<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>SYSCORE: A Coarse Grained Reconfigurable Array Architecture for Low Energy Biosignal Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Authors(s)</strong></td>
<td>Patel, Kunjan; McGettrick, Séamas; Bleakley, Chris J.</td>
</tr>
<tr>
<td><strong>Publication date</strong></td>
<td>2011-05-03</td>
</tr>
<tr>
<td><strong>Conference details</strong></td>
<td>19th IEEE International Symposium on Field-Programmable Custom Computing Machines (FCCM), Salt Lake City, Utah, USA, 1 - 3 May, 2011</td>
</tr>
<tr>
<td><strong>Publisher</strong></td>
<td>IEEE</td>
</tr>
<tr>
<td><strong>Item record/more information</strong></td>
<td><a href="http://hdl.handle.net/10197/7033">http://hdl.handle.net/10197/7033</a></td>
</tr>
<tr>
<td><strong>Publisher's statement</strong></td>
<td>© 2011 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.</td>
</tr>
<tr>
<td><strong>Publisher's version (DOI)</strong></td>
<td>10.1109/FCCM.2011.38</td>
</tr>
</tbody>
</table>
SYSCORE: A Coarse Grained Reconfigurable Array Architecture for Low Energy Biosignal Processing

Abstract—The promise of 24/7 patient monitoring and online diagnosis using wearable and implantable biomedical devices has engendered significant research interest in the development of low power biosignal processing platforms. Herein, a novel Coarse Grained Reconfigurable Array (CGRA) architecture is presented for low power, real time processing of biomedical signals. The proposed architecture differs from previously proposed CGRAs in that it is designed for low power, rather than for high performance. The proposed architecture was implemented in a software modeler and simulator and in Verilog. The architecture was shown to provide savings in energy consumption of up to 99% and speed up of up to 64 times compared to a conventional DSP processor for typical biosignal processing functions.

Keywords—coarse grained reconfigurable architecture, systolic, low energy, biosignal

I. INTRODUCTION

Biosignals, for example the Electroencephalogram (EEG), Electromyogram (EMG), Electrocardiogram (ECG) and Magnetoencephalogram (MEG), are used by clinicians for diagnosis of medical conditions and diseases. Current wearable devices, for example the EEG halter monitor, allow recording of biosignals but do not provide real time analysis of the data. All processing and analysis occurs offline, after the device has been returned to the clinician. These devices typically have battery lives of 2-3 days.

At present, there is significant research and commercial interest in developing wearable and implantable biomedical devices which can perform online, continuous biosignal processing allowing for automated real time symptom detection and diagnosis [1]. Real time analysis for important for many conditions, including, for example, seizure detection in subjects with epilepsy. Although sampled at a comparatively low rate (hundreds of Hz), the multi-channel nature of biosignals (possibly tens of channels) and the difficulty of the detection and classification problem make signal analysis task computationally expensive [2]. In addition, most wearable and implant applications require that devices are compact and have battery lifetimes of weeks or even months [1]. Previous work has shown that, for most biosignal processing applications, on-chip processing consumes significantly less energy than wireless transmission for the purposes of server-based processing [1]. These application requirements have led to research interest in low power on-chip biosignal processing.

General purpose processors cannot meet the high performance and low power requirements of biosignal processing applications at the same time [1]. Custom hard-wired Application Specific Integrated Circuits (ASICs) can achieve the performance and power consumption requirements. However, ASIC solutions lack flexibility. Designing a different ASIC for every medical condition is not cost effective given the low production volumes. Hence, a flexible low-power processor platform that can support a wide range of applications within the biosignal processing domain is required.

Herein, we propose a novel Coarse Grained Reconfigurable Array (CGRA) architecture, named SYSCORE, for low-power on-chip biosignal processing. A Coarse Grained Reconfigurable Array (CGRA) architecture consists of a grid of interconnected reconfigurable processing units which can perform logical or arithmetic operations. Unlike Field Programming Gate Arrays (FPGAs), the processing units are reconfigurable at the operation level rather than at the bit-level. This significantly reduces power consumption related to FPGAs while maintaining flexibility and increasing performance [3]. A host processor performs decision and control while computationally expensive tasks are offloaded to the CGRA.

