Accelerating Advanced MRI Reconstructions on GPUs

Sam S. Stone
Center for Reliable and
High-Performance Computing
University of Illinois at
Urbana-Champaign
Urbana, IL
ssstone2@crhc.uiuc.edu

Justin P. Haldar
Department of Electrical and
Computer Engineering
University of Illinois at
Urbana-Champaign
Urbana, IL
haldar@uiuc.edu

Stephanie C. Tsao
Center for Reliable and
High-Performance Computing
University of Illinois at
Urbana-Champaign
Urbana, IL
stsao3@crhc.uiuc.edu

Wen-mei W. Hwu
Center for Reliable and
High-Performance Computing
University of Illinois at
Urbana-Champaign
Urbana, IL
hwu@crhc.uiuc.edu

Zhi-Pei Liang
Department of Electrical and
Computer Engineering
University of Illinois at
Urbana-Champaign
Urbana, IL
z-liang@uiuc.edu

Bradley P. Sutton
Bioengineering Department
Biomedical Imaging Center,
Beckman Institute for
Advanced Science and
Technology
University of Illinois at
Urbana-Champaign
Urbana, IL
sutton@uiuc.edu

ABSTRACT

Computational acceleration on graphics processing units (GPUs) can make advanced magnetic resonance imaging (MRI) reconstruction algorithms attractive in clinical settings, thereby improving the quality of MR images across a broad spectrum of applications. At present, MR imaging is often limited by high noise levels, significant imaging artifacts, and/or long data acquisition (scan) times. Advanced image reconstruction algorithms can mitigate these limitations and improve image quality by simultaneously operating on scan data acquired with arbitrary trajectories and incorporating additional information such as anatomical constraints. However, the improvements in image quality come at the expense of a considerable increase in computation.

This paper describes the acceleration of an advanced reconstruction algorithm on NVIDIA's Quadro FX 5600. Optimizations such as register allocating the voxel data, tiling the scan data, and storing the scan data in the Quadro's constant memory dramatically reduce the reconstruction's required bandwidth to off-chip memory. The Quadro's special functional units provide substantial acceleration of the trigonometric computations in the algorithm's inner loops, and experimentally-tuned code transformations increase the reconstruction's performance by an additional 20%.

The reconstruction of a 3D image with 128^3 voxels ultimately achieves 150 GFLOPS and requires less than two minutes on the Quadro, while reconstruction on a quad-core CPU is thirteen times slower. Furthermore, relative to the true image, the error exhibited by the advanced reconstruction is only 12%, while conventional reconstruction techniques incur error of 42%. In short, the acceleration afforded by the GPU greatly increases the appeal of the advanced reconstruction for clinical MRI applications.

Categories and Subject Descriptors

C.1.4 [**Computer Systems Organization**]: Processor Architectures—*Parallel Architectures*; I.3.1 [**Computing Methodologies**]: Computer Graphics—*Hardware Architecture*; I.4.5 [**Computing Methodologies**]: Image Processing and Computer Vision—*Reconstruction*

General Terms

Algorithms, Performance

Keywords

CUDA, GPGPU, GPU computing, MRI, reconstruction

1. INTRODUCTION

Mainstream microprocessors such as the Intel Pentium and AMD Opteron families have driven rapid performance increases and cost reductions in science and engineering applications for two decades. These commodity microprocessors have delivered GFLOPS to the desktop and hundreds of GFLOPS to cluster servers. This progress, however, slowed in 2003 due to constraints on power consumption. Since that time, accelerators such as graphics processing units (GPUs)

CF'08, May 5–7, 2008, Ischia, Italy.
Copyright 2008 ACM 978-1-60558-077-7/08/05 ...$5.00.

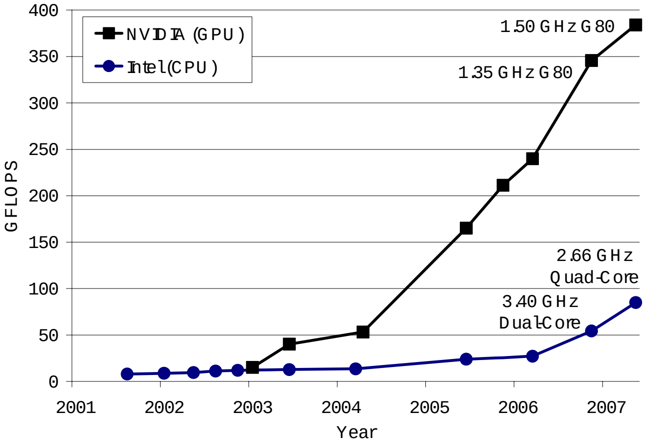


Figure 1: Peak throughput of programmable, floating-point, multiply-add operations on modern GPUs and CPUs. Adapted with permission from [24] © John Owens et al.

have led the advances in computational throughput for science and engineering applications. Figure 1 illustrates this trend.

Recent advances in architecture have also increased the GPU's attractiveness as a platform for science and engineering applications. Prior to 2006, GPUs found very limited use in this domain due to their limited support for both IEEE floating-point standards and arbitrary memory addressing. However, the recently released AMD R580 and NVIDIA G80 GPUs offer strong support for IEEE single-precision floating-point values (with double-precision soon to follow) and permit reads and writes to arbitrary addresses in memory [2, 23]. Furthermore, modern GPUs use massive multithreading, fast context switching, and high memory bandwidth to tolerate ever-increasing latencies to main memory by overlapping long-latency loads in stalled threads with useful computation in other threads [19].

Increased programmability has also enhanced the GPU's suitability for science and engineering applications. For example, the G80 supports the single-program, multiple-data (SPMD) programming model, in which each thread is created from the same program and operates on a distinct data element, but all threads need not follow the same control flow path. As the SPMD programming model has been used on massively parallel supercomputers in the past, it is reasonable to expect that many high-performance applications will port easily to the G80 [19, 38]. Furthermore, general-purpose applications targeting the G80 are developed using ANSI C with simple extensions, rather than the cumbersome graphics application programming interfaces (APIs) [32, 8] and high-level languages layered on graphics APIs [6, 4, 37] that have been used in the past.

A wide variety of magnetic resonance imaging (MRI) applications, ranging from quantitative imaging of the brain to dynamic imaging of the beating heart, can benefit greatly from these increases in computational resources and advancements in architecture and programmability. At present, many MRI experiments are specifically designed so that the image can be reconstructed quickly and efficiently on a standard CPU, often by acquiring the scan data on a uniform grid and applying a fast Fourier transform (FFT). However, in many applications the combination of tailored data acquisition and advanced image reconstruction significantly improves image quality. In particular, these techniques can increase signal-to-noise ratio, decrease scan time, and/or reduce imaging artifacts. However, advanced reconstruction algorithms often require several orders of magnitude more computation than conventional reconstruction algorithms. In this paper, we accelerate a reconstruction algorithm that can (1) generate MR images from arbitrary data sampling trajectories, and (2) incorporate prior anatomical knowledge into the reconstruction process, thereby increasing the signal-to-noise ratio.

For these advanced reconstructions to be viable in clinical settings, dramatic and inexpensive computational acceleration is required. We find that advanced reconstructions from arbitrary scan trajectories are very well suited to acceleration on modern GPUs. In particular, an advanced reconstruction of an image comprising 128^3 voxels completes in less than 2 minutes on the G80, while the same reconstruction requires 23 minutes on a quad-core CPU. Furthermore, relative to a conventional reconstruction, the advanced reconstruction reduces the error in the reconstructed image from 42% to 12%. The 13X acceleration achieved on the GPU makes the constrained reconstruction much more appealing in clinical settings.

The remainder of this paper is organized as follows. Section 2 first describes the architecture of the Quadro FX 5600 and its G80 GPU, then discusses the advantages of advanced MRI reconstructions. Section 3 presents the GPU-based implementation of the advanced reconstruction algorithm. Section 4 describes experimental methodology. Section 5 presents results and discusses features of the Quadro that enable the advanced reconstruction to achieve 150 GFLOPS in performance. Section 6 discusses related work in GPU-based medical imaging. Section 7 concludes.

2. BACKGROUND

2.1 The Quadro FX 5600 Graphics Card

The Quadro FX 5600 is a graphics card equipped with a G80 graphics processing unit (GPU). The Quadro has a large set of processor cores that can directly address a global memory. This architecture supports the single-program, multiple-data (SPMD) programming model, which is more general and flexible than the programming models supported by previous generations of GPUs, and which allows developers to easily implement data-parallel algorithms. In this section we discuss NVIDIA's Compute Unified Device Architecture (CUDA) and the architectural features of the G80 that are most relevant to accelerating MRI reconstructions. More complete descriptions are found in [23, 21, 26].

