Harnessing the GPU to Accelerate CT Image Reconstruction

Acceleware Whitepaper

Phone 403.249.9099
Web www.acceleware.com

Contents

Introduction **3**

Current Challenges **4**

Overview 4

Processing Power has Failed to Keep Pace with Other Advances in CT 4

Improving the Imaging Hardware 5

Improving the Algorithms 5

Improving the Processing 5

Surveying the Solution Landscape **7**

Backprojection on the GPU **9**

A Total CT Solution—Acceleware's AxRecon™ **10**

Removes the Backprojection Bottleneck 10

Easy Implementation 11

Reduces the Burden on IT 11

"Future-Proofing"—Allows Users to Keep Pace with Advances in Processor Technology 12

Flexible Workflow 12

Available at Reasonable Cost 12

"Battle-tested" Reliability 12

The AxRecon Advantage 13

About Acceleware **14**

Notes **15**

Introduction

New challenges for CT users:

- Advances in CT scanning hardware and supporting algorithms have increased the need for High Performance Computing (HPC) solutions.
- Solutions for processing power have in general failed to keep pace with advances in scanning technology, preventing CT users from realizing the full benefits of their investments in CT hardware.
- HPC solutions are typically intrusive, involving costly additional computer hardware, taxing the IT and development resources of medical organizations.
- Bottom line: medical organizations need an HPC solution that is cost-effective, non-intrusive, and does not unnecessarily burden their own IT and development resources.

The limitations on processing speed imposed by standard hardware and software solutions for image reconstruction constitute a significant bottleneck for Computed Tomography (CT) scanning performance. However, CT users are currently positioned to benefit from recent advances in processing technology—in particular the development of processing solutions called *accelerators* in the computer industry.

In this paper, we examine the challenges currently faced by hospitals, clinics, pharmaceutical companies and other organizations that employ or offer medical scanning services based on cone-beam CT scanners. We describe how processing power limitations can impose severe constraints on image reconstruction, leading to delays and significantly extended timelines for medical professionals who must analyze the data. And we explore how accelerators can be used to meet these challenges. We begin by surveying current options for accelerator technology. We show why recent innovations in GPU-based High Performance Computing (HPC) currently offer the greatest performance increase for the lowest cost, in addition to many other benefits.

GPU-based HPC allows users to leverage the performance of parallel processing for image reconstruction—using widely available commodity hardware. But while GPUs have gone mainstream, programming them for parallel processing can be difficult. Thus it is critical that users who want to harness the power of GPUs for CT find a solution vendor with as much expertise with the technology as possible.

Current Challenges

Overview

Clinicians and researchers employing CT scanning hardware are impacted by performance constraints inherent in the technologies used for image reconstruction. The image reconstruction process used in today's cone-beam micro CTs is complex and currently consumes much of the processing power available. In most cases, organizations are currently obtaining as much processing power as they can from single core and dual core PCs, and clusters. They are also using some acceleration technologies, such as Field Programmable Gate Arrays (FPGAs), Application-Specific Integrated Circuits (ASICs), and Cell Broadband Engine (CBE) boards. In all of these cases, performance falls well short of what available technology can deliver for an even lower cost, and in many cases the limitations are already exacting a toll on research, development, and clinical timelines. And although it is possible to further maximize speed with current computational resources, these gains have until now imposed a sacrifice in image quality.

Even worse, maintaining these sub-optimal hardware/software architectures often unnecessarily consume software development and/or IT resources that would ideally be better channelled to an organization's core business or research mandate.

Finally, with computer processing technology itself constantly changing, users are challenged to keep pace with new innovations in processing technology and require solution vendors that can help them transition to new technologies with minimal cost and disruption to workflows. This is often difficult, as hardware acceleration vendors are typically wedded to *specific* types of proprietary hardware. This may limit their agility and competitiveness when hardware advancements are made beyond their domain.

Processing Power Has Failed to Keep Pace with Other Advances in CT

In order to understand the factors that are currently impeding the performance of CT scanners, we need to take a closer look at how conventional CT scanners generate images that are useful in an experimental setting.

Generally speaking, there have been three ways in which to move CT technology forward: (1) improve the imaging hardware itself, (2) improve the supporting algorithms, and (3) increase processing power.

Improving the Imaging Hardware

Various performance and image quality advances have been achieved through hardware innovations. These include the development of helical, multi-detector, and dual-energy scanners. Improvements in scanner technology have brought increases in image resolution, increased scan lengths, and the development of time-resolved studies. They have also tended to increase the size of the datasets created by the image reconstruction process.

