Current Trends in Biomedical Imaging and the Role of the Reconfigurable Computing Platform of FPGA

Radha Guha
Pacific Lutheran University
guhara@plu.edu

***Abstract* - Tremendous advancement of technology in computer science and engineering discipline is recently aiming at improved patient care in a cost effective way by using more sophisticated techniques and patient health monitoring devices. Advancements in computer vision, virtual reality and robotics technology can be applied in rendering images of internal regions of the human body and for developing image-guided surgery by physicians or by medical robots. Detail visual representation of patient data helps doctors and researchers in analyzing the data, in tracing diseases and in quick decision making. With this goal of improved patient care, the newly emerged biomedical imaging field is facing the challenge of real time quantitative processing of large data sets of vital signals acquired from patient body by various compute intensive algorithms to extract important information, its visualization and efficient management of storing and retrieving of patient records. This research paper first overviews state-of-the-art technologies and terminologies evolved in the rapidly advancing biomedical imaging field. Secondly the paper explores the immense parallel processing power of the reconfigurable computing platform of field programmable gate arrays (FPGAs) as a hardware accelerator which can support improved patient care goal by designing faster, smaller, more lexible and less expensive medical imaging devices which can be brought to market quickly.**

Keywords – Bio Medical Imaging, High Performance Computing, Parallel Processing, FPGA

I. INTRODUCTION

Evolution of Biomedical Imaging Standards In this section, state-of-the-art technologies and terminologies evolved in the rapidly advancing biomedical imaging field is introduced [1][2][3][4]. The requirements of efficient digital imaging systems are analyzed and supports from different vendors of commercial applications are overviewed here.

Biomedical imaging is defined as taking images of internal aspects of human body i.e. images of anatomical or functional measurements in the scale varying from organ or tissue level to cellular and to molecular level by various non-invasive procedures. In the last decade use of various medical imaging procedures like X-ray, Ultra-sound, computed tomography (CT), magnetic resonance imaging (MRI) and positron emission tomography (PET) etc. have increased exponentially for taking images of the internal aspects of patient body by penetrating some kind of electromagnetic waves in the body. Now-a-days biomedical imaging is regularly used in cardiology, endoscopy, obstetrics, gynecology, brain and abdominal imaging for early detection of diseases and their treatments. Recently these imaging modalities have switched from film-based imaging to instant digital imaging on a display monitor.

Efficient digital imaging system requires instant data acquisition and recording in computer memory, high speed real time digital signal processing (DSP), rendering, and efficient and reliable image archiving and communication system for storing and fast retrieving of images. Biomedical image rendering or visualization needs accurate and high-speed modeling and simulation of huge amount of patient data of anatomical and functional measurements to produce high-resolution two-dimensional (2D) or three-dimensional (3D) images at real time. Also modeling and simulation of time variant four-dimensional (4D) dynamic physiological systems in human body such as glucoseinsulin dynamics, blood-alcohol dynamics, blood-drug dynamics and recovery stages of wounds etc. can help to accelerate biomedical research and training not only limited to disease diagnosis but also drug discovery and developments. The characteristics of this digital data processing are high-speed computation, high volume of data storage or memory requirement and high bandwidth of input and output (I/O) data communications to and from processors and memory.

For efficient and reliable storing and communication of images coming from different medical imaging scanners a new picture archiving and communication system (PACS) standard [1] has been created. PACS has the benefit of instant access and transmission of images over the internet and lower storage cost than the traditional film-based imaging system. Often time, images from multiple modalities like CT, PET, and MRI etc. are combined together to complement information for better and faster decision making. Thus digital imaging and communication in medicine (DICOM) standard is created in order to integrate different file formats in PACS and to exchange image files between any two machines which understand DICOM file formats and communication protocols. Efficient digital imaging system also requires high speed internet access and efficient indexing system to retrieve the correct set of images from huge medical database. Just like any other data, image data coding and decoding for security purposes and image compression and decompression for savings on storage and faster transmission over band-width limited network are applied for biomedical images also.

DICOM metadata of any image file, like patient id, physician id, imaging modality, image description, image size etc. are structured in XML file format which can be searched efficiently using XPath query language. DICOM standard provides about 2000 standard tags to describe an image which can be indexed for easy retrieval. DICOM unstructured image data is stored as a binary large object (BLOB) in a separate file. DICOM object model provides standard methods such as update, copy, delete etc. to make application programming easier. Oracle multimedia DICOM [3] supports efficient and secure storage of large-scale DICOM contents in a large database so that medical records can be shared efficiently and securely by authorized users at departmental, regional and national level in a cost effective way. Oracle 11 g supports DICOM data type to store DICOM content as a column in a table. Largescale DICOM contents can be populated in Oracle database for manipulation of data like select, update and delete with SQL query language. DICOM data from different tables can be joined to create views like any other data types.