From the application developer's perspective, the CUDA programming model consists of ANSI C supported by several keywords and constructs. CUDA treats the GPU as a coprocessor that executes data-parallel kernel functions. The developer supplies a single source program encompassing both host (CPU) and kernel (GPU) code. NVIDIA's compiler, nvcc, separates the host and kernel codes, which are then compiled by the host compiler and nvcc, respectively. The host code transfers data to and from the GPU's global memory via API calls, and initiates the kernel code by calling a function.

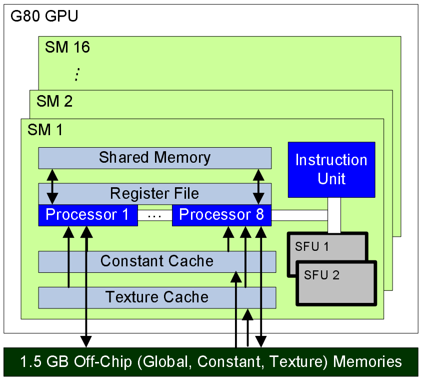


Figure 2: Architecture of Quadro FX 5600.

Figure 2 depicts the Quadro's architecture. The G80 GPU consists of 16 *streaming multiprocessors* (SMs), each containing eight *streaming processors* (SPs), or processor cores, running at 1.35 GHz. Each SM has 8,192 registers that are shared among all threads assigned to the SM. The threads on a given SM's cores execute in SIMD (single-instruction, multiple-data) fashion, with the instruction unit broadcasting the current instruction to the eight cores. Each core has a single arithmetic unit that performs single-precision floating point arithmetic and 32-bit integer operations. Additionally, each SM has two *special functional units* (SFUs), which perform more complex FP operations such as the trigonometric functions with low latency. Both the arithmetic units and the SFUs are fully pipelined. Thus, each SM can perform 18 FLOPS per clock cycle (one multiply-add operation per SP and one complex operation per SFU), yielding 388.8 GFLOPS (16 SM * 18 FLOP/SM * 1.35 GHz) of peak theoretical performance for the GPU.

The Quadro has 76.8 GB/s of bandwidth to its 1.5 GB, off-chip, global memory. Nevertheless, with computational resources supporting nearly 400 GFLOPS and each multiply-add instruction operating on up to 16 bytes of data, applications can easily saturate that bandwidth. Therefore, as depicted in Figure 2, the G80 has several on-chip memories that can exploit data locality and data sharing to reduce an application's demands for off-chip memory bandwidth. For example, the Quadro has a 64 KB, off-chip *constant memory*, and each SM has an 8 KB constant memory cache. Because the cache is single-ported, simultaneous accesses of different addresses yield stalls. However, when multiple threads access the same address during the same cycle, the cache broadcasts that address's value to those threads with the same latency as a register access. This feature proves quite beneficial for the MRI reconstruction algorithm studied in this paper. In addition to the constant memory cache, each SM has a 16KB *shared memory* for data that is either written and reused or shared among threads. Finally, for read-only data that is shared by many threads but not necessarily accessed simultaneously by all threads, the off-chip texture memory and the on-chip texture caches exploit 2D data locality to substantially reduce memory latency.

Threads executing on the G80 are organized into a three-level hierarchy. At the highest level, each kernel creates a single *grid*, which consists of many *thread blocks*. The maximum number of threads per block is 512. Each thread block is assigned to a single SM for the duration of its execution. Threads in the same block can share data through the shared memory and can perform barrier synchronization by invoking the **__syncthreads** primitive. Threads are otherwise independent, and synchronization across thread blocks is safely accomplished only by terminating the kernel. Finally, threads within a block are organized into *warps* of 32 threads. Each warp executes in SIMD fashion, with the SM's instruction unit broadcasting the same instruction to the eight cores on four consecutive clock cycles.

SMs can interleave warps on an instruction-by-instruction basis to hide the latency of global memory accesses and long-latency arithmetic operations. When one warp stalls, the SM can quickly switch to a ready warp in the same thread block or in some other thread block assigned to the SM. The SM stalls only if there are no warps with all operands available.

Tuning the performance of a CUDA kernel often involves a fundamental trade-off between the efficiency of individual threads and the thread-level parallelism (TLP) among all threads. This trade-off exists because many optimizations that improve the performance of an individual thread tend to increase the thread's use of limited resources that are shared among all threads assigned to an SM. For example, as each thread's register usage increases, the total number of threads that can simultaneously occupy the SM decreases. Because threads are assigned to an SM not individually, but in large thread blocks, a small increase in register usage can cause a correspondingly much larger decrease in SM occupancy [27, 28]. Section 5.6 examines this trade-off in the context of MRI reconstructions.

2.2 Advanced MRI Reconstruction

Magnetic resonance imaging (MRI) is commonly used by the medical community to safely and non-invasively probe the structure and function of biological tissues from all regions of the body, and images generated using MRI have a profound impact in both clinical and research settings. MR imaging consists of two phases, acquisition (*scan*) and reconstruction. During the scan phase, the scanner samples data in the k-space domain (*i.e.*, the spatial-frequency domain or Fourier transform domain) along a pre-defined trajectory. These samples are then transformed into the desired image during the reconstruction phase.

MRI is often limited by high noise levels, significant imaging artifacts, and/or long data acquisition (*scan*) times. In clinical settings, short scan times not only increase scanner throughput but also reduce patient discomfort, which tends to mitigate motion-related artifacts. High image resolution is equally important because it can enable earlier detection of pathology, leading to improved prognoses for patients. However, the goals of short scan time, high resolution, and high signal-to-noise ratio (SNR) often conflict; improvements in one metric tend to come at the expense of one or both of the others.

The sampling trajectory used by the MRI scanner can significantly affect the quality of the reconstruction. Figures 3(a) and 3(c) depict a Cartesian scan trajectory and a non-Cartesian (spiral) scan trajectory, respectively. The Cartesian trajectory samples k-space on a uniform grid, which allows image reconstruction to be performed quickly and efficiently by applying a fast Fourier transform (FFT) directly to the acquired data. Although the reconstruction of Cartesian scan data is computationally efficient, non-

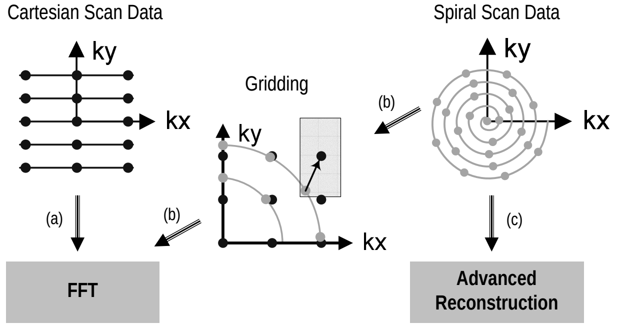


Figure 3: MRI reconstruction techniques. In (a) the scanner samples k-space on a uniform grid and reconstructs the image in one step via the FFT. In (b) the scanner samples k-space on a non-Cartesian (spiral) trajectory, then interpolates the samples onto a uniform grid and reconstructs the image in one step via the FFT. In (c) an advanced reconstruction algorithm is applied directly to the spiral scan data.

Cartesian scan trajectories can be preferable because they are often faster and less sensitive to imaging artifacts caused by non-ideal experimental conditions. For these reasons, non-Cartesian trajectories with radial [18] and spiral [1] sampling patterns are becoming increasingly common in MRI.

Image reconstruction from non-Cartesian scan data presents both challenges and opportunities. In the most common approach, *gridding*, the samples are first interpolated onto a uniform Cartesian grid and then reconstructed in one step via the FFT (see Figure 3(b)) [15, 31]. While gridding is computationally expedient, it satisfies no optimality criterion and cannot leverage prior information such as anatomical constraints. By contrast, optimal image reconstructions [25, 39, 10, 36, 11] can incorporate anatomical information [13, 12] to reduce noise while preserving the resolution of known image features. *Anatomically constrained reconstruction of non-Cartesian scan data enables brief scans to achieve high SNR, thereby decreasing imaging artifacts and increasing SNR simultaneously.* While such advanced reconstructions have been impractical for large-scale problems due to computational constraints, this paper shows that these reconstructions become viable in clinical settings when accelerated on GPUs.

We implemented the anatomically constrained reconstruction algorithm of [13, 12]. This algorithm finds the the solution to the following quasi-Bayesian estimation problem

$$\hat{\boldsymbol{\rho}} = \arg\min_{\boldsymbol{\rho}} \underbrace{\|\mathbf{F}\boldsymbol{\rho} - \mathbf{d}\|_2^2}_{\text{data fidelity}} + \underbrace{\|\mathbf{W}\boldsymbol{\rho}\|_2^2}_{\text{prior info}}, \tag{1}$$

where $\hat{\boldsymbol{\rho}}$ is a vector containing voxel values for the reconstructed image, $\mathbf{F}$ is a matrix that models the imaging process, $\mathbf{d}$ is a vector of data samples, and $\mathbf{W}$ is a matrix that can incorporate prior information such as anatomical constraints. The first term in the above cost function imposes that data simulated from the reconstructed image should match somewhat closely with the real acquired data; the second term is used to impose prior information regarding the image statistics.

