More IMPATIENT: A Gridding-Accelerated Toeplitz-based Strategy for Non-Cartesian High-Resolution 3D MRI on GPU

Jiading Gai[1], Joseph Lee Holtrop[2], Xiao-Long Wu[3], Fan Lam[3], Maojing Fu[3], Justin P. Haldar[4], Wen-mei W. Hwu[3], Zhi-Pei Liang[3], and Bradley P. Sutton[2]
[1]*Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, IL, United States,* [2]*Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL, United States,* [3]*Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, United States,* [4]*Electrical Engineering, University of Southern California, Los Angeles, CA, United States*

Introduction: We further accelerate the Illinois Massively Parallel Acceleration Toolkit for Image reconstruction with ENhanced Throughput in MRI (IMPATIENT MRI) package to approach clinically-acceptable times while still taking advantage of a variety of advanced image acquisitions and reconstruction techniques. The improved IMPATIENT implemented a faster Toeplitz-based iterative image reconstruction method, whose computation time is further reduced by an optimally tuned, GPU-accelerated gridding implementation. We demonstrate that the Toeplitz code running on a NVIDIA Tesla C1060 (field-corrected, SENSE) can reduce a one-week long, non-Cartesian 3D 1mm^3 high-resolution, whole brain DTI reconstruction (4-channel acquisition) to 4.3 hours.

Method: IMPATIENT MRI provides two strategies (brute force [1] and Toeplitz) to solve the following convex optimization problem:

$$\hat{\boldsymbol{\rho}} = \arg\min_{\boldsymbol{\rho}} \frac{1}{2}\left\|\mathbf{F}\boldsymbol{\rho}-\mathbf{d}\right\|_2^2 + \beta\left\|\mathbf{WC}\boldsymbol{\rho}\right\|_2^2, \quad (1)$$

where $\boldsymbol{\rho}$ is the image to be reconstructed, $\mathbf{F}$ is a matrix that models the acquisition physics, $\mathbf{d}$ is the vector of measured data, $\mathbf{C}$ is a regularization matrix, $\mathbf{W}$ is an optional diagonal weighting matrix, and β is the regularization parameter. Both techniques handle arbitrary non-Cartesian acquisitions, magnetic field inhomogeneity correction [3], multiple coil acquisition and SENSE reconstruction, and regularization (including both Tikhonov and l_1-based regularization strategies). The brute force strategy implements an exact evaluation of the system model, but has a high computational complexity [5]. The Toeplitz strategy takes advantage of the Toeplitz formulation for fast implementation of system matrix computations, and field inhomogeneity correction is performed via time segmentation [3]. In addition, for the Toeplitz-based strategy, we use gridding to speed up the computation of a convolution kernel resulting from the toeplitz structure of $\mathbf{F^H F}$. The operator product $\mathbf{F^H F}$ involves both forward ($\mathbf{F}$) and adjoint operators ($\mathbf{F^H}$) for non-Cartesian trajectories and field inhomogeneity compensation. To solve Eq. (1) by the conjugate gradient (CG) method, two dense matrix-vector multiplications are required by $\mathbf{F^H F}$ every iteration. With a time-segmented field map approximation, the product $\mathbf{F^H F}$ can be reduced to a discrete convolution using the Toeplitz structure as follows:

$$[F^H F\rho]_n = \sum_{k=0}^{L} e^{iw_n(\tau k + t[0])} \sum_{n'=1}^{N} \rho_{n'} \cdot e^{-iw_{n'}(\tau k+t[0])} Q_k(x_n), \text{ with } Q_k(x_n) = \sum_{m=1}^{M} a_k(t_m)\cdot e^{i2\pi k_m x_n}, \quad (2)$$

where $a_k(t_m)$ is the Hanning window interpolator for the kth time segment at time t_m of the mth k-space location, w_n is the field inhomogeneity measured at the nth voxel coordinate, Q_k is the convolution structure that permits the fast evaluation of $\mathbf{F^H F}$ via the FFT [6,7]. Normally, Q_k can be pre-computed before the scan since its computation involves only the k-space trajectory and the image size. However in the diffusion application, motion-induced phase results in random shifts to the trajectory for each shot. Prior to entering the CG step, the Toeplitz reconstruction algorithm also computes the vector $\mathbf{F^H d}$, whose definition is mathematically similar to Q_k. However, Q_k requires considerably more computation time and memory because it oversamples the input space two-fold in each dimension. Therefore, the complexity of the Toeplitz reconstruction far exceeds the complexity of reconstructions based on gridding and the FFT. Hence exact evaluation of Q_k on GPU has been impractical in clinical settings, especially for high-resolution, three-dimensional data. Alternatively, gridding provides an approximation of the exact computation of Q_k's. The GPU-accelerated gridding algorithm [2] in IMPATIENT is an optimized, output-driven algorithm, where every output pixel is computed by a single thread and the input k-space data is shared among all the threads. Input binning is used to ensure the output-driven gridding algorithm run in the same O(N) time as the traditional input-driven approach does on CPU. This is achieved by sorting the k-space points into bins with non-uniform capacity and regular k-space coordinates. Overlapping computations on CPU and GPU is used to further improve the load imbalance caused by varying bin sizes. Our reconstruction workstation is an Intel Xeon X5650 system, with a Tesla C1060 GPGPU card. The GPU optimization includes space compaction, input binning, data regularization, thread coarsening, scatter-to-gather transformation, memory tiling, page-locked memory allocation, common sub-expression elimination.