The SYSCORE architecture is specifically designed for low-power biosignal processing and so differs from previous CGRA architectures in a number of ways. Firstly, the architecture is systolic in that, for regular biosignal processing algorithms, functional unit input and output data is pumped simultaneously and synchronously between nearest neighbor processing units arranged in an n-dimensional pipeline manner. Hard-wired systolic architectures are popular for DSP applications since they afford high throughput and data reuse [4]. In contrast to hard-wired arrays, the systolic CGRA approach allows flexibility, supporting a range of DSP algorithms. Secondly, the architecture supports mapping of irregular algorithms, such as the Fast Fourier Transform (FFT), by means of a novel...
interconnection unit, Roundabout Interconnect (RAI), wherein non-nearest neighbor data transfer is supported without the area and power cost of a dense interconnect. Thirdly, in order to reduce power consumption, the SYSCORE architecture uses a minimal number of resources in each functional unit, 24-bit fixed-point with 2 operational units and 4 data registers. Overall, the architecture provides significant energy saving by eliminating the fetch-decode steps of traditional processors (by means of reconfiguration), by significantly reducing the number of intermediate data RAM accesses (by means of systolic data reuse), by reduced logic switching (by means of compact functional units) and by voltage scaling (by means of parallelism). To the authors’ knowledge no previous paper has studied the used of CGRAs for biosignal processing applications.

The remainder of the paper is structured as follows. Section II presents a discussions of previous biosignal processing platforms and CGRA architectures. A description of the proposed architecture is given in Section III. In Section V, the power management in proposed architecture is described. Section VI presents results. Finally, the paper is concluded in Section VII.

II. RELATED WORK

A. Architectures for Biosignal Processing Applications

Previous work on processor platforms for biosignal processing has focused on multi-core and ASIC architectures.

Multicore architectures allow parallel processing of multichannel data. The authors of [5] presented a multiprocessor system-on-chip for real-time human heart monitoring and analysis. An architecture with 12 DSP processors was proposed to process 12 channel ECG data. Since the DSP cores run concurrently, the architecture implements a semaphore and interrupt system for communication and resource sharing. The HiBRID-SoC [6] consists of three adaptable programmable cores integrated using an AMBA AHB bus. Each core is optimized for a particular application set. The architecture is not area efficient if only a single core is used for a particular applications. Multi-core architectures provide performances increases over single core architectures but do not typically provide energy consumption reductions, other than by voltage scaling. In fact, the resource sharing and communication overhead is often significant in terms of power consumption and area.

ASIC designs can achieve high performance with low power consumption. The authors of [7] presented an ASIC for heart rate variability parameter monitoring and assessment. The ASIC was used in conjunction with a microcontroller. Application specific tasks were offloaded on the ASIC in which separate a separate hardware blocks were dedicated to specific functions. The ASIC reduced power consumption by the factor of 7 compared to a standalone microcontroller. An energy-efficient ASIC for ultra low power wireless sensor nodes was presented in [8]. The ASIC was designed to perform the main functions of a proposed wireless Body Area Network sensor node. The main disadvantage of ASICs is that they lack flexibility, so targeting other low volume biomedical applications is not cost effective since ASIC redesign is a lengthy and costly process.

B. CGRA architectures

To the authors’ knowledge, only one previous publication has investigated the power consumption of a CGRA architecture. In [9], the authors reported the power consumption of a CGRA but didn’t propose, discuss or prove the effectiveness of particular power saving techniques. Most previous work on CGRAs has focused on increasing performance. An array level comparison and an architecture level comparison of SYSCORE with 12 previous CGRA architectures are presented in Tables I and II, respectively.

CGRAs have previously been proposed for multimedia, embedded and DSP applications. Biosignal processing applications are typically based on DSP algorithms. CGRAs for embedded applications (AMBER) do not support this functionality. Multimedia CGRA architectures (Morphosys, REMARC, ADRES, CGRA Express, PACT XPP) primarily provide support for vectorize-able two-dimensional image and video processing algorithms. Biosignal processing algorithms are typically one-dimensional, but multi-channel. Hence a DSP specific CGRA architecture offers the energy-delay product for biosignal processing applications. Minimum energy is achieved by means a fixed-point architecture rather than a floating-point architecture (REMARC, PolySA). For accurate EEG analysis, a fixed bit-width of minimum 12-bits for IO and 24-bits for internal processing is required [1]. Certainly architectures with less than 16-bits are insufficient (PipeRench).