Improving the Algorithms

As improvements in CT hardware were added, developers were challenged to develop algorithms that would allow the innovations to be fully utilized.

It was only with the development of systems based on the three-dimensional Feldkamp, Davis, and Kress (FDK) algorithm that the promise of cone-beam technology—high-quality 3D images—began to be realized. This algorithm, and certain modified versions of it, has become the fundamental algorithm used for all CT scanning. The downside to the FDK algorithm, and all its variants, is that the complexity of it is $O(N^4)$, so as data size sets increase the processing time also increases dramatically.

While the FDK algorithm is still the one most commonly used for CT, there is much research underway to improve image quality still further by developing better algorithms. For example, iterative algorithms have been developed with an improved ability to handle metal artifacts, limited-angle tomography, truncated projections, and noise.[i] Additional gains in quality are also possible through the use of statistics-based iterative algorithms.[ii] However, these algorithms are still very much in the development phase, and although some have the potential to produce more accurate images, they are simply too computationally intensive to be used today.

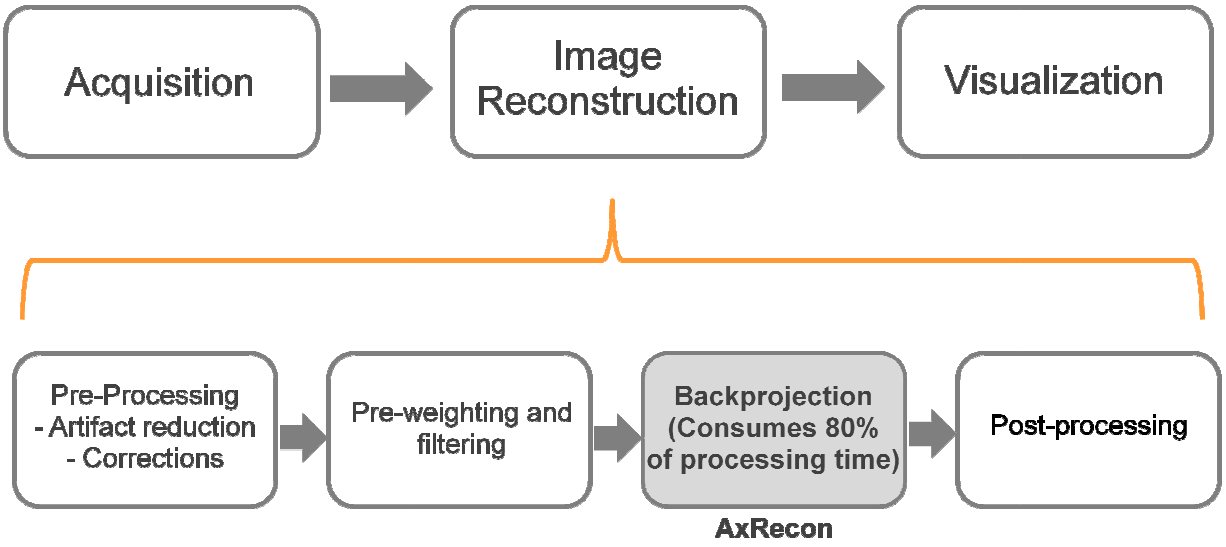
Improving the Processing

While they have led to impressive gains in image quality, both hardware innovations and algorithmic refinements have increased the need for more powerful processing.

The development of ever more sophisticated CT technology has led to dramatic increases in the size of data sets created in typical scanning sessions. Detector sizes have increased to the point where high resolution scans typically create multi-GB datasets. Indeed, recent improvements in hardware scanner technology mean that datasets of 32 GBs can now be obtained. Handling ever larger data sets requires increasing amounts of computing power.

And the rise in demand for higher quality results from complex 3D algorithms has also led to a greater demand for intensive computer processing.

Unfortunately, a lack of processing power currently impedes productivity in lab and clinical settings. The precise bottleneck in terms of CT image generation is the immense amount of processing required for backprojection, which traditionally consumes 80 per cent of all processing requirements for image reconstruction. The figure below situates backprojection in the context of the image reconstruction process.



To be of value in a laboratory or clinical setting, the processes that support high-quality CT scanning must be accomplished quickly. For settings in which multiple, repeated, CT scanning is conducted performance really matters—even small gains in processing performance add up over the course of many scans. Extending processing times by even a few minutes per scan can mean significant delays in a research program, missed deadlines, and an unacceptable

workflow. Researchers may find themselves frustrated by longer waits to view their data, due to the longer reconstruction times faced by everyone in the laboratory.

