Heterogeneous and Inexact: Maximizing Power Efficiency of Edge Computing Sensors for Health Monitoring Applications

Soumya Basu, Loris Duch, Miguel Peón-Quirós and David Atienza
Embedded Systems Laboratory (ESL)
Swiss Federal Institute of Technology Lausanne (EPFL)
Lausanne, Switzerland
{soumya.basu},{loris.duch},{miguel.peon},{david.atienza}@epfl.ch

Giovanni Ansaloni and Laura Pozzi
Faculty of Informatics
Università della Svizzera Italiana (USI)
Lugano, Switzerland
{giovanni.ansaloni},{laura.pozzi}@usi.ch

***Abstract*—In the Internet-of-Things (IoT) era, there is an increasing trend to enable intelligent behavior in edge computing sensors. Thus, a new generation of smart wearable devices for health monitoring is being developed, able to perform complex Digital Signal Processing (DSP) routines that extract features of clinical relevance from the acquired data. These new edge computing sensors for personalized healthcare must operate within a tight energy envelope; addressing the ensuing challenge, we herein introduce an inexact and heterogeneous edge computing architecture, specifically tailored to the bio-DSP domain. We observe that bio-signal analysis applications present task-level parallelism, intensive computational hotspots and a high degree of resilience towards errors. These characteristics drive our new bio-DSP edge node architecture design composed of multiple processing cores, a Coarse-Grained Reconfigurable Array (CGRA) accelerator, and hardware-software co-design support to become resilient to a non-zero probability of bit-flips at runtime. All these characteristics enable our new bio-DSP architecture to operate with an ultra-low voltage operating point. Indeed our results indicate that the energy benefits attained from the inclusion of all these characteristics in bio-DSP architectures are more than additive: task parallelism is harnessed both at the processor and the accelerator level, and the high tolerance of the CGRA towards voltage down-scaling is exploited to further decrease the IoT edge bio-DSP system energy envelope.**

I. Introduction

Embedded sensor devices that continuously acquire and analyze sensed data are opening novel and exciting opportunities for the future of healthcare in the context of Internet-of-Things (IoT), as they enable long-term monitoring of bio-signals outside the hospital environment with minimal medical supervision. These smart IoT sensors (also called edge computing sensors in the literature) perform complex bio-signal related Digital Signal Processing (bio-DSP) on the input bio-data to extract relevant features, which can then be wirelessly transmitted to a hub for decision making by specialized medical personnel. However, bio-DSP often involves complex functions and hence introduces high power requirements. To address this challenge, previous works have shown that task-level parallelism, typically present in bio-DSP applications, can be leveraged by multicore architectures with Single-Instruction Multiple-Data (SIMD) capabilities [1], while the execution of computational hotspots (*kernels*) can be effectively supported by Coarse-Grained Reconfigurable Array (*CGRA*) accelerators [2] [3] [4]. Thanks to the faster and more efficient execution achieved with the aforementioned techniques, smart bio-DSP IoT systems can remain longer in deep-sleep mode, consequently reducing their energy consumption.

An interesting and complementary paradigm is inexact computing [5], which waives the requirement of exact results, to reduce power requirements. Inexactness has been shown to be acceptable in some application domains where other considerations, such as energy efficiency, deadlines or throughput, are more important than precise results. Similarly, applications in the healthcare domain are amenable to some degree of errors because they are inherently subject to noise, while their outputs frequently have a qualitative or statistical nature. Therefore, they can tolerate a non-zero probability of runtime failures to improve the efficiency of the system, as long as the produced data retains appropriate clinical value and the system is never trapped in unrecoverable or erroneous states. This tolerance to inexactness can be exploited via Near-Threshold Computing (NTC), which enables aggressive power savings by operating electronic systems at sub-nominal voltage points.

