Rapid Computation of Sodium Bioscales Using GPU-Accelerated Image Reconstruction

Ian C. Atkinson,[1] Geng Liu,[2] Nady Obeid,[2] Keith R. Thulborn,[1] Wen-mei Hwu[2]

[1] Center for Magnetic Resonance Research, University of Illinois at Chicago, Chicago, IL

[2] Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL

Received 8 August 2012; revised 15 October 2012; accepted 29 October 2012

ABSTRACT: Quantitative sodium magnetic resonance imaging permits noninvasive measurement of the tissue sodium concentration (TSC) bioscale in the brain. Computing the TSC bioscale requires reconstructing and combining multiple datasets acquired with a non-Cartesian acquisition that highly oversamples the center of *k*-space. Even with an optimized implementation of the algorithm to compute TSC, the overall processing time exceeds the time required to collect data from the human subject. Such a mismatch presents a challenge for sustained sodium imaging to avoid a growing data backlog and provide timely results. The most computationally intensive portions of the TSC calculation have been identified and accelerated using a consumer graphics processing unit (GPU) in addition to a conventional central processing unit (CPU). A recently developed data organization technique called Compact Binning was used along with several existing algorithmic techniques to maximize the scalability and performance of these computationally intensive operations. The resulting GPU+CPU TSC bioscale calculation is more than 15 times faster than a CPU-only implementation when processing 256 × 256 × 256 data and 2.4 times faster when processing 128 × 128 × 128 data. This eliminates the possibility of a data backlog for quantitative sodium imaging. The accelerated quantification technique is suitable for general three-dimensional non-Cartesian acquisitions and may enable more sophisticated imaging techniques that acquire even more data to be used for quantitative sodium imaging. Int J Imaging Syst Technol, 23, 29–35, 2013; Published online in Wiley Online Library (wileyonlinelibrary. com). DOI 10.1002/ima.22033

Key words: quantitative sodium magnetic resonance imaging; bioscale; graphics processing unit processing

I. INTRODUCTION

Quantitative sodium magnetic resonance (MR) imaging enables noninvasive measurement of the in vivo tissue sodium concentration (TSC) of the brain that may be relevant to medical decision making in a number of clinical settings where tissue viability is in question (Thulborn et al., 1999; Boada et al., 2005; Lu et al., 2010a). The resultant quantitative map of TSC, termed a bioscale (Atkinson et al., 2010; Lu et al., 2010b), reflects sodium ion homeostasis that is tightly regulated in healthy tissue (Somjen, 2004). Although the clinical potential of ^{23}Na neuroimaging has been recognized for decades (Hilal et al., 1985), high- and ultra-high field MR scanners and optimized ultra-short echo time pulse sequences (Boada et al., 1997; Gurney et al., 2006; Qian et al., 2008; Nagel et al., 2009; Lu et al., 2010) have only recently made quantitative sodium MR imaging of the human brain possible with acceptable acquisition times of 8–10 min. TSC can be combined with a biochemical tissue model to compute the interstitial volume fraction (IVF) or tissue cell fraction (TCF) bioscales (Atkinson et al., accepted for publication). The IVF and TCF bioscales inform on cell packing and cell density which have been shown to be informative in clinical settings of tissue viability in stroke (Thulborn et al., 1999; Boada et al., 2005) and brain tumor response to radiation and chemotherapy (Thulborn et al., accepted for publication; Kline et al., 2000). Current investigations suggest that the TSC, IVF, and TCF bioscales may reflect regional changes in sodium concentration in early neurodegenerative diseases (Mellon et al., 2009; Thulborn et al., 2011).

Computing the TSC, IVF, and TCF bioscales requires processing and combining data from multiple MR acquisitions performed on the subject of interest and a calibration standard of known sodium concentration (Lu et al., 2010a). The correction of known sources of quantification error, including B_0- and B_1-inhomogeneities (Lu et al., 2010a) and interacquisition head movement (Atkinson et al., 2012A), requires repeated image reconstructions to compute an accurate bioscale and so require significant computational time under conventional approaches. For example, when processing resolution-optimized quantitative sodium imaging data into the TSC bioscale, the processing time can be approximately 60 min, greatly exceeding the 30 min required to collect all of the high-resolution human data required for quantification with corrections (Atkinson et al., 2011). Even ignoring the additional time

Correspondence to: Ian C. Atkinson; e-mail: ian@uic.edu

Grant sponsor(s): The authors gratefully acknowledge financial support from the PHS RO1 CA CA1295531. This work was funded in part by the Chicago Biomedical Consortium with support from the Searle Funds at The Chicago Community Trust, the CUDA Center of Excellence at the University of Illinois, and the FCRP Gigascale Systems Research Center.

