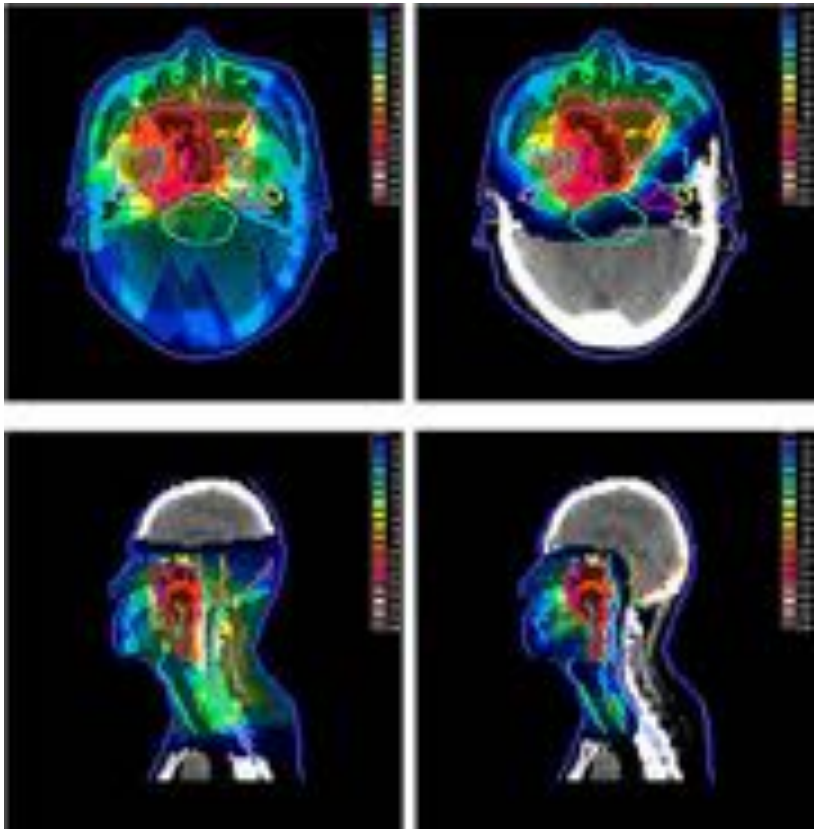
Using Parallel Monte Carlo to Validate Radiotherapy Calculations

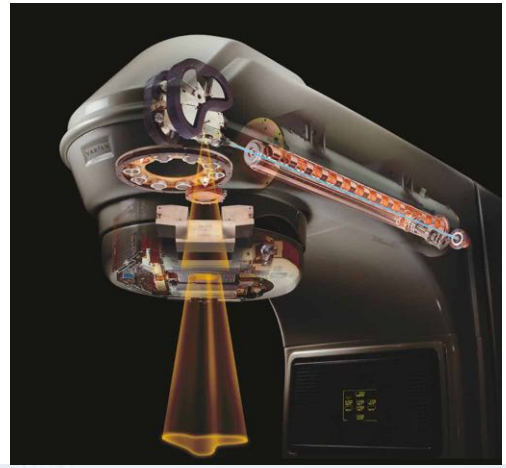
Hywel Owen
Andrew Green, Ran Mackay, Adam Aitkenhead

University of Manchester
Christie Hospital
Cockcroft Institute for Accelerator Science and Technology

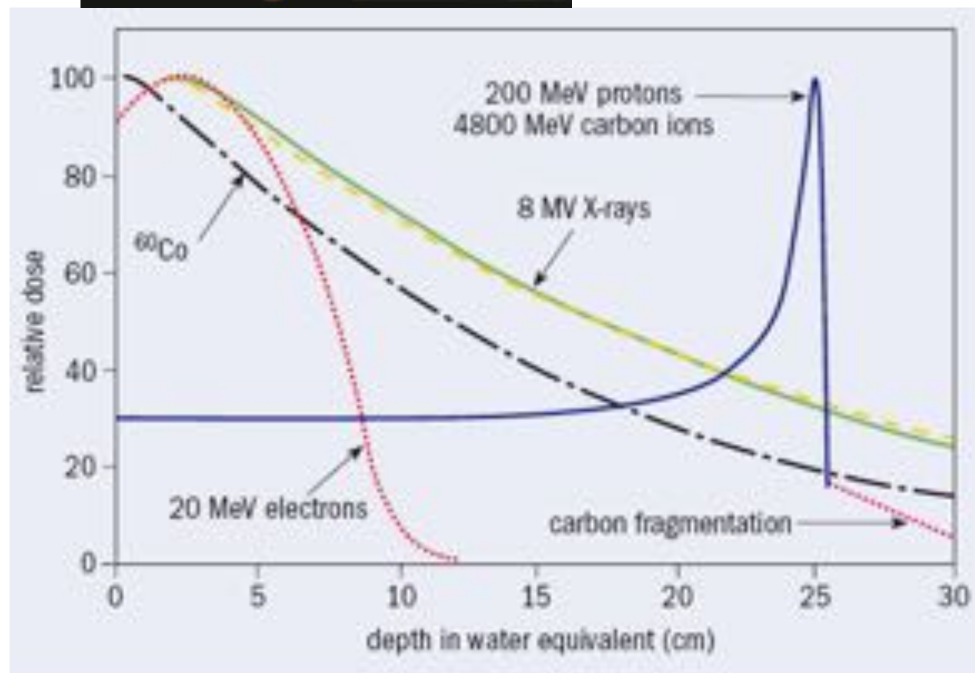
EMiT Emerging Technology Conference 30 June/1 July 2015