Because Eq. 1 defines a linear least squares problem, the solution is

$$\hat{\boldsymbol{\rho}} = \left(\mathbf{F}^H\mathbf{F} + \mathbf{W}^H\mathbf{W}\right)^{-1}\mathbf{F}^H\mathbf{d}. \tag{2}$$

However, the size of the matrix $\left(\mathbf{F}^H\mathbf{F} + \mathbf{W}^H\mathbf{W}\right)$ makes direct matrix inversion impractical for high-resolution reconstructions. For the 128^3-voxel reconstructions examined in this paper, the inverted matrix contains well over four trillion complex-valued elements (the number of elements in the inverted matrix equals the square of the number of voxels in the reconstructed image). An iterative method for matrix inversion, such as the conjugate gradient (CG) algorithm [14], is therefore preferred.

The conjugate gradient algorithm reconstructs the image by iteratively solving Eq. 2 for $\hat{\boldsymbol{\rho}}$. During each iteration, the CG algorithm updates the current image estimate $\boldsymbol{\rho}$ to improve the value of the quasi-Bayesian cost function (Eq. 1). The computational efficiency of the CG technique is largely determined by the efficiency of matrix-vector multiplication operations involving $\mathbf{F}^H\mathbf{F}$ and $\mathbf{W}^H\mathbf{W}$, as these operations are required during each iteration of the CG algorithm. Fortunately, matrix $\mathbf{W}$ often has a sparse structure that permits efficient multiplication by $\mathbf{W}^H\mathbf{W}$, and matrix $\mathbf{F}^H\mathbf{F}$ has a convolutional structure [39] that enables efficient matrix multiplication via the FFT.

The advanced reconstruction algorithm described in this paper therefore consists of three primary computations. First, the algorithm computes each element of $\mathbf{Q}$, given by

$$Q\left(\mathbf{x}_n\right) = \sum_{m=1}^{M} \left|\phi(\mathbf{k}_m)\right|^2 e^{\left(i2\pi\mathbf{k}_m\cdot\mathbf{x}_n\right)}, \tag{3}$$

where $\mathbf{Q}$ is the convolution kernel that facilitates multiplication operations involving $\mathbf{F}^H\mathbf{F}$, and $\phi(\cdot)$ is the Fourier transform of the voxel basis function. There are M k-space sampling locations, with $\mathbf{k}_m$ denoting the location of the m^{th} sample. Likewise, there are N voxel coordinates, with $\mathbf{x}_n$ denoting the coordinates of the n^{th} voxel. Because $\mathbf{Q}$ depends only on the scan trajectory (not the scan data) and the size of the image, it can be computed before the scan occurs and can be reused during any reconstruction that shares the same scan trajectory and image size. Second, the algorithm computes the vector $\mathbf{F}^H\mathbf{d}$, defined as

$$\left[\mathbf{F}^H\mathbf{d}\right]_n = \sum_{m=1}^{M} \phi^*(\mathbf{k}_m)\mathbf{d}(\mathbf{k}_m)e^{\left(i2\pi\mathbf{k}_m\cdot\mathbf{x}_n\right)}. \tag{4}$$

Although Eq. 3 and Eq. 4 are quite similar, the former necessitates significantly more computation because the $\mathbf{Q}$ algorithm oversamples the image space by a factor of two in each dimension. Therefore, during a 3D reconstruction, Eq. 3 is evaluated at 8N values of $\mathbf{x}_n$, while the Eq. 4 is evaluated at only N values of $\mathbf{x}_n$. Finally, the CG solver performs iterative matrix inversion to solve Eq. 2.

The complexity of the advanced reconstruction far exceeds the complexity of a conventional, gridded reconstruction. Given a reconstruction problem of N voxels and M scan data points, the computations of $\mathbf{Q}$ and $\mathbf{F}^H\mathbf{d}$ have O(MN) complexity, compared to O(N log N) complexity for reconstructions based on gridding and the FFT. For this reason, advanced reconstruction of high-resolution, three-dimensional images has been impractical in clinical settings, despite the technique's clear advantages over conventional reconstructions. Our work demonstrates that these advanced reconstructions can be performed quickly and efficiently on modern GPUs, increasing their viability in clinical settings.

3. ADVANCED MRI RECONSTRUCTION

The advanced MRI reconstruction algorithm described in Section 2.2 consists of three steps: computing the data structure $\mathbf{Q}$ (which depends only on the scan trajectory), computing the vector $\mathbf{F}^H\mathbf{d}$ (which depends on the scan trajectory and the scan data), and finding the image iteratively via a conjugate gradient linear solver. As Figure 4 shows, the algorithms for $\mathbf{F}^H\mathbf{d}$ and $\mathbf{Q}$ are quite similar, The most significant difference is that the $\mathbf{Q}$ algorithm requires more computation because its outer loop executes 8N iterations, compared to N iterations for $\mathbf{F}^H\mathbf{d}$. Otherwise, $\mathbf{Q}$ suffers from the same bottlenecks and benefits from the same code transformations as $\mathbf{F}^H\mathbf{d}$.

Because $\mathbf{Q}$ can be computed prior to acquiring an image's scan data, the critical path for a given reconstruction consists only of computing $\mathbf{F}^H\mathbf{d}$ and executing the linear solver. Therefore, the remainder of this section describes the algorithms for $\mathbf{F}^H\mathbf{d}$ and the linear solver, focusing on the implementation of the $\mathbf{F}^H\mathbf{d}$ algorithm on the GPU. The interested reader may refer to [34] for more detailed discussion of $\mathbf{Q}$.

3.1 $\mathbf{F}^H\mathbf{d}$

As Figure 4(b) shows, the algorithm for $\mathbf{F}^H\mathbf{d}$ is an excellent candidate for acceleration on the GPU because it contains substantial data-parallelism. The algorithm first computes the real and imaginary components of $\boldsymbol{\mu}$ at each sample point in the trajectory space (k-space), then computes the real and imaginary components of $\mathbf{F}^H\mathbf{d}$ at each voxel in the image space. The value of $\mathbf{F}^H\mathbf{d}$ at any voxel depends on the values of all sample points, but no elements of $\mathbf{F}^H\mathbf{d}$ depend on any other elements of $\mathbf{F}^H\mathbf{d}$. Therefore, all elements of $\mathbf{F}^H\mathbf{d}$ can be computed independently and in parallel.

Despite the algorithm's inherent parallelism, potential performance bottlenecks are evident. First, in the loop that computes the elements of $\mathbf{F}^H\mathbf{d}$, the ratio of floating-point operations to memory accesses is at best 3:1 and at worst 1:1. The best case assumes that the **sin** and **cos** operations are computed using five-element Taylor series that require 13 and 12 floating-point operations, respectively. The worst case assumes that each trigonometric operation is computed as a single operation in hardware. In either case, the GPU-based implementation of the algorithm must conserve memory bandwidth and tolerate memory latency. Second, the ratio of FP arithmetic to FP trigonometry is only 13:2. Thus, GPU-based implementation must tolerate or avoid stalls due to long-latency **sin** and **cos** operations.

The GPU-based implementation of the $\mathbf{F}^H\mathbf{d}$ algorithm (see Figure 4(c)) uses the G80's constant memory caches to shatter the potential bottleneck posed by memory bandwidth and latency. To overcome the memory bottleneck, the scan data is divided into many tiles, with each tile containing a distinct subset of sample points. For each tile, the host CPU loads the corresponding subset of sample points into constant memory before executing the **cmpFhD** function. Each thread then computes a partial sum for a single element of $\mathbf{F}^H\mathbf{d}$ by iterating over all the sample points in the tile. This optimization increases the ratio of FP operations to global memory accesses dramatically.

Likewise, the G80's special functional units (SFUs) enable the algorithm to avoid the potential bottleneck of long latency trigonometric operations. When the **use_fast_math** compiler option is invoked, the **sin** and **cos** operations are not linked to long-latency library calls, but rather are executed as individual, low-latency instructions on the SFUs. The speed of the SFU comes at the expense of some loss in accuracy when the argument to the **sin** or **cos** is very small, but, as we show in Section 5, this optimization does not necessarily decrease on the overall accuracy of the algorithm.

