ICT-ENERGY LETTERS

An Ultra-Low Power NVM-Based Multi-Core Architecture for Embedded Bio-Signal Processing

Rubén Braojos, David Atienza
Embedded Systems Laboratory, École Polytechnique de Lausanne
Switzerland

***Abstract*— Healthcare delivery is evolving towards new Wireless Body Sensor Nodes (WBSN), which are miniaturized devices able to acquire, process and transmit subjects' bio-signals in real time within a tiny energy budget. Recent efforts on AD converters and transmission schemes have enabled a major power consumption reduction of these components, thus leaving the embedded processing stage as the dominant power-hungry component. In this context, new multi-core architectures designed with smaller CMOS devices and aggressive voltage scaling greatly improve the energy efficiency of WBSNs, but originate reliability operation concerns. In this work we present a novel WBSN architecture equipped with a completely re-designed memory subsystem (including a low-voltage low-latency non-volatile partition), which operates in combination with an advanced code synchronization management to reduce the platform power consumption by up to 82%.**

I. INTRODUCTION AND MOTIVATION

ONGOING lifestyle changes are increasing the prevalence of chronic disorders, which are now the major sources of death worldwide [1]. These ailments require extensive monitoring, which represent a major financial burden for healthcare providers. Wireless Body Sensor Nodes (WBSNs) can lower these costs by allowing to acquire and analyze the bio-signals of patients even outside of a hospital environment and with little intervention from the medical staff. These devices must autonomously sense, process and wirelessly transmit body signals (such as electrocardiograms) for extended periods of time, while relying on small batteries. Thus, energy-efficiency (from acquisition to transmission) is fundamental for their ubiquitous use.

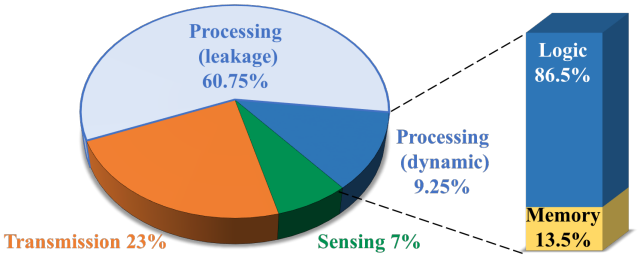

Fig. 1. Power consumption breakdown on a WBSN executing a multi-channel biosignal processing application.

With the reduction of the energy required by signal transmission, the efficient implementation of the digital signal processing (DSP) stage is key in order to minimize the power consumption of WBSNs. As shown in Fig. 1, most of the power dissipated by these devices is due to the processing of the acquired samples. In fact, the leakage power dominates the consumption of the overall system, which, based on our analysis, reaches up to 86% of the power devoted for DSP. To overcome this problem, voltage-frequency scaling has been proposed [2] [3], but aggressive reduction of supply voltage is unfeasible under certain levels and leads to undesired memory and logic reliability issues.

In this context, herein we proposed a novel WBSN architecture equipping a completely re-designed 2-level memory subsystem, which combines low-voltage, low-latency non-volatile memories (NVMs) with tiny volatile banks, to obtain superior energy-efficiency while meeting real-time constraints.

II. PROPOSED ARCHITECTURE

Typical bio-signal processing architectures based on volatile memories are designed to minimize the idle time by employing the lowest possible supply voltage and a clock frequency that allows to barely meet real time constraints [4] [2]. Conversely, our NVM-based architecture performs short computing bursts at a higher frequency in order to minimize active time. In this way, during long idle periods the full digital architecture can be power gated, while new samples are acquired in what we term "deep-sleep sensing". This strategy is possible thanks to the availability of persistent memory provided by the NVM.

The proposed architecture, depicted in Fig. 2, is similar to the one introduced by in [3]. It features eight low-power RISC cores interfaced to 16 data memory banks and 8 instruction memory banks. The cores have access to the memory banks through a logarithmic interconnect [5], that provides single-cycle read/write operations and perform arbitration in case of conflict among several memory requests.