Increasingly sophisticated pre- and post-processing, such as artifact reduction, have created additional demands on computing resources. Optimal acceleration requires accelerating multiple components of the image reconstruction process.

Of the three types of advances in CT scanning technology—hardware, algorithms, and processing—it is increasingly clear that limited computer processing power has emerged as the most important single fetter on throughput. As a result, users are not realizing the productivity gains they could be obtaining from their investments in new scanning machines. Nor are they likely to fully realize the benefits of any future advances in scanning technology or the supporting algorithms without further increases in processing power.[iii] Users of scanning technology are therefore challenged to find computing solutions that can reduce scanning times by reducing the time required for image reconstruction, and help them obtain maximum benefit from their investments in scanning technology.

Moreover, faster processing solutions need to be more than fast—they must have minimal or negligible impact on the existing processes in the laboratory or clinical setting, with form factors of reasonable size.

Finally, they must be based on architectures that do not draw unnecessarily on the native IT and development resources of medical organizations—forcing them to channel these precious resources on efforts that are divorced from their core business or research mandates.

Surveying the Solution Landscape

What processing options are available for CT scanning that can be quickly and easily implemented in actual medical settings without impacting the way medical professionals actually work, or imposing unwarranted burdens on the IT and development resources of medical organizations?

CPU Clusters (multi-core): This solution involves a standard distributed-computing architecture, in which multiple CPUs are each tasked with a part of a processing problem. Adding multiple cores increases the overall amount of processing power and aggregated memory bandwidth available, which can increase speed. However, although each processor in such a configuration may be intrinsically fast, performance is compromised as each processor must wait for data to be processed. CPUs also bear the burden of supporting the computer's operating system, which draws processing power away from the calculations.

Field-Programmable Gate Arrays (FPGAs): These are a semiconductor devices consisting of programmable logic components and interconnects. FPGAs have a flexible architecture, and can complete specific calculations, very quickly once the data is on the chip. However, their relatively low memory bandwidth impedes their performance, they are relatively expensive, and are not highly scalable. As well, with FPGAs a large portion of real estate is used for interconnections, not processing. And their floating point precision consumes a lot of resources. FPGAs run at an average of approximately **21 GBPS**.

Cell Broadband Engines (CBEs or "Cell processors"): The Cell processor was jointly developed by Sony Computer Entertainment, Toshiba, and IBM. Cells combine a general-purpose core and co-processing elements designed for multimedia applications. Cell technology achieves higher performance through a greater number of streaming processors (more cores) and a unique memory bus architecture. Nonetheless, its memory bandwidth is not as high as GPUs—still averaging approximately only **21 GBPS**—and it is relatively expensive compared to GPUs. They are also widely recognized as being difficult to program.

Graphics Processing Units (GPUs): GPUs are a type of processor that has been developed by NVIDIA® and other companies to support computer animation and gaming software. Generally speaking, graphics processing involves applying the same calculation steps to many cores in parallel. GPUs operate at a peak of **76.8 GBPS**, which is far faster than other technologies. GPUs have a maximum speed of 500 GFLOPS, achieved through *parallel processing*—performing each calculation simultaneously using multiple streaming processors. GPUs feature memory bandwidths that are three to four times as large as those of other technologies, which reduces the need for data to be moved to and from the card and minimizes the latency this can produce. They are also relatively inexpensive and highly scalable.

Backprojection on the GPU

Because of the advantages GPUs present in terms of cost, scalability and performance, many industries have been drawn to them for their HPC needs. In fact, GPUs have already been successfully used in many areas that demand HPC levels of computing performance and whose computing challenges can be met through parallel processing—for example, electromagnetic field simulation and seismic data processing.

And since CT scanning is both computing intensive and amenable to parallel processing—here too with great advantage in terms of performance—using GPUs for CT represents the natural extension of a proven technology to a familiar processing challenge.

In fact, the GPU is an optimal tool for CT reconstruction.[iv] GPUs work best for parallelizable calculations, and backprojection—the processing bottleneck—is highly parallelizable. Backprojection is a "gathering" operation—it involves collecting many input values to map each voxel to a specific location on the projection plane and calculate a value for that voxel. Although calculating voxel values is itself computing intensive, each of these calculations is largely independent of the others, allowing for simultaneous—as opposed to more time-consuming sequential—processing.

GPUs are designed for exactly this type of computing task. Backprojection is in many respects similar to the texture-mapping operations that support graphics-intensive applications like computer games. Both require many lookups into large sets of data and many mathematical calculations. The large memory bandwidths on GPUs are particularly suited to maximize the speed of this type of computing task.

The following table compares GPU-based performance with that of the other possible architectures.