Adobe Photoshop creative suite (CS3) extended [4] can help the researchers in high quality biomedical image visualization and fast image analysis as it supports high resolution 16/32 bits color images, 300,000*300,000 pixels spatial resolution, DICOM file formats along with other file formats like JPEG, TIFF, GIF, PNG, BMP etc. and has a range of tools for image analysis like filtering, image enhancement, measurement, counting, image overlays, annotating image with text etc. Adobe CS3 also integrates to MATLAB which can process data with any complex algorithm and send the data back to Adobe Photoshop for visualization and for publishing. Using MATLAB and Simulink [3] dynamic physiological system such as diabetes diagnosis, arterial circulation, respiratory system, cardiovascular system, blood pressure etc. can be modeled and simulated easily and analyzed quantitatively.

Mathematical model of the underlying system can be built with differential equations and solved using MATLAB and Simulink; which helps engineers and engineering students to understand a complex system.

II. HIGH PERFORMANCE NEED FOR BIOMEDICAL IMAGE PROCESSING ALGORITHMS

For furthering the research of biomedical imaging field, high performance need of biomedical image processing is analyzed in this section. Biomedical image processing algorithms can be divided into two groups:

1. Image reconstruction algorithms and
2. image analysis algorithms.

Tomographic (cross sectional) image reconstruction algorithms are used for generating images from signals acquired by various imaging instruments like X-Ray, Ultrasound, CT, MRI and PET by penetrating or projecting many rays in the body varying uniformly in the angle of 1800 or 3600. Image reconstruction from the acquired data is an inverse problem and as the number of rays increased it gives more accurate image of the original object. Image reconstruction algorithms are of two types: 1. filtered back projection (FBP) algorithms and 2. iterative reconstruction (IR) algorithms. Filtered back projection algorithm uses inverse Radon transform method. In 1917 Australian mathematician Johan Radon first discovered projection of a 2D image f(x, y) as a set of line integrals by projecting parallel beam from multiple sources in a certain direction. He also devised the inverse of creating the 2D image from a number of projections from different angles. But it was not used until 1972 when Hounsfield introduced computerized tomography using Radon’s reconstruction formula.

Image reconstruction is very compute intensive because the amount of data acquired from a single cross section scan is huge for CT, MRI and PET modalities and image reconstruction requires several steps like back projection, filtering, fast Fourier transform (FFT), interpolation, convolution, noise reduction, and segmentation etc. Iterative reconstruction (IR) algorithms are more compute intensive than filtered back projection algorithms but produce more clear and accurate images. Obviously 3D volume visualization from a number of cross sectional images is much more compute-intensive than reconstructing 2D cross sectional image. 3D volume visualization goes through compute intensive ray tracing, alpha blending and correlation of 2D images coming from different cross sections or from different modalities for video image stabilization and registration (VISAR) purposes. Time variant 4D volume visualization is obviously much more compute intensive than 3D volume visualization.

The second type of biomedical image processing algorithms is for quantitative image analysis by digital manipulation of the original image to produce a modified image in multiple steps. These are for classification of images, pattern recognition, feature extraction, segmentation, image enhancement, restoration, texture, shape, motion measurements and spectral analysis etc. based on geometrical, statistical, physical or functional models.

Various compute intensive techniques of computer vision and machine intelligence like transforming time domain data to frequency domain data, filtering by convolution, principle component analysis (PCA), hidden Markov model, anisotropic diffusion, self-organizing maps (SOM), partial differential equations (PDE) and artificial neural network (ANN) etc. are applied for biomedical image analysis.

All image analysis procedures are very compute intensive. As for example continuous Fourier transform is very compute intensive. When complex algorithm like

Fourier transform is applied to data intensive application the computational work load becomes very high. Computers do not have the processing power to work with infinitely long continuous signals. Discrete Fourier Transform (DFT) is the approximation of the continuous Fourier transform and works on a finite length of signal sampled at discrete points. DFT produces discrete number of frequencies of the sampled signal. Though DFT is an O(n2) algorithm where n is the number of sample points, but to process a signal with 109 data elements and with the 30 frames/sec video display requirements, the total number of data operation required is [(109)2]*30 = 30 exo flops. Exo flops are not yet conceivable in computer speed terminology.