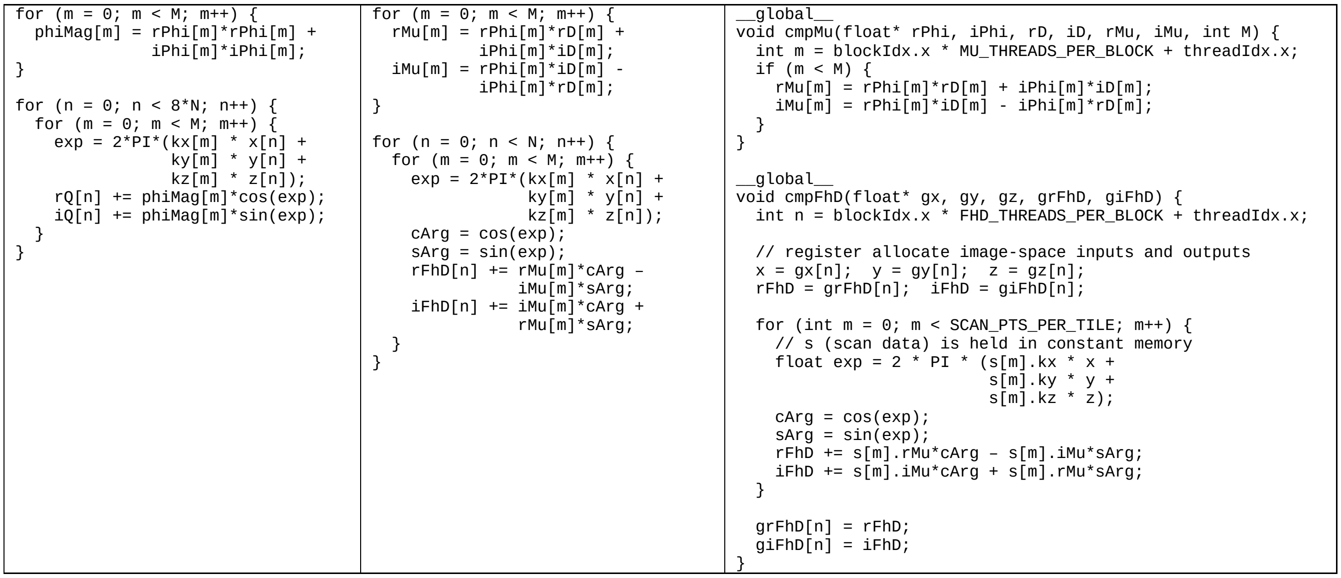
3.2 Conjugate Gradient Linear Solver

As described in Section 2, the CG solver iteratively solves Eq. 2 to find the desired image $\hat{\boldsymbol{\rho}}$. When the iterations converge or the number of iterations exceeds a threshold, the solver terminates. During each iteration, the solver performs a large FFT and inverse FFT, several BLAS and sparse BLAS operations (including multiplication of sparse matrices and vectors, as well as addition, scaling, and scalar multiplication of vectors), and several other computations (such as summation reduction, shifting, and sampling).

We ported the linear solver from MATLAB to CUDA, using NVIDIA's CUDA CUFFT Library [22] for the FFT and inverse FFT operations, and implementing the other operations by hand. Complex-valued objects were represented using CUDA's **cufftComplex** data type, as required by the CUFFT Library. Sparse matrices were stored in compressed row format [9] to facilitate efficient GPU-based execution of the expression $\mathbf{A} * \mathbf{x}$, where $\mathbf{A}$ is a sparse matrix and $\mathbf{x}$ is a vector. Although we have made only small efforts to optimize the CUDA-based solver, it is roughly an order of magnitude faster than the MATLAB-based solver. We use the CUDA-based solver for all experiments presented in this paper and view its performance as acceptable.

4. METHODOLOGY

To quantify the effects of the Quadro's architectural features on the performance and quality of the reconstruction, we implemented seven versions of the algorithm for $\mathbf{F}^H\mathbf{d}$, five of which are depicted in Figure 5. The base version (GPU.Base, see Figure 5(a)) simply executes in data-parallel fashion on the GPU, without using even the simplest optimizations to conserve memory bandwidth or tolerate long latency loads and trigonometric operations. The second version (GPU.RegAlloc, see Figure 5(b)) register allocates the voxel data, thereby conserving some memory bandwidth and reducing the latency of all voxel accesses. GPU.Coalesce (Figure 5(c)) register allocates the voxel data and changes the layout of the scan data in the Quadro's global memory so that accesses to the scan data are coalesced, thereby making more efficient use of the memory bandwidth. GPU.ConstMem (Figure 5(d)) register allocates the voxel data and places the scan data in the Quadro's constant memory so that accesses to the scan data are cached. The fifth version (GPU.FastTrig, see Figure 5(e)) additionally uses the G80's special functional units to compute fast, approximate versions of the trigonometric operations. The sixth version, GPU.Tune, also uses experimentally-tuned settings for three code transformations: loop unrolling, data tiling (scan points per thread), and number of threads per block. The tuned settings balance allocation of GPU re-



```
for (m = 0; m < M; m++) {
  phiMag[m] = rPhi[m]*rPhi[m] +
              iPhi[m]*iPhi[m];
}

for (n = 0; n < 8*N; n++) {
  for (m = 0; m < M; m++) {
    exp = 2*PI*(kx[m] * x[n] +
                ky[m] * y[n] +
                kz[m] * z[n]);
    rQ[n] += phiMag[m]*cos(exp);
    iQ[n] += phiMag[m]*sin(exp);
  }
}
```

```
for (m = 0; m < M; m++) {
  rMu[m] = rPhi[m]*rD[m] +
           iPhi[m]*iD[m];
  iMu[m] = rPhi[m]*iD[m] -
           iPhi[m]*rD[m];
}

for (n = 0; n < N; n++) {
  for (m = 0; m < M; m++) {
    exp = 2*PI*(kx[m] * x[n] +
                ky[m] * y[n] +
                kz[m] * z[n]);
    cArg = cos(exp);
    sArg = sin(exp);
    rFhD[n] += rMu[m]*cArg –
               iMu[m]*sArg;
    iFhD[n] += iMu[m]*cArg +
               rMu[m]*sArg;
  }
}
```

```
__global__
void cmpMu(float* rPhi, iPhi, rD, iD, rMu, iMu, int M) {
  int m = blockIdx.x * MU_THREADS_PER_BLOCK + threadIdx.x;
  if (m < M) {
    rMu[m] = rPhi[m]*rD[m] + iPhi[m]*iD[m];
    iMu[m] = rPhi[m]*iD[m] - iPhi[m]*rD[m];
  }
}

__global__
void cmpFhD(float* gx, gy, gz, grFhD, giFhD) {
  int n = blockIdx.x * FHD_THREADS_PER_BLOCK + threadIdx.x;

  // register allocate image-space inputs and outputs
  x = gx[n];  y = gy[n];  z = gz[n];
  rFhD = grFhD[n];  iFhD = giFhD[n];

  for (int m = 0; m < SCAN_PTS_PER_TILE; m++) {
    // s (scan data) is held in constant memory
    float exp = 2 * PI * (s[m].kx * x +
                          s[m].ky * y +
                          s[m].kz * z);
    cArg = cos(exp);
    sArg = sin(exp);
    rFhD += s[m].rMu*cArg – s[m].iMu*sArg;
    iFhD += s[m].iMu*cArg + s[m].rMu*sArg;
  }

  grFhD[n] = rFhD;
  giFhD[n] = iFhD;
}
```

(a) Q algorithm

(b) F^Hd algorithm

(c) F^Hd algorithm in CUDA

Figure 4: Data-parallel phases of advanced MRI reconstruction. Panels (a) and (b) show simplified C code for the algorithms that compute Q and $\mathbf{F}^H\mathbf{d}$, respectively. Panel (c) depicts the $\mathbf{F}^H\mathbf{d}$ algorithm in CUDA.

sources to improve hardware utilization and thread efficiency. Finally, GPU.Multi executes the tuned version on multiple Quadros.

To obtain a reasonable baseline, we implemented two versions of $\mathbf{F}^H\mathbf{d}$ on the CPU. Version CPU.DP uses double-precision for all floating-point values and operations, while version CPU.SP uses single-precision. Both CPU versions are compiled with Intel's icpc (version 10.1) using flags -O3 -msse3 -axT -vec-report3 -fp-model fast=2, which (1) vectorizes the algorithm's dominant loops using instructions tuned for the Core 2 architecture, and (2) links the trigonometric operations to fast, approximate functions in the math library. Based on experimental tuning with a smaller data set, the inner loops are unrolled by a factor of four and the scan data is tiled to improve locality in the L1 cache.

