Born	10 eV - 500 keV Miller Green	1 keV - 400 MeV Miller Green
	Charge change	–	100 eV - 10 MeV Dingfelder	100 eV - 10 MeV Dingfelder	1 keV - 400 MeV Dingfelder
	Ionization	10 eV - 10 keV Emfietzoglou 10 keV - 1 MeV Born	100 eV - 500 keV Rudd 500 keV - 100 MeV Born	100 eV - 100 MeV Rudd	1 keV - 400 MeV Rudd
	Vibrational excitation	2 - 100 eV Michaud et al.	–	–	–
	Disociative attachment	4 - 13 eV Melton	–	–	–

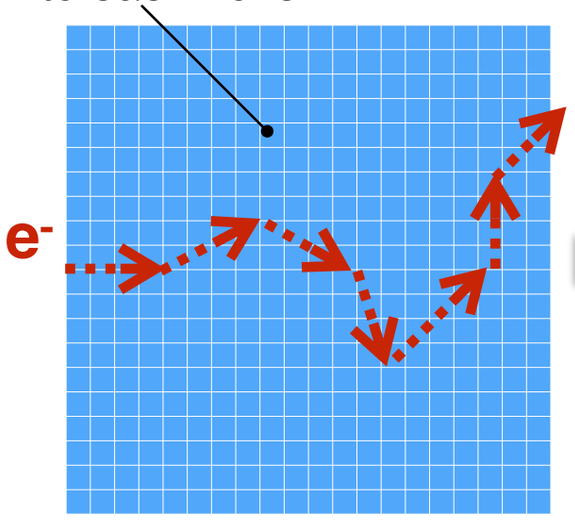
Physics coverage, same as Geant4-DNA

Overall Physics Verification

- Prior to performance check, validate implemented models in CUDA.
  - Incident particles into voxel water phantom: e^-, p, He^{++}
  - All physics processes are turned on in simulation.
  - Score energy deposition in each voxel, and compare depth dose distributions.

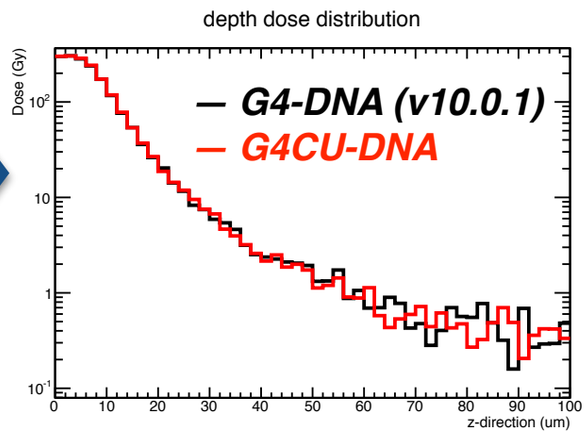
particle	initial energy	phantom size	# of voxel cells (voxel size)
e-	100 keV	100 x 100 x 100 um	50 x 50 x 50 (2 x 2 x 2 um)
p	1 MeV	25 x 25 x 25 um	50 x 50 x 50 (0.5 x 0.5 x 0.5 um)
He++	1 MeV	10 x 10 x 10 um	50 x 50 x 50 (0.2 x 0.2 x 0.2 um)