Fast Fourier transform (FFT) is a technique which uses divide and concur paradigm of problem solving to solve the problem in O(nlogn) time and the same problem will take 270 Giga flops of operations. If each data is represented by 3 colors, red, green and blue and each having resolution of 14 bits then data operation rate required is 270*109*14*3 = 12*1012 = 12 Tera bits per second. Currently available processing speed of 3 GHz, quad core Intel-Xeon processor which can perform two operations in one cycle can process only 3*2*4*109=24 Giga flops per second and is not sufficient for real-time simulation need of volume visualization. Thus compute-intensive biomedical image processing which requires high quality images and high speed rendering is not possible by a single microprocessor and requires parallel processing.

III. PARALLEL PROCESSING

Unless the algorithms or applications exhibit substantial amount of parallelism they are not suitable for parallel processing as parallel processing also introduces communication and synchronization overhead which degrades the performance measures of the applications. Communication overhead is due to control instructions communication between the controller and the processing elements and also for data transfer from memory to processing elements.

Usually application program exhibits two kinds of parallelism: 1. fine-grained instruction level parallelism and 2. coarse grained task level parallelism. Fine-grained instruction level parallelism can be exploited efficiently by deep pipelining whereas coarse-grained task level parallelism can be exploited efficiently by parallel processors. Also applications are never hundred percent data-parallel or task-parallel as there are also sequential control intensive instructions which are suitable for a general purpose microprocessor or CPU. Thus design of a hybrid system with CPU controlling and synchronizing the operations between the CPU, HW accelerator board and external memory is required.

Algorithms which need more mathematical operations per data element can compensate the communication overhead of round trip data transfer cost from external memory to processor and back to external memory over the low data rate, bandwidth limited bus and thus are more suitable for parallel processing on hardware accelerator. Also applications which has high arithmetic density i.e. the same operation is applied on multiple data elements in parallel can reduce control flow data path cost. For these high density algorithmic or arithmetic data streaming applications, hardware platform can be designed with more number of parallel processors and less amount of cache memory and control hardware.

Parallel processing can be supported by general purpose multi-core processors design to a great extent. Though contemporary Intel's general purpose quad-core can support high-performance need of biomedical image processing by its quad cores to a great extent, for the insatiable demand of more compute intensive real-time high-definition image processing, dedicated graphics processor unit (GPU) is designed (Figure 1). GPU has specialized parallel circuitry which can manipulate large block of image data in a frame buffer more efficiently than the general purpose CPU. From the year 1999, Nvidia, ATI like GPU vendors are making 3D visualization of the real world pervasive by performing the compute intensive rendering tasks of transformation, lighting, texture mapping, triangle setup/clipping etc. of more than 10 million polygons per second.

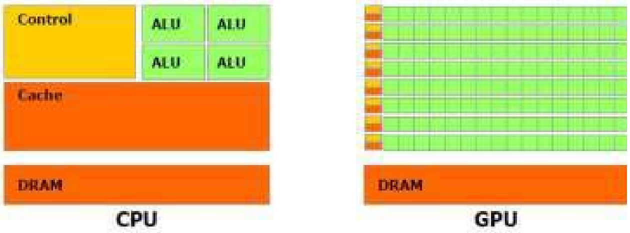

Figure 1: GPU devotes more transistors to data processing than cache and control hardware [6]

Graphics processor unit (GPU) has evolved to general purpose graphics processor unit (GPGPU) [6] with parallel data architecture with many core processors and high memory bandwidth for multithreaded programming. GPGPU has programmable stages which is suitable for streaming non-graphic data also. Thus CPU can offload some of its compute intensive tasks to GPGPU. In 2006, NVIDIA introduced CUDA [6], a general purpose parallel computing architecture for GPGPU by adding few extensions to the standard high level programming language C to map available application parallelism to hardware parallel architecture transparently. CUDA provides three abstractions for hierarchy of thread groups, shared memory and barrier synchronization which helps the programmer to partition the problem into coarse grained task parallel sub problems first and then fine grained data parallel sub-sub problems.

A compiled CUDA program can run on any number of parallel processors which is only discovered at the run time of the program.

GPGPU is more suitable for exploiting task level parallelism than fine-grained data instruction level parallelism.

Moreover competition among medical imaging vendors are fierce to manufacture smaller, portable, cart held or hand held medical monitoring devices with higher performances than before [7][8]. Portable devices have the advantage that they can reach patients at their homes instead of patients being brought in to the hospital and this idea is revolutionizing the health care industry. Portable medical devices are battery operated and needs low power consumption processors. Digital signal processors (DSP) are ideal for portable medical imaging devices as they can support high work load of the medical images and has low power consumption. DSPs can lengthen battery life and reduce battery size and the device size. Recently Texas Instruments [8] has started focusing on medical imaging field and has a suite of 15 DSPs of single core, multi core and system on chip (SOC) variety for tackling high performance, reliable and low power consumption need of the portable devices.