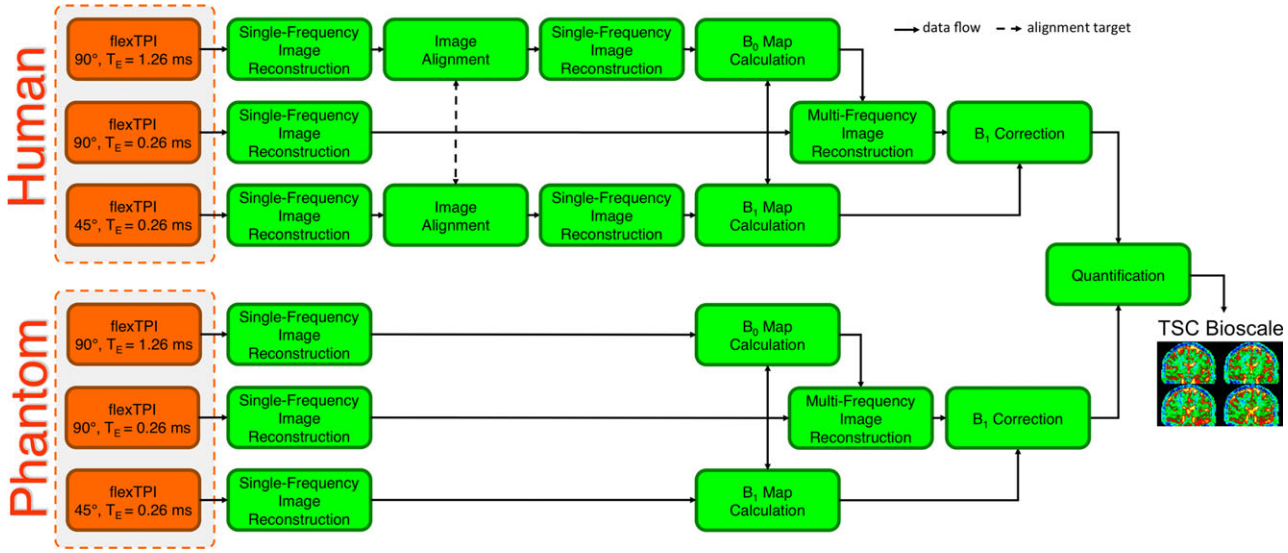

Figure 1. Processing algorithm for computation of the TSC bioscale from ^{23}Na *k*-space data. The orange boxes denote the raw input datasets and the green boxes are processing steps. Eight single-frequency image reconstructions and two multifrequency reconstructions are necessary to produce the final TSC bioscale. A typical multifrequency reconstruction requires approximately 20–30 times more processing than a single-frequency reconstruction. TCD and IVF are computed directly from TSC using basic (noncomputationally intensive) operations. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

required to transfer data from the acquisition system to the processing workstation, the TSC bioscale computational time is longer than that of the data acquisition. Such a mismatch between acquisition and processing times presents a challenge for routine sodium imaging to avoid a backlog of data and provide timely results.

This report describes the results of using commercial consumer-grade graphics processing units (GPUs) to accelerate bioscale computation and greatly reduce the total processing time. GPUs are well suited to this task as several of the algorithms used to process the sodium *k*-space data into the final sodium bioscales are highly parallelizable. GPUs offer a low-cost path to thousands of parallel computing cores. General-purpose computation on GPUs (GPGPU) is now readily accessible with vendor-supported application programming interfaces and consumer GPU-based products dedicated to economically practical high-performance computing. High-performance computing with GPUs is an active research area in many areas of science and medicine. GPUs have previously been used for computationally intensive MR imaging applications including image reconstruction (Stone et al., 2008; Roujol et al., 2009), radio frequency pulse design (Deng et al., 2001), gridding (Schiwietz et al., 2006), image alignment (Huang et al., 2011), and more. This work adds to these results the use of a GPU to accelerate the most computationally intensive parts of computing a sodium bioscale. A primary challenge of GPU-accelerated bioscale calculation is efficiently gridding the highly nonuniform sample density *k*-space data produced by non-Cartesian acquisition schemes suitable for quantitative sodium MR imaging (Boada et al., 1997; Gurney et al., 2006; Qian, 2008; Nagel et al., 2009; Lu et al., 2010a,b). For example, in flexible twisted projection imaging (flexTPI), the sample density in the center of *k*-space can be 1000 times higher than that at the edge of *k*-space. Naïve partitioning of *k*-space into even sized three-dimensional (3D) sections for parallel gridding would result in a significant workload imbalance among execution threads, which would greatly limit the achievable performance gains. This challenge is addressed through the use of a gridding algorithm designed for nonuniform data that dynamically partitions the data to balance workload and collaboratively use a central processing unit (CPU) and GPU for gridding (Obeid, 2010). The final result of this work is a processing pipeline that uses a conventional CPU in combination with a GPU to greatly reduce the time required to compute a sodium bioscale from quantitative sodium MR imaging data.

II. MATERIALS AND METHODS

The high-level processing required for conversion of sodium *k*-space data into a TSC bioscale is shown in Figure 1. This processing includes the necessary corrections for B_0 and B_1 inhomogeneties as well as interacquisition alignment of the human imaging data. The B_0 and B_1 maps can be calculated in a few seconds and the typical time required for computing the alignment transform for two sodium MR images is less than 1 s (Atkinson et al., 2012). Applying these corrections, however, requires significant time since an entire single-frequency reconstruction is required to optimally apply the alignment and a subsequent multifrequency reconstruction is performed for the B_0 correction (Man et al., 1997). B_1 inhomogeniety correction (Insko et al., 1993) is rapid as it can be performed directly on the aligned and B_0 corrected image.

