Power/Performance Exploration of Single-core and Multi-core Processor Approaches for Biomedical Signal Processing

Ahmed Yasir Dogan[1], David Atienza[1], Andreas Burg[2],
Igor Loi[3], and Luca Benini[3]

[1] Embedded Systems Lab. (ESL) - EPFL; Lausanne - 1015, Switzerland
{ahmed.dogan,david.atienza}@epfl.ch
[2] Telecommunications Circuits Lab. (TCL) - EPFL; Lausanne - 1015, Switzerland
andreas.burg@epfl.ch
[3] UNIBO-Micrel Lab, Viale Risorgimento 2, 40136, Bologna, Italy
{igor.loi,luca.benini}@unibo.it

Abstract. This study presents a single-core and a multi-core processor architecture for health monitoring systems where slow biosignal events and highly parallel computations exist. The single-core architecture is composed of a processing core (PC), an instruction memory (IM) and a data memory (DM), while the multi-core architecture consists of PCs, individual IMs for each core, a shared DM and an interconnection cross-bar between the cores and the DM. These architectures are compared with respect to power vs performance trade-offs for a multi-lead electrocardiogram signal conditioning application exploiting near threshold computing. The results show that the multi-core solution consumes 66% less power for high computation requirements (50.1 $MOps/s$), whereas 10.4% more power for low computation needs (681 $kOps/s$).

Keywords: WBSN, ECG, Parallel Processing, Near Threshold Computing.

1 Introduction

Personal health systems monitor metabolic functions, such as heart and respiratory rates, blood oxygen, and carbon dioxide levels to detect and diagnose potential health problems. Wireless body sensor networks (WBSNs) are the enabling technology for such personal health systems [1]. A WBSN for health monitoring consists of a number of light-weight sensor nodes attached to the human body, where each node is responsible for processing a specific low rate physiological signal. For instance, one of the most important physiological signals is the electrocardiogram (ECG), which is typically acquired at sampling rates between 125 Hz and 1 kHz to capture the often important details of the waveform. In order to monitor the heart rate for extended periods of time (up to multiple days or weeks), an ultra low power design with embedded biomedical

signal processing for feature extraction on the sensor node is necessary [2] to reduce the costly signal storage or transmission to the essence. The corresponding algorithms, consist mostly of low-effort computations and can thus be optimized to run in real time on typical embedded low-power microcontroller. For example, Rincon et al. [3] showed how delineation of ECG signals using the relatively complex wavelet transform algorithm can be realized on a commercially available WBSN sensor node with limited computation capability.

Unfortunately, despite the reasonable required computational effort, achieving low-power consumption remains a pressing issue since devices are expected to operate on a single battery for long periods of time. An effective technique to reduce power consumption is supply voltage scaling, potentially all the way to sub-threshold operation. In the literature, voltage scaling and its limitations and disadvantages such as performance loss, the risk of functional failure, performance variability, etc., have been extensively analyzed [4,5,6,7] and various low-power architectures have been presented. For example, Chen et al. [8] proposed a sensor platform capable of nearly-perpetual operation by using harvesting from solar cells. The proposed single processor architecture has an ARM Cortex M3 core with both retentive and non-retentive SRAM and a power management unit which controls the active and ultra low power sleep modes. In another work, Hanson et al. [9] presented a new ultra low energy processor with low voltage operations for wireless monitoring systems. They optimized the standby power consumption of the processor with the help of a new low leakage memory, memory size and instruction set adjustments, and power gating. However, the main issue with low-voltage operation is the performance loss, which, for a given processing requirement, can limit the degree of use of voltage-scaling. Parallel computing using multiple cores can alleviate this issue, provided that the algorithms to be executed can be parallelized. To this end, Dreslinski et al. [10] proposed a near threshold computing (NTC), cluster-based multi-processor architecture with a shared cache that operates at a higher supply voltage to be able to serve multiple cores at the same time. In another study, Yu et al. [11] introduced a sub/near threshold co-processor for low energy mobile image processing using architecture level parallelism to compensate the performance loss. Finally, Krimer et al. [12] proposed a massively parallel stream processor operating in NTC to achieve 1 Giga-operations per second with 1 mW of total power consumption.