Protons vs X-Rays





IMRT
Intensity
Modulated
Radiotherapy

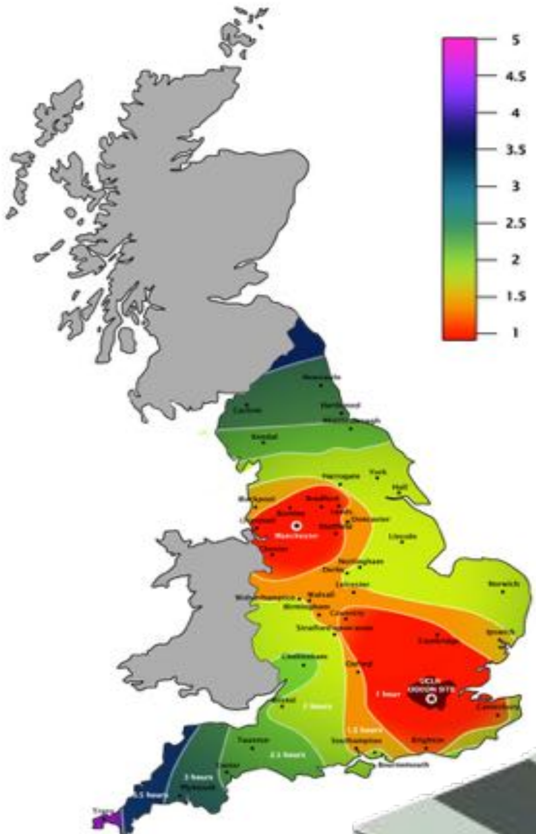


$$-\frac{dE}{dx} = \frac{4\pi}{m_e c^2} \cdot \frac{n z^2}{\beta^2} \cdot \left(\frac{e^2}{4\pi\varepsilon_0}\right)^2 \cdot \left[\ln\left(\frac{2 m_e c^2 \beta^2}{I \cdot (1-\beta^2)}\right) - \beta^2\right]$$

The Cockcroft Institute
of Accelerator Science and Technology

MANCHESTER
1824
The University of Manchester

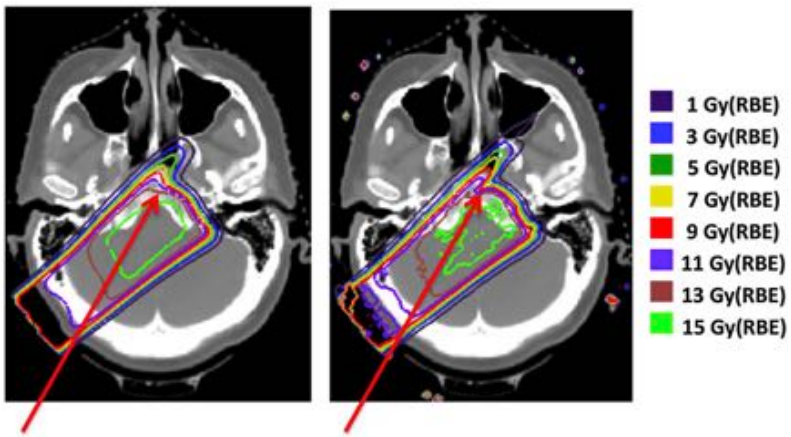
These centres are big...



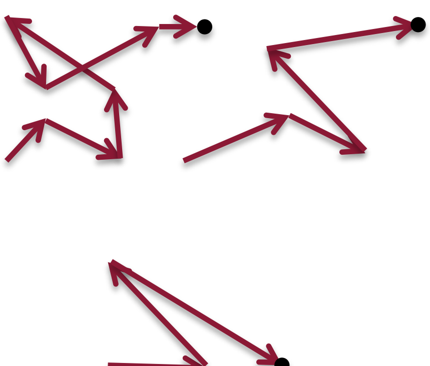
£125M each
750 patients per year
(~30 treatments per patient)
£25k per patient
18 hours running per day
Running from 2018 onwards

