Inexact-Aware Architecture Design for Ultra-Low Power Bio-Signal Analysis

Soumya Basu[1],*, Pablo G. Del Valle[1], Georgios Karakonstantis[2],
Giovanni Ansaloni[3], Laura Pozzi[3] and David Atienza[1]

[1] Embedded Systems Laboratory, École Polytechnique Fédérale de Lausanne, Switzerland
[2] School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, U.K
[3] Universitá della Svizzera Italiana, Lugano, Switzerland
*Corresponding author: soumya.basu@epfl.ch

Abstract: This paper introduces an inexact, but ultra-low-power, computing architecture devoted to the embedded analysis of bio-signals. The platform operates at extremely low voltage supply levels to minimize energy consumption, especially in the memory subsystem, which accounts for a major part of it. However, a high reliability of memories cannot be guaranteed at ultra-low voltages, when using conventional 6-transistor SRAMs. While error correction codes and dedicated SRAM implementations can ensure correct operations in this near-threshold regime, they incur in significant area and energy overheads, and should be therefore employed judiciously.

In this scenario, we propose a novel scheme to design inexact computing architectures that selectively protects memory regions based on their significance, i.e., their impact on the end-to-end quality of service, as dictated by the bio-signal application characteristics. Herein, we illustrate our scheme on an industrial benchmark application performing the Power Spectrum Analysis (PSA) of electrocardiograms. Experimental evidence showcases that a significance-based memory protection approach leads to a small degradation in the output quality, while resulting in substantial increase in energy efficiency for embedded signal processing, with respect to an exact computing implementation. This approach thereby augments ultra-low voltage scaling.

Keywords: Ultra-Low Power, embedded systems, wearable health monitors, error tolerance, power spectral analysis.

1. Introduction

Modern society is witnessing changes in lifestyle more than ever before. Busy and unhealthy lifestyles are becoming common, resulting in a rise in the number of people developing or living with cardiovascular conditions. Moreover, a significant part of the world population is aging, and hence becoming in danger of contracting cardiac diseases. This scenario calls for increased levels of medical supervision and management, which are resulting in high costs, and traditional health care infrastructures are finding it increasingly difficult to cope with these demands [1].

Emerging Wireless Body Sensor Network [12] technologies can offer large-scale and cost-effective solutions to this problem. These wearable devices for bio-signal monitoring are bringing about a revolutionary change in healthcare systems by allowing long-term monitoring of chronic patients, while providing a low-cost and unobtrusive solution. Wireless Body Sensor Nodes (WBSNs) [28], represented in Fig. 1, are the building blocks of such a network, being able to provide real-time and personalized monitoring of patients and coordinate with medical staff depending on the patients' medical records. They are designed to monitor different organs of the human body, including the heart. Such devices involve sensing of bio-signals and then transmitting them wirelessly to receiver devices, for further analysis. The analysis of the received sensed data generally consists of labour-intensive manual inspection or offline execution on a server infrastructure.