Unfortunately, even though researchers focused on low energy solutions in both multi-core and single-core approaches individually, the two approaches have not been compared in terms of energy efficiency for the moderate workloads that are typical for biomedical applications. Thus, in this paper we propose as a main contribution a single-core and a multi-core architecture for embedded biomedical signal processing on WBSNs, where algorithms have a limited, yet, at near-threshold voltage, non-negligible complexity and where a significant part of the processing can be done in parallel. We explore the power/performance trade-offs between these proposed architectures for an ECG signal conditioning application while exploiting near threshold computing.

The rest of this paper is organized as follows: First Section 2 presents the multi-lead ECG signal conditioning application. Next, Section 3 introduces the single-core and multi-core processor architecture. Then, Section 4 gives the experimental results and, finally, we summarize the main conclusions of this work in Section 5.

2   ECG Signal Conditioning Application

ECG entails the analysis of electrical changes, sensed by electrodes attached to the body, occurring when the heart muscle depolarizes during each heartbeat. In single-lead ECG, the voltage difference between two electrodes placed at both sides of the heart indicates the heart rate (i.e., 60 to 100 beats per minute for an adult with a normal resting heart) and allows to identify weaknesses in different parts of the heart muscle. To obtain a better and more complete picture of the heart muscle activities, up to 12 leads can be used [3]. Each lead shows the activity of the heart from a different point of view.

Unfortunately, raw ECG signals, even when recorded in a controlled environment, contain various types of noise and baseline drifts. ECG signal conditioning is therefore a fundamental application for a sensor node in WBSNs for automated ECG analysis or for signal compression for recording [1]. Hence, our benchmark application is an ECG signal conditioning algorithm based on morphological filtering given in [13]. This algorithm performs baseline correction and noise suppression on ECG signals and operates on multiple leads in parallel and independently. For our case study, we assume 8 leads, which is a typical configuration. The average processing requirements for each lead amount to 681 operations (Ops) per sample, as the algorithm always processes blocks of 1024 samples. To investigate different processing requirements related to the application, we consider ECG sampling rates between $f_s = 125\,\mathrm{Hz}$ and $f_s = 1\,\mathrm{kHz}$ for capturing signals with quality levels from "barely acceptable" to "excellent".
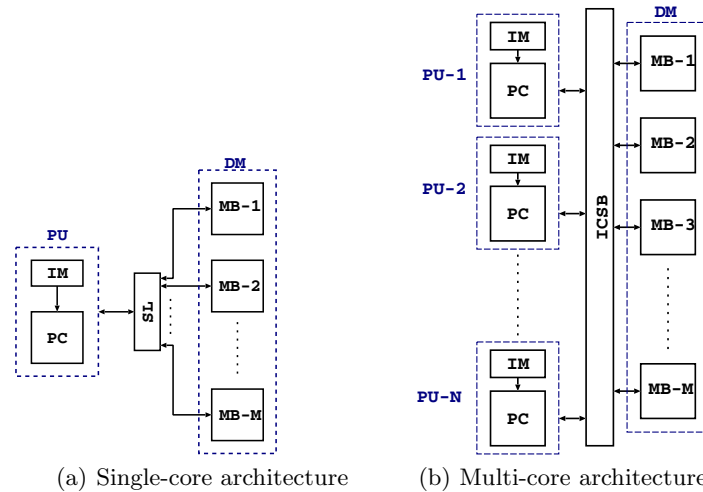
3   Processing Platform Architecture

To focus on the comparison between the single-core and multi-core configuration, we build both reference designs using the same processing unit (PU) and a data memory (DM). The designs are implemented in a 90 nm low leakage process technology trading peak performance for significant leakage power reduction, especially in the memories.

Processing Unit: A PU comprises a processing core (PC) and a 24-bit wide instruction memory (IM) for 4k instruction words (12 kBytes) which is sufficient for many typical biomedical applications on WBSNs such as delineation and compressed sensing data compression [1], [3]. The PC is a 16-bit Reduced Instruction Set Computer (RISC)-like architecture with sixteen working registers, and a Harvard memory model. The simple two-stage pipeline matches the low

to moderate performance requirements of the application and reduces the number of registers that need to be clocked. In the first pipeline stage, instructions are decoded and read addresses are generated. In the second pipeline stage, the operations are executed and the results are stored in a data memory location or in a working register. Among others, the instruction set comprises arithmetic and logic operations, multi-bit shifts and single-cycle multiplications to support the energy efficient execution of signal processing algorithms. Most instructions occupy only a single (24-bit) instruction-word and are executed in only one clock cycle with a latency of two cycles.

