ISSN: 2582 - 6379
Orange Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2021, Dec , Volume 2 Issue 4
www.ijisea.org

Low-Power Many Core Accelerator PENC for Personalized Biomedical Engineering Applications

V Madhuri Sowmya[1],Ch.Suresh[2]
PG Scholar[1] , Assistant Professor[2]
Department of Electronics and Communications Engineering
Amrita Sai Institute of Science and Technology
Paritala , Kanchikacherla ,Krishna District , Andhra Pradesh , India

ABSTRACT

Wearable personalized health monitoring systems can over a cost effective solution for human health-care. These systems must constantly monitor patient's physiological signals and provide highly accurate and quick processing and delivery of the vast amount of data within a limited power and area footprint. These personalized biomedical applications require sampling and processing multiple streams of physiological signals with a varying number of channels and sampling rates. The processing typically consists of feature extraction, data fusion, and classification stages that require a large number of digital signal processing and machine learning kernels. In response to these requirements, in this paper, a tiny, energy-efficient and domain-specific many core accelerators referred to as Power Efficient Nano Clusters (PENC) is proposed to map and execute the kernels of these applications.

Keywords: PENC , Bio medical Applications , Core accelerators

I.INTRODUCTION

Recent innovations in the semiconductor industry made it possible to integrate various sensors and computing components in an embedded system on a chip (SoC) processing platform. Wearable mobile platforms use embedded SoCs to process sophisticated and computationally intensive applications. With the rapid advances in small, low-cost wearable computing technologies, including smart phones and smart watches, there is a tremendous opportunity to develop ubiquitous personalized biomedical embedded systems capable of continuous vigilant monitoring of physiological signals. These systems have the potential to reduce the morbidity, mortality, and economic cost associated with many chronic diseases by enabling early intervention and preventing costly hospitalizations. In addition, recent advances in non-invasive sensor technologies enable the possibility that these systems can potentially monitor and analyse several modalities, including acceleration, pressure, temperature, electrocardiography (ECG), electromyography (EMG), electroencephalography (EEG), ultrasound, audio, and image signal streams. Embedded biomedical applications primarily consist of three basic stages: 1. A sensor front-end to capture and digitize physiological signals, 2. A processing stage to analyse, classify, and potentially store the sensors data, 3. AN RF module stage to transmit the data, classification, and/or diagnostics to the user or medical personnel. There has been an incredible amount of innovation and improvement in sensor design that has dramatically reduced power while maintaining high accuracy. This is the result of technologies such as Micro electromechanical systems (MEMS) sensors and specialized Analog-Front-End (AFEs) products targeted for physiological signals, such as Texas Instruments medical AFEs like ADS129x and AFE44xx. There has also been a tremendous amount of work done on wireless RF modules ranging from specialized research modules to commercial modules such as Blue-tooth Smart.

ISSN: 2582 - 6379
Orange Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2021, Dec , Volume 2 Issue 4
www.ijisea.org

II. LITERATURE REVEIW

Heterogeneous Processors: Heterogeneous architecture platforms have shown to provide significant advantages in enabling energy-efficient or area-efficient computing. Integrating heterogeneous core in a multi-core (such as ARM+MIPS), CPU+GPU, or heterogeneous CPU+GPU+FPGA, has been investigated in various studies. In more complex heterogeneous architecture, multi-core, GPU and even FPGA have been integrated to solve the ILP and TLP challenges. An example for FPGA+CPU+GPU is the Axel system and Nvidia Tegra K1 and X1 that combines the benefits of the specialization of FPGA, the parallelism of GPU and the scalability of a multi-core architecture. These examples show that heterogeneous architecture can over significant improvement for high computing demand applications. In general, in these systems, the overall performance can be improved by smart scheduling, allowing various heterogeneous computing components to work collaboratively on different parts of the program. In spite of all the performance benefit of integrating heterogeneous architectures, the challenge of high power consumption and high operating temperature remains an obstacle for deploying these designs in an embedded, wearable, and power constrained environment, including mobile devices. Particularly for many-cluster DSP and GPU platforms, while it's been shown that these architectures are capable of providing the performance requirements of many computing intensive applications, they still super from high power consumptions and high operating temperatures. Thus, these systems are impractical for resource constrained embedded portable environments.

Domain Specific Accelerator Processors: In the domain specific platforms, several research works have been carried on implementation of simpler cores for optimization rather than having application specific processors. There has been work on simple programmable processors used for application specific mapping.