In NTC, the reliability bottleneck of digital platforms resides in SRAM banks, which contain instructions and data [6]. On the contrary, combinational elements, such as Arithmetic Logic Units (ALUs), and registers (flip-flops) are less sensitive to the voltage supply level, particularly at low frequencies (few MHz). The robustness of SRAMs when operating at Near-Threshold Voltages (NTVs) is further hampered when runtime supply voltage fluctuations (*droops*) are taken into account.

In this work, we therefore focus on memories (data and instruction) as a source of failures, manifested as bit-flips in IoT edge systems. First, we analyze their application-level effect on two real-world bio-DSP benchmarks. Then, we introduce

This work has been partially supported by the E4Bio (grant no. 200021-159853) and the MagicISEs (grant no. 200021-156397) projects funded by the Swiss NSF, and by the HSD project (grant no. 15048) and the MyPreHealth project (grant no. 16073) funded by Hasler Stiftung.

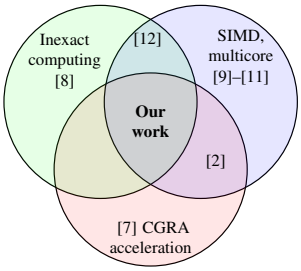

Fig. 1: Previous works have explored, in isolation, multicore processing, CGRA acceleration and inexact computing in the scenario of ultra-low power bio-signal processing. Here, we explore their ensuing synergies.

architectural techniques, such as memory protection, input and intermediate buffer surveillance, and monitoring of runtime synchronization between computing elements, to minimize the impact of failures and guarantee system recovery. While these techniques result in some (bounded) data loss produced by system resets, they ensure that computation resumes correctly afterwards, confining errors to small time windows without permanent effects.

Moreover, our results show that the energy gains derived from the combined application of CGRA acceleration, SIMD execution and inexact computing are more than additive. In fact, SIMD reduces the amount of memory accesses, while during the execution of kernels on CGRAs no instruction memory accesses are performed. Thus, both strategies increase the system robustness toward an aggressive scaling of the voltage supply. Indeed our study of the benefits deriving from this synergistic (hardware-software) co-design strategy (as depicted in Fig. 1) is the main contribution of this work. We embodied its conclusions in the design of an ultra-low power system, which features a CGRA mesh as a computational resource shared by multiple processors, and, at the same time, is able to operate at extremely low supply levels, countering the impact of ensuing memory bit-flips. Our results showcase energy efficiency gains of up to 70 % with respect to an equivalent multi-core bio-DSP IoT architecture that does not consider hardware acceleration and inexact computing.

II. Architecture

A. Multi-core system

As depicted in Fig. 2, and similarly to [2] and [12], the target platform features multiple RISC processors, which are interconnected through combinational crossbars to separate data and instruction memory banks. Each processor implements a Harvard architecture [13], supported by a three-stage pipeline, which can be clock-gated by a synchronizer unit while waiting for another processor to finish its task or when a kernel acceleration is running on the CGRA.

At the application level, the processors rely on several synchronization instructions to support SIMD execution modes and producer-consumer relationships between cores, as proposed in [1]. Then, an additional special instruction enables the request for a kernel acceleration in the CGRA [2]. Two further architectural modules define our bio-DSP IoT edge system. First, the *CGRA Controller* handles acceleration requests from processors, enabling both concurrent and sequential access to the reconfigurable resource. Requests are stored in a dedicated queue and are mapped on the CGRA mesh (whose structure is described in Section II-B) as soon as enough resources are available. Second, the *Execution Monitor* resets the platform if an incorrect system state (derived from a bit-flip in memory) is detected. Its architecture is detailed in Section II-C.

B. CGRA accelerator

The CGRA is composed of a mesh of Reconfigurable Cells (RCs) connected in a torus configuration. Each of them embeds a Configuration Register (CR, 16 words of 32 bits), which stores the instructions to be executed by an RC when processing a kernel. RCs support SIMD execution, by featuring two datapaths governed by the same control logic. Thus, the same kernel requests originating from different processors (and hence operating on different input data) can be efficiently processed in SIMD mode [2]. When not operating in SIMD mode, kernels can be concurrently mapped on separate CGRA columns. To support this feature, each column of the mesh has its own program counter, that selects the configuration word in the CR that should be executed at each clock cycle.