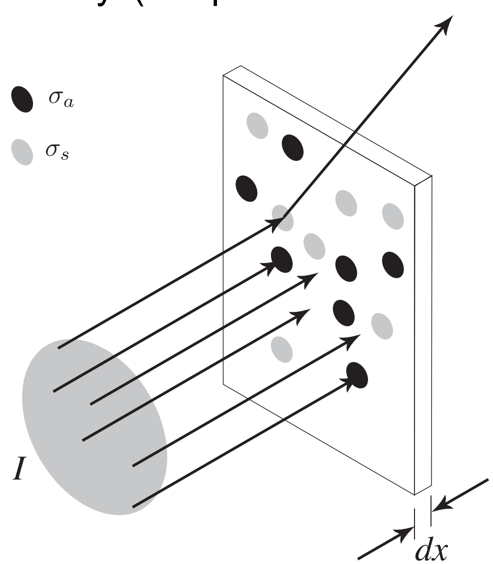
UCLH Centre

The Role of Monte Carlo in Proton Therapy



Pencil beam algorithm (left) vs Monte Carlo (right); arrow indicates a range error due to MCS in heterogenous boundary. (Adapted from Paganetti (2008))



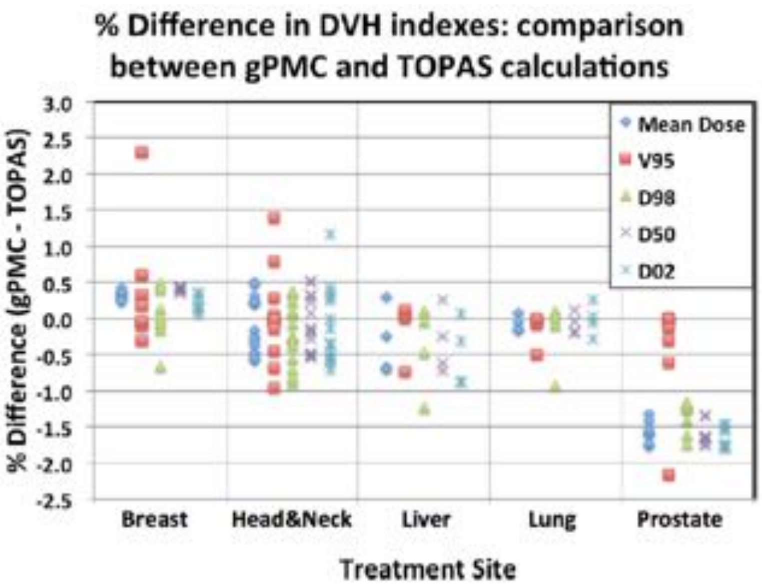


- Algorithms used in planning tools can have drawbacks.
- Monte Carlo is too slow to be used directly for planning (maybe not, see later…)
- Use Monte Carlo to validate plans before delivery
- Must fit planning workflow:
  - Import patient CT file as a geometry of voxels.
    - Usually a few million, more in head and neck
    - c. 40 hours on single CPU
  - Track protons through the voxels recording where the energy goes.
    - Must simulate enough protons to get good uncertainty (< 1%).
    - Must be confident in the physics implementation.
    - Needs to fit within a clinical workflow – minutes not hours.

gPMC – Simplified Physics using GPU

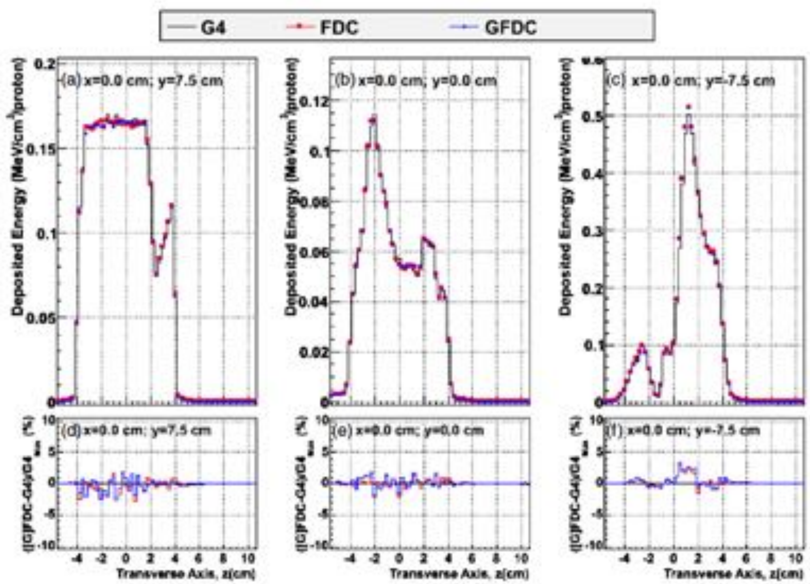
- Image grid 1 x 1 x 1.5 mm
- Dose grid 2 x 2 x 2.5 mm
- Simulation rate
  - gPMC: ~2.6s/MP
  - TOPAS: 4h/MP
- Simulation time
  - 50-500 MP (histories)
  - gPMC: ~130s to 1300s
- Random uncertainties (depends on site):
  - gPMC: 0.5-2.4%
  - TOPAS: 1-2%
- Dose difference gPMC/TOPAS:
  - Around 1% on D98,D50,D02
- This problem is in the simulation of nuclear interactions, exacerbated by the higher energies required in prostate treatment