Biomedical Processors: In the domain of general-purpose platforms for biomedical applications, recent work has shown how multi-core architectures over significant efficiency advantage over single core architecture when running various biomedical applications. This is mainly motivated by the inherent parallelism exist in biomedical applications with multichannel signal analysis requirements, where multi-core architectures can bring significant energy efficiency compared to a single core.

III. PROPOSED SYSTEM

PENC Many core Overview and Key Features:

PENC many core accelerator is a homogeneous multiple instruction, multiple data (MIMD) architecture that consists of in-order tiny processors with a 6 stage pipeline, a RISClike DSP instruction set and a Harvard Architecture model. The core operates on a 16-bit data path with minimal instruction and data memory suitable for task-level and data-level parallelism. Furthermore, these cores have a low complexity, minimal instruction set to further reduce area and power footprint. The lightweight cores also help to ensure that all used cores execute an application without an idle state, which can further reduce overall energy consumption. These light cores have simplified data memory, instruction memory and instruction set architecture ensuring full utilization of their resources when used. The processor can support up to 128 instructions, 128 data memory, and provides 16 quick-access registers. In the network topology, a cluster consists of three cores that can perform intra-cluster communication directly via a bus and inter-cluster communication through a hierarchical routing architecture. Each cluster also contains a shared memory. Each core, bus, shared memory and router was synthesized and fully placed and routed in a 65 nm CMOS technology using Cadence SoC Encounter and results for one cluster are summarized in Fig 1 .E. The Processing core contains additional bowering on the input in the form of a 32-element contentaddressable memory (CAM). It is used to store packets from the bus and allow a finite state machine (FSM) to find a word where the source core field corresponds to that in the IN instruction itself, where the IN instruction is used to communicate between the cores. For example, if the core is executing IN 3, the FSM searches

ISSN: 2582 - 6379
Orange Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2021, Dec , Volume 2 Issue 4
www.ijisea.org

through the CAM to find the first word whose source core is equal to three. This word is then presented to the processing core and processing continues. PENC many core architecture has 3 light-weight processing cores and a shared memory in a single cluster. Our initial many core architecture designs had 4 processing cores and a hierarchical router within a cluster, which was ideal for DSP kernels for minimal data storage and localized processing. Since personalized biomedical applications use ML kernels which often require large amount of memory for their model data, the previous architecture resulted in memory access time bottleneck. Hence the proposed PENC many core architecture replaces the 4 core implementation with 3 cores and a shared SRAM memory of 3K words and low latency bus-based architecture for inter-cluster communications, while maintaining the efficiency of low area and power consumption. Our initial results showed that performance benefit of bringing additional cores within the cluster diminishes given the increase in total area, power consumption and network congestion. Below are the key characteristics of the PENC many core platforms. 1) Bus based Cluster: Cores use the IN and OUT instructions to communicate with each other. When a core executes an OUT instruction, the data and relevant addressing information is packetized and sent to its output FIFO through a bus. When data is present in a core's output FIFO, it requests to use the cluster bus. The bus then arbitrates between requests, only granting those whose transactions can be completed.