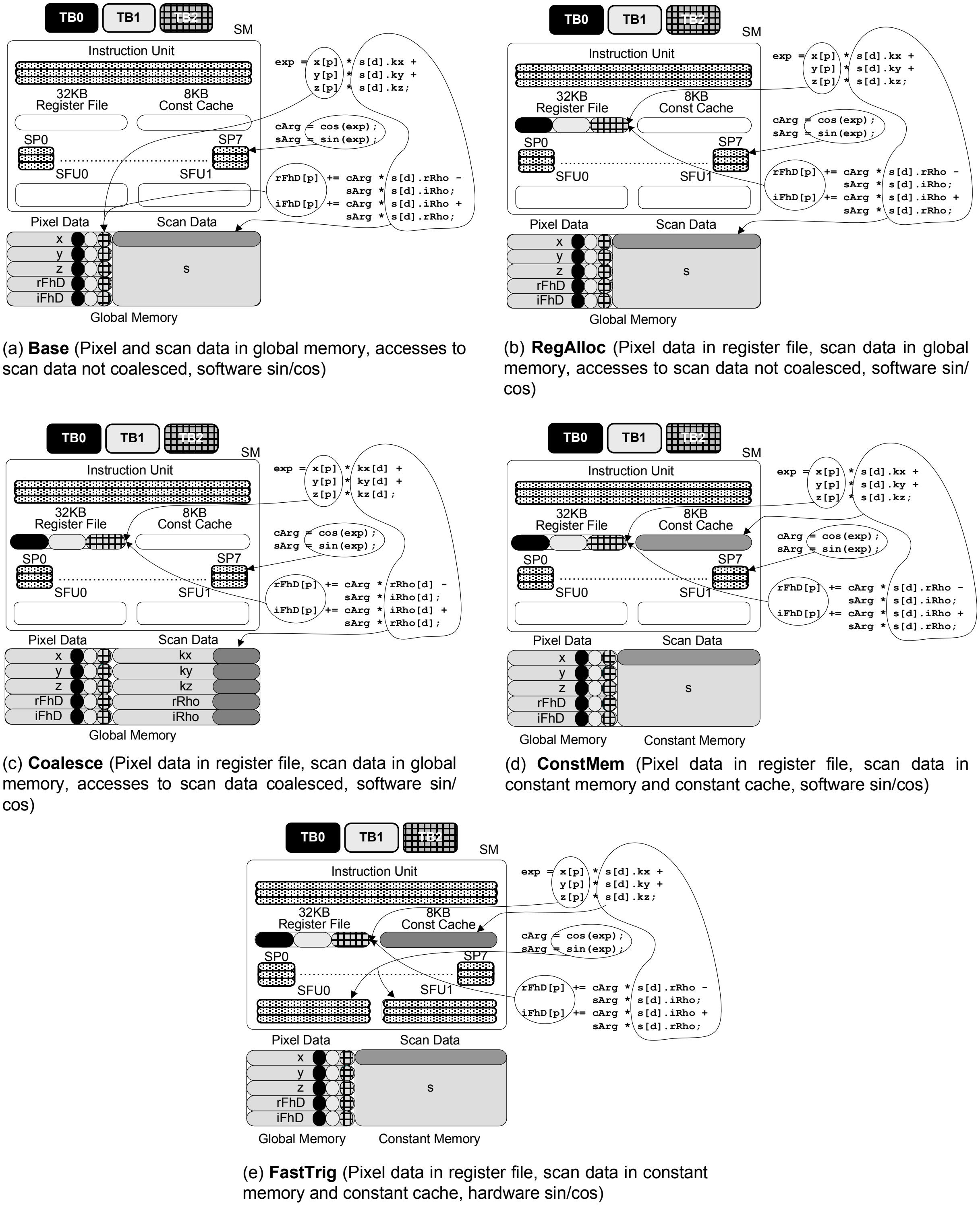
Each GPU version of $\mathbf{F}^H\mathbf{d}$ is compiled using nvcc -O3 (CUDA version 1.0) and executed on a 1.35 GHz Quadro FX 5600. The Quadro card is housed in a system with a 2.4 GHz dual-socket, dual-core Opteron 2216 CPU. Each core has a 1 MB L2 cache. The CPU versions use pthreads to execute on all four cores of 2.66 GHz Core 2 Extreme quad-core CPU, which has peak theoretical capacity of 21.2 GFLOPS per core and a 4 MB L2 cache. The CPU versions perform substantially better on the Core 2 Extreme quad-core than on the dual-socket, dual-core Opteron.

All reconstructions use the GPU version of the linear solver, which executes 60 iterations on the Quadro FX 5600. Two versions of $\mathbf{Q}$ were computed on the Core 2 Extreme, one using double-precision and the other using single-precision. The single-precision $\mathbf{Q}$ was used for all GPU-based reconstructions and for the reconstruction involving CPU.SP, while the double-precision $\mathbf{Q}$ was used only for the reconstruction involving CPU.DP. We have implemented the $\mathbf{Q}$ computation on the GPU and observed that it runs roughly five to six times longer than the GPU version of $\mathbf{F}^H\mathbf{d}$, as expected. As the computation of $\mathbf{Q}$ is not on the reconstruction's critical path, we give $\mathbf{Q}$ no further consideration.

To facilitate comparison of the advanced reconstruction with a conventional reconstruction, we also evaluated a reconstruction based on gridding and the FFT [15]. Our version of the gridded reconstruction is not optimized for performance, but it is fair to assume that an optimized implementation would execute in several seconds [34].

All reconstructions are performed on sample data obtained from a simulated, three-dimensional, non-Cartesian scan of a phantom image [17]. There are 284,592 sample points in the scan data set, and the image is reconstructed at 128^3 resolution, for a total of 2^{21} voxels. In the first set of experiments, the simulated data contains no noise. In the second set of experiments, we added complex white Gaussian noise to the simulated data. When determining the quality of the reconstructed images, the *percent error* and *peak signal-to-noise ratio* metrics are used. The percent error is the root-mean-square (RMS) of the voxel error divided by the RMS voxel value in the true image (after the true image has been sampled at 128^3 resolution). To permit fair comparison of the gridded and advanced reconstructions, we adjusted the scale of each gridded image to match the scale of the true image before computing the gridded image's percent error and PSNR.

Finally, the advanced reconstruction leverages two optimizations that are not evident elsewhere in our discussion. First, the scan trajectory is symmetric, and the advanced reconstruction uses prior knowledge of that symmetry to mitigate the effects of numerical imprecision on the accuracy of the reconstruction. Second, we manually balanced the resolution and the noise in the advanced reconstruction by performing the reconstruction multiple times while adjusting a regularization parameter. Regularization can be performed automatically or analytically prior to acquiring



```
exp = x[p] * s[d].kx +
      y[p] * s[d].ky +
      z[p] * s[d].kz;

cArg = cos(exp);
sArg = sin(exp);

rFhD[p] += cArg * s[d].rRho -
           sArg * s[d].iRho;
iFhD[p] += cArg * s[d].iRho +
           sArg * s[d].rRho;
```

(a) **Base** (Pixel and scan data in global memory, accesses to scan data not coalesced, software sin/cos)

```
exp = x[p] * s[d].kx +
      y[p] * s[d].ky +
      z[p] * s[d].kz;

cArg = cos(exp);
sArg = sin(exp);

rFhD[p] += cArg * s[d].rRho -
           sArg * s[d].iRho;
iFhD[p] += cArg * s[d].iRho +
           sArg * s[d].rRho;
```

(b) **RegAlloc** (Pixel data in register file, scan data in global memory, accesses to scan data not coalesced, software sin/cos)

```
exp = x[p] * kx[d] +
      y[p] * ky[d] +
      z[p] * kz[d];

cArg = cos(exp);
sArg = sin(exp);

rFhD[p] += cArg * rRho[d] -
           sArg * iRho[d];
iFhD[p] += cArg * iRho[d] +
           sArg * rRho[d];
```

(c) **Coalesce** (Pixel data in register file, scan data in global memory, accesses to scan data coalesced, software sin/cos)

```
exp = x[p] * s[d].kx +
      y[p] * s[d].ky +
      z[p] * s[d].kz;

cArg = cos(exp);
sArg = sin(exp);

rFhD[p] += cArg * s[d].rRho -
           sArg * s[d].iRho;
iFhD[p] += cArg * s[d].iRho +
           sArg * s[d].rRho;
```

(d) **ConstMem** (Pixel data in register file, scan data in constant memory and constant cache, software sin/cos)

