GPU Accelerated Real Time kV/MV Dose Computation

Robert Jacques[1], Daniel Smith[2], Erik Tryggestad[1], John Wong[1], Russell Taylor[2], Todd McNutt[1]

[1]School of Medicine, Johns Hopkins University, Baltimore, MD 21231-2410, MD USA
[2]Johns Hopkins University, Baltimore, MD 21218, MD USA

Abstract

We address GPU-accelerated dose computation using superposition/convolution with both a modern dual-source MV model and an analytical kV source model. We achieved a speed-up of 80-142x over a highly optimized CPU implementation. The kV source calculates the analytical fluence at each voxel. The primary MV source supports both focused and rounded MLC leaf ends. The extra-focal MV source is modelled as a discretized, isotropic area source and incorporates leaf height effects. The spectral and attenuation effects of static beam modifiers were integrated into each source's spectral function. The total energy released per unit mass (TERMA) computation used back-projection with optional exact multi-spectral attenuation. Superposition/convolution was implemented using the inverse cumulative-cumulative kernel and exact radiological path ray-tracing with optional kernel tilting. Two superposition variants were implemented and benchmarked. Multi-resolution superposition approximates true, solid angle ray-tracing. Arc superposition increases the relative temporal TERMA sampling, which increases computation efficiency.

Keywords

GPU, Beam Modeling, Radiation Therapy, Superposition/convolution, Dose computation, Small Animal

Introduction

Traditionally, improvements in the speed of treatment planning have been realized by faster hardware. However, instead of getting faster, computers are gaining more cores. Simultaneously, the many-core architectures of graphic processing units (GPUs) have become able to run general purpose algorithms. In order to realize the promised performance gains of this hardware we need new parallel algorithms. We address the comprehensive conversion of radiation therapy dose computation, from fluence generation to dose deposition, to the graphics processing unit (GPU) for both clinical MV linear accelerators (linac) and a kV small animal research radiation system.

Fast, accurate dose computation is important to radiation therapy planning as an estimation of the dose delivered to a patient. It is a major bottleneck for the inverse planning of intensity modulated radiation therapy (IMRT) and, more recently, volumetric modulated radiation therapy [1] (VMAT) and adaptive radiation therapy [2] (ART).

Furthermore, the need in radiation therapy research for a modern bench top model has resulted in kV small animal treatment machines. [3] These devices require fine resolution, very large volumes and customized source models, all packaged in a simple cost-effective solution. Customization of an existing clinical treatment planning system resulted in a slow, inaccurate and expensive solution.

Material and methods

Dose computation consists of two parts: a source model and a transport model. The source model computes the incident fluence exiting the head of the accelerator and the transport model the resultant dose deposition. The three main transport algorithms in order of increasing accuracy/decreasing performance are pencil beam, superposition/convolution and Monte Carlo. Superposition/convolution is the current clinical standard. [4]

We have implemented our algorithm using a combination of NVIDA's Compute Unified Device Architecture (CUDA) software development environment for GPU routines and the Digital Mars' D programming language for CPU setup and analysis routines.

Source Modeling

The incident fluence is divided up into separate spectral and intensity components. We've integrated the spectral component into the TERMA computation as a discretized radial (MV without wedge) or Cartesian (kV and MV with wedge) spectral function. This allows for a great deal of modeling flexibility; the kV spectra were set using Monte Carlo simulations, while the MV spectra were set using an analytical off-axis softening function. We support up to 21 spectral bins.

MV Source Model

For the primary source intensity, we model the jaws using transmission factors and precisely calculate the

MLC attenuation using the ray path-length and material properties. Both focused and rounded leaf ends are allowed. We allowed the rounded leaf ends to be vertically offset to better model the leaves of certain linear accelerators. We use a standard Gaussian source-size blur. Back-projection TERMA methods require the integration of the incident fluence with the projection of each voxel. We implement by approximating the projection as a Gaussian function; a standard computer graphics approximation.