Each datapath is composed of an ALU, a local register file for storing intermediate data, and multiplexers required to select input operands (either from the local register file, or from the output generated by the RC itself, or by any of its 4 neighbors). Then, the ALU is capable of executing arithmetic operations (addition, subtraction, multiplication), arithmetic and logical shifts, and bitwise operations (AND, OR, XOR,

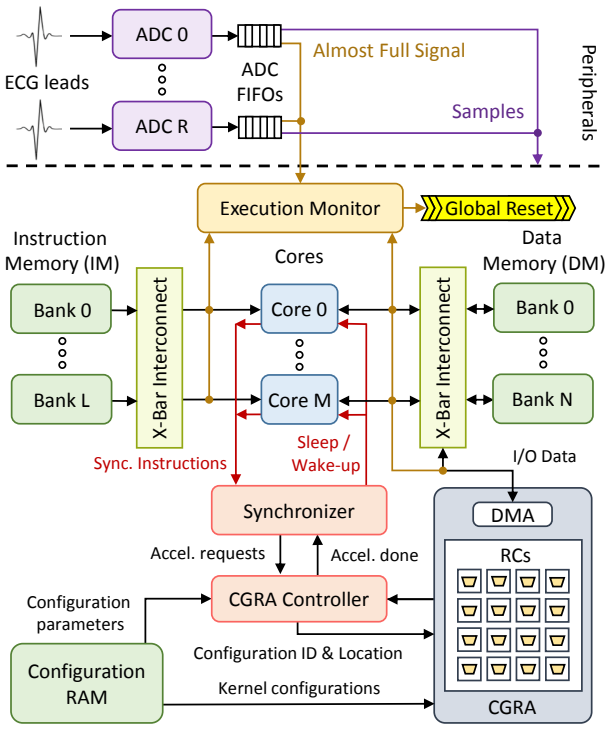

Fig. 2: High-level block scheme of the heterogeneous and inexact bio-DSP IoT edge system architecture.

XNOR). It also generates 1-bit condition flags (zero, sign and overflow), which are used in the MUX operation to implement if-conversion, thus allowing the execution of kernels with conditional statements. Moreover, one ALU per CGRA column is equipped with a square root calculator.

At runtime, each kernel must be configured (mapped) on the CGRA before its execution. In the configuration phase, the parameters related to the kernel (number of iterations, number of columns employed and modulo scheduling parameters, instruction bit-streams) are fetched from the CGRA Configuration RAM to configure the required CGRA column(s).

C. Execution Monitor (EM)

Even small voltage supply fluctuations can result in memory access failures when operating in the NTV region. Instead of employing large (and costly) voltage guardbands, our platform features a hardware monitor that checks system consistency by monitoring few critical elements: First-In First-Out (FIFO) buffers used for inter-processor I/O communications, the sequence of synchronization instructions issued by the cores, and the legality of performed instruction memory (IM) and data memory (DM) accesses. In particular, the EM (Fig. 2) supports the following error mitigation strategies:

- *Runtime coherence:* Detects inconsistencies in the synchronization sequence and in the execution flow of the cores. It ensures synchronization events are processed in a Last-In First-Out (LIFO) fashion for each core. A violation of this constraint indicates an erroneous execution flow (e.g., due to a jump to an incorrect location).
- *FIFO surveillance:* Monitors overflows and underflows in FIFOs, which indicate that a processor is stuck (e.g., due to a bit-flip causing the execution of an infinite loop).
- *DM protection:* Detects illegal data writes to read-only memory locations.
- *IM footprint violation:* Detects erroneous jumps to non-initialized IM addresses.

If an error is detected, the EM activates a global reset signal to restart the platform from a safe initial state. The above-mentioned consistency checks can be realized with small hardware overhead ($\sim 1\,\%$ of the total area for the considered system), and no performance penalties.