Of the steps shown in Figure 1, the single- and multifrequency image reconstructions are by far the most time-consuming. Figure 2 details the core internal operations of these gridding-based image reconstructions that convert the flexTPI *k*-space data into image space. A conventional CPU software implementation of the processing outlined in Figures 1 and 2 was developed in c++ and subsequently optimized over the past 5 years of routine use. The fast Fourier transform library FFTW (Frigo and Johnson, 2005) was used for computing all Fourier transforms with in place transforms used for memory efficiency. All Fourier transforms are planned with the FFTW_PATIENT flag and planning information, termed wisdom by FFTW, is stored to disk so that a given transform size (e.g., 256 × 256 × 256) is only planned once. The FFTW is

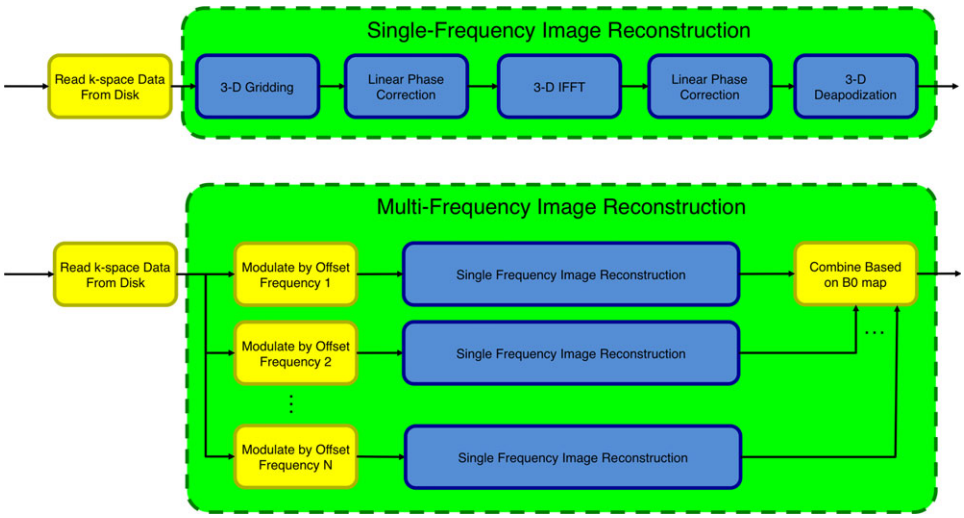

Figure 2. Core operations of a single-frequency (top) and multifrequency (bottom) image reconstruction. The blue boxes are the operations that were accelerated using GPU processing. The yellow boxes always performed on the CPU because they are not computationally expensive or require disk input (e.g. Read *k*-space Data from Disk). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

multithreaded and uses multiple CPU cores. All other operations are custom implementations that were optimized through profiling and execution timing.

The most computationally intensive reconstruction operations were identified (Figure 2, blue boxes) for GPU acceleration. Operations requiring file input/output (I/O) were excluded from GPU acceleration as the graphics card does not have direct access to the file system and disk I/O is rarely computationally bound. The operations shown as blue boxes in Figure 2 were accelerated using GPU processing performed on an NVIDIA Tesla C2050 GPU with 4GB of DRAM. The NVIDIA CUDA environment was used for accessing the general purpose processing capabilities of the Tesla C2050 GPU. The 3D IFFT was accelerated by replacing the calls to the CPU Fourier transform library, FFTW, with calls to GPU Fourier transform library, CUFFT, provided by NVIDIA as part of CUDA. The 3D gridding was performed using a recently developed hybrid gridding algorithm that utilized both the GPU and the CPU (Obeid, 2010) to efficiently process the highly nonuniform sample density found in many 23Na imaging sequences (Boada et al., 1997; Gurney et al., 2006; Qian, 2008; Lu et al., 2010a,b). This algorithm uses the GPU to perform a two-pass spatial sorting, or binning, operation on the nonuniformly sampled *k*-space data. The first pass determines the size needed for each bin using atomic operations on an array of counters, with a limit on the bin size to control load balance. A prefix-scan operation is then used to allocate the exact amount of storage for each bin in a dense array. The second pass places the *k*-space sample data into the bins. Once a bin is full, any extra sample points are placed into a CPU bin. This two-step spatial sorting algorithm is called compact binning (Obeid, 2010), in that it requires much less DRAM and on-chip memory space as well as bandwidth compared with traditional spatial sorting algorithms when the sample density of highly nonuniform. Once the spatial sorting is complete, all samples within a bin are processed by a thread of execution to guarantee no interference among parallel threads. The excess sample data from all bins are shipped to CPU for processing. The gridded data from the GPU and CPU are combined afterward on the GPU. The time required to transfer the CPU-gridded data back to the GPU for combination can be overlapped with the GPU processing and thus does not impose a true time penalty. The complete technical details on compact binning can be found elsewhere (Obeid, 2010). Although there exist other GPU gridding implementations [e.g., (Schiwietz et al., 2006)], compact binning was selected for the required 3D gridding as it is well suited to the highly nonuniform sample density of sodium MR imaging performed with the flexTPI pulse sequence. Custom GPU processing kernels were developed for the remaining Linear Phase Correction and 3D Deapodization operations so as to perform the required processing for these two steps solely on the GPU. The 3D Deapodization removes the appodization due to gridding using a 3D IFFT followed by a point-wise complex multiplication. Linear phase correction is a basic point-wise operation that performs complex multiplication to modify the phase of each sample point. The flexTPI sequence samples *k*-space symmetrically about the origin from $-K_{MAX}$ to $+K_{MAX}$. As a result, an even-sized gridding matrix will not have a grid cell that is centered at the *k*-space origin (i.e., $K_X = K_Y = K_Z = 0$). The DFT is defined with a bin centered at the origin (DC bin). Thus, when gridding onto an even-sized matrix the ''DC'' bin is centered at $K_X = K_Y = K_Z = 1/(2*\text{FOV})$ rather than $K_X = K_Y = K_Z = 0$, where FOV is the field-of-view in the image domain. This half bin displacement in the *k*-space domain results in an unwanted linear phase term in the image domain that must be corrected. The linear phase correction block after the 3D IFFT performs this correction. Similarly, to have the final image symmetrically represented over the spatial range $-\text{FOV}/2$ to $+\text{FOV}/2$, a linear phase correction is also performed before the 3D IFFT. Without this first linear phase correction block, the final image would be displaced by a half voxel in the *X*, *Y*, and *Z* directions.