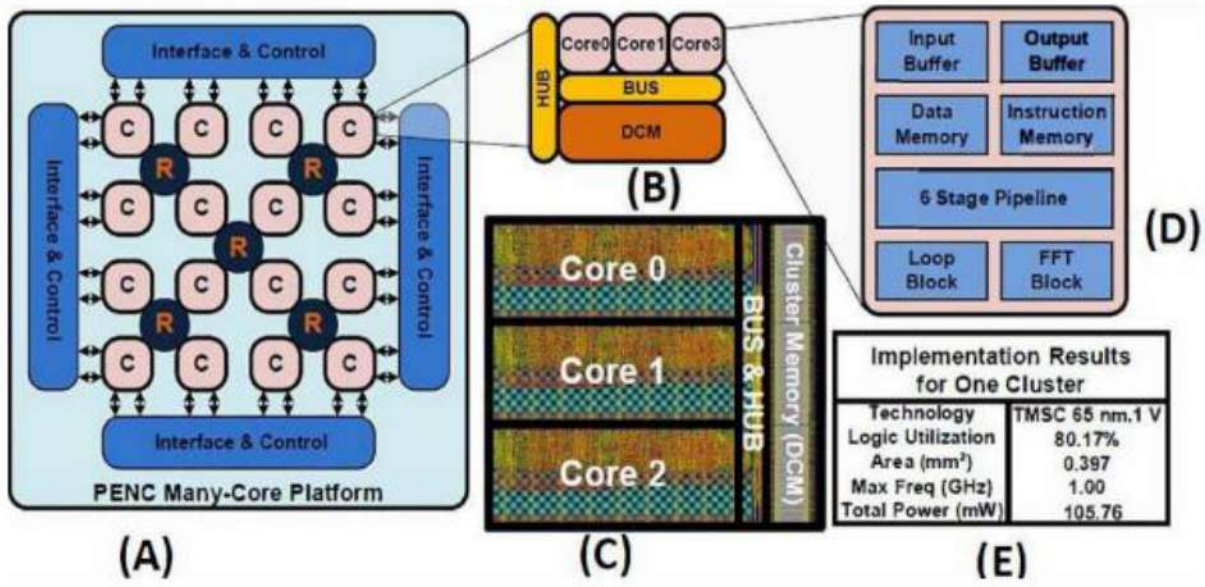

Figure 1 : Shows (A) Power Efficient Nano Clusters (PENC), Manycore Architecture (B) Bus-based Cluster Architecture (C) Post-layout view of bus based cluster implemented in 65nm, 1V TSMC CMOS technology (D) Block Diagram of core architecture (E) Post Layout implementation results of optimized bus-based cluster

The bus treats each transmission of data as a single transaction since it behaves with a simple push or data-driven protocol. The bus is used for intra-cluster communication. This includes a round robin arbiter which chooses the next node to grant access based on roundrobin scheme. Once the node gets access, it wraps the processing core pipeline with layers of buffering and is the main level in the PENC architecture that interacts with the bus. The destination core is used by the bus to forward the packet to the appropriate location, and the source core is used by the requesting node to satisfy its corresponding IN instruction. Based on the destination address and the data fields, the recipient core stores the address of the data. 2) Domain Specific Customization of Instruction Sets: Customizing a processor's instruction set for a particular computing domain is an efficient way of improving the processors performance. Designing an application-specific hardware for each given application is expensive; hence a customized instruction set in the many core can have a remarkable electron power and area. The PENC architecture is optimized to best suited for machine learning kernels. There are lightweight processing cores containing a limited instruction set for efficiency with a handful of specialized instructions such as absolute distance calculation and sorting. 3) Efficient Cluster Memory Access Architecture: While the lightweight cores are ideal for DSP

ISSN: 2582 - 6379
Orange Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2021, Dec , Volume 2 Issue 4
www.ijisea.org

kernels that require minimal static data, ML kernels often require larger amounts of memory for their model data. This is addressed with the distributed cluster-level shared memory (DCM) that is interfaced to the bus. The shared memory within a cluster consists of 3 instances of SRAM cells of memory size 1024x16 bits making up a total of 3072 words and can be accessed within the cluster using the bus and from other clusters through the router. To access the memory, cores use two memory instructions: LD and ST. The maximum depth of the cluster memory is 216 words since registers and data memory are both 16-bits wide and can therefore supply a 16-bit memory address. Using data memory as operands for instructions is still beneficial to using LD and ST from an efficiency standpoint because of the one-cycle read/write capability. Referencing data from the cluster memory has latency and requires a separate instruction, which reduces the overall instructions per cycle that the pipeline can complete. However, the LD and ST instructions enable the use of a much larger addressable space, which allows the PENC to support many applications. PENC architecture is ideally suited for personalized biomedical applications which require to compute a variety of multiphysiological signals in real time within limited power budget. The proposed PENC features including lightweight processing cores, domain specific customization of instructions (i.e sort, distance calculation, FFT, MAC, as well as low latency memory and IO access instructions), and enhanced bus-based cluster architecture for low latency shared memory access make this MIMD platform address the needs of this class of applications. Next section provides empirical results showing how these many core specific features are well suited for personalized biomedical applications.

V. RESULTS AND DISCUSSION

RTL SCHEMATIC In digital circuit design, register-transfer level (RTL) is a design abstraction which models a synchronous digital circuit in terms of the flow of digital signals (data) between hardware registers, and the logical operations performed on those signals. Register-transfer-level abstraction is used in hardware description languages (HDLs) like Verilog and VHDL to create high-level representations of a circuit, from which lower level representations and ultimately actual wiring can be derived. Design at the RTL level is typical practice in modern digital design.

RTL DESCRIPTION A synchronous circuit consists of two kinds of elements: registers (Sequential logic) and combinational logic. Registers (usually implemented as D flip-flops) synchronize the circuit's operation to the edges of the clock signal, and are the only elements in the circuit that have memory properties. Combinational logic performs all the logical functions in the circuit and it typically consists of logic gates. For example, a very simple synchronous circuit is shown in the figure. The inverter is connected from the output, Q, of a register to the register's input, D, to create a circuit that changes its state on each rising edge of the clock, clk. In this circuit, the combinational logic consists of the inverter. When designing digital integrated circuits with a hardware description language, the designs are usually engineered at a higher level of abstraction than transistor level (logic families) or logic gate level. In HDLs the designer declares the registers (which roughly correspond to variables in computer programming languages), and describes the combinational logic by using constructs that are familiar from programming languages such as if-then-else and arithmetic operations. This level is called “register transfer level”. The term refers to the fact that RTL focuses on describing the flow of signals between registers describing the flow of Signals between registers.