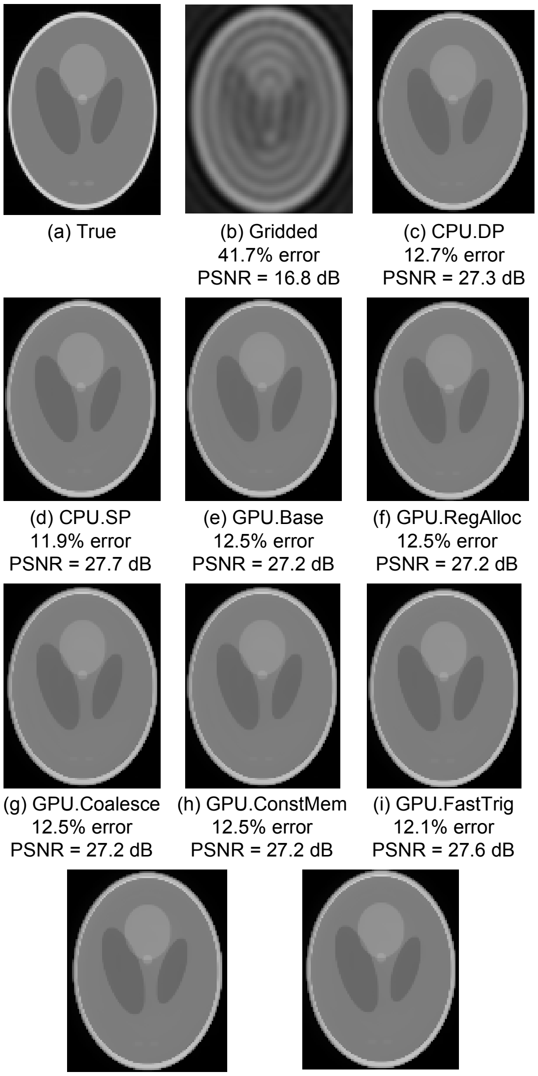
```
exp = x[p] * s[d].kx +
      y[p] * s[d].ky +
      z[p] * s[d].kz;

cArg = cos(exp);
sArg = sin(exp);

rFhD[p] += cArg * s[d].rRho -
           sArg * s[d].iRho;
iFhD[p] += cArg * s[d].iRho +
           sArg * s[d].rRho;
```

(e) **FastTrig** (Pixel data in register file, scan data in constant memory and constant cache, hardware sin/cos)

Figure 5: Versions of the F^{H}d algorithm on the GPU.



(a) True

(b) Gridded
41.7% error
PSNR = 16.8 dB

(c) CPU.DP
12.7% error
PSNR = 27.3 dB

(d) CPU.SP
11.9% error
PSNR = 27.7 dB

(e) GPU.Base
12.5% error
PSNR = 27.2 dB

(f) GPU.RegAlloc
12.5% error
PSNR = 27.2 dB

(g) GPU.Coalesce
12.5% error
PSNR = 27.2 dB

(h) GPU.ConstMem
12.5% error
PSNR = 27.2 dB

(i) GPU.FastTrig
12.1% error
PSNR = 27.6 dB

(j) GPU.Tune
12.1% error
PSNR = 27.6 dB

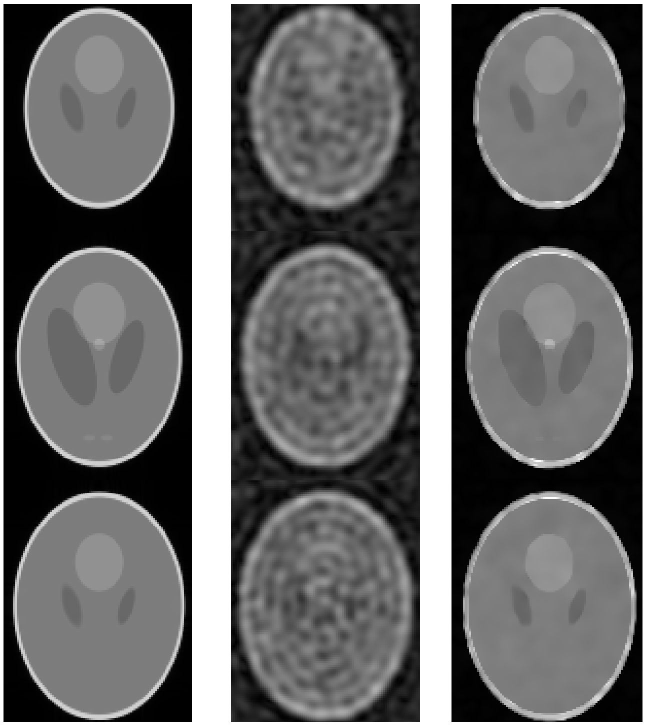
(k) GPU.Multi
12.1% error
PSNR = 27.6 dB

Figure 6: Noiseless data: One 2D slice of the 3D image. The percent error and PSNR values in each sub-figure caption are calculated over the entire 3D image.

the sample data, given the sampling trajectory, noise levels, and other readily available prior information [12].

5. EVALUATION

To be useful in clinical settings, the advanced reconstruction must satisfy two criteria. First, the quality of an image obtained via the advanced reconstruction should significantly exceed the quality of an image obtained via a gridded reconstruction. Second, the reconstruction must complete quickly. After image acquisition, the patient typically remains in the scanner during image reconstruction. The scanner operator then decides whether the image is acceptable or whether it should be acquired again. Any delays therefore increase patient discomfort and decrease scanner throughput. Also, when the administration of a medical treatment depends on the MR images, any delay is at best frustrating and at worst harmful to the patient's health.



(a) True

(b) Gridded
46.6% error
PSNR = 15.8 dB

(c) GPU.Tune
16.3% error
PSNR = 24.9 dB

Figure 7: Noisy data: Three 2D slices of the 3D image. The percent error and PSNR values in each sub-figure caption are calculated over the entire 3D image.

Our experiments indicate that the advanced reconstruction definitely satisfies the first criterion. As Figure 6 shows, advanced reconstruction of the noiseless data yields significantly better images than gridded reconstruction. Relative to the true image (Figure 6(a)), the advanced reconstructions (Figure 6(c-k)) exhibit 12% to 13% error and 27 dB to 28 dB PSNR, compared to 42% error and 17 dB PSNR for the gridded reconstruction (Figure 6(b)). There are no significant differences among the images obtained from the advanced reconstruction, despite the use of single-precision floating-point in Figures 6(d-k) and approximate trigonometric operations in Figure 6(c, d, and i-k).

The images reconstructed from the noisy data (Figure 7) further demonstrate the superiority of the advanced reconstruction. Relative to the true image, the advanced reconstruction exhibits 16% error and 25 dB PSNR, while the error and PSNR for the gridded reconstruction are 47% and 16 dB, respectively. Again, there are no significant differences among the images obtained from the various versions of the advanced reconstruction.

In addition to producing significantly better images than the gridded reconstruction, the GPU-accelerated advanced reconstruction arguably satisfies the second criterion for clinical use: speed. As Figure 8 shows, the fastest single-GPU version of the advanced reconstruction completes in less than 2 minutes (99 seconds, to be precise). This reconstruction

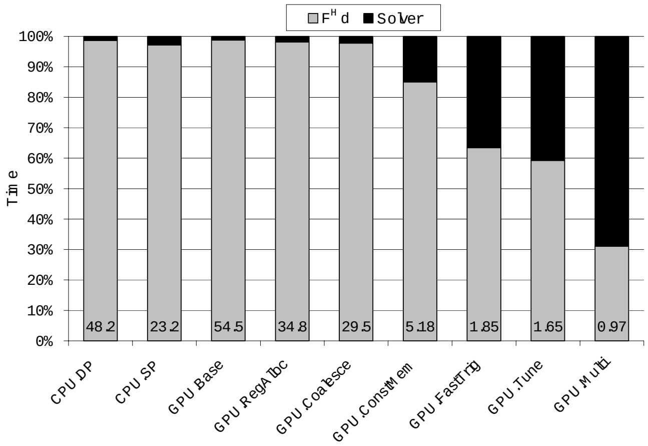


Figure 8: Performance of advanced MRI reconstruction. The reconstruction time includes the time to compute $\mathbf{F}^H\mathbf{d}$ and the time to run 60 iterations of the linear solver.

time is clearly much more appealing for clinical applications than the fastest CPU-based reconstruction, which completes in 23 minutes.

The fastest single-GPU version of the advanced reconstruction computes $\mathbf{F}^H\mathbf{d}$ in 59 seconds, compared to 22.5 minutes for the fastest CPU-based reconstruction. The remainder of this section describes how the advanced reconstruction leverages the GPU's resources to achieve such impressive acceleration when computing $\mathbf{F}^H\mathbf{d}$. We find that the constant memory caches are quite effective in reducing the number of accesses to global memory, while the special functional units provide substantial acceleration for the trigonometric computations in the algorithm's inner loops. We also find that experimentally-tuned code transformations have a significant impact on the algorithm's performance. Specifically, the algorithm's performance increases by 20% when the tiling factor, the number of threads per block, and the loop unrolling factor are experimentally tuned.

5.1 GPU.Base

As Figure 9 shows, GPU.Base is significantly slower than CPU.SP, the optimized, single-precision, quad-core implementation of $\mathbf{F}^H\mathbf{d}$. In GPU.Base (see Figure 5(a)), the inner loops are not unrolled. There are 256 threads per block and 256 scan points per tile. Given these parameters, each thread uses 13 registers. Therefore, up to 8192/13 = 630 threads can execute on each SM simultaneously, which represents 82% utilization of the G80's thread contexts.

Because GPU.Base leverages neither the constant memory nor the shared memory, memory bandwidth and latency are significant performance bottlenecks. With one 4-byte global memory accesses for every three FP operations, and with memory bandwidth of 76.8 GB/s, the upper limit on the kernel's performance is only 57.6 GFLOPS. Due to other performance bottlenecks, the kernel actually achieves only 7.0 GFLOPS, less than half of the 16.8 GFLOPS achieved by CPU.SP.