Similarly to [12], we classify the transient errors (i.e., bit-flips induced by voltage droops) affecting the execution of the system or the computed results into three categories:

- *Masked errors:* A bit-flip with no visible effects on the execution flow of the application or on its outputs.
- *Silent Data Corruption errors (SDC):* A bit-flip which has no visible effects on the execution flow of the application, but degrades the output results.
- *Unrecoverable errors:* A bit-flip leading to a critical runtime change in the execution flow of the application (e.g., memory footprint violation, execution stuck in an infinite loop), triggering the assertion of a system reset from the EM to recover a consistent state.

Crucially, consistency checks allow us to counter all unrecoverable errors and most SDCs (cf. Section III).

III. Experimental evaluation

A. Experimental setup

To evaluate our inexact and heterogeneous bio-DSP IoT architecture, we have developed a hybrid framework combining a post-synthesis model and a cycle-accurate simulator of the system represented in Fig. 2. The RTL, cycle-accurate (SystemC) model and compiler of the processors are derived using Synopsys ASIP Designer [14]. The complete system includes 8 processors, a data memory of $64\,\mathrm{KiB}$ ($32\,\mathrm{K}$ of $16\,\mathrm{bit}$ words) and an instruction memory of $96\,\mathrm{KiB}$ ($32\,\mathrm{K}$ of $24\,\mathrm{bit}$ words). The Configuration RAM has a size of $6\,\mathrm{KiB}$ ($1.5\,\mathrm{K}$ of $32\,\mathrm{bit}$ words), which is sufficient to store all the configurations of the considered kernels.

The CGRA, the EM and the memory subsystem have been developed as HDL modules (using pre-synthesized models from the memory vendor), and the whole system has been synthesized using Synopsys Design Compiler [15] with a $65\,\mathrm{nm}$ UMC low-leakage library. The kernels were simulated with Mentor Graphics ModelSim [16] and the corresponding switching activities were then used to accurately derive the power consumption of the CGRA, considering complete power-gating of unused columns at runtime. Similarly, the performance and power consumption of the various components of the inexact system were derived by running small synthetic benchmarks (both with and without SIMD support) on a post-synthesized design. The obtained energy profiles, along with the kernel run-times on the CGRA, were then used in the SystemC simulator of the system, thus enabling faster simulations.

To validate the experimental setup we employed two bio-signal processing application benchmarks, written in C:

1) Eight-lead Compressed Sensing (CS) [2], which is a highly parallel application to perform lossy encoding of eight ECG leads, where each lead is processed by a different core.

2) Four-lead Morphological Filtering [17], combined with CS (MF-CS). The MF stage processes four ECG leads in parallel (one core per lead), removing high- and low-frequency noise with morphological operators. The filtered ECGs are then compressed by the remaining four cores, producing a filtered and compressed ECG as output. Both benchmarks receive as input ECG signal windows of 1024 samples acquired at $500\,\mathrm{Hz}$ and extracted from the T-Wave Alternans Challenge database [18]. After the $50\,\%$ CS compression, signal windows of 512 samples are delivered as output.

We rely on VARIUS-NTV [19] to obtain SRAM failure rates due to stability and timing violations. Furthermore, we model voltage droops as random fluctuations, following a normal distribution centered on the nominal supply value Vdd and with variance σ_n. Three types of memory access corruptions are considered: data read, data write and instruction read. For each benchmark and for each type of memory access failure, a representative set of 1000 input windows was processed, each incurring in a bit-flip at run-time. Then, we observed the bit-flip impact at the application level (i.e., if it caused an SDC, an unrecoverable error, or if its effect was masked).