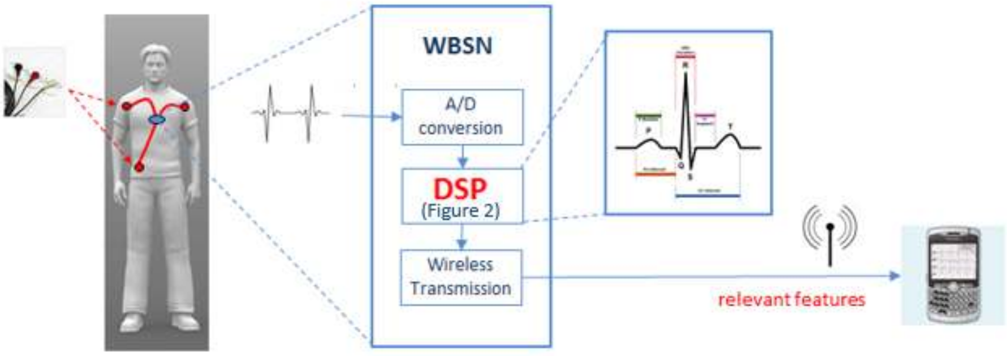

Fig 1. *Schematic representation of a simple WBSN node*

Recently, however, a new generation of smart WBSNs has emerged which are able to perform digital signal processing (DSP) directly on-board to analyse the acquired bio-signals and extract clinically-relevant features, in addition to data acquisition and transmission [13]. These devices pave the way for truly autonomous and versatile health monitoring devices. They run various bio-signal processing applications, which are useful for doctors for quick analysis of important signal data. A basic scheme of the digital processing unit of a state-of-the-art WBSN is depicted in Fig. 2, which typically comprises low power processors along with supporting memories. It is assumed that the system loads the instructions for processing into its instruction memory (typically an SRAM) from a non-volatile memory (NVM), like flash, etc., at start-up, or bootstrap phase. It can then autonomously run the required applications.

Energy efficiency is of paramount importance in these battery operated WBSNs, as they work under tight energy constraints defined by battery-based power supplies on the device. Performing on-chip signal analysis results in increased computation, thereby increasing the energy consumption. For this reason, there is an urgent need for efficient energy management in WBSNs, which has fuelled significant research interest. In this context, inexact computation or approximate computation is a new method to achieve higher energy efficiency in WBSNs. It involves *trading off the accuracy of logic circuits in order to save energy*, by applying techniques like voltage scaling [14] and circuit pruning [15], among others.

In this paper, we present an architecture that follows a paradigm shift in the processing of bio-signals from exact computation to an inexact one. Bio-signal processing applications normally acquire noisy input and produce qualitative outputs, thereby being error-resilient in nature. Also, they frequently involve storing data that is *sparse* in nature [25], with the memory components accounting for a key part of the energy consumed [18].

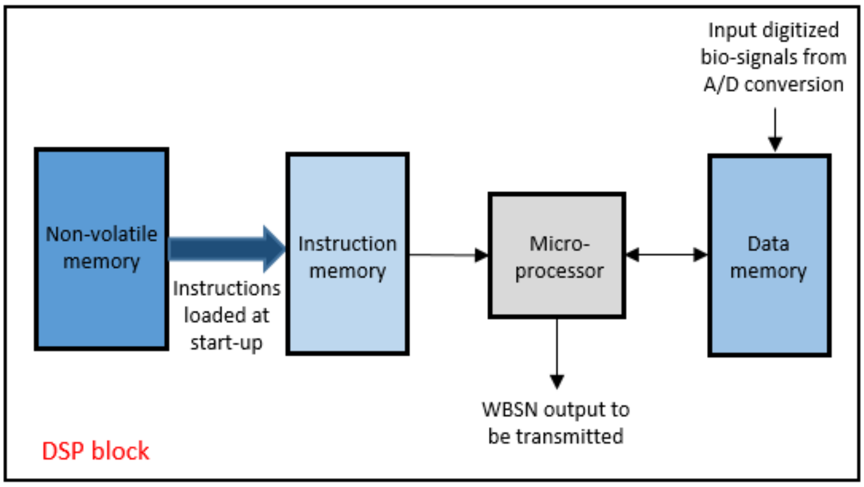
Input digitized bio-signals from A/D conversion
Non-volatile memory
Instructions loaded at start-up
Instruction memory
Micro-processor
Data memory
WBSN output to be transmitted
DSP block

Fig 2. *Block scheme of a typical WBSN's processing unit*

An efficient method to achieve large energy savings in the execution of these applications is voltage-scaling. However, at near-threshold voltage supplies, errors are introduced in the memory subsystem. But the noise-resilient properties of bio-signal processing applications motivate the application of voltage scaling. In this context, we introduce the concept of *data significance*, and we propose a significance-based heterogeneous protection of the memory working at near-threshold supplies. We explore the effects of the proposed scheme on the energy savings obtained w.r.t protecting the whole memory subsystem, and the effect of errors introduced on the final output.

To evaluate our proposed methodology, we selected the power spectral analysis (PSA) application of the heart rate variability (HRV) out of the many applications that have been proposed to predict heart diseases, which range from the automated detection of epileptic seizures [4], to the predictive risk assessment of atrial fibrillations [16]. This choice is based on the fact that PSA is one of the most widely used strategies for predicting cardiac failures, as it allows the monitoring of various health condition associated with the heart, as well as other organs [31] [17].