Single Instruction, Multiple Data (SIMD) architectures (Morphosys, REMARC, ADRES, CGRA Express, PACT XPP) require that all processing units execute the same operation. This is efficient for algorithms which can be vectorized, for example in multimedia algorithms, but is inefficient when processing steps can be concatenated, as in biosignal processing algorithms. VLIW architectures (Montium) include processing units which can perform a large number of operations in parallel. This is an unnecessary overhead in systolic architectures as address calculation is not needed since intermediate RAM accesses are eliminated. Uniquely, SmartCell can execute operations in
<table>
<thead>
<tr>
<th>Architecture</th>
<th>Bitwidth</th>
<th>Configuration bitwidth</th>
<th>Supported operations</th>
<th>Operations per cycle</th>
<th>Data passing</th>
<th>Data bypassing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morphosys[10]</td>
<td>28</td>
<td>32</td>
<td>General ALU operations, MAC, absolute difference, conditional ADD/SUB</td>
<td>1</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Montium[11]</td>
<td>16</td>
<td>16</td>
<td>General ALU operations</td>
<td>5</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PipeRench[12]</td>
<td>8</td>
<td>80</td>
<td>General ALU operations</td>
<td>1</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>REMARC[13]</td>
<td>16</td>
<td>32</td>
<td>ADD, SUB, shift, all logical operations (30)</td>
<td>1</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>RaPiD[14]</td>
<td>16</td>
<td>100</td>
<td>General ALU operations</td>
<td>1</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>ADRES[15]</td>
<td>32</td>
<td>53</td>
<td>General ALU (20) Operations including branch operations</td>
<td>1</td>
<td>-</td>
<td>No</td>
</tr>
<tr>
<td>CGRA Express[16]</td>
<td>32</td>
<td>55</td>
<td>20 Operations including branch operations</td>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>PPA[17]</td>
<td>32</td>
<td>-</td>
<td>ADD, SUB, MUL, and other common operations</td>
<td>1</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>AMBER[18]</td>
<td>32</td>
<td>32</td>
<td>Logical operations or ADD/ SUB/ COMPARE or SHIFT</td>
<td>1</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PACT XPP[19]</td>
<td>32</td>
<td>-</td>
<td>General ALU operations</td>
<td>1</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PolySA[20]</td>
<td>32</td>
<td>4</td>
<td>ADD, MUL, DIV</td>
<td>1</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>SmartCell[9]</td>
<td>16</td>
<td>64</td>
<td>General ALU operations, MAC, abs sum</td>
<td>3</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>SYSCORE</td>
<td>24</td>
<td>22</td>
<td>MUL, ADD, SUB, MUL-ADD, MUL-SUB, MAC, NOP</td>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
SIMD, MIMD (Multiple Instruction, Multiple Data) or systolic fashion due to the inclusion of dense interconnections and instruction memory. However, the rich interconnect and associated instruction memory consume 53% of the total power, which is not appropriate for low power applications.

Two previous proposals claim to operate in a systolic fashion - PolySA and SmartCell. Neither are designed for low power. PolySA is floating-point and SmartCell is rich in interconnect logic. SYSCORE also differs from these architectures in that its functional units have the capability of passing input data in parallel with outputting the results of the computation. This feature greatly facilitates systolic mapping of the applications (see later). Previously proposed architectures (e.g. RaPiD, PPA) are capable of passing data to nearest neighbor processing units but cannot pass input data and output data in parallel with operation execution.

Due to it’s data access pattern, the Fast Fourier Transform (FFT) cannot be easily mapped to a systolic architecture. Hence more complex interconnect must be provided to allow for efficient computation of the FFT. This support is provided in some CGRAs, such as PolySA, Montium. However this requires dense interconnections which increases chip area and power. The SYSCORE architecture uses a reconfigurable RAI scheme to provide interconnect flexibility at low overhead.

Most previous CGRA functional units are large in terms of area. For example, REMARC using a floating-point number format and AMBER contains 64 registers. Based on publicly available information, we estimate that the functional units of all previous architectures are more than double that of SYSCORE except for PACT XPP and PolySA which are 20% and 50% larger, respectively.

### III. Proposed Architecture

**A. Overview**

A 8x4 SYSCORE architecture is shown in Figure 1. There are two main elements: Configurable Function Units (CFUs) and RoundAbout Interconnect (RAI) units. The designer can use as many units as desired, according to the application performance targets and area constraints. Two Direct Memory Access (DMA) units inject data into the architecture from the West and North and one DMA unit collects data from the architecture. A column of RAI elements is inserted after every second column of CFUs to facilitate FFT computation. Array configuration and DMA operations are controlled by the host processor.

**B. CFU Design**

Figure 2 shows the architecture of a CFU. The CFU has 4 input ports (In0-In3) and 3 output ports (Out0-Out2). It has a Computation Unit (CU) that can perform computational operations. The CU differs from a conventional ALU/MAC in terms of the Set of Operations (SoOs) it can support. The CU can perform Addition (ADD), Subtraction (SUB), Multiplication (MUL), Multiply Accumulation (MAC), Multiply-Addition (MUL-ADD) and Multiply-Subtraction (MUL-SUB). These last operations, MUL-ADD and MUL-SUB, are more useful than the traditional MAC operation for systolic algorithms mapping [21]. Because of the feedback from CU_reg to the CU, the CFU can be configured to perform a MAC operation without extra hardware cost. All the operations can be performed in a single cycle. Two data can be passed in parallel with the result computed by CU through 3 output ports.