Data Memory: The PC can access the DM for reading and writing in the same clock cycle. Therefore the DM requires two separate access ports, one for reading and another one for writing. The 64 kByte of DM, required for 8-lead ECG real time processing, is split into M (i.e, 16) memory banks (MBs) with 2k words per bank. This configuration corresponds to the maximum available from our 2-port memory generator and it allows partial shutdown for leakage power reduction with applications with reduced memory requirements.

(a) Single-core architecture   (b) Multi-core architecture

Fig. 1. Processing platforms

3.1   Single-Core Processor Architecture

The single-core WBSN sensor node reference architecture is shown in Fig. 1(a). A simple selection logic (SL) connects the single PU to the individual MBs and multiplexes the data. The system processes the 8-lead ECG signals sequentially by using 5580 cycles per sample.

When optimized for speed, our single-core design could operate up to 147 MHz at nominal 1.2 V supply voltage, which is much higher than what is required for most biomedical signal processing applications, even for very high sampling rates. Since the cost for this high speed is a reduced energy efficiency, we optimize the

reference design for minimum area instead to lower the active and leakage power consumption. To this end, we request the EDA design tools choose logic gates with weak driving capability. The corresponding circuit is still capable of working up to 50 MHz at nominal voltage, which easily meets the timing requirements of our reference application, given in Section 2.

The second column of Table 1 shows the distribution of the power consumption of the area-optimized design running at 16 MHz clock frequency at 1.2 V. It is interesting to note that the power consumed by the SL and the interconnect network (routing and buffers) between the PU and MBs is almost 15% (0.55 mW) of the total power consumption. A more detailed analysis shows that much of this power is due to glitches on the address and data bus. To alleviate the impact of these glitches we place 48 low-transparent latches at the output ports of the PU (read- and write-addresses, and write-data). The result of this simple measure is a reduction of the overall power consumption of the single-core architecture by 6.7%, as shown in the third column of Table 1.

3.2   Multi-core Processor Architecture

The multi-core processor design, shown in Fig. 1(b) consists of N (i.e., 8) PUs with individual IMs. Each PU accesses the 16 shared MBs through a central crossbar interconnect [14] to enable full access to the entire memory space for each PU. This architecture is different from the one proposed by Dreslinski et al. [10] in which several slower cores share a cache that is proportionally faster and thus requires a higher supply voltage. Compared to their single, which relies on a fully shared memory-block configuration, our proposed architecture simplifies the clock-network design[1] and neither requires an additional faster clock, nor level-shifters between the cores and the shared cache. Furthermore, the ability to operate with only a single supply voltage considerably simplifies the overall system design and can result in additional energy savings, because multiple weakly loaded DC/DC-converters can be avoided. The drawback of our approach are the occasional access conflicts when two or more PUs access the same MB on the same port. In this case, the conflicting requests are served one after another based on PU priorities, while the waiting PUs are stalled using clock gating to avoid unnecessary active power consumption.

The multi-core design, which is also optimized for minimum area, is capable of operating up to 48 MHz. For our 8-lead ECG application, all cores are active to process one lead per core in 761 clock cycles per sample. When accounting for the 8x parallel processing, this corresponds to a 12% penalty in terms of timing due to stall-cycles compared to the number of cycles required for a single lead in the single-core architecture. To compensate for this penalty when comparing the two architectures in terms of power consumption, we always adjust the clock frequency of the multi-core design to correspond to the same sampling frequency (throughput) as in the single-core reference architecture. In particular, we provide results at nominal 1.2 V supply voltage for a frequency of 2.3 MHz

---

[1] As seen in Table 1, the clock tree in the proposed architecture consumes only 5.0% of the whole power consumption.

which ultimately corresponds to the same throughput as the single-core design running at 16 MHz. The corresponding power consumption figures are provided in the two rightmost columns of Table 1. The results show that the overhead of the crossbar in terms of power consumption is insignificant, only 13% of the entire multi-core design. This overhead can be further reduced by applying the same technique for the glitch reduction as in the single-core design. After placing latches in the PUs, the power consumption of the crossbar interconnect is reduced significantly, resulting in the 8.3% power improvement in overall power consumption shown in the rightmost column of Table 1,