Data transfers between the CPU and GPU, which can be time-consuming, were minimized to maximize processing performance. The end result was a single software program for image reconstruction that could perform standard (CPU-only) or GPU-accelerated (GPU+CPU) processing by compiling the software source code with appropriate flags.

The CPU-only and GPU+CPU versions of the image reconstruction software were compiled to executables using the GNU Compiler Collection (GCC; version 4.4.5) and the NVIDIA C-compiler for the GPU with level three optimization (-O3 compile option) and available architecture optimizations (e.g., SSE/SSE2).

Fifteen quantitative sodium MR imaging data sessions were performed on 10 unique volunteers (six men, four women) including both healthy and for-cause subjects (e.g., brain tumor, migraine, electrical trauma, etc) ranging in age from 29 to 73 years (average age of 56.4 years). These individuals represent a cross-section of typical subjects on whom sodium imaging is performed at this institution. Quantitative sodium MR imaging was performed using a custom developed 9.4 Tesla (T) human scanner optimized for brain imaging (Thulborn, 2007). As this scanner has a static field above the current Food and Drug Administration insignificant risk guideline of 8 T, the data were collected under investigational device exemption with institutional review board approval and written informed consent of each subject. Human safety testing performed on adults over the past 7 years has not revealed any significant adverse effects on vital signs or cognitive function at 9.4 Tesla (22). All imaging was performed using a custom birdcage radiofrequency coil tuned to 105.92 MHz and the flexTPI acquisition (Lu et al., 2010a). Each flexTPI acquisition was performed in 10 min over a 22 cm field-of-view using a 0.5 ms hard pulse, $T_R = 175$ ms, $T_E = 0.26$ ms, a maximum gradient amplitude of 5.47 mT/m, and a readout time of 27.3 ms (0.203 radial fraction, effective matrix of $76 \times 76 \times 76$, nominal resolution of $2.89 \times 2.89 \times 2.89$ mm^3) or 10 ms (0.305 radial fraction, effective matrix of $62 \times 62 \times 62$, nominal resolution of $3.54 \times 3.53 \times 3.54$ mm^3).

Subjects underwent up to 60 min of 23Na imaging at 9.4T, with most subjects completing approximately 30 min of imaging to collect the necessary data. Acquisitions for B_0 and B_1 mapping varied the echo time and excitation power as described elsewhere (Lu et al., 2010a). All acquisitions performed on human subjects were repeated on one or more calibration phantoms of known concentration, a requirement for calibration of the sodium bioscale. The experimental human and phantom data provided a total of 37 unique TSC bioscales to be computed.

All data were processed on a 2.67 GHz Core i5 Ubuntu 10.10 workstation with 8 GB of RAM and an NVIDIA Tesla C2050 GPU. The compact binning was performed with a bin size of 32, which made the executing time of the CPU portion of the GPU + CPU gridding less than that of the GPU portion (Obeid, 2010). The average time of each processing block was measured using timing traces executed on the CPU and recorded. CPU-only and GPU+CPU processing of each dataset was performed twice with grid matrix sizes of $128 \times 128 \times 128$ and $256 \times 256 \times 256$, respectively, to investigate the impact of smaller and larger memory requirements and working data sizes on the overall processing of quantitative sodium MR imaging data into TSC bioscales. In both cases, the required sizes of the 3D IFFT, linear phase correction, and 3D deapodization are equal to the grid matrix size. Three-dimensional gridding was performed using a Kaiser-Bessel kernel with parameters selected to minimize aliasing (Beatty et al., 2005) and a length equal to $|2 \cdot GOF$, where GOF is the gridding oversample factor. GOF was ~1.7 (effective matrix of $76 \times 76 \times 76$ data matrix gridded onto $128 \times 128 \times 128$ grid matrix), ~2.1 (effective matrix of $62 \times 62 \times 62$ data matrix gridded onto $128 \times 128 \times 128$ grid matrix), ~3.4 (effective matrix of $76 \times 76 \times 76$ data matrix gridded onto $256 \times 256 \times 256$ grid matrix), or ~4.1 (effective matrix of $62 \times 62 \times 62$ data matrix gridded onto $256 \times 256 \times 256$ grid matrix). Although having GOF > 2 is unlikely to be necessary in practice, it allowed the effect of the larger memory requirements from the larger gridding matrix size to be evaluated.