- *gPMC can already calculate good accuracy in a few minutes!*



Giantsoudi *et. al.* PMB60, 2257 (2015)

Track Repeating



Dose distribution projected into each plane. Differences between Geant4 and the GPU track repeating implementation are around 2 %

- Create a database with 1 million proton histories.
- Stores step length, angle relative to previous step, energy lost and energy deposited for a 251 MeV beam in water and 41 other biological materials.
- This data is then scaled for different energies.
- Validated against Geant4 – accurate to ~2%
- 5.4s/MP (on older hardware)

What about Intel?

- XeonPhi MIC – Many Integrated Core
- 61 cores/244 threads (today)
- Only 16GB on-card RAM (today)
- Challenging target of <67MB/thread.
- BUT: x86 architecture
  - Same source code
  - Same compilers
  - Same development cycle
  - Less debugging!
  - Keep all physics!

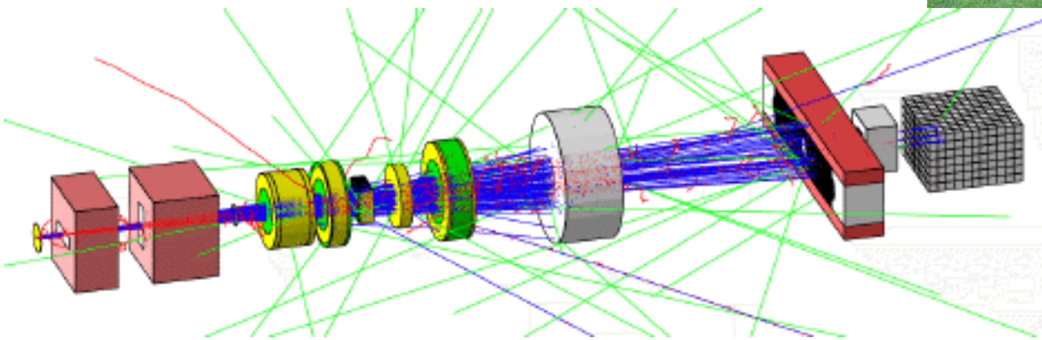
Model	3120A	7120P
# Cores (threads)	57 (228)	61 (224)
RAM	6 GB	16 GB
Clock Speed	1.100 GHz	1.238 GHz
Cost	$1695-$1960	$4129

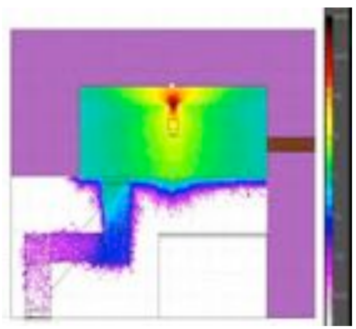
Used Manchester Research IT test system (thanks!)

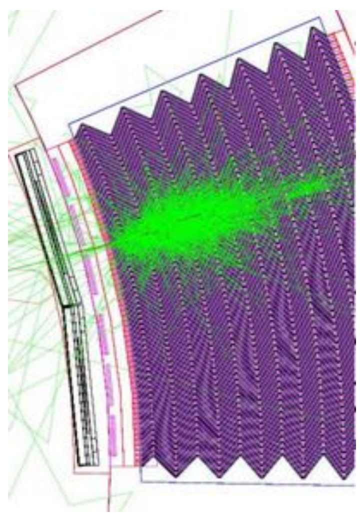
The Cockcroft Institute
of Accelerator Science and Technology

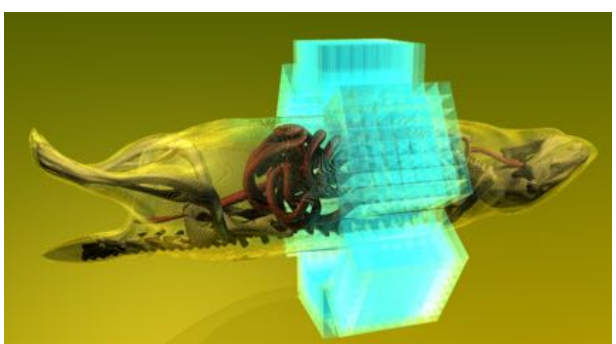
GEANT4MT

MANCHESTER
1824
The University of Manchester









- Well known in HEP, well validated (see e.g. (Yarba, 2012))
- Lots of front ends for Medical Physics (TOPAS, GATE, GAMOS etc)
- Validation in medical applications is good (e.g. (Testa *et. al.* 2013))
- Recent addition of multithreading opens the possibility of running on Xeon Phi
  - Per process memory required is ~100s MB
  - Per thread can be ~10s MB
- Useful to have a common code base on all systems

