Fast dose recalculation for proton beam therapy using GPU

Joakim da Silva, Richard Ansorge, Rajesh Jena
Cavendish Laboratory and Department of Oncology, University of Cambridge
Scanned proton beam therapy

![Diagram showing dose distribution and Bragg peak for different proton energies and tissue sparing regions](image)

- **18 MeV photons**
- **118 - 150 MeV protons**
- **150 MeV protons**

**Depth [mm]:**
- **Proximal tissue sparing**
- **Tumour**
- **Distal tissue sparing**

**Dose [%]:**

- First magnet horizontal scanning
- Second magnet vertical scanning
- Last layer minimum energy
- First layer maximum energy
Fast dose calculation

- Adaptive treatment (ART) necessary to take advantage of better dose conformity offered by proton therapy
- Aim: on-line 4D treatment
  - Recalculate dose between energy layers/spills
  - Map dose back to reference CT
  - Compare with expected dose and act if different
- Requires fast dose calculation
- No more exponential increase in single-core performance → need to parallelise dose calculation algorithms
Graphic processing units (GPUs)

- “Graphics card”: dedicated hardware for rendering 3D graphics in real-time
GPU considerations

- Slow CPU-GPU transfers $\rightarrow$ minimise traffic
- Global memory access costly $\rightarrow$ coalesce accesses
- Hide memory latency $\rightarrow$ oversubscribe GPU $\rightarrow$ limit use of resources (trade-off)
- Lock step execution $\rightarrow$ avoid branching
- Race conditions $\rightarrow$ use atomics and synchronisation, but only where necessary
- Use built-in hardware functionality where possible
- Single-precision faster than double-precision arithmetic
For parallelisable algorithms that fits the GPU architecture, a speedup of up to 20x compared to a multi-threaded and optimised single-CPU implementation can be expected.

Many algorithms are not a perfect match for the GPU and might see a smaller speedup, or none at all.

Efficient GPU implementations tend to become bandwidth limited.
Pencil beam algorithm
Initial implementation approach

- Run all steps on GPU
  - Limit CPU-GPU transfers by keeping intermediates on the GPU
- Perform calculations in beam’s eye view (BEV) system
  - Coalesced memory access
  - Fast kernel superposition (widening) step
- Calculate dose per energy layer rather than per spot
  - Not suitable for optimisation but fast and useful for on-line calculation between delivery of energy layers
- Employ single-Gaussian beam model, cut off at $3\sigma$
  - Ignoring contributions from nuclear interactions and large-angle scattering
Implementation overview

Start
- Load CT and plan data
- Allocate GPU memory CT, dose, plan data
- Copy data to GPU and bind to textures
- Copy dose grid back to CPU

Finish

All beams?
- Yes: Allocate GPU memory BEV intermediates
- No: All energies?
  - Yes: Calculate scaled IDDs and CPB widths
  - No: All radii?
    - Yes: Find tile cut-off radii and batch tiles
    - No: Perform KS

Ray trace to find WEPL and density
Convolve entry fluence maps (all energies)
Ray trace to add BEV dose to grid

GPU
Water tank simulations

- Difference between GPU pencil beam and Fluka MC results in percent of max dose
- Contours represent MC iso-dose lines
- Single-Gaussian model underestimates low-dose halo, compensated by excess dose to central axis

Skull base test case

- Two beam directions, 38+45 energy layers, 6776 spots in total
- 2%/2 mm $\gamma$-index compared to Fluka MC
- Passing rate: 96.7\% (voxels receiving $>$10\% of max dose)
- Corresponding value for Syngo: 96.8\%

Timings

- Benchmarking on Tesla K40 GPU donated by Nvidia
  - 2880 cores @ 875 MHz
- Resolution
  - 2x2x2 mm$^3$ dose grid, 1x1x1 mm$^3$ computational pencil beams (CPBs)
- Total calculation time: 0.22 s
- Individual energy layers: 2.2–6.4 ms
- 13 times faster than previously presented (partial) GPU implementation in cubic target test case¹
- Fast enough for on-line dose calculation!

Low-dose halo
Low-dose halo model

- Use double-Gaussian beam model
- Problem: halo Gaussian about three times wider $\rightarrow$ nine times longer calculation time
- Solution\(^1\): use only one CPB per spot for the halo
- Rough numbers:
  - Inter-spot distance $\approx 3\text{mm}$
  - $\sigma_{\text{halo}} \approx 3\sigma_{\text{primary}}$
  - Additional relative calculation time: $3^2/3^2 = 1$
  - $1/3^2$ fewer CPBs
- Estimate: including nuclear interactions calculation would increase the calculation time by 10–20%

Preliminary results

- Clear improvement in single spot dose:

- Patient case calculation time up by ~15% to 0.26 seconds
- Similar increase in energy layer calculation time
- 2%/2 mm $\gamma$-index passing rate: 97.8%
The pencil beam algorithm is well suited for parallel implementation

By properly adapting each step of the algorithm to the GPU considerably better performance than for a partial implementation can be achieved

A double-Gaussian beam model can be implemented without prohibitively prolonging the calculation time

The calculation times presented are short enough to enable on-line dose calculation between energy layers on a single workstation
Acknowledgements

- Mario Ciocca, Giuseppe Magro, Andrea Mairani & Silvia Molinelli @ CNAO
- Till Böhlen @ MedAustron
- Andrea Attili & Simona Giordanengo @ INFN
- European Commission FP7 ENTERVISION project

Thank you

- Questions / Feedback?