Table 1. Power distribution of the single-core and the multi-core design with/without latches in the PU at 1.2 V supply voltage and 16 MHz and 2.3 MHz operating frequency, respectively

	single-core		multi-core	
	w/o latches	with latches	w/o latches	with latches
Total	3.56 mW	3.32 mW	3.72 mW	3.41 mW
PUs	2.53 mW	2.53 mW	2.81 mW	2.81 mW
MBs	0.24 mW	0.24 mW	0.24 mW	0.24 mW
SL-ICSB	0.55 mW	0.33 mW	0.48 mW	0.19 mW
Clock Tree	0.24 mW	0.22 mW	0.19 mW	0.17 mW
Reduction	-	6.7%	-	8.3%

The occupied silicon area of the single- and multi-core design is given in Table. 2. As expected, the total area of PUs in the multi-core design is almost 8 times the area of the PU in the single-core design. However, the total area of the multi-core design is only 1.76 times of the total area of the single-core design since the shared MBs are responsible for most of this area.

Table 2. Area results for the multi-core and single-core designs ($1\,GE = 3.136\ \mu m^2$)
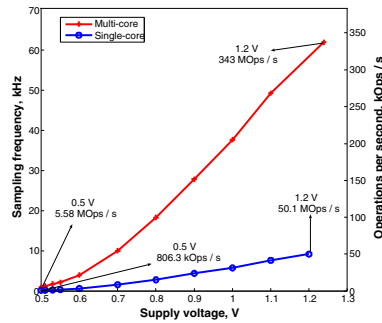
	Single-core	Multi-core
Topmodule	644.7 kGE	1138.1 kGE
PUs	68.0 kGE	541.4 kGE
MBs	576.7 kGE	576.7 kGE
ICSB	-	20.0 kGE

4   Experimental Results

The setup we used for the experiments is as follows: 1024 samples of an 8-lead ECG signal are pre-stored in the DM of the single-core and multi-core designs. Each sample occupies 16 bits of memory, which results in 16 kBytes of total storage for the pre-stored ECG samples. The single-core design processes the leads sequentially while in the multi-core design each core processes one lead. The results of each lead are stored individually in the data memory, and the total memory requirement for storing the results is 16 kBytes for each design.

We run our reference application on the two architectures for various workload requirements to explore the power/performance trade-offs between the architectures. A workload requirement in our experiments corresponds to a number of operations per second (Ops/s). This exploration allows us not only to examine the architectures for our reference application, but also to generalize the results and trends to other applications. In addition, we also analyze the architectures with respect to the ECG sampling frequencies corresponding to our application requirement.

We limit the scaling of the operating voltages to the transistor-threshold level (0.5 V) to avoid the performance variability and functional failure issues occurring mainly at sub-threshold regions. Fig. 2 shows the processing capabilities of the two approaches with respect to the supply voltage. At the nominal voltage (1.2 V) the single- and multi-core approaches achieve 50.1 $MOps/s$ and 343 $MOps/s$, respectively. As expected, these processing capabilities decrease with voltage scaling. When the supply voltage of the designs reaches the threshold level, the single-core accomplishes only up to 806.3 $kOps/s$ while the multi-core design still achieves up to 5.58 $MOps/s$.

Fig. 2. Single-core and multi-core designs: Maximum allowed ECG sampling rate and corresponding number of operations for various supply voltages

Fig. 3(a) shows the total power consumption of the single- and multi-core design for various workload requirements. As can be seen from the figure, the multi-core approach is the only viable solution for workloads between 50.1 $MOps/s$ and 343 $MOps/s$. Moreover, when the workload requirement is between 1356.5 $kOps/s$ and 50.1 $MOps/s$, the multi-core is more energy efficient than the single-core design, because the multi-core design can meet the workload requirements at a lower operating voltage compared to the single-core design. In particular, to meet a high workload requirement (50.1 $MOps/s$) the single-core design operates at 1.2 V and consumes 10.4 mW, whereas the multi-core design operates at 0.7 V and consumes only 3.5 mW. Thus, the multi-core solution consumes almost 66% less power than the single-core design. On the contrary, if the required workload is light (lower than 1356.5 $kOps/s$) the single-core design consumes less power than the multi-core design, because the multi-core design is able to reach the