Our back-projected TERMA algorithm samples the incident fluence at a single, linearly interpolated point. As each voxel is physically exposed to multiple fluence pixels, we add a Gaussian voxel-size blur so that the back-projected TERMA algorithm may use point sampling.

We have chosen to use only one extra-focal source based on prior experiments [5], though the system is capable of more. For the extra-focal intensity, we use an isotropic, discretized area source model with no MLC or jaw transmission. We used a sum-area-table [6] instead of analytical integrals to compute the visible source area, allowing MLC leaf height effects to be taken into account. The integration is performed by walking the MLC projection boundary in two parts; first below the calculation point, then above. Accuracy is maintained by computing the lower area first, pairing additions and subtractions, limiting the sum area table in extent to the non-zero source values, limiting the sum area table size to 256x256 pixels and using separate field and control point accumulators. Experiments with double precision implementations and open fields computations found the truncation error to be under the machine epsilon.

kV Source Model

We found that a traditional primary or extra-focal source model to be insufficient for modeling the small animal collimators, which were both very small and close to the isocenter. Instead, we use an analytical source model to determine the fluence exposed to each voxel. The source was modeled as a uniform, rectangular area source, whose size was determined using pin-hole measurements. The fluence was found from the area of intersection between the source and the collimator's top and bottom projected back to the source plane. The collimator transmission was negligible and was not modeled.

Superposition/Convolution

Superposition/convolution [7,8,9] consists of two parts: First, fluence is transported through the patient to compute the Total Energy Released per unit Mass (TERMA) in the volume. Then superposition spreads the TERMA by a dose deposition kernel to determine the final dose at each location. To allow the dose deposition kernel to scale realistically with tissue inhomogeneities, the radiological distance is used, differentiating superposition from convolution.

Total Energy Released per unit Mass (TERMA)

The TERMA for photon energy E at point r',

$$T_E(r') = \frac{\mu_E(r')}{\rho(r')}\Psi_E(r')$$

is defined as the energy's fluence, Ψ, from source s weighted by the density relative to water, ρ, and the linear attenuation constant, μ_E, at point r':

$$\Psi_E(r') = \frac{\Psi_{E,s}(r')}{\|r'-s\|^2} e^{\int_s^{r'} -\mu_E(t)\,dt}$$

Although, traditionally a homogeneous approximation of the attenuation is used:

$$A_E(r') = e^{\int_{r'}^{s} -\mu_E(t)\,dt} \cong e^{\frac{-\mu_E(r')}{\rho(r')}\int_{r'}^{s}\rho(t)\,dt}$$

Both back-projection [10] and forward-projection were used to compute TERMA. Forward-projection used ray divergence to identify sets of rays to run in parallel. This forward method exhibited discretization artifacts, had serialization overhead, had a 25-50% memory efficiency and was $O(n^3)$. The back-projection method was artifact free, overhead free, $O(n^4)$ and highly cache efficient; the performance of back-projection exceeded forward-projection.

The back-projection method also anables the storing of an attenuation volume, which increases the performance of multiple MV planning operations, such as when only an MLC or jaw position is changed. However, this strategy was harmful for kV planning; the fixed collimator of the kV system, combined with the ignoring collimator transmission, allows the number of computed attenuation voxels to be drastically reduced, thereby increasing the performance of field specific attenuation.

Superposition

Superposition spreads the TERMA at point r' by a dose-deposition kernel, K, [11,12] to determine the final dose, D, at a point r. The kernel is indexed by angle, ω, and radiological distance.

$$D(r) = \iiint \sum_E \frac{T_E(r')}{\|r-r'\|^2} K_E\left(\int_r^{r'} \rho(t)\,dt, \omega(s,r,r')\right) dr'$$

However, typically only the voxels along a finite set of rays, v, are used in the poly-energetic approximation:

$$D(r) \cong \sum_v \oint T(r+tv)\, K\left(\int_r^{r+tv} \rho(t')\,dt', \omega(v)\right) dt$$

The GPU implementation allows these rays to be tilted to match the primary ray axis [13] with minimal computational cost, whereas CPU implementation orient the rays with the beam axis for speed. Direct use of the kernel is numerically unstable; [14] instead the cumulative (CK) [15] or cumulative-cumulative kernel (CCK) [14] are used. We use the CCK for its greater accuracy at coarser resolutions; halving the resolution gives 16x the performance.