- We used stock Geant4MTv10

The Cockcroft Institute
of Accelerator Science and Technology

MANCHESTER 1824
The University of Manchester

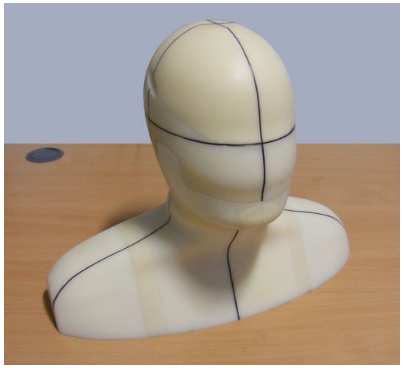
Benchmark Calculation

- Used a radiotherapy phantom (Aitkenhead *et. al.* 2013) to avoid data protection issues
- CT scan image converted to materials 16M voxels
- Image grid 1 x 1 x 2 mm
- Dose grid 2 x 2 x 2 mm

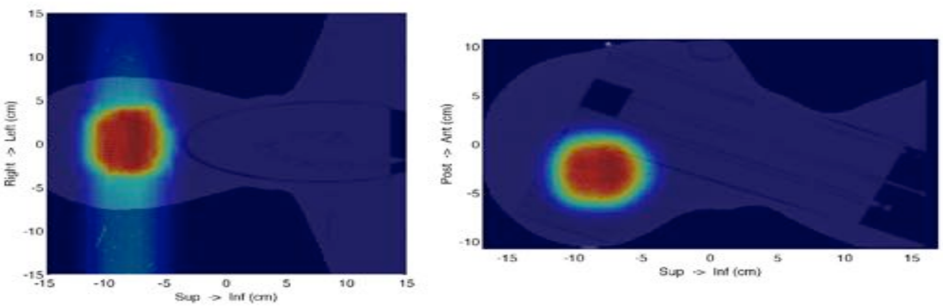
- Simulation time
  - 50-500 MP (histories)
  - gPMC: ~130s to 1300s

- Two beam angles, each with c.1500 spots;
  ~3000 GPS sources

- Simulation of 10MP:
  - ~1% uncertainty in high dose region



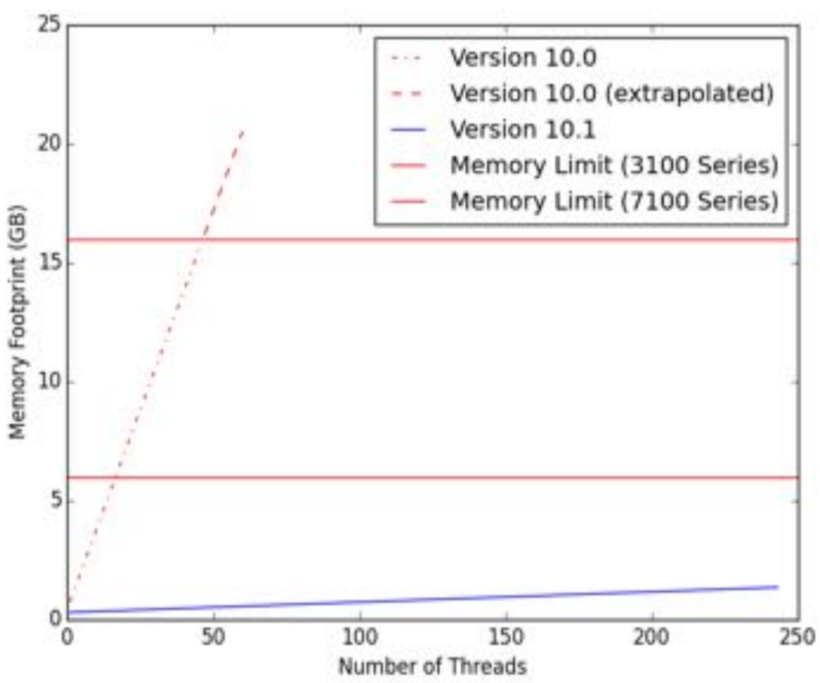
- The Geant4 General Particle Source is the obvious choice for simulating spots:
  - Any energy distribution.
  - Any source shape/size.
  - Any particle.
  - Weighted sampling.
- However, not previously optimized for multithreading – lots of memory required.