But as the medical imaging field is rapidly advancing many novel algorithms and techniques are required to be brought to the market quickly. To lower the cost of patient care medical devices need to be flexible so that new algorithms or modified version of the same algorithms can be uploaded in the same device. This flexibility can be accomplished only by the reconfigurable computing platform of field programmable gate arrays (FPGAs). Also finegrained instruction level parallelism and coarse-grained task level parallelism both can be exploited more efficiently in this reconfigurable FPGA platform. Designing circuits in FPGA technology eliminates the non-recurring design cost and longer time to market characteristics of application specific integrated circuits. In comparison to other options of parallel computing i.e. by supercomputers, PC clusters, multi-core processors, GPUs, DSP and ASIC hardware, the reconfigurable FPGAs are the most cost effective platform for low latency, low power and high performance applications.

IV. BIOMEDICAL IMAGE PROCESSING ON FPGA PLATFORM

Advancement in VLSI technology can pack billions of transistors on field programmable gate array (FPGA) chip [9][10][11][12][13][14][15] providing larger real estate and more opportunity for parallelism. FPGA comes with programmable configurable logic blocks (CLBs), input output blocks (IOBs), block random access memory (BRAM), interconnects, clock resources and configuration circuitry [9], [10], [11]. CLBs are lookup table (LUT) based where a k input LUT can store any logic function of k-binary variables. The interconnections between the CLBs are a two dimensional network of connectors and routers called connection network. Today Actel, Altera, Atmel, Cypress, Lattice, Xilinx are popular FPGA manufacturers. Recently, Xilinx's Virtex-7 [10] chip, which is built on 28 nm (nano meter) triple oxide process technology, comes with 1955,000 logic cells, 1200 I/O pins, 22 Mbits distributed memory, 47 Mbits block memory, 427 Mbits configuration memory, low power Rocket I/O serial transceiver, built-in PCI Express endpoint and Ethernet MAC blocks. PCI Express slot provides 8 Gbps (giga bits per second) per line data transfer rate to and from the FPGA chip. This chip provides complete system level integration capability with external memory, CPU and other h/w boards which shrinks design cycle time and system cost. System on chip (SOC) board can easily accommodate DSP, RISC processor and FPGA accelerator which can be portable and has low power consumption.

In a fine-grained data parallel programming model the distributed small memories are used as registers, register files or level 1 (L1) cache. Whereas in coarse grained data parallel programming model larger block memories called embedded memory blocks (EMBs) are used as level 2 (L2) cache. One EMB may be of a few CLBs in height and width. For optimal use of the chip area these EMBs must be fully utilized by the applications. Modern FPGAs also provide fast and dedicated ASIC quality arithmetic operation units such as multiply and accumulate (MAC) units as DSP blocks for computation intensive algorithms. The CLBs, IOBs, DSPs and EMBs are all run-time reprogrammable on the order of milliseconds to create an efficient interface between the logic data path and the memory. FPGA provides high processing power, high memory band width and high connectivity between parallel processing elements. According to the requirement of the parallel algorithm a number of processing elements can be configured to operate concurrently with optimal control and data paths and the compute intensive part of an algorithm can be off-loaded to this hardware accelerator.

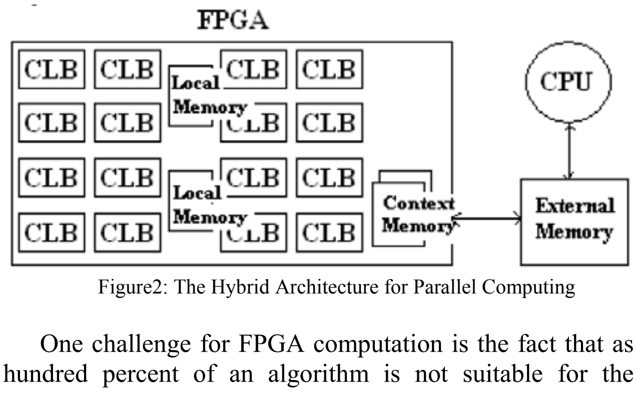

Figure2: The Hybrid Architecture for Parallel Computing

One challenge for FPGA computation is the fact that as hundred percent of an algorithm is not suitable for the parallel architecture of the hardware, a microprocessor is still needed for sequential execution of the control intensive parts (Figure 2). The microprocessor is also needed for hosting a real-time operating system (RTOS) for reconfiguration, hardware resource management and task scheduling and mapping on FPGA. The main challenge is to segregate the sequential part of the algorithms for general-purpose microprocessor and parallel part for parallel architecture of FPGA, providing sufficient communication bandwidth between CPU and FPGA and within FPGA between multiple processing elements.