Multi-resolution superposition [10] approximates true, solid angle ray-tracing by increasing voxel size, and therefore ray width, with geometric distance. This has the additional advantages of changing the complexity from $O(n^4)$ to $O(n^3 \log n)$ and of reducing the star

artifacts generated by the relatively coarse azimuth sampling of the kernel. However, larger step sizes reduce kernel accuracy, resulting in a systematic under-dosage with MV beams. kV kernels, with their steep fall-off, maintained accuracy. We used volumetric mip-maps and limited resolution changes to the coarser resolution's voxel boundaries.

Arc superposition separates the temporal sampling of the TERMA from that of superposition during dynamic arc computations. As the TERMA computation is faster than superposition, the performance for a given eccuracy is increased. Logically, this can be thought of as moving the ~5° of angular error implicit in non-tilted kernels to the linac gantry. The finer fluence sampling also better captures the effects from MLC leaf motion.

Optimizing CUDA performance

CUDA's execution model is a 2D grid of 3D blocks of threads. We found that maximizing the block size increased cache reuse and therefore performance; this effect was substantial enough to warrant a refactoring to less efficient TERMA and tilted superposition routines in order to achieve the maximum block size. The multi-resolution algorithms were the only case where we didn't reach the maximum block size. We found cube-like block sizes increased spatial cohesion and therefore performance. Generally we used a 1:1 mapping of threads to elements in the x and y directions; the z direction was looped over with a stride of the z block size.

Generally, all data was cached in textures. The TERMA volume was transferred to the superposition algorithm using 16-bit floats. This increased performance by ~13% with a truncation error of 1.8×10^{-5}% of D_{max}.

Shared memory, a small per block memory area, was used to store the MLC leaf positions for the extra-focal MV source and the array of mip-map volumes in multi-resolution superposition.

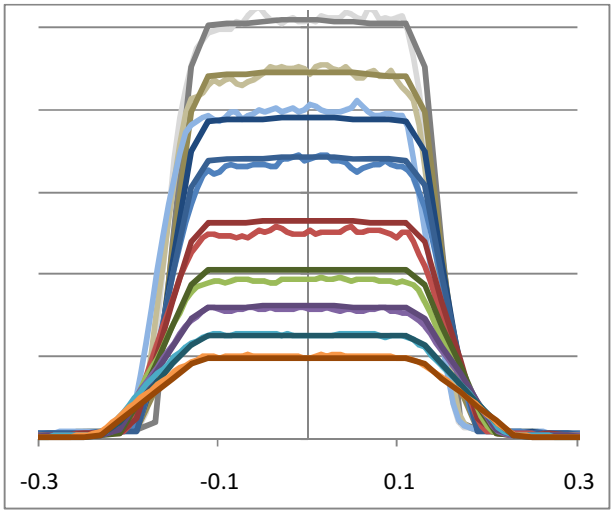

Figure 1: 3mm square small animal collimator measured (light) and computed (dark) X profiles for depths of 5.1, 10.3, 15.5, 20.7, 30.9, 41.1, 51.3, 61.6, 71.7 mm.

Results and discussion

All superposition accuracy results are compared against an ideal superposition computation. The MV source model parameters were based on a commissioned Varian 6EX linac. The kV source model only has a single fitted parameter: the source intensity per second, which was manually fitted to measured data, see figure 1.

All performance results are reported in absolute time and/or absolute speedup, S_p, relative to the optimum serial CPU superposition implementation in Pinnacle[3] (Philips - Madison, WI). Pinnacle[3] times were measured on Opteron 254 (2 cores, 2.8 GHz).