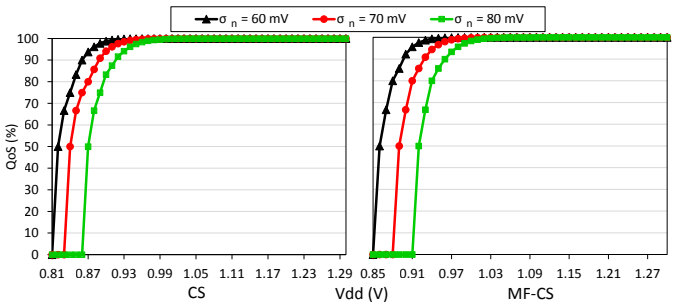

Fig. 3: QoS for different voltage supply values and droop variations for the inexact bio-DSP IoT architecture.

B. Experimental results

To assess the system performance from an inexact computing perspective, we define, as metric for the Quality-of-Service (QoS) of the system, the ratio between the windows that terminate the required processing and the total number of signal windows. In our experiments, we have pessimistically assumed that an entire ECG window is discarded upon the detection of any error by the EM. The resulting QoS of the system depending on the supply voltage in presence of voltage droops is presented in Fig. 3. The QoS remains high ($> 90\,\%$) till the supply voltage is reduced to $0.89\,\mathrm{V}$ and $0.93\,\mathrm{V}$ (at $\sigma_n = 70\,\mathrm{mV}$), for CS and MF-CS, respectively. These supply voltage values correspond to withstanding bit-flip rates in the order of thousands per hour. In comparison, an exact system that guarantees less than one bit-flip per year needs a supply voltage of $1.3\,\mathrm{V}$, meaning that the inexact system reduces the supply voltage requirements by $31.5\,\%$ and $28.5\,\%$ for CS and MF-CS, respectively.

Fig. 4 shows the energy consumed per valid window of computed ECG (i.e., the ones that are not discarded due to a reset assertion by the EM). Using this graph, the optimal supply voltage can be inferred as the point in which the amount of energy per valid window is minimized. Considering an average $\sigma_n = 70\,\mathrm{mV}$, the optimal Vdd values are $0.9\,\mathrm{V}$ and $0.95\,\mathrm{V}$, which guarantees a very high QoS of $94.1\,\%$ and $96.7\,\%$ for CS and MF-CS, respectively. The number of bit-flips per hour at these supply voltages are 6500 and 1050, respectively.

To assess the effect of undetected SDC errors on the received signal quality, we calculate the Signal-to-Noise Ratio (SNR) and the Percentage Root-mean-square Difference (PRD) of the received signals [20]. We compare them to error-free executions, according to the expected amount of correct windows and windows affected by SDCs (computed by our system depending on the voltage supply level). Our results show that, when working at the optimal Vdd, the system's SNR remains very high ($65.8\,\mathrm{dB}$ for CS and $79.3\,\mathrm{dB}$ for MF-CS), with a net PRD falling in the range of very good signal quality [21] (0.36 for CS and 0.11 for MF-CS).

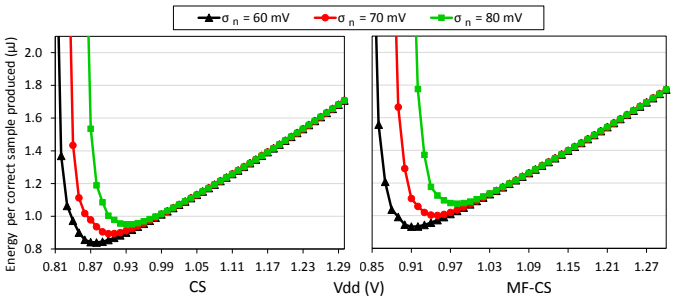

Fig. 4: Evolution of the energy consumption per compressed sample produced, for different voltage supply values and droop variations for the inexact bio-DSP IoT architecture.