Technology	Performance*
*Reference size: 512 projections onto a 512 x 512 x 512 volume (backprojection only)	
Single-core CPU	3.21 minutes[v]
FPGA	25 seconds[vi]
Cell Processors	17 seconds[vii]
Acceleware's AxRecon Solution	8 seconds

A Total CT Solution—Acceleware's AxRecon™

In addition to the sheer performance of the underlying hardware, there are other factors that should be considered when assessing the viability of technologies to accelerate CT reconstructions. CT users need more than just accelerated reconstruction times. They need a solution that combines high performance with high image quality—one that is easy to use, implement, and maintain without physically intruding on the work setting.

Acceleware, a leading developer of HPC solutions for a variety of industrial, business and research applications, has recently sought to address these key requirements with their AxRecon total solution for CT processing. AxRecon is a complete GPU-powered combined hardware and software solution that is easy to add to a CT system with no disruption in workflow. The AxRecon solution includes a GPU supercomputer, with a high-level library exposed through an API for easy programming and configuration.
AxRecon provides the following benefits.

Removes the Backprojection Bottleneck

AxRecon leverages the power of the NVIDIA GPU to create a deskside supercomputer that dramatically reduces reconstruction times—typically by a factor of 20 to 35x. And with a choice of one, two, or four card solutions, there is an AxRecon solution to help you sustain that magnitude of performance gain on even the largest datasets.

NVIDIA's current line of computing GPUs operate at a peak of 500 GFLOPS, effectively providing cluster-sized compute capability in a card-sized form factor. The Acceleware ClusterInABox™ solution combines four GPUs to deliver two Teraflops of peak performance. By comparison, a state-of-the-art quad core processor delivers only approximately 100 GFLOPS.

> "A volume that takes approximately 15 min (on a good day) to reconstruct, takes only three minutes on Acceleware's solution."
>
> *Steven Pollmann,*
>
> *Robarts Research Institute*

Easy Implementation

A standard implementation of AxRecon involves connecting NVIDIA hardware to an existing desktop, and integrating the AxRecon software platform. The underlying hardware has a physically small footprint: typically it is a deskside unit containing two, or four NVIDIA GPUs (depending on the needs of the user), connected by a cable to the desktop via the desktop's PCI Express x16 slot. A single GPU is installed directly inside the workstation, sometimes with no need for an additional box.

Acceleware offers high level libraries which are easily integrated with existing software. This allows integrators to keep their optimized data conditioning, customized data handling, and user interfaces.

The AxRecon library has a standard C interface, which makes it compatible with all of the popular programming languages and software applications, and the AxRecon API allows for quick and easy integration, allowing users to take full advantage of the additional performance provided by the GPU.

Reduces the Burden on IT

AxRecon reduces the burden on an organization's in-house IT resources, allowing them to channel those resources to areas that are more germane to their core business or research goals. And as one IT professional at a leading pharmaceutical company recently observed, their AxRecon implementation had other ongoing benefits for their organization, including a reduced foot print, lower power consumption, lower cooling requirements, and more effective system administration.

> "From an IT perspective, ClusterInABox is advantageous in terms of reduced foot print, power consumption, cooling requirements, and system administration."

"Future-Proofing"—Allows Users to Keep Pace with Advances in Processor Technology

GPUs and other acceleration hardware are constantly evolving. Today's cutting edge hardware will inevitably become outdated, supplanted by succeeding generations of hardware with steadily increasing performance capabilities. Acceleware takes this into account by providing an application platform that abstracts the ability to process scans in parallel from the acceleration hardware. Although Acceleware currently favours the GPU as the best choice for acceleration, the company is committed to providing an advanced software platform to allow customers to continue to take advantage of ongoing innovations in processing technology.

Flexible Workflow

AxRecon is designed to make your workflow easier and simpler—with far fewer intrusions. AxRecon's high level libraries are designed to give integrators maximum flexibility, with a configurable solution whereby you can choose which accelerated features you want to implement.

Available at Reasonable Cost

AxRecon is not based on expensive "boutique" hardware but on a widely available commodity device—the NVIDIA GPU, which helps to reduce the purchase and upgrade cost. The large economies of scale in the gaming market have acted to reduce the cost of GPUs, with consumer demand essentially supporting their development.

"Battle-tested" Reliability

AxRecon is based on the NVIDIA GPU, the world's largest manufacturer of GPUs. NVIDIA spends an average of $400 million each year on R&D alone to develop GPUs and subjects all its products to exhaustive functional and stress testing. NVIDIA's latest products are being designed for high-duty cycle scientific applications.