The primary MV source performance was 0.15ms per control point. The extra-focal performance was under 5.3ms (data dependent). Total performance, including blurring was under 8.3ms. Field size was 400^2 pixels.

Performance of our TERMA implementation is listed below in table 1. The back-projection algorithm exhibited an empirical complexity of $O(n^{\sim 3.7})$ which is a slight improvement over its theoretical complexity of $O(n^4)$.

	Volume Size		
Attenuation Method	64^3	128^3	256^3
Forward & Multi.	44 ms	150 ms	1869 ms
Multi-spectral	10.0 ms	117.9 ms	1615.1 ms
Homogeneous	1.5 ms	26.3 ms	359.3 ms
Cached	0.2 ms	1.2 ms	9.8 ms

Table 1: TERMA method performance. Back-projection used except where noted.

The performance of our superposition implementation is listed below in table 2. We found that using kernels with 12 azimuth angles was more accurate than kernels with 8 azimuth angles, even if the zenith or total ray sampling was higher. We found that for kernels with 8 azimuth angles, a tilted kernel with 4 zenith angles had similar high dose and better gradient and low dose accuracy than a non-tilted kernel with 10 zenith angles. We found that the ray divergence in tilted kernels to affect cache performance with large volumes.

				64^3		128^3	
Method	*Type*	*Tilt*	*Rays*	*Time(s)*	S_p	*Time (s)*	S_p
Standard	CCK	✓	10x8	0.160	52x	3.140	30x
Standard	CCK	✓	6x12	0.134	62x	2.714	35x
Standard	CCK	×	10x8	0.121	68x	2.292	41x
Standard	CCK	✓	4x8	0.058	142x	1.186	80x
Multi-Res.	CCK	✓	4x8	0.061	N/A	0.517	N/A
Multi-Res.	CCK	×	4x8	0.053	N/A	0.458	N/A
Pinnacle[3]	CK	×	10x8	8.268	1x	94.508	1x

Table 2: Performance of multiple superposition methods.

We found the CCK to be 4.8% slower than the CK on the GPU; on the CPU, the CCK is 50% slower than the CK. We found performance cost of kernel tilting to be

~29% for standard and ~14% for multi-resolution superposition. Comparatively, CPU kernel tilting costs 300%. [13] However, due to its greater accuracy, kernel tilting results in a net performance gain of ~50%. Multi-resolution superposition was up to 2.6x faster than standard superposition and scaled better: an empirical $O(n^{3.1})$ vs an $O(n^{4.3})$.

Preliminary results using arc superposition are listed below in table 3. We investigated the error in dose deposition half-way between two calculation points, where the angular error is maximal. These experiments indicate a reasonable accuracy can be achieved with as little as 9 superposition calculations, which represents an orders of magnitude performance improvement for arc therapies, such as VMAT.

Method	*Arc*			*Standard*	
Rays	*6x12*	*4x8*	*4x8*	*10x8*	*4x8*
Tilt	✓	✓	✓	×	×
Multi-res.	×	×	✓	×	✓
∠ Error	*High Dose Region*				
0°	0.12%	0.28%	0.99%	0.27%	1.12%
0.5°	0.12%	0.28%	0.99%	0.32%	1.14%
1°	0.12%	0.28%	0.99%	0.40%	1.18%
5°	0.27%	0.31%	1.02%	2.73%	3.18%
10°	0.50%	0.44%	1.12%	7.05%	7.30%
20°	1.07%	0.92%	1.48%	14.05%	14.19%
	Gradient Region ($\|\nabla D\| > 0.3D$)				
0°	0.14%	0.31%	0.88%	0.47%	1.12%
0.5°	0.16%	0.31%	0.88%	1.30%	1.81%
1°	0.18%	0.32%	0.89%	2.16%	2.56%
5°	0.63%	0.61%	1.07%	7.63%	7.69%
10°	1.24%	1.08%	1.42%	11.86%	11.83%
20°	2.31%	2.25%	2.30%	17.08%	17.01%

Table 3: Arc superposition accuracy comparison for an IMRT head and neck patient with different angular samplings. Error reported as average mean absolute error, relative to D_{max}.