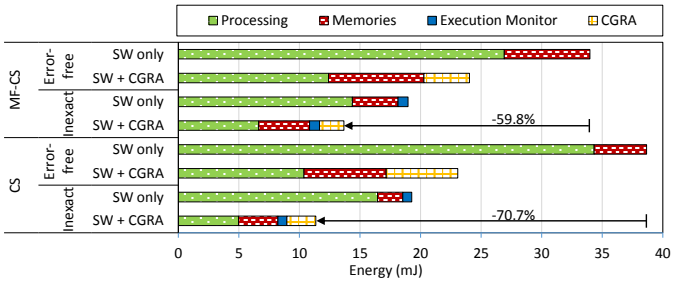

Fig. 5: Energy consumption breakdown for the studied bio-DSP IoT system implementations.

Finally, Fig. 5 compares the energy consumption of different platforms, for the two considered applications. The reference point is the exact bio-DSP IoT system (running at $1.3\,\mathrm{V}$) without CGRA acceleration. The first comparison is made with a system working at the same Vdd but with CGRA support. By transferring the execution of kernels to the accelerator, the time to process a sample of signal is reduced by $36.3\,\%$ and $41.7\,\%$ on average for CS and MF-CS, respectively, leading to energy savings of $40.3\,\%$ and $29.2\,\%$ for the two applications. Thanks to inexact computing, further energy savings can be obtained, due to voltage overscaling. When running at the energy-optimal Vdd obtained from Fig. 4, the inexact system without CGRA increases its efficiency by $50.2\,\%$ and $44.2\,\%$, for the two considered benchmarks. By combining an inexact system with CGRA acceleration, $70.7\,\%$ and $59.8\,\%$ energy consumption savings are obtained for CS and MF-CS, respectively, over an exact system without CGRA acceleration.

IV. Conclusion

Low-power edge computing platforms are fundamental components in advanced IoT ecosystems for healthcare applications. In this paper we have explored how the combination of several techniques, namely, multicore processing, CGRA acceleration and inexact computing, can synergistically reduce the energy consumption required for bio-DSP IoT systems.

Our proposed IoT system, embedding these characteristics, achieves savings in energy consumption of up to $70.7\,\%$ and $59.8\,\%$ for two real-world workloads (standalone compressed sensing or in combination with morphological filtering), in comparison with traditional exact architectures without CGRA accelerators. Thus, experimental evidence showcases that considerable energy benefits can be obtained by leveraging the high-level characteristics of edge computing applications for healthcare in the design of ultra-low power, domain-specific smart bio-DSP IoT platforms.

REFERENCES

[1] R. Braojos, A. Dogan, I. Beretta, G. Ansaloni, and D. Atienza, "Hardware/software approach for code synchronization in low-power multi-core sensor nodes," in *Design, Automation and Test in Europe Conference and Exhibition (DATE), 2014*. IEEE, Mar. 2014, pp. 1–6.

[2] L. Duch, S. Basu, R. Braojos, G. Ansaloni, L. Pozzi, and D. Atienza, "HEAL-WEAR: An ultra-low power heterogeneous system for bio-signal analysis," *IEEE Transactions on Circuits and Systems I*, vol. 64, no. 9, pp. 2448–2461, Sep. 2017.

[3] M.-H. Lee, H. Singh, G. Lu, N. Bagherzadeh, F. J. Kurdahi, E. M. Filho, and V. C. Alves, "Design and Implementation of the MorphoSys Reconfigurable Computing Processor," *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, vol. 24, no. 2, pp. 147–164, Mar. 2000.

[4] N. Ozaki, Y. Yasuda, M. Izawa, Y. Saito, D. Ikebuchi, H. Amano, H. Nakamura, K. Usami, M. Namiki, and M. Kondo, "Cool Mega-Arrays: Ultralow-Power Reconfigurable Accelerator Chips," *IEEE Micro*, vol. 31, no. 6, pp. 6–18, Nov. 2011.

[5] S. Mittal, "A survey of techniques for approximate computing," *ACM Computing Surveys (CSUR)*, vol. 48, no. 4, p. 62, May 2016.

[6] S. Khare and S. Jain, "Prospects of Near-Threshold Voltage Design for Green Computing," in *International Conference on VLSI Design and International Conference on Embedded Systems (VLSID)*, Jan. 2013, pp. 120–124.