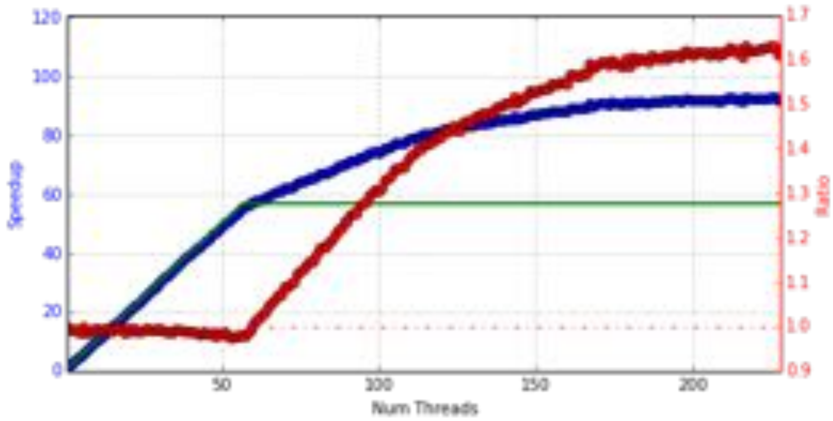
The dose distribution:
50 cm³ sphere roughly 7 cm below surface of head

The Geant4 General Particle Source

- In active scanned proton therapy, need to model many beamlike sources.
  - Simulated by Gaussian distributions in space and energy.
  - The location and shape of the Gaussians define the treatment plan.
- Share data between threads:
  - Things that don't change during a run
  - Source positions, energies, intensities etc.
- Use an approximation to neutron cross sections – negligible impact on accuracy (see Asai *et. al.* 2014)



Calculation Speed



Speedup as a function of number of threads. Ratio shows the speedup over the number of physical cores – it should be 1 until hyper-threading starts.

- Now the simulation can run, how fast does it go?
  - One 7120P card – 1h 54m for 10MP
  - Two 7120P cards – 1h 1m
  - Best performance – 366 s/MP
  - (cf. 2.6s/MP for gMPC)
  - 2 x XeonPhi is $13,500
- Speed Comparisons:
  - AMD Opteron machine (48 threads, 2.6 GHz) gives similar speed at $6,000
  - Single card gPMC about 100x faster

Bottom Line: Lose about 100x speed cf. GPU, but keeps all physics/accuracy

More details: https://epubs.stfc.ac.uk/work/12298866

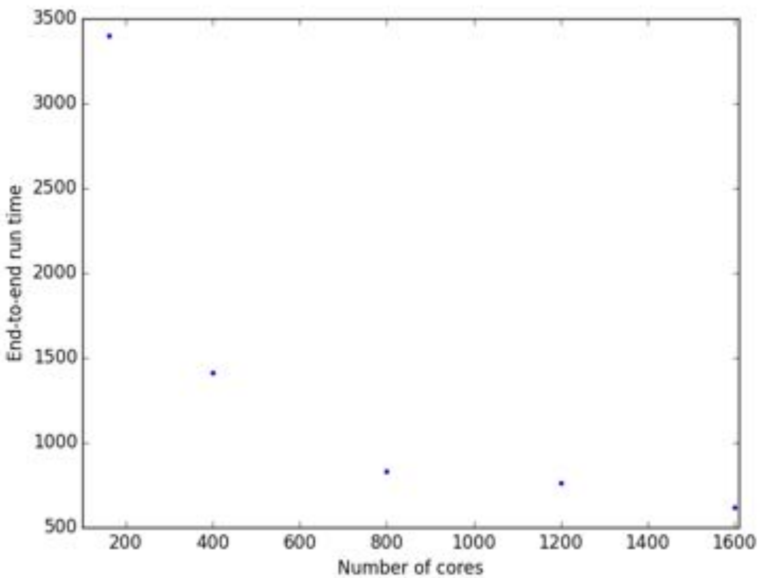
Cloud Computing

- Been done before using Microsoft Azure cloud
- Previous work has looked at cost
  - We also looked at speed

- We used Google Cloud's Google Compute Engine (GCE)
  - Competitive pricing
  - Powerful machines
  - Extremely fast network

- MC codes are almost ideal problem for cloud computing:
  - Independent calculations combined at end