**C. Registers**

There are 2 General Purpose Registers (GPRs), 2 Coefficient Registers (CERs) and 1 CU register (CU_reg)
in a CFU. GPRs are used to store input data from input ports, CERs are used to store coefficients for functions such as FIR and FFT. The CU reg is used to store results from the CU unit. All data registers are 24 bits.

D. Configuration Register

Each CFU has one 32 bits configuration register (Config_reg). It stores the configuration passed via port In2 when the Config_en signal is high. Table III shows the settings and purpose of different bit fields of the configuration register. The size of the configuration is 22 bits, the remaining bits are left for future use.

<table>
<thead>
<tr>
<th>Bitfield</th>
<th>Abbreviation</th>
<th>Purpose</th>
<th>Selection list</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to 2</td>
<td>OP</td>
<td>Operation selection</td>
<td>ADD, SUB, MUL, MUL-ADD, MUL-SUB</td>
</tr>
<tr>
<td>3 to 5</td>
<td>ALU0</td>
<td>ALU selection line 0</td>
<td>In0-In3, 2 GPR</td>
</tr>
<tr>
<td>6 to 8</td>
<td>ALU1</td>
<td>ALU selection line 1</td>
<td>In0-In3, 2 CER, 2 GPR</td>
</tr>
<tr>
<td>9 to 11</td>
<td>ALU2</td>
<td>ALU selection line 2</td>
<td>In0-In3, CER, 2 GPR, ALU_reg</td>
</tr>
<tr>
<td>12 to 13</td>
<td>REG0_sel</td>
<td>Input selection for GPR0</td>
<td>In0, In1</td>
</tr>
<tr>
<td>14 to 15</td>
<td>REG1_sel</td>
<td>Input selection for GPR1</td>
<td>In2, In3</td>
</tr>
<tr>
<td>16 to 17</td>
<td>OP0_sel</td>
<td>Input selection for Out0</td>
<td>ALU_reg, GPR</td>
</tr>
<tr>
<td>18 to 19</td>
<td>OP1_sel</td>
<td>Input selection for Out1</td>
<td>ALU_reg, GPR</td>
</tr>
<tr>
<td>20 to 21</td>
<td>OP2_sel</td>
<td>Input selection for Out2</td>
<td>ALU_reg, GPR</td>
</tr>
</tbody>
</table>

E. Control Signals

SYSCORE can operate in three different modes: configuration mode, execution mode and flush mode. The mode of operation is set using the following global control signals:

1) Config_en: When this signal is high, SYSCORE operates in configuration mode. The data from port In2 is stored in Config_reg and the data from port In0 is stored in the CERs.

2) Flush_en: When this signal is high, SYSCORE operates in flush mode. In this mode, each CFU passes data to the CFU in the east direction. This mode is used to transfer results which are stored in CFUs which is not directly accessible by DMA.

3) Coeff_sel: This signal controls switching between different CERs for input to the CU in both configuration and execution modes.

4) Global_en: This signal enables/disables voltage supply to a row of CFUs to save power when the CFUs are not in use. For simplicity of control, CFUs can be turned off row wise, not individually.

F. Interconnections

As shown in Figure 1, all CFUs are connected to their nearest neighbor to the East and West. To avoid dense interconnections, cross interconnections are only introduced at odd numbered columns. Cross interconnections are useful for performing non-systolic functions, such as the FFT butterfly. Cross interconnection functionality is provided by RAI elements that allow data to pass from any Westerly CFU to any Easterly CFU. The conceptual structure of a RAI element is shown in Figure 3. Each RAI element has 6 input ports (I0-I5), 6 output ports (O0-O6) and a 16 bit configuration register. As in a CFU, the RAI element can be reconfigured when SYSCORE is in the configuration mode. The output ports of the RAI element can be configured to take data from the input ports. Figure 3(a) shows the available output port options in RAI. There are no global interconnections, except control signals (as described in the previous section), which saves chip area and reduces power consumption and control overheads.