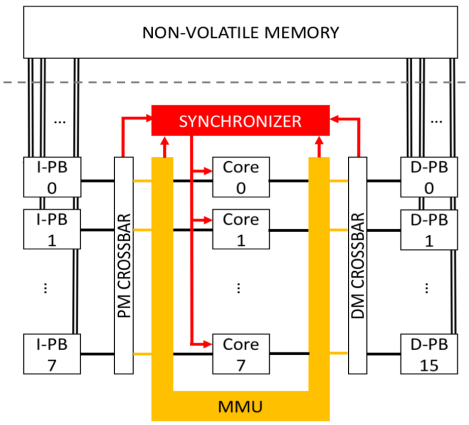

Fig. 2. Proposed multi-core architecture featuring the 2-level memory subsystem consisting on a low-latency large NVM partition and a set of small instruction and data page buffers (I-PBs, D-PBs respectively).

In the architecture of [3], the entire instructions and data contents (96 KB and 64 KB respectively) reside in volatile SRAM banks while in our proposed architecture those volatile memories are realized as tiny full-custom banks, termed "page buffers", that collectively act as a cache for the unified non-volatile storage (160 KB). These buffers have been implemented as arrays of latches that incorporate a direct input line connected to each bit cell allowing a single-cycle massive page storage or readout. For the

non-volatile partitions, low-voltage STTRAM [6] structures have been used as they significantly reduce the access energy with respect to standard solutions, such as FLASH-based NVMs.

In addition, the architecture supports advanced code synchronization [3] to manage efficient core-to-core notifications and single-instruction-multiple-data (SIMD) code execution. SIMD increases the efficiency when the same algorithm is applied on multiple streams, by dramatically reducing the instruction memory accesses [2]. A hardware synchronizer unit (see Fig. 2) orchestrates the run-time behavior of the system keeping track of data-dependent branches and producer-consumer relationships among cores [3]. This unit has been extended to manage the system transitions to and from deep-sleep sensing and to stall cores that experience a miss in the data or instruction buffers. Finally, the decision of issuing a page transfer is taken by a light-weight memory management unit (MMU, see Fig. 2), which signals the synchronizer the cores to stall while transfers are in progress. This unit works as a simple content-addressable-unit (CAM) first making the translation between the address of a request and the location of the corresponding page, and second loading the corresponding page, if it is not available, evicting an existing one if needed.

III. EXPERIMENTAL SET-UP

We comparatively evaluate the proposed architecture (hereafter **TARGET**) against the state-of-the-art SRAM-based WBSN architecture proposed in [3] (hereafter **SOA**). To this end, we performed a full physical design (through place and routing) of the system components using a 28nm process design kit (PDK) (1.0V VDD) to extract area, power and performance characteristics setting the operating frequency of the TARGET architecture to 20MHz. The obtained parameters were used to back-annotate a SystemC simulator of the architectures with which we extracted all the necessary run-time statistics for our study. We employed 4 representative benchmarks from the field of embedded electrocardiogram processing [7] [8]:

- **8L-CS**: Lossy compression of 8 ECG channels.
- **3L-MF**: Morphological filtering of 3 ECG channels.
- **3L-MMD**: Multi-scale Morphological Derivate delineation of a multi-channel ECG signal.
- **RPCLASS**: Selective multi-channel ECG delineation based on a heartbeat classifier.

Table 1: Runtime metrics of the analyzed benchmarks using 8-word instruction and 8-word data page buffers

	8L-CS	3L-MF	3L-MMD	RP-CLASS
Active time (%)	**5.5**	**4.7**	**8.2**	**7.0**
→ *Page exchange (%)*	*2.3*	*5.4*	*4.3*	*5.8*
→ *Processing (%)*	*97,7*	*94.6*	*95.7*	*94.2*
Deep-sleep Sensing (%)	**94.5**	**95.3**	**91.8**	**93.0**

IV. RESULTS

First, we explored the configuration of the NVM-based memory subsystem of our architecture (TARGET) to determine the optimal sizes of the page buffers, which result to be 8 words each. Even though small page buffers induce an increase in the amount of page transfers, the exchange timing overhead remains below 6% as shown in Table 1. The table also shows that the platform can meet the required real-time constraints while allowing for long deep-sleep sensing periods (>90%). Such amount of inactivity leads to a considerable decrease of the platform power consumption as depicted by Fig 3. As a result, the proposed TARGET architecture can obtain up to 82% reduction (3L-MF) with respect to the SOA architecture.