Table 1. Processing time in seconds for computing TSC bioscale as outlined in Figures 1 and 2

$128 \times 128 \times 128$ Matrix			
Operation	CPU-Only (s)	CPU+GPU (s)	Speedup
3-D Gridding	3.17 ± 0.73	0.29 ± 0.12	10.93
3-D IFFT	0.21 ± 0.01	0.02 ± <0.01	10.50
Deapodization	1.01 ± 0.04	0.02 ± <0.01	50.50
Linear phase correction	0.67 ± <0.01	<0.01 ± <0.01	>67
Single frequency reconstruction	5.65 ± 0.55	2.34 ± 1.23	2.41
Multifrequency reconstruction	162.04 ± 41.03	35.17 ± 16.88	4.60
Total processing for TSC bioscale	492.38 ± 81.75	204.70 ± 34.00	2.41
$256 \times 256 \times 256$ Matrix			
Operation	CPU-Only (s)	CPU+GPU (s)	Speedup
3-D Gridding	35.25 ± 5.15	1.51 ± 0.74	23.36
3-D IFFT	2.06 ± 0.05	0.02 ± <0.01	103.00
Deapodization	8.58 ± 0.07	0.16 ± <0.01	53.63
Linear phase correction	5.45 ± 0.03	<0.01 ± <0.01	>545
Single frequency reconstruction	57.19 ± 5.08	2.45 ± 1.49	23.34
Multifrequency reconstruction	1512.84 ± 295.37	46.52 ± 16.82	32.52
Total processing for TSC bioscale	3565.39 ± 605.50	233.35 ± 40.52	15.28

All processing performed on a 2.67 GHz Core i5 Ubuntu 10.10 workstation with 8 GB of RAM and a NVIDIA Tesla C2050 with 4 GB of RAM. CPU-only processing only utilized the CPU while the GPU+CPU processing used the GPU for acceleration of computationally intensive operations (see Figures 1 and 2). Execution time was measured by timing traces that called the standard function ''gettimeofday()'' on the CPU. The execution time reported was rounded to the nearest 10 ms.

The speedup of each operation and the entire reconstruction was computed using the formula:

$$\text{Speedup} = T_{\text{CPU-only}}/T_{\text{GPU+CPU}} \quad (1)$$

where $T_{\text{CPU-only}}$ and $T_{\text{GPU+CPU}}$ are the processing time of the GPU-only and GPU+CPU implementations, respectively. The time of all GPU+CPU operations included any necessary data transfers between the CPU and GPU.

To examine the equivalency of the computed results, the GPU+CPU and CPU-only results were compared by computing the absolute and percentage difference of the final bioscale computed from the CPU-only implementation and the GPU+CPU implementation. The percent difference was computed relative to the to CPU-only result. Only voxels with a TSC of at least 5 mM in the CPU-only implementation were included in the difference calculations.

III. RESULTS

Table I shows the time (rounded to the nearest 10 ms) required for each of the GPU accelerated operations and the entire TSC bioscale calculation when performed with conventional CPU processing (CPU-only) and when accelerated using the GPU (GPU + CPU) for both matrix sizes tested. The mean absolute (percent) difference between the final bioscale computed using each of the two reconstructions was 0.01 ± 0.02 mM (0.04 ± 0.06%) for the $256 \times 256 \times 256$ matrix and 0.02 ± 0.01 mM (0.08 ± 0.06%) for the $128 \times$

