nuFFTW: A Parallel Auto-Tuning Library for Performance Optimization of the nuFFT

Mark Murphy[1], Michal Zarrouk[2], Kurt Keutzer[2], and Michael Lustig[2]
[1]Google, Mountain View, CA, United States, [2]EECS, UC Berkeley, Berkeley, CA, United States

Target audience: Image reconstruction and signal processing engineers.

Purpose: We present a fast, autotuned, Gridding-based non-uniform FFT (nuFFT)[1,2] library with parallel implementions on CPUs and GPUs for reconstructing from non-Cartesian data. Existing nuFFT algorithms lie on a spectrum of precomputation levels ranging between none, partial[3,4] and full, with a corresponding tradeoff between arithmetic throughput and memory usage/transfer rates. The influence of a nuFFT implementation and parameter selection on the resulting runtime is non-trivial. Using the maximum aliasing amplitude (MAA)[5] as an accuracy measure, one can define a space of error-equivalent pairs of kernel-widths and grid oversampling ratios. Our auto-tuning approach empirically selects an optimal implementation per trajectory by searching over algorithms and parameters, and saves it for future reconstructions (i.e. parallel imaging). Auto-tuning has proven effective in a variety of numerical libraries such as the FFTW library for computation of FFTs. We show that the optimal implementation depends also on the target platform and the sampling pattern itself. As exhaustive search is prohibitively expensive in many cases, we propose a simplified heuristic where only runtime of the FFT phase is minimized. Since FFT runtimes don't depend on the gridded trajectory, this heuristic requires only a one-time FFT benchmark during system installation.

Methods and results: Currently two algorithms representing both extremes of the precomputation spectrum are implemented: direct (non-precomputed) convolution and Sparse-Matrix (SpM) -based precomputed convolution. Performance results were measured on a 12-core 2.67 GHz Intel Westmere CPU and Nvidia GTX580 GPU. We use the Kaiser-Bessel kernel[1] with parameters chosen to satisfy a MAA=1e-2. We rely on external optimized and autotuned libraries for SpM (OSKI[6] and CUSparse) and FFT (FFTW and CUFFT). We measured runtimes of direct and SpM-based convolution, and FFT of a 14M-sample Cones trajectory, on CPU and GPU for a range of oversampling ratios (fig. 1). Results show that SpM is generally faster than direct convolution. However, due to the substantially higher throughput of GPUs over CPUs (relative to memory bandwidth), the performance acceleration on GPUs is lower than for CPUs. SpM-based convolution on GPU was infeasible for small grid oversampling ratios which resulted sparse matrices larger than the available memory on the GPU. Best FFT performance is achieved when the grid size factors into small prime numbers. We also auto-tuned the direct convolution-based and SpM-based gridding, as well as FFT-heuristic tuned gridding for Cones trajectories with isotropic 1mm spatial resolution and a range of FOVs (fig. 2,3). Results show that different oversampling ratios are selected for various trajectories. Figure 3 shows that our suggested heuristic achieves near-optimal performance.

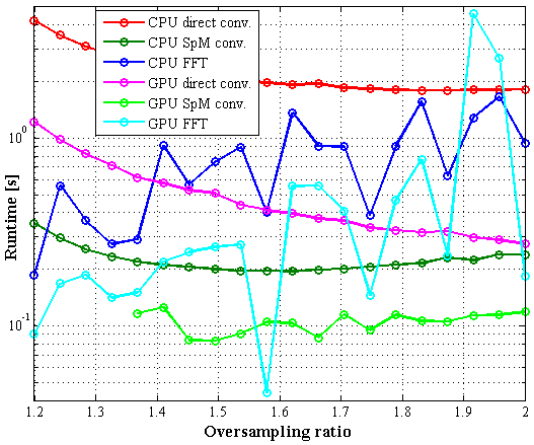

Figure 1: Runtimes of direct convolution, SpM-based convolution and FFT on CPU and GPU for different selections of α.

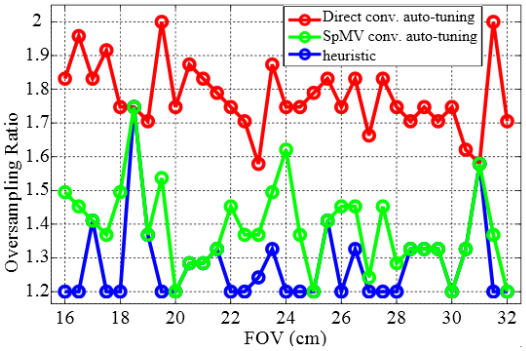

Figure 2: Oversampling ratios selected by auto-tuning the direct convolution-based and SpM-based gridding, and by FFT-heuristic gridding.

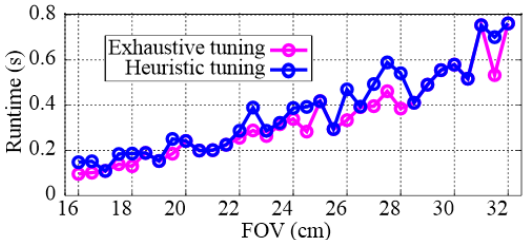

Figure 3: Optimized runtimes of gridding via exhaustive tuning, FFT-heuristic tuning.

Discussion and conclusions: Our results show that the optimal implementation depends not only on the implementation parameters, but also on the underlying architecture, the available memory and the sampled trajectory. Consequently, auto-tuning the nuFFT will speed up MR image reconstruction times. Moreover, the high degree of flexibility afforded by the empirical optimization approach is able to account for micro-architectural changes in future processor generations, for example if calculations become faster than memory transfer rates. Complete parameter tuning is particularly beneficial when a trajectory is to be used many times for image reconstruction (multiple receivers, dynamic imaging), however when offline tuning is impractical our suggested heuristic achieves near-optimal performance.

References: 1. Jackson et al. *IEEE Trans Med Imag.* 1991 Sep;10(3):473–478. 2. M. Murphy. PhD Thesis, EECS, UC Berkeley. 2011. 3. Sorensen et al. *IEEE Trans Med Imag.* 2008 Apr;27(4):538-547. 4. Obeid et al. *Proc ISMRM.* 2011;2547. 5. Beatty et al. *IEEE Trans Med Imag.* 2005 Jun;24(6):799-808. 6. Vuduc et al. *Proc. SciDAC, J. Physics: Conf. Ser.* 2005 Jun;16:521-530.