record energy deposit into each voxel

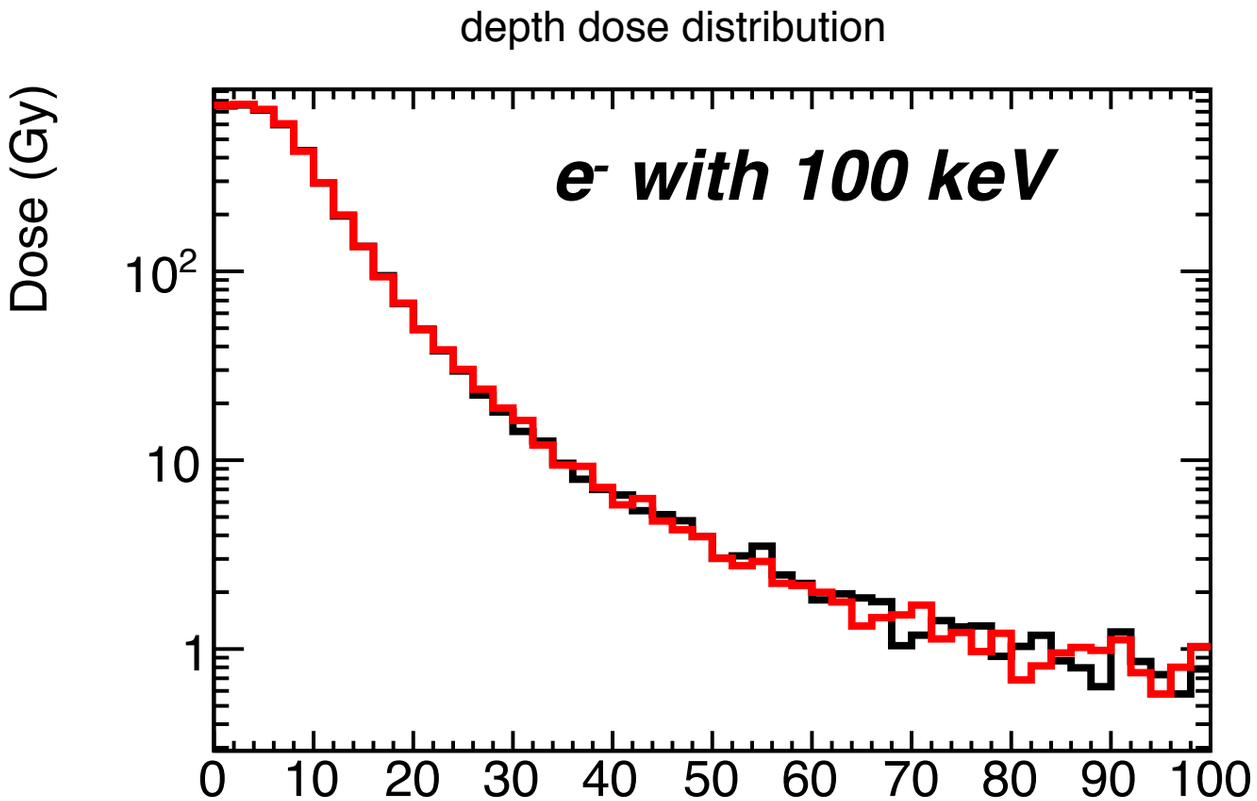


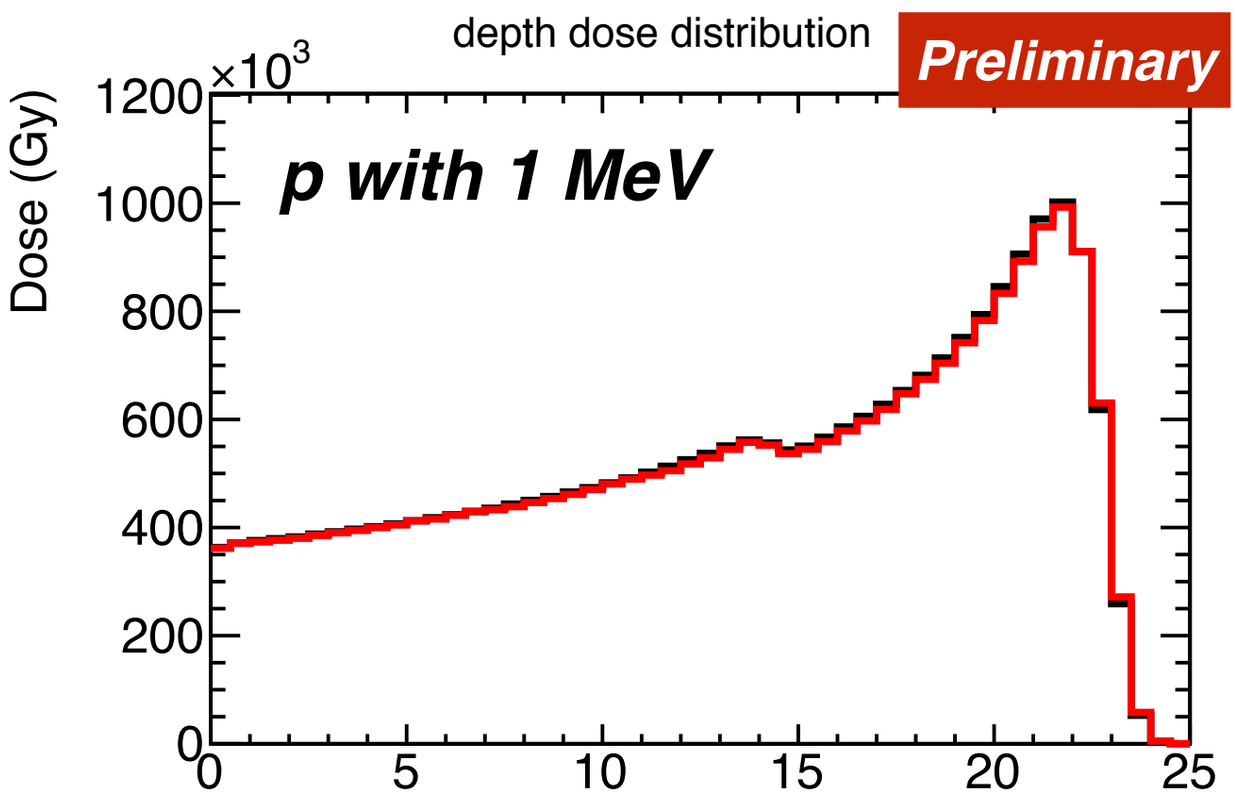