Results: Table 1 shows the performance comparison. Although the data would normally be amenable to gridding in-plane and a separate FFT across the slice direction, due to motion-induced phase errors in diffusion, the resulting k-space trajectories are truly 3D and change for each data set. Using a Tesla C1060 GPU, we tested three 3D datasets coming from a multi-shot stack of constant density spirals with ~65ms readouts per shot with matrix size in the x and y directions increased by increasing the number of spiral interleaves for each phase encode and the z dimension increased by increasing the number of phase encoding steps. All reconstructions execute 10 conjugate gradient iterations and 8 time segments. The 'Time (min)' row shows the execution timings, comparing Toeplitz-based iterative regularized SENSE reconstructions using single-precision, floating-point arithmetic with a hand-tuned, single-thread Matlab code performing the similar calculation on a CPU. In summary, we observed a speedup of more than 48x in single precision mode. A resulting example image from a high resolution diffusion imaging scan reconstructing with GPU can be seen in Figure 1. For reconstruction of a full 3D multi-slab DTI reconstruction, twenty-eight 240x240x32 volumes must be reconstructed for the whole data set. Access to nodes with multiple GPUs can provide trivial acceleration over the multiple volumes to be reconstructed.

Dataset	128x128x16 (SENSE,4coils)			240x240x32 (SENSE,4coils)			512x512x32 (SENSE,4coils)		
	Matlab	Toeplitz	Speedup	Matlab	Toeplitz	Speedup	Matlab	Toeplitz	Speedup
Time (min)	62.31	1.28	48.68x	380.28	9.30	40.89x	1520 (est.)	83.41	N/A

Discussion: This paper describes the second-generation IMPATIENT MRI reconstruction toolkit. We implemented two reconstruction strategies to provide a choice of the tradeoff between accuracy and speed. Brute force, aiming for accuracy, evaluates the system matrix exhaustively with no approximations. Conversely, Toeplitz emphasizes speed more than accuracy. The large increase in reconstruction speed provided by the Toeplitz strategy enables higher resolution 3D data and more coils. These improvements will enable advances in 3D non-Cartesian sequences, such as cones and stacks of spirals, by reconstructing images in reasonable amounts of time.

References: [1] Wu, *et al. ISMRM* 2011, pp. 4396. [2] Obeid, *et al. ISMRM* 2011, pp. 2546. (2):412-5. [3] Sutton, *et al. IEEE Trans. Med. Imaging*, vol. 22, pp. 178-188, 2003. [4] Fessler, *et al. IEEE Trans. Med. Imaging*, vol. 53, pp. 3393-3402, 2005. [5] Wu, *et al. ISBI* 2011, pp.69-72. [6] Wajer, *et al. ISMRM* 2001, pp. 767. [7] Stone, *et al. J. of Para. and Distr. Comp.*, vol. 68, pp. 1307-1318, 2008.

Acknowledgements: This work was supported by NIH grant 1R21EB009768-01A1.

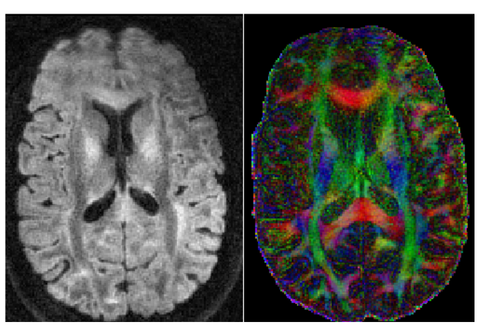

Figure 1: Left: 1mm isotropic diffusion weighted image from a 3D multi-slab multi-shot acquisition reconstructed on a Tesla C1060 GPU using SENSE and field correction. **Right:** An FA map, with color indicating the primary direction of diffusion, resulting from 6 diffusion weighted images that took 4 hrs to reconstruct with GPU compared with one week for reconstruction on a CPU.