threshold voltage at 5.58 $MOps/s$ workload while the single-core design reaches at the threshold level at 806.3 $kOps/s$ workload requirement. More precisely, to meet a low workload requirement (681 $kOps/s$), both designs operate at 0.5 V and the single-core design consumes 25.9 μW while the multi-core design consumes 28.6 μW. Hence the multi-core design consumes 10.4% more power than the single-core design.
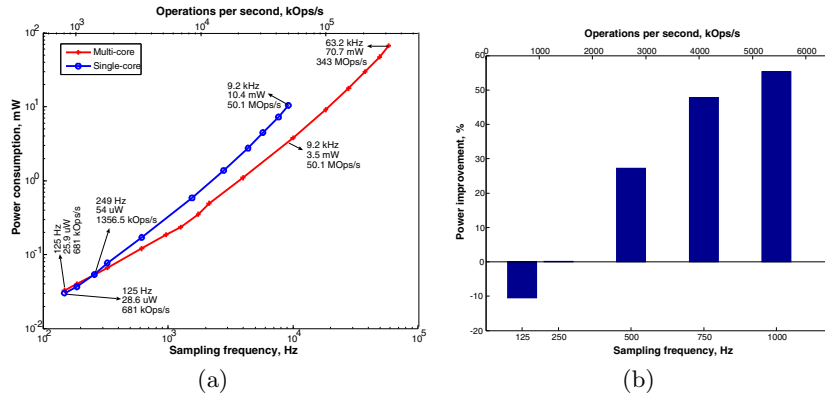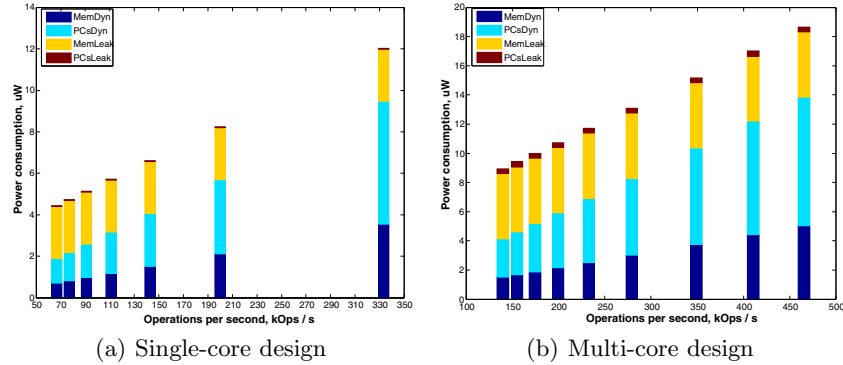

Fig. 3. (a) Total power consumptions for various workloads (b) Power efficiency of multi-core design respect to single-core design for ECG signal conditioning application

The corresponding workload to our application ranges from 681 $kOps/s$ to 5448 $kOps/s$ with an ECG sampling rate (f_s) from 125 Hz to 1 kHz. Fig. 3(b) shows the power efficiency of the multi-core design with respect to the single-core design for our application. As the sampling rate increases, the multi-core becomes more and more energy efficient. At the highest sampling rate, f_s=1 kHz, the multi-core design is 55% more power efficient. However if the sampling rate is reduced down to 250 Hz, the multi-core design becomes less power efficient. At the lowest ECG sampling rate in our range, f_s=125 Hz, the multi-core consumes 10.4% more power than the single core design.

Another interesting point is the comparison between dynamic and leakage power consumptions in the two designs. For our case study, where the lightest workload requirement is 681 $kOps/s$ (f_s=125 Hz) the leakage power consumption of the single- and multi-core design is 2.6 μW and 5 μW, respectively. Thus the leakage power consumption represents 10% and 17% of the total power consumptions for the single-core and multi-core designs, respectively. Fig. 4(a) and Fig. 4(b) show the dynamic and leakage power consumptions of the PCs and the memories, including both IM and MBs, for various workload requirements for the single-core and multi-core designs. In both figures the dynamic power consumption of the memories is labeled as `MemDyn` while the leakage is labeled as `MemLeak`. Similarly, the dynamic and leakage power consumptions of the PCs are indicated as `PCsDyn` and `PCsLeak`, respectively. As shown in the