voxelized water phantom

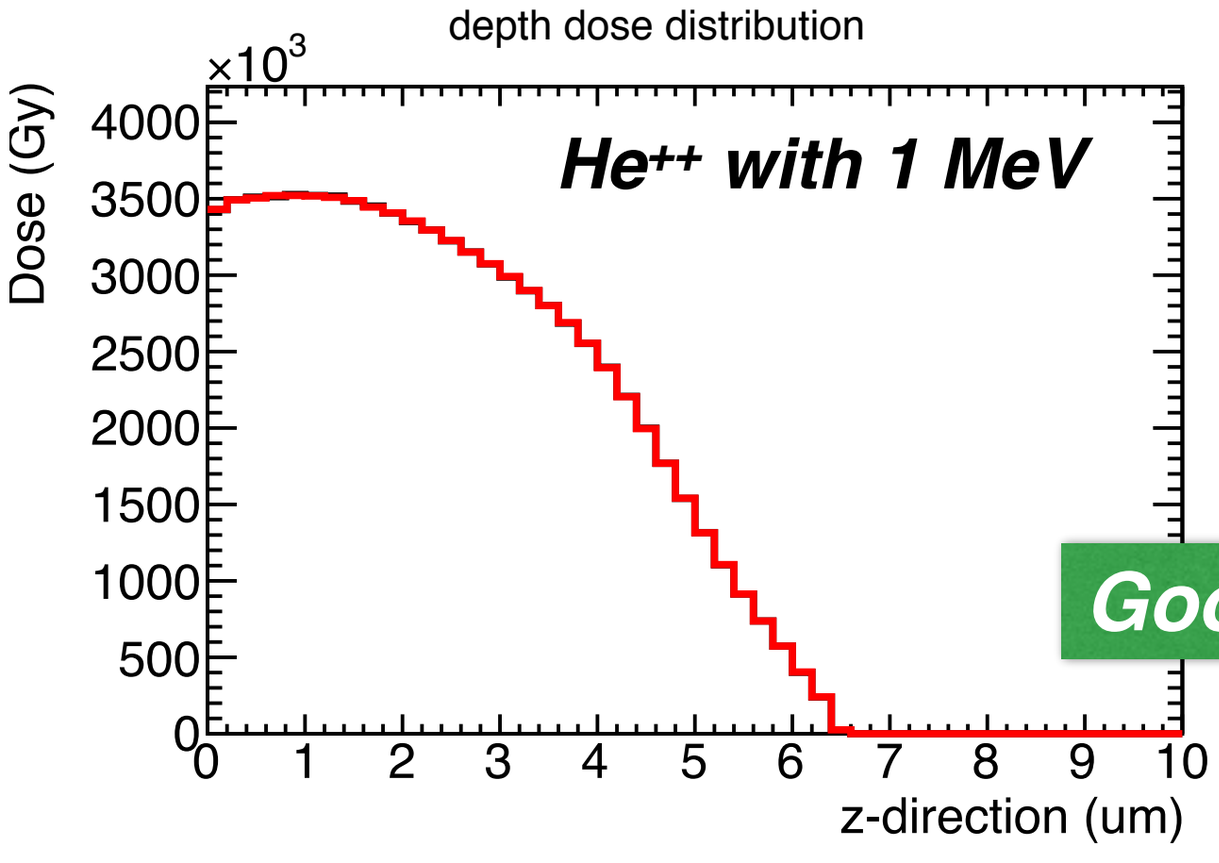
compare depth dose distributions between G4-DNA and G4CU-DNA.