Conclusion

We have implemented a GPU accelerated dose engine with near real-time performance based on the superposition/convolution algorithm. We have developed modern, deterministic GPU-accelerated source models for both MV and kV treatment machines. The MV extra-focal fluence model was enhanced with arbitrary fluence profiles and MLC leaf height modeling. The TERMA calculation was enhanced with physically correct multi-spectral attenuation and back-projection. We found several improvements to the superposition algorithm to be substantially more efficient on the GPU than on the CPU, warranting the main stream use of kernel tilting, the cumulative-cumulative kernel and exact radiological path ray-tracing. We explored separating the temporal sampling of the TERMA and superposition computations in the form of arc superposition and found it to increase both performance and accuracy. We have used our kV source model and multi-resolution superposition routines to implement a simple, forward planning tool for small animal radiation research.

References

[1] Otto K. Volumetric modulated arc therapy: IMRT in a single gantry arc. Med. Phys. 2008 January; 35(1): p. 310-317.

[2] Yan D, Vicini F, Wong J, Martinez A. Adaptive radiation therapy. Phys Med Biol. 1997; 42(1): p. 123-132.

[3] Tryggestad E, Armour M, Iordachita I, Verhaegen F, Wong JW. A comprehensive system for dosimetric commissioning and Monte Carlo validation for the small animal radiation research platform. Phys.Med.Biol. 2009 August; 54: p. 5341–5357.

[4] Ahnesjo A, Aspradakis M. Dose calculations for external photon beams in radiotherapy. Phys. Med. Biol. 1999; 44: p. R99–R155.

[5] Yan G, Liu C, Lu B, Palta JR, Li JG. Comparison of analytic source models for head scatter factor calculation and planar dose calculation for IMRT. Phys. Med. Biol. 2008; 53: p. 2051–2067.

[6] Crow F. Summed-area tables for texture mapping. In SIGGRAPH; 1984. p. 207-212.

[7] Mackie TR, Scrimger JW, Battista JJ. A convolution method of calculating dose for15-MV x-rays. Med. Phys. 1985; 12: p. 188-196.

[8] Mackie TR, Ahnesjo A, Dickof P, Snider A. Development of a convolution/superposition method for photon beams. Use of Comp. In Rad. Ther. 1987;: p. 107-110.

[9] Mackie TR, Reckwerdt PJ, McNutt TR, Gehring M, Sanders C. Photon dose computations. Teletherapy: Proceedings of the 1996 AAPM Summer School. 1996.

[10] Jacques R, Taylor R, Wong J, McNutt T. Towards Real-Time Radiation Therapy: GPU Accelerated Superposition/Convolution. In High-Performance MICCAI Workshop; 2008.

[11] Ahnesjo A, Andreo P, Brahme A. Calculation and application of point spread functions. Acta. Oncol. 1987; 26: p. 49-56.

[12] Mackie TR, Bielajew AF, Rogers DWO, Battista JJ. Generation of photon energy deposition kernels using the EGS Monte Carlo code. Phys. Med. Biol. 1988; 33: p. 1–20.

[13] Liu HH, Mackie TR, McCullough EC. Correcting kernel tilting and hardening in convolution/superposition dose calculations for clinical divergent and polychromatic photon beams. Med. Phys. 1997 November; 24(11): p. 1729-1741.

[14] Lu W, Olivera GH, Chen M, Reckwerdt PJ, Mackie TR. Accurate convolution/superposition for multi-resolution dose calculation using cumulative tabulated kernels. Phys. Med. Biol. 2005; 50: p. 655-680.

[15] Ahnesjo A. Collapsed cone convolution of radiant energy for photon dose calculation in heterogeneous media. Med. Phys. 1989 July; 16(4): p. 577-592.