Cloud Computing

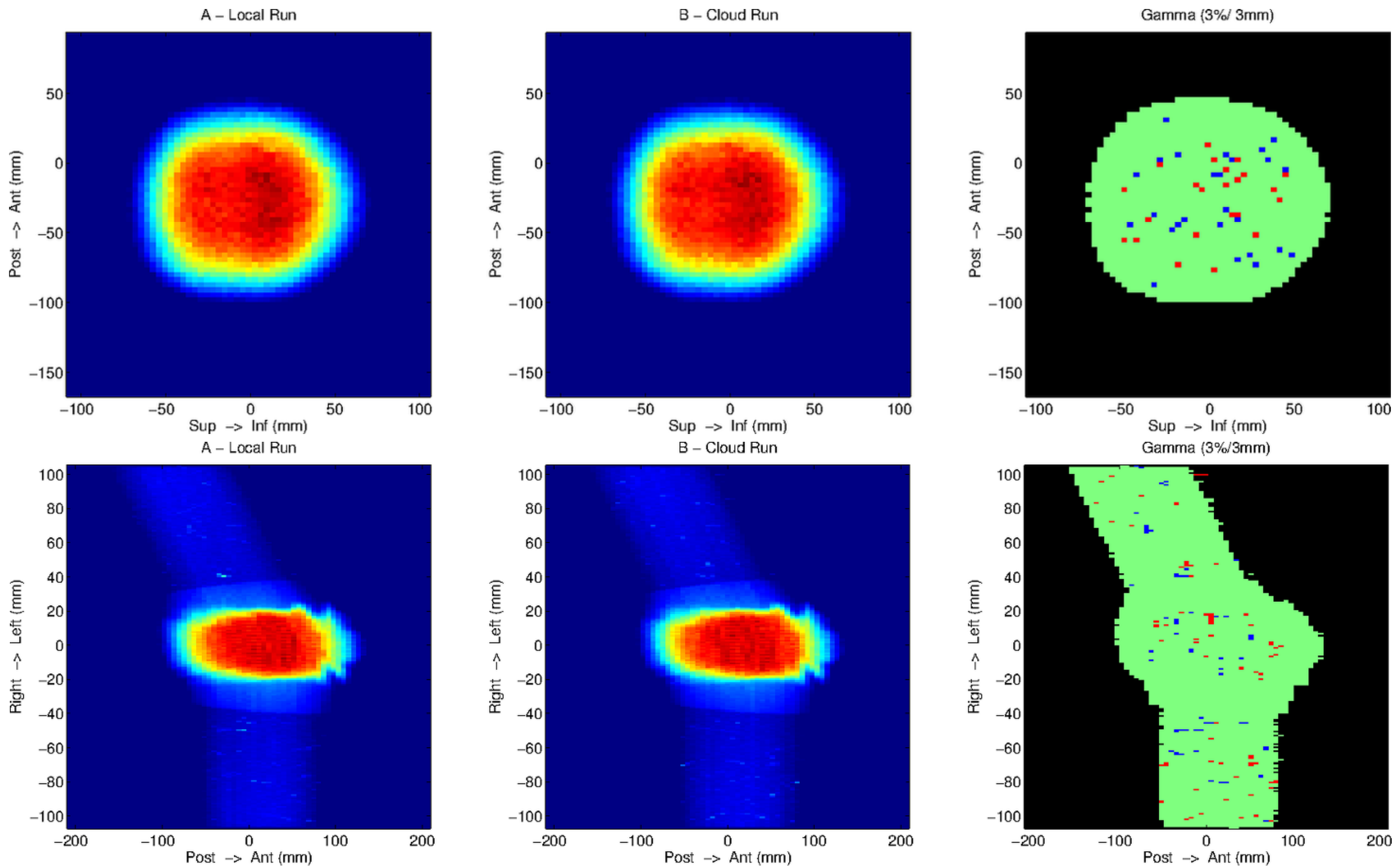


Scaling with number of cloud cores used. Includes calculation of uncertainty and collation and combination of data. Download of the result adds roughly 10 seconds.

- A simple plan requires 10^7 proton primaries
  - 36 hours on a desktop
  - 1 hour on 2x Xeon Phi
  - 10 minutes on GCE
  - Roughly on par with gPMC
- Rent a small supercomputer for 10 minutes
  - 1600 cores
- Scales well, and very cost competitive
  - <<$10 per validation
  - Price dropping fast