ISSN: 2582 - 6379
Orange Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2021, Dec , Volume 2 Issue 4
www.ijisea.org

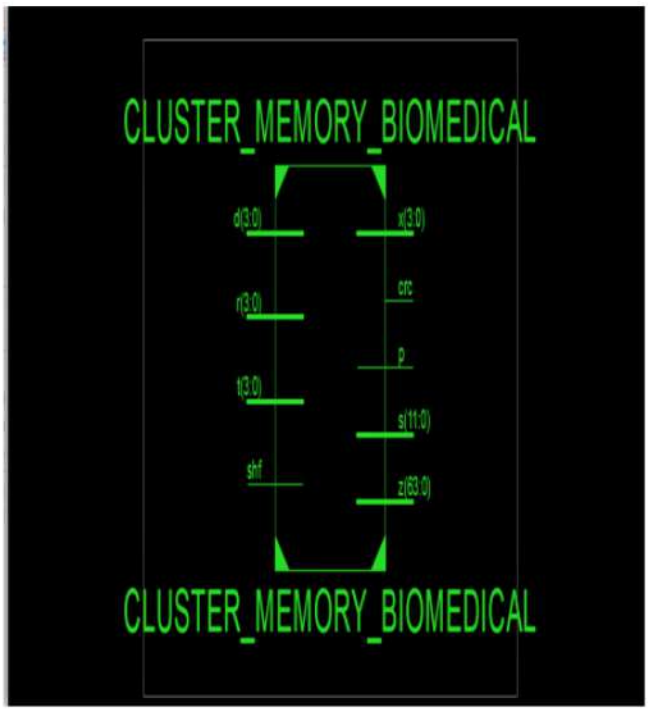

Figure 2: Shows RTL Schematic

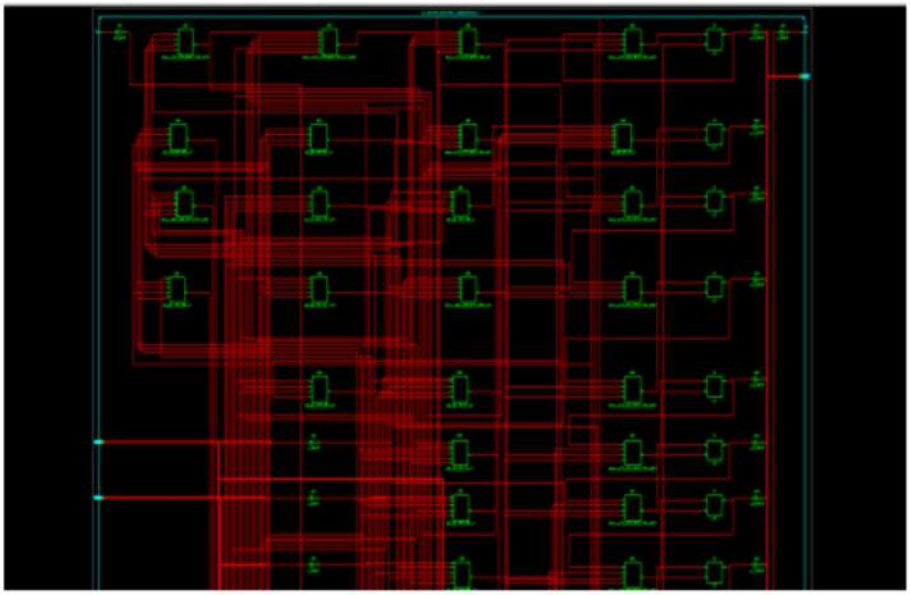

Figure 3: Shows Technological Schematic

ISSN: 2582 - 6379
Orange Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2021, Dec , Volume 2 Issue 4
www.ijisea.org