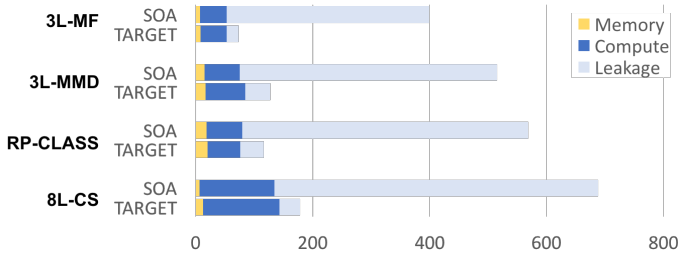

Fig. 3. Average power consumption (in µW) of the TARGET and SOA architectures for the studied benchmarks.

Finally, area-wise, the novel 2-level memory subsystem is more compact than a traditional SRAM-based structure. However, as depicted in Fig. 4, the routing of the latch-based page buffers incurs in a non-negligible area overhead, increasing the footprint of TARGET by 1.27x with respect to the SOA architecture.

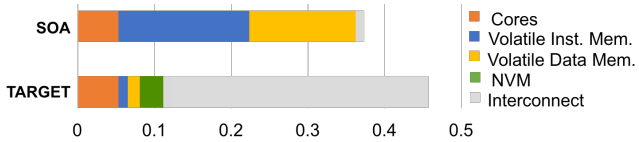

Fig. 4. Area breakdown (in mm^2) of the TARGET and SOA architectures.

V. CONCLUSIONS

In this paper we have proposed a novel WBSN architecture featuring a completely re-designed 2-level NVM-based memory subsystem that allows for new power management strategies resulting in up to 82% power savings with respect to state-of-the-art alternatives. Moreover, the new memory design of this architecture enables further benefits by capitalizing on new nano-scale manufacturing technologies, but this is out of the scope of this paper. We refer to the interested reader to [9] for more details.

ACKNOWLEDGMENTS

The authors would like to thank G. Ansaloni (from USI Lugano) and T. Wu, M. Sabry and S. Mitra (from Stanford University) for their insights and support in the architecture and used NVM technology.

REFERENCES

[1] World Health Organization., “Cardiovascular diseases,” 2015. [Online]. Available: www.who.int/topics/cardiovascular_diseases/en/

[2] A. Dogan et al. “Multi-Core Architecture Design for Ultra-Low-Power Wearable Health Monitoring Systems” In Proc. DATE, pp. 988-993, 2012.

[3] R. Braojos et al. “Hardware/software approach for code synchronization in low-power multi-core sensor nodes” In Proc. DATE, pp. 1-6, 24-28, 2014.

[4] M. Seok et al. “The Phoenix Processor: A 30pW Platform for Sensor Applications,” VLSI Circuits, pp. 188–189, 2008.

[5] A. Rahimi et al. “A fully-synthesizable single-cycle interconnection network for Shared-L1 processor clusters,” In Proceedings of DATE, 2011.

[6] A. D. Kent and D. C. Worledge, “A new spin on Magnetic Memories.” Nature Nanotechnology, vol. 10, pp. 187-191, 2015.

[7] F. Rincon et al. “Development and Evaluation of Multi-lead Wavelet-Based ECG Delineation Algorithms for Embedded Wireless Sensor Nodes,” Info. Tech. in Biomedicine, vol.15, no.6, pp. 854–863, 2011.

[8] H. Mamaghanian et al. “Compressed Sensing for Real-Time Energy-Efficient ECG Compression on Wireless Body Sensor Nodes”, In IEEE Trans. on Biomedical Engineering vol 58(9), pp.2456-2466, 2011

[9] R. Braojos et al. “Nano-Engineered Architectures for Ultra-Low Power Wireless Sensor Nodes”, In Proceedings of ESWeek 2016