More details: https://epubs.stfc.ac.uk/work/12273478

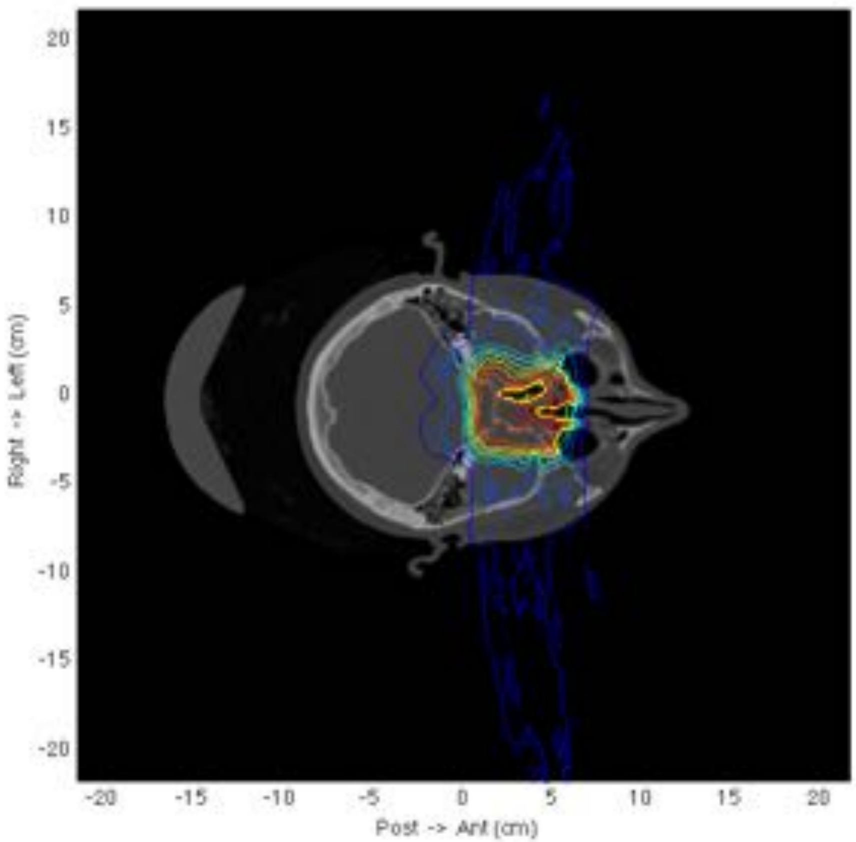
Results – Dose Comparison



Gamma analysis shows identical calculation; limited by expected statistical fluctuations

Cloud Computing in Radiotherapy

- Issues of patient data control
  - Probably surmountable
  - Needs interpretation/change in NHS data policy
- Cloud will get faster + cheaper
- Especially good fit for hospitals
  - No specialist local skills
  - No specialist local infrastructure
  - Continuous upgrading
  - Highly scalable
- A number of groups/ companies are now looking at this



Future Work

- Further work on GCE
  - Keep optimising the launch code to minimise overheads
  - Develop MPI version of the code?
  - Completely general: applicable to any other use of GEANT4

- Develop tailored proton therapy application optimised for Xeon Phi
  - Should be faster, but much less general

The Cockcroft Institute
of Accelerator Science and Technology

MANCHESTER
1824
The University of Manchester

Summary

- Geant4 version 10.1 onwards will be able to run proton therapy treatment plans on Xeon Phi
  - Largely due to our memory reductions in the code, especially the GPS.
- A computer with 2 Xeon Phi cards is roughly equal to a 'normal' cluster machine
  - The cluster machine was not latest generation – expect better performance on Intel E5s
  - Expect the next Xeon Phi generation to be about 3x faster
  - This is with essentially no optimisation other than in memory use
- Algorithm changes in the code, and variance reduction techniques may have big benefits
  - Variance reduction is under-studied
- Computation in the cloud is likely to be a big growth area for hospitals
  - Cheaper to let Google do the hosting and maintenance
  - **As fast as gPMC, but with full (well validated) physics**

The Cockcroft Institute
of Accelerator Science and Technology

The University of Manchester

References

Aitkenhead et. al. 2013, "Marvin: an anatomical phantom for dosimetric evaluation of complex radiotherapy of the head and neck." *Physics in medicine and biology 58*(19), 6915.

Asai et. al. 2014, "Recent developments in Geant4", *Annals of Nuclear Energy*.

Giantsoudi, D et. al. 2015, "Validation of a GPU-based Monte Carlo code (gPMC) for proton radiation therapy: clinical cases study." *Physics in medicine and biology, 60*(6), 2257.

Jan, S. et. al. 2004, "GATE: a simulation toolkit for PET and SPECT." *Physics in medicine and biology, 49*(19) 4543.

Kohno, R. et. al. 2011, "Clinical implementation of a GPU-based simplified Monte Carlo method for a treatment planning system of proton beam therapy." *Physics in medicine and biology 56*(22), N287.

Kohno, R. et. al. 2002, "Simplified Monte Carlo dose calculation for therapeutic proton beams." *Japanese journal of applied physics, 41*(3A), L294

Paganetti, H 2008, "Clinical implementation of full Monte Carlo dose calculation in proton beam therapy." *Physics in medicine and biology, 53*(17), 4825.

Perl, J et. al. 2012, "TOPAS: an innovative Monte Carlo platform for research and clinical applications." *Medical Physics, 39*(11) 6818-6837.

Testa, M. et. al. 2013 "Experimental validation of the TOPAS Monte Carlo system for passive scattering proton therapy." *Medical physics, 40*(12) 121719.

Yarba, J. 2012, "Recent developments and validation of Geant4 hadronic physics." In *Journal of Physics: Conference series* (Vol. 396, No. 2, p 022060).

Yepes, P et. al. 2010, "A GPU implementation of a track-repeating algorithm for proton radiotherapy dose calculations." *Physics in medicine and biology, 55*(23), 7107.

Usolids library - http://aidasoft.web.cern.ch/USolids