[7] K. Patel, S. McGettrick, and C. J. Bleakley, "Syscore: A Coarse Grained Reconfigurable Array Architecture for Low Energy Biosignal Processing," in *Field-Programmable Custom Computing Machines (FCCM), 2011 IEEE 19th Annual International Symposium on*. IEEE, 2011, pp. 109–112.

[8] D. Bortolotti, H. Mamaghanian, A. Bartolini, M. Ashouei, J. Stuijt, D. Atienza, P. Vandergheynst, and L. Benini, "Approximate Compressed Sensing: Ultra-Low Power Biosignal Processing via Aggressive Voltage Scaling on a Hybrid Memory Multi-Core Processor," in *Proceedings of the 2014 international symposium on Low power electronics and design*. ACM, 2014, pp. 45–50.

[9] R. Braojos, D. Bortolotti, A. Bartolini, G. Ansaloni, L. Benini, and D. Atienza, "A Synchronization-Based Hybrid-Memory Multi-Core Architecture for Energy-Efficient Biomedical Signal Processing," *IEEE Transactions on Computers*, vol. 66, no. 4, pp. 575–585, 2017.

[10] A. Y. Dogan, R. Braojos, J. Constantin, G. Ansaloni, A. Burg, and D. Atienza, "Synchronizing Code Execution on Ultra-Low-Power Embedded Multi-Channel Signal Analysis Platforms," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2013*. IEEE, 2013, pp. 396–399.

[11] Y. He, Y. Pu, R. Kleihorst, Z. Ye, A. A. Abbo, S. M. Londono, and H. Corporaal, "Xetal-Pro: An Ultra-Low Energy and High Throughput SIMD Processor," in *Proceedings of the 47th Design Automation Conference*. ACM, 2010, pp. 543–548.

[12] S. Basu, L. Duch, R. Braojos, G. Ansaloni, L. Pozzi, and D. Atienza, "An inexact ultra-low power bio-signal processing architecture with lightweight error recovery," *ACM Trans. Embed. Comput. Syst.*, vol. 16, no. 5s, pp. 159:1–159:19, Sep. 2017.

[13] J. L. Hennessy and D. A. Patterson, *Computer Architecture: A Quantitative Approach*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2003, ch. 1.

[14] "ASIP Designer," https://www.synopsys.com/dw/ipdir.php?ds=asip-designer, 2017.

[15] Synopsys, Design Compiler Software, https://www.synopsys.com/implementation-and-signoff/rtl-synthesis-test/design-compiler-graphical.html, Nov. 2017.

[16] Mentor Graphics, ModelSim Software, https://www.mentor.com/products/fv/modelsim/, Nov. 2017.

[17] Y. Sun, K. Chan, and S. Krishnan, "ECG signal conditioning by morphological filtering," *Computers in Biology and Medicine (CBM)*, vol. 32, no. 6, pp. 465–479, Nov. 2002.

[18] "PhysioBank," http://www.physionet.org/physiobank/, Aug. 2016.

[19] U. R. Karpuzcu, K. B. Kolluru, N. S. Kim, and J. Torrellas, "VARIUS-NTV: A Microarchitectural Model to Capture the Increased Sensitivity of Manycores to Process Variations at Near-Threshold Voltages," in *Dependable Systems and Networks (DSN), 2012 42nd Annual IEEE/IFIP International Conference on*. IEEE, Jun. 2012, pp. 1–11.

[20] H. Mamaghanian, N. Khaled, D. Atienza, and P. Vandergheynst, "Compressed sensing for real-time energy-efficient ecg compression on wireless body sensor nodes," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 9, pp. 2456–2466, Sep. 2011.

[21] Y. Zigel, A. Cohen, and A. Katz, "The weighted diagnostic distortion (wdd) measure for ecg signal compression," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 11, pp. 1422–1430, Nov. 2000.