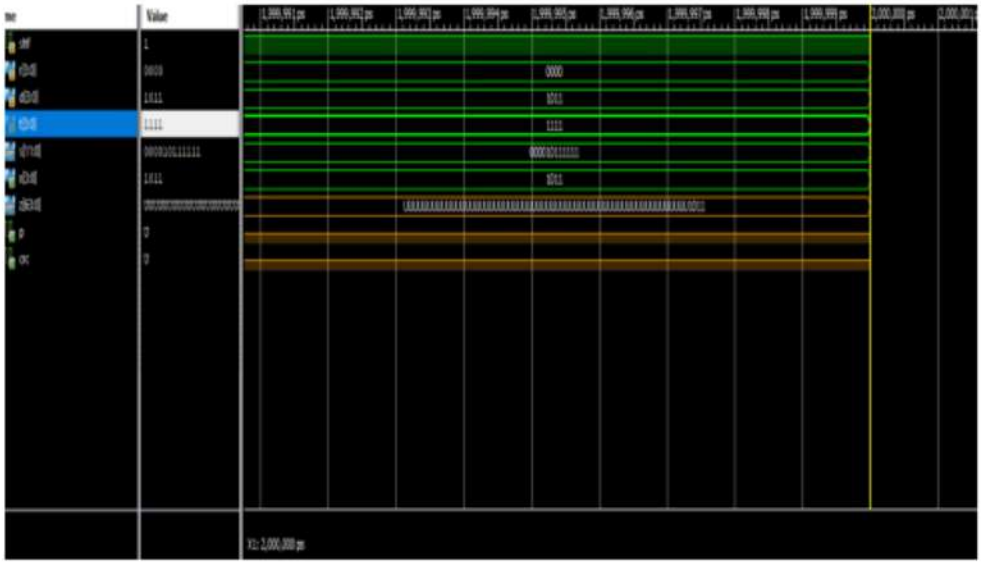

Figure 4 : Shows Timing Diagram

This paper explores the choice of embedded architectures for energy-efficient processing of personalized biomedical applications. Biomedical applications share strong commonalities requiring sampling from a number of physiological signals and processing that contains various digital signal processing and machine learning kernels. The software, as well as hardware implementations of machine learning personalized biomedical applications, are compared. For the choice of software, state-of-the-art commercial off-the-shelf embedded processing platforms such as ARM and Atom CPUs along with K1 GPU are compared with the hardware implementation of these kernels on embedded low-power FPGA. To further push the energy-efficiency, a custom lightweight, symmetric many core architecture is proposed that enables exploiting task level and data-level parallelism within biomedical kernels, dynamic frequency scaling, and specialized instructions and memory architecture to significantly reduce the energy usage.

REFERENCES:

[1]Page, C. Sagedy et al., "A flexible multichannel eeg feature extractor and classifier for seizure detection," Circuits and Systems II: Express Briefs, IEEE Transactions on, vol. 62, no. 2, pp. 109–113, 2015.

[2]S. Viseh, M. Ghovanloo, and T. Mohsenin, "Towards an ultra low power on-board processor for tongue drive system," Circuits and Systems II: IEEE Transactions on, accepted, vol. 62, no. 2, pp. 174–178, Feb 2015.

[3] J. Yoo, L. Yan, D. El-Damak, M. A. B. Altaf, A. H. Shoeb, and A. P. Chandrakasan, "An 8-channel scalable eeg acquisition soc with patient specific seizure classification and recording processor," IEEE journal of solid-state circuits, vol. 48, no. 1, pp. 214– 228, 2013.

[4] K. H. Lee and N. Verma, "A low-power processor with configurable embedded machine-learning accelerators for high-order and adaptive analysis of medical-sensor signals," IEEE Journal of Solid-State Circuits, vol. 48, no. 7, pp. 1625–1637, 2013.

ISSN: 2582 - 6379
Orange Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2021, Dec , Volume 2 Issue 4
www.ijisea.org

[5] Jafari and T. Mohsenin, "A low power seizure detection processor based on direct use of compressively-sensed data and employing a deterministic random matrix," in IEEE Biomedical Circuits and Systems (Biocas) Conference, Oct 2015.

[6] M. Malik and H. Homayoun, "Big data on low power cores: Are low power embedded processors a good fit for the big data workloads?" in Computer Design (ICCD), 2015 33rd IEEE International Conference on. IEEE, 2015, pp. 379–382.

[7] M. Malik, S. Rafatirah, A. Sasan, and H. Homayoun, "System and architecture level characterization of big data applications on big and little core server architectures," in Big Data (Big Data), 2015 IEEE International Conference on. IEEE, 2015, pp. 85–94.

[8]M. K. Tavana, M. H. Hajkazemi, D. Pathak, I. Savidis, and H. Homayoun, "Elastic core: enabling dynamic heterogeneity with joint core and voltage/frequency scaling," in Proceedings of the 52nd Annual Design Automation Conference. ACM, 2015, p. 151.

www.ijisea.org