5.2 GPU.RegAlloc

Relative to GPU.Base, GPU.RegAlloc (see Figure 5(b)) decreases the time required to compute $\mathbf{F}^H\mathbf{d}$ from 53.9 minutes to 34.1 minutes. In short, register allocating the voxel data increases the computation intensity (the ratio of FP operations to off-chip memory accesses) from 3:1 to 5:1. This substantial reduction in required off-chip memory bandwidth translates into increased performance. Eliminating the two stores to global memory during every loop iteration is particularly beneficial.

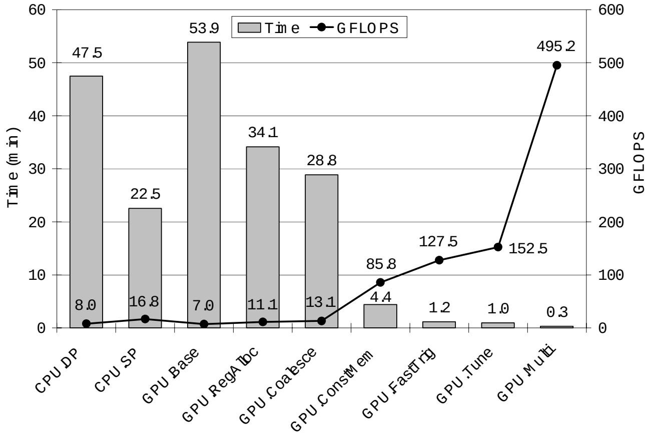


Figure 9: Performance of $\mathbf{F}^H\mathbf{d}$ computation. The first six configurations (CPU.DP - GPU.ConstMem) compute the trigonometric functions in software, using approximately 13 and 12 FLOPS for the sin and cos operations, respectively. The remaining configurations compute the trigonometric operations in hardware; therefore, each sin or cos accounts for a single FLOP.

5.3 GPU.Coalesce

On the Quadro, accesses to an aligned, continuous range of global memory addresses are coalesced into a single access, thereby conserving off-chip bandwidth. By changing the layout of the scan data in global memory (see Figure 5(c)), GPU.Coalesce enables memory coalescing, because the scan data addresses accessed by each warp of SIMD threads are always contiguous. Thus, GPU.Coalesce achieves an additional speedup of nearly 20% over GPU.RegAlloc. Nevertheless, GPU.Coalesce is still slower than CPU.SP.

5.4 GPU.ConstMem

GPU.ConstMem (Figure 5(d)) achieves speedup of 6.5X over GPU.Coalesce by placing each tile's scan data in constant memory rather than global memory. GPU.ConstMem therefore benefits from each SM's 8 KB constant memory cache. At 4.4 minutes and 85.8 GFLOPS, this version of $\mathbf{F}^H\mathbf{d}$ is 5X faster than the optimized CPU version.

We now analyze the off-chip memory accesses on a single SM during the execution of three thread blocks. With 7 global memory accesses per thread, 256 threads per thread block, and 3 thread blocks per SM, there are 5,376 accesses to global memory. Assuming no constant memory cache evictions due to conflicts, there are also 1,280 accesses to constant memory (256 data points per tile, with 5 floating-point values per data element), yielding a total of 6,656 off-chip memory accesses. The number of floating-point computations performed by the 3 thread blocks is 3*256*256*38 = 7,471,104. Thus, the ratio of FP operations to off-chip memory accesses has increased by over two orders of magnitude, from 3:1 to 1100:1. However, GPU.ConstMem still achieves only 85.8 GFLOPS (roughly 20% to 25% of the Quadro's peak theoretical throughput), which implies the existence of another bottleneck.

5.5 GPU.FastTrig

GPU.FastTrig (Figure 5(e)) achieves acceleration of nearly 4X over GPU.ConstMem by using the special functional units (SFUs) to compute each trigonometric operation as a single operation in hardware. When compiled without the **use_fast_math** compiler option, the algorithm uses implementations of **sin** and **cos** provided by an NVIDIA math library. Assuming that the library computes **sin** and **cos** using a five-element Taylor series, the trigonometric operations require 13 and 12 floating-point operations, respectively. By contrast, when compiled with the **use_fast_math** option, each **sin** or **cos** computation executes as a single floating-point operation on an SFU. The SFU achieves low latency at the expense of some accuracy. In our experiments (not shown), the images reconstructed by GPU.FastTrig often had lower percent error and higher PSNR than images reconstructed by GPU.ConstMem. In one reconstruction, however, the approximate trigonometric operations introduced significant additional error. Thus, while the SFU's approximate implementations of **sin** and **cos** often have negligible impact on the reconstruction's accuracy, further experimentation is necessary to determine the conditions under which these instructions may decrease the quality of a reconstruction.

5.6 GPU.Tune

While GPU.FastTrig overcomes the potential bottlenecks related to off-chip memory accesses and trigonometric computations, the algorithm still performs at only 127.5 GFLOPS, which is roughly one-third of the Quadro's peak theoretical performance. There are two culprits: instruction mix and resource utilization. When the inner loop is not unrolled, the ratio of overhead instructions (such as memory accesses, address calculations, and branches) to FP instructions is far too high. Unrolling the main loop decreases the ratio of overhead-to-FP ratio. However, the per-thread register usage also increases as the loop unrolling factor increases. Because the number of threads that can execute simultaneously is inversely proportional to the number of registers per thread, the loop unrolling optimization must carefully balance the competing goals of increasing the percentage of FP instructions and maintaining high utilization of the G80's cores [27, 28].

To determine the potential performance impact of experiment-driven code transformations, we conducted an exhaustive search that varied the number of threads per block from 32 to 512 (by increments of 32), the tiling factor from 32 to 2,048 (by powers of 2), and the loop unrolling factor from 1 to 8 (inclusive). Recent work has demonstrated that this type of experimental tuning can be performed quickly and accurately using static analysis techniques, as long as the code is parameterized correctly [28]. For reference, all previous configurations (GPU.Base - GPU.FastTrig) performed no loop unrolling and set both the number of threads per block and the tiling factor to 256. The exhaustive, experiment-driven search selects 320 threads per block, a tiling factor of 2,048, and a loop unrolling factor of 5. This configuration increases the algorithm's performance by 20%, with the runtime decreasing to 59 seconds and the throughput increasing to 152.5 GFLOPS.

5.7 GPU.Multi

In this final experiment, the voxels are divided into four distinct subsets, with one of four Quadros computing $\mathbf{F}^H\mathbf{d}$ for each subset. This optimization decreases the time required to compute $\mathbf{F}^H\mathbf{d}$ to 18 seconds and increases the throughput to nearly 500 GFLOPS. The acceleration is slightly sub-linear because the 3.5 second overhead required to marshal the data represents 25% of the time required to compute $\mathbf{F}^H\mathbf{d}$ for each subset of voxels. With $\mathbf{F}^H\mathbf{d}$'s runtime reduced to just 18 seconds, Amdahl's law is beginning to assert itself.

6. RELATED WORK

General-purpose computing on graphics processing units (often termed *GPGPU* or *GPU computing*) supports a broad range of scientific and engineering applications, including physical simulation, signal and image processing, database management, and data mining [24]. Medical imaging was one of the first GPU computing applications, with computed tomography (CT) reconstruction achieving a speedup of two orders of magnitude on the SGI RealityEngine in 1994 [5]. A wide variety of CT reconstruction algorithms have since been accelerated on graphics processors [19, 40, 7, 20], and the Cell Broadband Engine [3, 29]. In [20] the GPU is used to accelerate Simultaneous Algebraic Reconstruction Technique (SART), an algorithm that increases the quality of image reconstruction relative to the conventional filtered backprojection algorithm under certain conditions. SART, which requires significantly more computation than backprojection, becomes a viable clinical option when executed on the GPU.

By contrast, MRI reconstruction on the GPU has not been studied extensively. Research in this area has focused on accelerating the fast Fourier transform (FFT), which is a key component of many MRI reconstruction algorithms. Speedups on the order of 2x-9x have been reported [35, 30, 16]. In [33], Sørensen et al. use a GPU to accelerate a gridding algorithm for MRI reconstruction, achieving a substantial speedup over the baseline implementation. Finally, the acceleration of the advanced reconstruction algorithm described in this paper builds on our earlier work with the same algorithm [34].

7. CONCLUSIONS AND FUTURE WORK

In many applications, magnetic resonance imaging is limited by high noise levels, imaging artifacts, and long scan times. Advanced image reconstruction, which can operate on arbitrary scan trajectories and incorporate anatomical constraints, can mitigate these limitations at the expense of substantial computation. The computational resources, architectural features, and programmability of the Quadro FX 5600 reduce the time for an advanced reconstruction of non-uniform MRI scan data from 23 minutes on a quad-core CPU to less than 2 minutes, making the reconstruction practical for many clinical applications.