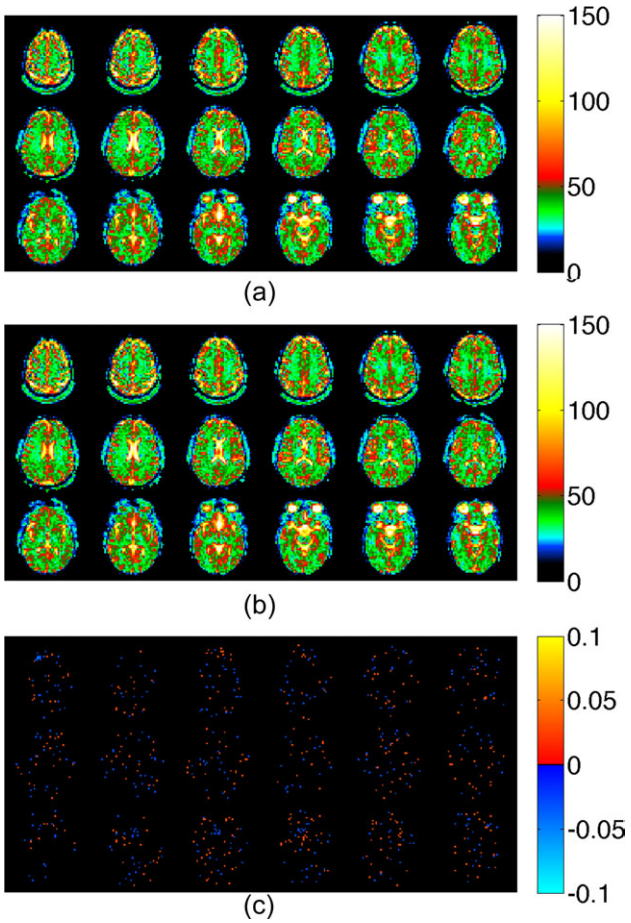

Figure 3. Selected axial cross-sections of the TSC bioscale in an adult volunteer computed with CPU-only processing (a) and GPU+CPU processing (b). The sodium MR imaging data were acquired using a flexTPI acquisition (22 cm field-of-view using a 0.5 ms hard pulse, T_R = 175 ms, T_E = 0.26 ms, a maximum gradient amplitude of 5.47 mT/m, 0.203 radial fraction, nominal resolution of 2.89 × 2.89 × 2.89 mm^3). Image reconstruction was performed using a 256 × 256 × 256 gridding matrix and a length-9 Kaiser-Bessel kernel with parameters selected to minimize aliasing (Beatty et al., 2005). The minor differences between these two final results (c) are likely a result of numerical precision. The maximum absolute difference is less than 0.1 mM. All color scales are in mM of sodium. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

128 × 128 matrix. Figure 3 shows selected slices from a computed TSC bioscale using the CPU-only and the GPU+CPU implementations. The difference image (Fig. 3c) visually presents the difference in the final TSC bioscale for the selected data. Other datasets (not shown) were similar.

The possessing times in Table I show that the GPU+CPU implementation is faster for each individual operation and for the overall processing of a TSC bioscale from sodium *k*-space data. In all cases, the larger working memory size of 256 × 256 × 256 showed greater speedup in processing than the 128 × 128 × 128 matrix. The largest improvements in absolute time came in the 3D gridding (35.25 ± 5.15 s to 1.51 ± 0.74 s, 23.36 times speedup) and deapodization (8.58 ± 0.07s to 0.16 ± <0.01 s, 53.63 times speedup). The largest speedup (>545 times) was achieved in linear phase adjustment, which is a point-wise operation that parallelized very efficiently.

IV. DISCUSSION

Overall, accelerating the operations of the reconstruction led to a 23.34 times speedup for a standard single-frequency gridding-based reconstruction. The majority of the improvement was from the 3D gridding operation. For a multifrequency reconstruction to reduce B_0 inhomogeneities, where the data are processed at several offset frequencies, the improvement is even greater with a 32.52 times speedup. Again the primary gain is from the gridding, which is performed once for each offset frequency. For the entire bioscale calculation, which requires the complete processing shown in Figure 1, use of the GPU resulted in a speedup of 15.28 times (from 3565.39 ± 605.50 s to 233.35 ± 40.52 s, inclusive of all disk I/O). This reduction in time is significant as a TSC bioscale can be computed with the GPU+CPU in less than 4 min rather than the roughly 60 min required for the CPU-only approach. For the smaller 128 × 128 × 128 matrix size similar trends were seen, although the absolute speedups were not as great. This is likely due to the lower compute-to-bandwidth ratio associated with the smaller matrix size (the actual dataset was the same number of samples in both cases). Nonetheless, an overall speedup of 2.41 was found (from 492.38 ± 81.75 s to 204.70 ± 34.00 s, inclusive of all disk I/O) even for the smaller matrix size. The overall speedup is lower than for the larger matrix size due to the higher amount of disk I/O relative to the overall computations. The speedup of all individual operations (3D gridding, 3D IFFT, deapodization, linear phase corrected) remained at least a factor of 10. Although a smaller 128 × 128 × 128 matrix size is sufficient for processing many sodium MRI datasets, the larger matrix size allows for a larger grid-oversampling ratio and may be required as advanced data acquisition techniques continue to advance sodium imaging.

The absolute difference in the final CPU-only and the GPU+CPU bioscales was small, 0.08 ± 0.06% in regions of at least 5 mM for a 256 × 256 × 256 gridding matrix and 0.04 ± 0.06%) for a 128 × 128 × 128 gridding matrix. The existence of the small difference is not surprising given that the CPU and the GPU implementations use separate trigonometric and FFT implementations, which are not guaranteed to produce bit-exact results. However, this small difference is well within the experimental error of quantitative sodium imaging (Lu, et al., 2010) making it of little practical concern.

Rapid sodium bioscale calculation is essential for quantitative sodium imaging of stroke (Thulborn et al., 1999; Boada et al., 2005). The GPU-accelerated sodium image reconstruction described in this report significantly reduces the time required to compute a sodium bioscale from raw *k*-space data, potentially enabling near real-time monitoring of stroke progression and intervention using sodium MR imaging. The reduction in processing time also means that it is now practical for subjects to have the results of the imaging before leaving their appointment. This is of particular importance for subjects in longitudinal studies, such as those who are imaged weekly during ongoing treatment for brain tumors (Thulborn et al., accepted for publication). Immediate review of results with the subject is vital minimizing attrition from longitudinal research studies and will ultimately be required if sodium imaging becomes a clinical tool. The GPU-accelerated sodium bioscale calculation described in this report facilitates this immediate result feadback.