Depth Dose Curves







— ***Geant4-DNA (v10.2.b1)***
— ***G4CU-DNA***

- x-axis: Depth (um)
- y-axis: Dose (Gy)

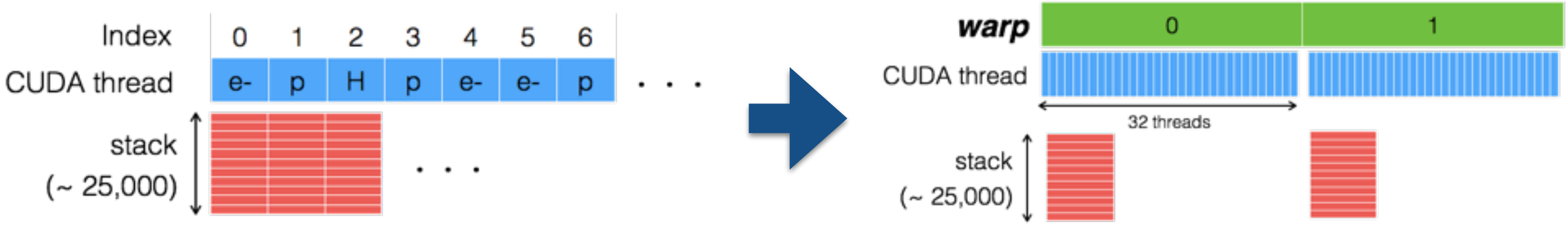
Good agreement with Geant4-DNA

Before Performance Optimization

	Incident particle	Initial energy	Typical # of secondaries generated	Total active CUDA thread numbers (N_{blk} x $N_{thr/blk}$)	Speedup factor (= G4/G4CU)
DNA Physics	e^-	100 keV	2,058	32,768 (256 x 128)	***72.9***
	p	1 MeV	17,784	4,096 (64 x 64)	***19.4***
	He^{++}	1 MeV	20,279	4,096 (64 x 64)	***18.6***
Standard EM Physics	e^-	20 MeV	< 40	524,288 (4,096 x 128)	***< 250***

Large Gap!!

- The large performance gap existed due to the difference of active thread number.
  - Active thread number is limited by GPU memory usage for stacks.
    - DNA physics processes "do not" apply production threshold.
    - Memory usage of stack per one thread significantly increases.
      -> *Many CUDA threads are vacant.*
- ***Stack assignment to CUDA threads has been changed.***
  - Reducing memory usage, and increasing active thread number.



“Current” Computational Performance

- Currently, G4CU-DNA achieves ***~ 200 times*** speedup against single CPU core simulation.
  - Comparable with Standard EM Physics simulation by G4CU.
  - ex.) Process time of shooting ~16k protons with 1 MeV into water phantom
    - ~ 53 hours (single CPU core)
    - ***~ 16 minitues (GPU)***

Preliminary

	Incident particle	Initial energy	G4CU-DNA		Geant4-DNA	speedup factor (=G4/G4CU)
			Total thread numbers (N_{blk} x $N_{thr/blk}$)	process time (sec/particle)	process time (sec/particle)	
DNA Physics	e^-	100 keV	524,288 (4,096 x 128)	3.53×10^{-3}	0.764	***216***
	p	1 MeV	524,288 (4,096 x 128)	5.97×10^{-2}	11.8	***198***
	He^{++}	1 MeV	524,288 (4,096 x 128)	6.10×10^{-2}	12.3	***202***
Standard EM Physics	e^-	20 MeV	524,288 (4,096 x 128)	8.81×10^{-6}	1.84×10^{-3}	***208***

- GPU (NVIDIA, Tesla K20c, 2496 cores, 706 MHz)
- CPU single core (Intel, Xeon E5-2643 v2, 3.50 GHz)

Summary

- Geant4-DNA physics processes / models are implemented in CUDA, and simulations for biological effects of radiation can be run on GPU.

- Currently, G4CU-DNA achieves ***~ 200 times*** speedup against Geant4-DNA with single CPU core.

- ***Drastically improving computational speed of radiation simulation at DNA scale will bring a breakthrough in radiological science!!***