The single-precision floating-point arithmetic and approximate trigonometric operations that help accelerate the advanced reconstruction may also, under certain conditions, degrade the quality of the reconstructed image. We view further investigation of the advanced reconstruction algorithm's sensitivity to numerical approximations as important future work.

8. ACKNOWLEDGMENTS

The authors wish to thank Keith Thulborn and Ian Atkinson of the Center for MR Research at the University of Illinois at Chicago for assisting with an earlier version of this paper and for providing the scan trajectory used in some of our experiments. We acknowledge the support of the Gigascale Systems Research Center, one of five research centers funded under the Focus Center Research Program, a Seminconductor Research Corporation program. This material is based on work supported under two National Science Foundation Graduate Research Fellowships (Sam Stone, Justin Haldar). Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NSF. Experiments were made possible by generous donations of hardware from NVIDIA and Intel and by NSF CNS grant 05-51665. This work was supported in part by research grants NIH-P41-EB03631-16 and NIH-R01-CA098717.

9. REFERENCES

[1] C. B. Ahn, J. H. Kim, and Z. H. Cho. High-speed spiral-scan echo planar NMR imaging. *IEEE Trans. Med. Imag.*, 5(1):2–7, 1986.

[2] AMD Stream Processor. http://ati.amd.com/products/streamprocessor/index.html.

[3] O. Bockenbach, M. Knaup, and M. Kachelrieß. Implementation of a cone-beam backprojection algorithm on the Cell Broadband Engine processor. In *SPIE Medical Imaging 2007: Physics of Medical Imaging*, 2007.

[4] I. Buck. *Brook Specification v0.2*, October 2003.

[5] B. Cabral, N. Cam, and J. Foran. Accelerated volume rendering and tomographic reconstruction using texture mapping hardware. In *1994 Symposium on Volume Visualization*, 1994.

[6] Cg. http://developer.nvidia.com/page/cg_main.html.

[7] K. Chidlow and T. Möller. Rapid emission tomography reconstruction. In *Int'l Workshop on Volume Graphics*, 2003.

[8] DirectX Developer Center. http://www.msdn.com/directx/.

[9] J. Dongarra. Compressed Row Storage (CRS). http://netlib.org/utk/papers/templates/node91.html.

[10] J. A. Fessler, S. Lee, V. T. Olafsson, H. R. Shi, and D. C. Noll. Toeplitz-based iterative image reconstruction for MRI with correction for magnetic field inhomogeneity. *IEEE Trans. Signal Process.*, 53(9):3393–3402, 2005.

[11] J. A. Fessler and B. P. Sutton. Nonuniform fast Fourier transforms using min-max interpolation. *IEEE Trans. Signal Process.*, 51(2):560–574, 2003.

[12] J. Haldar, D. Hernando, S.-K. Song, and Z.-P. Liang. Anatomically-constrained reconstruction from noisy data. *Magnetic Resonance in Medicine (in press).*

[13] J. P. Haldar, D. Hernando, M. D. Budde, Q. Wang, S.-K. Song, and Z.-P. Liang. High-resolution MR metabolic imaging. In *Proc. IEEE EMBS*, pages 4324–4326, 2007.

[14] M. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436, 1952.

[15] J. I. Jackson, C. H. Meyer, D. G. Nishimura, and A. Macovski. Selection of a convolution function for Fourier inversion using gridding. *IEEE Trans. Med. Imag.*, 10(3):473–478, 1991.

[16] T. Jansen, B. von Rymon-Lipinski, N. Hanssen, and E. Keeve. Fourier volume rendering on the GPU using a split-stream FFT. 9th International Fall Workshop on Vision, Modeling, and Visualization, 2004.

[17] C. Koay, J. Sarlls, and E. Ozarslan. Three dimensional analytical magnetic resonance imaging phantom in the Fourier domain. *Magn. Reson. Med.*, 58:430–436, 2007.

[18] P. C. Lauterbur. Image formation by induced local interactions: Examples employing nuclear magnetic resonance. *Nature*, 242:190–191, 1973.

[19] K. Mueller, F. Xu, and N. Neophytou. Why do commodity graphics hardware boards (GPUs) work so well for acceleration of computed tomography? In *SPIE Electronic Imaging 2007, Computational Imaging V Keynote*, 2007.

[20] K. Mueller and R. Yagel. Rapid 3-D cone-beam reconstruction with the simultaneous algebraic reconstruction technique (SART) using 2-D texture mapping hardware. *IEEE Transactions on Medical Imaging*, 19(12):1227–1237, 2000.

[21] J. Nickolls and I. Buck. NVIDIA CUDA software and GPU parallel computing architecture. Microprocessor Forum, May 2007.

[22] NVIDIA Corporation. *CUDA CUFFT Library, version 1.1*, 2007.

[23] NVIDIA Corporation. *NVIDIA CUDA Programming Guide, version 1.1*, 2007.

[24] J. Owens, D. Luebke, N. Govindaraju, M. Harris, J. Krüger, A. Lefohn, and T. Purcell. A survey of general-purpose computation on graphics hardware. *Computer Graphics Forum*, 26(1):80–113, March 2007.

[25] K. P. Pruessmann, M. Weiger, P. Börnert, and P. Boesiger. Advances in sensitivity encoding with arbitrary k-space trajectories. *Magn. Res. Med.*, 46(4):638–651, 2001.

[26] S. Ryoo, C. Rodrigues, S. Baghsorkhi, S. Stone, D. Kirk, and W. m.W. Hwu. Optimization principles and application performance evaluation of a multithreaded GPU using CUDA. In *Symposium on Principles and Practice of Parallel Programming (PPOPP)*, 2008.

[27] S. Ryoo, C. Rodrigues, S. Stone, S. Baghsorkhi, S. Ueng, and W. Hwu. Program optimization study on a 128-core GPU. First Workshop on General Purpose Processing on Graphics Processing Units (GPGPU), 2007.

[28] S. Ryoo, C. Rodrigues, S. Stone, S. Baghsorkhi, S.-Z. Ueng, J. Stratton, and W. m.W. Hwu. Optimization space pruning for a multithreaded GPU. In *International Symposium on Code Generation and Optimization (CGO)*, 2008.

[29] M. Sakamoto and M. Murase. Parallel implementation for 3-D CT image reconstruction on Cell Broadband Engine. In *International Conference on Multimedia and Expo*, 2007.

[30] T. Schiwietz, T. Chang, P. Speier, and R. Westermann. MR image reconstruction using the GPU. In *SPIE Medical Imaging 2006*, 2006.

[31] H. Schomberg and J. Timmer. The gridding method for image reconstruction by Fourier transformation. *IEEE Trans. Med. Imag.*, 14(3):596–607, 1995.

[32] M. Segal and K. Akeley. *The OpenGL Graphics System: A Specification (Version 2.0)*. Silicon Graphics, Inc., October 2004.

[33] T. Sørensen, T. Schaeffter, K. Noe, and M. Hansen. Accelerating the non-equispaced fast Fourier transform on commodity graphics hardware. *IEEE Transactions on Medical Imaging (in press)*.

[34] S. Stone, H. Yi, J. Haldar, W. Hwu, B. Sutton, and Z. Liang. How GPUs can improve the quality of magnetic resonance imaging. First Workshop on General Purpose Processing on Graphics Processing Units (GPGPU), 2007.

[35] T. Sumanaweera and D. Liu. Medical image reconstruction with the FFT. In M. Pharr, editor, *GPU Gems 2: Programming Techniques for High-Performance Graphics and General-Purpose Computation*, pages 765–784. Addison-Wesley, March 2005.

[36] B. P. Sutton, D. C. Noll, and J. A. Fessler. Fast, iterative image reconstruction for MRI in the presence of field inhomogeneities. *IEEE Trans. Med. Imag.*, 22(2):178–188, 2003.

[37] D. Tarditi, S. Puri, and J. Oglesby. Accelerator: Using data parallelism to program GPUs for general-purpose uses. In *Int'l Conference on Architectural Support for Programming Languagues and Operating Systems (ASPLOS-XII)*, 2006.

[38] P. Trancoso and M. Charalambous. Exploring graphics processor performance for general purpose applications. In *Euromicro Symposium on Digital System Design, Architectures, Methods, and Tools (DSD 2005)*, 2005.

[39] F. T. A. W. Wajer. *Non-Cartesian MRI Scan Time Reduction through Sparse Sampling*. PhD thesis, Technische Universiteit Delft, Delft, Netherlands, 2001.

[40] X. Xue, A. Cheryauka, and D. Tubbs. Acceleration of fluoro-CT reconstruction for a mobile C-Arm on GPU and FPGA hardware: A simulation study. In *SPIE Medical Imaging 2006*, 2006.