One key for achieving the significant performance improvement was to minimize data transfers between the host CPU and the GPU. The compact binning-based gridding simultaneously used both the GPU and CPU for gridding and combined the results on the GPU to

obtain the complete gridded data. All other operations shown in blue on Figure 2 are all performed solely on the GPU. This means that the only data transfers were before the gridding (CPU to GPU, all *k*-space data), after the CPU portion of the gridding (CPU to GPU, only the CPU gridded result), and after the deapodization (GPU to CPU, complete 128 × 128 × 128 or 256 × 256 × 256 gridded *k*-space). Furthermore, the time required to transfer the CPU portion of the gridded result to the GPU for combination with the GPU gridded results can be overlapped with the GPU gridding to effectively remove its execution cost.

Although the CPU-only code used for comparison is largely single threaded (FFTW is multi-threaded), it has been thoroughly optimized over several years of use and extensive execution profiling. Although it would be possible to implement multithreaded versions of the other CPU operations, even prominent CPU vendors state that the performance of a proper GPU implementation is expected to be at least an order of magnitude better than a multithreaded CPU implementation for operations with a high compute-to-bandwidth ratio 4 (Lee et al., 2010). For the data sizes required to compute a TSC bioscale using either a 128 × 128 × 128 or 256 × 256 × 256 internal working matrix size (e.g., the grid matrix size), all data can fit within the memory on the GPU to maximize the compute-to-bandwidth of the reconstruction operations.

The GPU-accelerated reconstruction for a 3D non-Cartesian acquisition was presented in the setting of quantitative sodium MR imaging and bioscale computation. However, because the actual image reconstruction algorithm outlined in Figures 1 and 2 is a generic gridding-based reconstruction, the same techniques can be used to accelerate image reconstruction of other datasets such as high-resolution projection imaging of protons (Lu et al., 2005). For example, the same GPU+CPU reconstruction presented in this work can achieve a 6.7 times speedup (inclusive of all disk I/O) when reconstructing a high-resolution (384 × 384 × 384 matrix) proton projection imaging dataset when processing with a 512 × 512 × 512 grid matrix size.

V. CONCLUSIONS

Consumer GPUs can accelerate computation of sodium bioscales by more than 15 times and reduce the overall processing time from approximately 1 h to less than 4 min. The resultant bioscale is equal (to within a fraction of a percent) to the traditional result computed using only a CPU. The low cost and pervasive existence of GPUs makes GPU-based processing an attractive option for rapid high-performance calculation of projection imaging as represented in sodium bioscales derived from quantitative sodium MR imaging data.

REFERENCES

I.C. Atkinson, A. Lu, and K.R. Thulborn, Clinically constrained optimization of flexTPI acquisition parameters for the tissue sodium concentration bioscale, Magn Reson Med 66 (2011), 1089–1099.

I.C. Atkinson, A. Lu, and K.R. Thulborn, Preserving the accuracy and resolution of the sodium bioscale from quantitative sodium MRI during intrasubject alignment across longitudinal studies, Magn Reson Med 68 (2012A), 751–761.

I.C. Atkinson, A. Lu, and K.R. Thulborn, Quantitative metabolic MR imaging using ^{17}O and ^{23}Na, Encyclopedia magnetic resonance, R.K. Harris and R.E. Wasylishen (Editors), Wiley, Chichester; DOI:10.1002/9780470034590.emrstm1284. 2012B.

I.C. Atkinson, R. Sonstegaard, N.H. Pliskin, and K.R. Thulborn, Vital signs and cognitive function are not affected by 23-Sodium and 17-Oxygen MR imaging of the human brain at 9.4 Tesla, J Magn Reson Imaging 32 (2010), 82–87.

P.J. Beatty, D.G. Nishimura, and J.M. Pauly, Rapid gridding reconstruction with a minimal oversampling ratio, IEEE Trans Med Imaging 24 (2005), 799–808.

F.E. Boada, J.S. Gillen, G.X. Shen, S.Y. Chang, and K.R. Thulborn, Fast three dimensional sodium imaging, Magn Reson Med 37 (1997), 706–715.

F.E. Boada, G. LaVerde, C. Jungreis, E. Nemoto, C. Tanase, and I. Hancu, Loss of cell ion homeostasis and cell viability in the brain: What sodium MRI can tell us, Curr Top Dev Biol 70 (2005), 77–101.

W. Deng, C. Yang, and V.A. Stenger, Accelerated multidimensional radiofrequency pulse design for parallel transmission using concurrent computation on multiple graphics processing units, Magn Reson Med 65 (2011), 363–369.

P.T. Gurney, B.A. Hargreaves, and D.G. Nishimura, Design and analysis of a practical 3-D cones trajectory. Magn Reson Med 55 (2006), 575–582.