(a) Single-core design   (b) Multi-core design

Fig. 4. Leakage and dynamic power consumption comparison for various workload requirements

figures, MemDyn becomes comparable with MemLeak when the workload is 200 $kOps/s$ and 410 $kOps/s$ for the single-core and the multi-core designs, respectively. As expected, MemLeak in the multi-core design becomes comparable with the MemDyn power at an earlier point, because the total memory leakage power is higher in the multi-core design. Furthermore, the overall leakage and dynamic power consumptions become comparable around at 80 $kOps/s$ for the single-core design while around 140 $kOps/s$ for the multi-core design.

5   Conclusion

Embedded biomedical signal processing on WBSNs involves relatively low complex and highly parallel computations on a low-rate physiological data, which creates the opportunity of low voltage operations as well as parallel processing. In this paper we present a single- and a multi-core processor architecture for biomedical signal processing on WBSNs where both energy efficiency and real-time processing are crucial design goals. To address the energy efficiency and data throughput requirements, we explored the power/performance tradeoffs between the two architectures, including near threshold voltage computing, for different workloads using an 8-lead ECG signal conditioning application. Our results show that the multi-core approach consumes 66% less power than the single-core approach for high biosignal computation workloads (i.e., 50.1 $MOps/s$). However, the multi-core architecture becomes more power consuming for relatively lighter workloads and it consequently consumes 10.4% more power (681 $kOps/s$).

References

1. Mamaghanian, H., et al.: Compressed Sensing for Real-Time Energy-Efficient ECG Compression on Wireless Body Sensor Nodes. IEEE Transactions on Biomedical Engineering 12, 120–129 (2011)

2. Hanson, M.A., et al.: Body Area Sensor Networks: Challenges and Opportunities. IEEE Computer 42, 58–65 (2009)
3. Rincon, F., et al.: Multi-Lead Wavelet-Based ECG Delineation on a Wearable Embedded Sensor Platform. Computers in Cardiology, 289–292 (2010)
4. Hanson, S., et al.: Exploring variability and performance in a sub-200 mV processor. IEEE J. Solid-State Circuits 43, 881–891 (2008)
5. Zhai, B., et al.: A 2.60 pJ/Inst subthreshold sensor processor for optimal energy efficiency. In: Symposium on VLSI Circuits. Digest of Technical Papers, Honolulu (2006)
6. Wang, A., Chandrakasan, A.: A 180 mV FFT processor using sub- threshold circuit techniques. In: IEEE Int. Solid-State Circuits Conference. Dig. Tech. Papers (2004)
7. Dreslinski, R.G., et al.: Near-Threshold Computing: Reclaiming Moore's Law Through Energy Efficient Integrated Circuits. Proceedings of the IEEE 98, 253–266 (2010)
8. Chen, G., et al.: Millimeter-scale nearly perpetual sensor system with stacked battery and solar cells. In: Solid-State Circuits Conference. Digest of Technical Papers, San Francisco (2010)
9. Hanson, S., et al.: A Low-Voltage Processor for Sensing Applications With Picowatt Standby Mode. IEEE J. Solid-State Circuits 44, 1145–1155 (2009)
10. Dreslinkski, R.G., et al.: An Energy Efficient Parallel Architecture Using Near Threshold Operation. In: 16th International Conference on Parallel Architecture and Compilation Techniques, Brasov, pp. 175–188 (2007)
11. Yu, P., et al.: An Ultra-Low-Energy Multi-Standard JPEG Co-Processor in 65 nm CMOS With Sub/Near Threshold Supply Voltage. IEEE J. Solid-State Circuits 45, 668–680 (2010)
12. Krimer, E., et al.: Synctium: a Near-Threshold Stream Processor for Energy-Constrained Parallel Applications. Computer Architecture Letters 9, 21–24 (2010)
13. Sun, Y., et al.: ECG signal conditioning by morphological filtering. Computers in Biology and Medicine 32(6), 465–479 (2002)
14. Rahimi, A., et al.: A fully-synthesizable single-cycle interconnection network for Shared-L1 processor clusters. In: Design, Automation Test in Europe Conference Exhibition (DATE), pp. 1–6 (2011)