The goals of biomedical imaging field for designing more powerful smaller portable and flexible medical devices in a cost effective way can be satisfied by this parallel architecture of FPGA because it is also reconfigurable statically or dynamically at run time for application specific optimization and has higher resource utilization. The reprogrammable FPGA chip can be configured as many times as required statically or dynamically during its run time after it is manufactured. The huge capacity of FPGAs can configure thousands of parallel processing elements (PEs) especially for low precision computations with customized data-path widths of 1, 2, 3, 4, 5, 6, 7, 8 bits etc. in contrast to the 32 or 64 bits fixed data-path widths in general-purpose microprocessor architecture. Though the FPGA clock frequency is order of magnitude slower than the microprocessor clock frequency, the huge parallel computing option of FPGA can solve the high performance computing (HPC) needs much faster and with less area and power consumptions.

Several FPGA vendors like Altera are providing key video and image processing (VIP) tool sets IP blocks for scaling, de-interlacing, alpha-blending, noise reduction, FIR filter, median filter etc. which can accelerate development and implementation of newer, more sophisticated high-resolution medical imaging algorithms for streaming video applications. Researchers in biomedical imaging field can use this FPGA platform to prototype their novel parallel algorithms and can determine amount of performance improvement much faster. A lot of research opportunity exists in order to exploit the full potential of FPGA platform to support high-performance parallel processing needs of biomedical image processing.

V. CONCLUSION

With recent advances of computer science and engineering field, newly emerged biomedical engineering field is aiming at designing and manufacturing sophisticated medical imaging devices for improved patient health monitoring and drug discovery. The primary objectives of these devices are they have to be small, portable, cartheld or hand-held, fast, more accurate, flexible and less expensive. This paper first reports the current trend in rapidly advancing medical imaging technologies and terminologies to meet all the desired objectives of the medical devices with the emphasis on parallel processing needs on the reconfigurable computing platform of field programmable gate arrays (FPGAs). This article is intended to serve as a useful overview of the biomedical imaging field to the students and researchers so that they can carry their future research by quickly grasping the state-of-theart technologies and terminologies in biomedical imaging field.

REFERENCES

[1] Frost and Sullivan. Maximum Return on Investments in Medical Imaging IT with IBM's Grid Medical Archive Solution. Feb. 2006.

[2] Matlab Digest. Using Modeling and Simulation to Teach Dynamic Systems Concepts in the Context of Physiology. Feb. 2011.

[3] Oracle. Oracle Database 11g DICOM Medical Image Support. Sept. 2009.

[4] Adobe Photoshop. Optimum Strategies for Using Adobe Photoshop CS3 Extended in Biomedical Imaging.

[5] Wikipedia entry on GPUs. http://en.wikipedia.org/wiki/GPU. Nov. 2011.

[6] NVIDIA. NVIDIA CUDA Programming Guide, Version 2.3.1. Aug. 2009.

[7] Radisys. Power Up: Moving Toward Parallel Processing in Medical Imaging Compute Systems. Jul. 2009.

[8] Texas Instruments. DSP in Medical Imaging. Nov. 2008.

[9] Y. Chen, N. Gamble, M. Zettler and L. Maki. FPGA-Based Algorithm Acceleration for S/W Designers. SBS Technologies (Canada) Inc. March 2004.

[10] Xilinx.Virtex-6 FPGA Configuration User Guide UG360 (v3.4) Nov. 18, 2011.

[11] Xilinx. Virtex-7 Overview DS180 (v1.8) Sep. 13, 2011.

[12] A. Cosoroaba. Memory Interfaces Made Easy with Xilinx FPGAs and the Memory Interface Generator. Xilinx WP260 (V1.0) February 16, 2007.

[13] M. Gokhales and P.Graham. Reconfigurable Computing: Accelerating Computation with Field-Programmable Gate Arrays (Springer, 2005).

[14] David Dye. Partial Reconfiguration of Xilinx FPGAs with ISE Design Suite. WP374(v1.1) Jul. 6, 2011.

[15] Altera. Medical Imaging Implementation Using FPGAs. Jul. 2010.