M. Frigo and S.G. Johnson, The design and implementation of FFTW3, Proc IEEE 93 (2005), 216–231.

S.K. Hilal, A.A. Maudsley, J.B. Ra, H.E. Simon, P. Roschmann, S. Wittekoek, Z.H. Cho, and S.K. Mun, In vivo NMR imaging of sodium-23 in the human head, J Comput Assist Tomogr 9 (1985), 1–7.

T.-Y. Huang, Y.-W. Tan, and S.-Y. Ju, Accelerating image registration of MRI by GPU-based parallel computation, Magn Reson Imaging 29 (2011), 712–716.

E.K. Insko and L. Bolinger, Mapping of the radiofrequency field, J Magn Reson 103 (1993), 82–85.

R.P. Kline, E.X. Wu, D.P. Petrylak, M. Szabolcs, P.O. Alderson, M.L. Weisfeldt, P. Cannon, and J. Katz, Rapid in vivo monitoring of chemotherapeutic response using weighted sodium magnetic resonance imaging, Clin Cancer Res 5 (2000), 2146–2156.

V.W. Lee, C. Kim, J. Chhugani, M. Deisher, D. Kim, A.D. Nguyen, N. Satish, M. Smelyanskiy, S. Chennupaty, P. Hammarlund, R. Singhal, and P. Dubey, Debunking the 100X GPU vs. CPU Myth: An evaluation of throughput computing on CPU and GPU, In Proc ISCA'10, Saint-Malo, France, 2010.

A. Lu, I.C. Atkinson, T. Claiborne, F. Damen, and K.R. Thulborn, Quantitative sodium imaging with a flexible twisted projection pulse sequence, Magn Reson Med 63 (2010a), 1583–1593.

A. Lu, I.C. Atkinson, and K.R. Thulborn, Sodium magnetic resonance imaging and its bioscale of tissue sodium concentration, In: Encyclopedia of magnetic resonance, R.K. Harris and R.E. Wasylishen (Editors), Wiley, Chichester; DOI:10.1002/9780470034590.emrstm1171. 2010B.

A. Lu, E. Broadsky, T.M. Grist, and W.F. Block, Rapid fat suppressed isotropic steady-state free precession imaging using true 3D multiple-half-echo projection reconstruction. Magn Reson Med 53 (2005), 692–699.

L.C. Man, J.M. Pauly, and A. Macovski, Multifrequency interpolation for fast off-resonance correction. Magn Reson Med 37 (1997), 785–792.

E.A. Mellon, D.T. Pilkinton, C.M. Clark, M.A. Elliott, W.E. Witschey, A. Borthakur, R. Reddya, Sodium MR imaging detection of mild Alzheimer disease: Preliminary study. AJNR Am J Neuroradiol 30 (2009), 978–984.

A.M. Nagel, F.B. Laun, M.A. Weber, C. Mattheis, W. Semmler, and L.R. Schad, Sodium MRI using a density-adapted 3D radial acquisition technique. Magn Reson Med 62 (2009), 1565–1573.

N. Obeid, Compact binning for parallel processing of limited-range functions, Masters Thesis, University of Illinois at Urbana Champaign, 2010.

Y. Qian and F.E. Boada, Acquisition-weighted stack of spirals for fast high-resolution three-dimensional ultra-short echo time MR imaging, Magn Reson Med 60 (2008), 135–145.

S. Roujol, B.D. de Senneville, E. Vahala, T.S. Sørense, C. Moonen, and M. Ries, Online real-time reconstruction of adaptive TSENSE with commodity CPU/GPU hardware. Mag Reson Med 62 (2009), 1658–1664.

T. Schiwietz, T. Chang, P. Speier, and R. Westermann, MR image reconstruction using the GPU, In Proceedings of SPIE, San Diego, California USA, 2006.

G.G. Somjen, Ions of the brain, normal function, seizures and stroke, Oxford University Press, New York, 2004.

S.S. Stone, J.P. Haldar, S.C. Tsao, W.-M.H. Hwu, Z.-P. Liang, and B.P. Sutton. Accelerating Advanced MRI Reconstructions on GPUs, In Proceedings of the 2008 International Conference on Computing Frontiers, Ischia, Italy, 2008.

K.R. Thulborn, I.C. Atkinson, D. Fleischman, and R. Shah, Feasibility of detecting preclinical hippocampal neuronal cell loss in subjects destined to develop Alzheimer's disease, In Proceedings of the International Society for Magnetic Resonance in Medicine, Montreal, Quebec Canada, 2011.

K.R. Thulborn, A. Lu, I.C. Atkinson, F. Damen, and J.L. Villano, Quantitative sodium MR imaging and sodium bioscales for the management of brain tumors. Neuroimag Clin N Am, 19 (2009), 615–624.

K.R. Thulborn, T.S. Gindin, D. Davis, and P. Erb, Comprehensive MRI protocol for stroke management: Tissue sodium concentration as a measure of tissue viability in a non-human primate model and clinical studies. Radiology 139 (1999), 26–34.

K.R. Thulborn, The challenges of integrating a 9.4T MR scanner for human brain imaging. In: Ultra high field magnetic resonance imaging, P.M. Robitaille and L. Berliner (Editors), Springer, Berlin; 2007. p 105–126.