IV. MAPPING OF ALGORITHMS

Since no benchmark application suite is available for biosignal processing, we selected a set of algorithms from those generally used in biosignal processing applications. The algorithms listed in Table IV were manually mapped using a methodology that is based on Synchronous Data Flow (SDF) and Control and Data Flow (CDF) graphs. FIR filtering is a common function in biosignal processing. Figure 4 shows the SDF graph, the CDF graph and topology matrix for a 5 tap FIR filter. Once data is fetched, it is injected to the input ports of CFUs using DMA and the partial results are passed directly to the next CFU in the following cycle, and so on, until the final result is
obtained. No intermediate RAM access are needed, which significantly reduces power consumption. More details on the algorithms mapping approach can be found in [21].

\[
\Gamma = \begin{bmatrix}
1 & -2 & 0 & 0 & 0 & 0 \\
0 & 1 & -2 & 0 & 0 & 0 \\
0 & 0 & 1 & -2 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & -2 \\
0 & 0 & 0 & 0 & 0 & 1 \\
-1 & -1 & -1 & -1 & -1 & -1 \\
\end{bmatrix}
\]

(a)

Data input

1 2 1 -2 3 4 1 -2 5 6 1 2 3

(b)

Input Data Coefficient

Store data in output register

Execution cycle

To next CFU

(c)

Figure 4. a) FIR filter topology matrix; b) 5 taps FIR filter SDF graph and c) FIR filter CDF graph for a single CFU.

V. POWER MANAGEMENT

A. Dynamic Voltage and Frequency Scaling (DVFS)

DVFS can significantly reduce the power consumption of processing architectures [22]. The power consumption of an architecture is directly proportional to the clock frequency \(f\) and the square of the supply voltage \(V_{dd}\). Delay is proportional to \(1/f\). Hence, the inherent parallelism in SYSCORE can be traded for reduced energy consumption by minimizing the clock frequency and supply voltage. Section VI-D, shows how energy varies with respect to \(V_{dd}\) for different applications.

B. Turning off Unused CFUs

The Global_en line allows unused CFUs to be disconnected from supply on a row-by-row basis. Energy savings arising are clearly dependent on CFU utilization.

VI. RESULTS

A. Implementation

An 8x8 SYSCORE array was built use two array blocks, each as shown in Figure 1. The architecture was implemented and application mapping was performed using a software model and simulator called RaCAMS [23]. RaCAMS was used to obtain performance results. Simulation outputs were verified by comparison with Matlab. The hardware architecture was implemented in Verilog and the algorithms were mapped using SystemVerilog. The Synopsys tool chain was used for synthesis, RTL simulation, gate-level simulation and power estimation. A 90nm CMOS technology library was used. The area of SYSCORE was 7750 cells and operating frequency was 100 MHz. Because of the differences in technology libraries, it was not possible to directly compare SYSCORE’s hardware metrics with those reported for previously proposed architectures. So, for comparison purposes, a hypothetical DSP processor was implemented. The processor architecture had 1 MAC unit, 24 registers, a fetch and decode unit, Program RAM and Data RAM. The ISA of this DSP can execute the instructions that a typical DSP processor can execute.

B. Performance and Energy Consumption

The performance and energy consumption comparison between SYSCORE and the DSP is shown in Table IV. The clock frequency of SYSCORE and the DSP are assumed to be equal. Reconfiguration time and energy are included in the CGRA figures. DVFS was not taken into consideration in the analysis. It can be seen that SYSCORE architecture gives up to 99% of energy savings and provides speed up factor of 64 depending on the algorithm.

C. RAM Data Reuse (RDR)

When DVFS is not used, the majority of the power saving in the CGRA case is due to reduction in the number of RAM accesses. Figure 5 shows a comparison of RAM Data Reuse (RDR) between DSP processor and SYSCORE architecture. RDR is given by:

\[
RDR = \frac{\text{Number of unique RAM addresses accessed}}{\text{Number of RAM accesses}}
\]

A RDR value close to 1 indicates that RAM locations are only accessed once. A value close to 0 indicates that same RAM locations are accessed many times.

It is clear from the results that data reuse of SYSCORE architectures is considerably higher than that of DSP processor.

D. Dynamic Voltage and Frequency Scaling

Figure 6 shows energy saving by reducing the voltage for the different algorithms.
techniques were studied to further reduce the energy consumption. A novel CGRA architecture, SYSCORE, is proposed herein for these applications. The architecture allows systolic mapping of DSP algorithms to reduce memory accesses and so reduce power consumption. RAI interconnect elements were introduced to increase the flexibility in the architecture in supporting algorithms which cannot be easily mapped systolically. A number of power savings techniques were studied to further reduce the energy consumption. The SYSCORE architecture gives up to 99% of energy savings and up to 64 times speed up compared to a conventional DSP processor architecture implemented using the same technology. Planned future work includes assessment of the architecture for a wearable biomedical monitoring application.

ACKNOWLEDGMENT

Removed for blind review

REFERENCES


[23] “Removed for blind review.”