The AxRecon Advantage

Problem	Solutions			
	CPU Cluster	Cell	FPGA	AxRecon
Reconstruction takes too long	120 to 400 mega-backprojections/ second (mBP/s).	Faster than CPU clusters, but much slower than AxRecon due to slower processing.	Faster than CPU clusters, but much slower than AxRecon due to slower processing and lower memory bandwidth.	**Up to 12000 mega-backprojections/second (mBP/s)** **Far faster than any competing solution—near real-time!**
Cost	Relatively low hardware cost.	Slow market adoption means higher hardware costs.	Higher cost of implementation.	**Large economies of scale, gaming and visualization market drives low cost.**
Maintaining the software and hardware consumes precious IT/developer resources	Difficult to integrate. Integrators assume programming burden, which requires MPI parallelization expertise and a large IT management commitment.	Difficult to integrate. Integrators assume programming burden. Difficult to program due to a very unique architecture and a relatively immature programming model.	Difficult to integrate. Integrators assume programming burden. Difficult to program due to different hardware architectures, program flow, and development tools.	**Acceleware virtualizes hardware from integrators and users—no need for integrators to worry about programming hardware.** **Easy to integrate—allows organizations to allocate technical resources to support core business and research goals.**
Hardware takes up too much space and consumes too much power	Largest form factor. Highest power and cooling costs.	Smaller form factor than CPU. Reduced cooling cost.	Smaller form factor than CPU. Reduced cooling cost.	**Low TCO, small footprint, lowest power consumption per compute capacity—consumes about half the power of a CPU to compute a specific task.**
Difficult and expensive upgrade path	Users must usually research and implement solutions on their own.	Hardware specific vendors increase reliance on specific hardware platform.	Hardware specific vendors increase reliance on specific hardware platform.	**Acceleware "future proofs" to allow customers to achieve the best performance possible, using the best available technology.**
Vendors offer incomplete solution	Limited support for software, card programming, drivers, etc. Your application is not guaranteed to work on any cluster that you or your IT department decides to build.	Limited support for software, card programming, drivers, etc.	Limited support for software, card programming, drivers, etc.	**Acceleware includes all hardware, APIs, libraries, drivers, for a total solution.**

About Acceleware

With over seven years of domain experience in acceleration technology, Acceleware is at the forefront of solutions to help medical and research organizations work more quickly and efficiently. Our customer list spans several industries, and we are able to draw on this extensive record of success to develop the solutions for accelerated reconstruction that CT users have been waiting for. Our experience positions us for constant innovation, and gives us the ability to evaluate new technologies as they become available to determine their suitability for our customers' needs.

Acceleware's Customers Include

- Philips
- Boston Scientific
- St. Jude Hospital
- Medtronic
- Sony Ericsson
- Fujifilm
- LG Electronics
- Hitachi
- Nokia

Our current "state of the art" GPU-based solutions for accelerated cone-beam CT reconstruction are just the beginning. We're working to continuously improve the FDK algorithm, as well as the overall AxRecon platform for performance, image quality, and ease of use. Our efforts don't stop at backprojection. You can trust Acceleware for solutions based on rigorous analyses of every aspect of CT technology. Even now our researchers are working on new technology to accelerate forward projections, on cutting-edge iterative algorithms for improved image quality, and on various pre-processing steps to further enhance both clarity and speed.

We're already looking beyond the GPU to the next generation of acceleration technologies, striving to develop solutions to make high quality real-time CT scanning a reality, and ensure that our customers continue to take advantage of the best technology available.

Notes

[i] Jian Hsieh, “Computed Tomography Principles, Design, Artifacts, and Recent Advances,” 2003, SPIE Press, Bellingham, Washington, USA.

[ii] Ibid.

[iii] Steve Hamm, Business Week, April 12, 2007. http://www.businessweek.com/technology/content/apr2007/tc20070412_406858.htm?chan=technology_technology+index+page_computers.

[iv] K. Mueller, F. Xu, and N. Neophytou, “Why do Commodity Graphics Hardware Boards (GPUs) Work so Well for Acceleration of Computed Tomography?” SPIE Electronic Imaging 2007, Computational Imaging V Keynote.

[v] Marc Kachelrieß and Olivier Bockenbach, “High Performance 3D image Reconstruction Platforms.” Boards and Solutions, July 2007.

[vi] Ibid.

[vii] Ibid.

For more information about how Acceleware can help you dramatically reduce your image reconstruction times, please contact Acceleware at:

Email: imaging@acceleware.com
Tel: +1.403.249.9099 **Fax:** +1.403.249.9881
Address: 1600 37th St SW, Calgary AB, T3C 3P1