The implementation of PSA on ultra-low power embedded devices requires a carefully tailored digital architecture. Lowering supply voltages might result in quadratic energy savings, but memory subsystems using conventional 6-transistor (6T) SRAM cells become unreliable in nanometre technologies [5]. Soft errors like bit-flips start occurring at near-threshold voltages, and their probability of occurrence increases as the supply voltage is further lowered [21]. These issues regarding reliability of memories become more prominent with modern technology scaling, as we conceive transistors with smaller lengths. In this context, larger memory cells have been proposed in previous research work, which consist of 8-transistors (8T) or 10- transistors (10T), because they ensure reliable operation of the memories at much lower supply voltages compared to 6-transistor (6T) cells [6]. However, 8T and 10T cells come with a high area overhead, thus limiting the amount of memory space that can be included in the already area-constrained WBSNs. It has been shown that the majority of the silicon real estate of typical WBSN processors is devoted to the memory subsystem, with this trend set to become even more emphatic in the near future, where memories are forecasted to occupy as much as 95% of the entire chip area [27]. Thus, the use of 6T cells over 8T or 10T cells is preferable in terms of area. Another proposed approach in the literature to deal with errors in memories is to use error correction code (ECC) [8], in the 6T SRAM memories. The negative aspect of this method is that they present significant area and energy overheads when all or large parts of the memories need to be protected. This fuels the need of an effective memory protection system at low supplies, which presents low area and energy overheads, and at the same time guarantees reliable operation of the system.

In this article we introduce a novel memory protection scheme which takes advantage of the *sparsity* of data in the targeted application. We advocate the application of inexact computation scheme based on significance of data rather than based just on significant bits, to achieve high energy savings, while binding the error introduced in the output of the application within permissible limits.

The main contributions of this paper are the following:

1) We analyse the PSA application's software code in order to explore its statistical properties. This includes analysis of the data elements in the intermediate steps of the application and the classification of the data into more significant and less significant, depending on their contribution to the output quality of the system.
2) We introduce a novel hybrid memory protection scheme, which involves a significance-based protection in hardware of the data memory for the PSA system, using ECC bits. This enables us to explore a system-wide application of ultra-low voltage scaling. This is an extension of our previous research work [24], and includes the effects of voltage scaling in the instruction memory and the processor system, in addition to the data memory, as done before. Thus, in this paper we study the energy consumption of the whole of the WBSN's processing system using our proposed scheme.

3) We estimate the effects of using voltage scaling with the proposed memory protection scheme on the performance of the PSA application and the energy savings that the scheme results in.

The rest of the paper proceeds as follows. In Section 2, we introduce our proposed scheme of memory protection driven by data significance. Section 3 presents a case study where we have analysed the power spectral analysis application. In section 4 we explain our experimental setup, followed by the results we have obtained, before finally concluding the paper.

2. Data-significance driven criticality approach in WBSNs

The impact of adopting ultra-low voltage supplies is not homogeneous across architectural blocks. Combinational logic circuits such as Arithmetic Logic Units (ALU) are the least affected, because they are stateless and do not present internal feedback connections. Indeed, combinational circuits have been proposed operating at a supply voltage (Vdd) in the range of few hundreds of milliwatts [19]. Closely coupled with ALUs is the register file, which embodies the first level of the memory hierarchy. In a load/store architecture such as ARM, target of the present work, all instructions except explicit loads and stores operate with data residing in registers. The implementation of the register file (the Cortex M3 has 16 architecturally-visible registers) is usually based on Standard Cell Memories (SCMEMs), which can reliably operate at extremely low-voltage levels [20]. From these two observations, we conclude that the computing core, integrating the ALU and the register file, is not the resiliency bottleneck of the system.

Conversely, the Data and Instruction Memories (DM and IM, respectively) of WBSNs are commonly implemented as 6T Static RAMs (6T-SRAMs), which are more prone to random bit-flips when operated in the near-threshold regime. In [21], a non-negligible probability of error of $1.3e^{-5}$ is reported for a Vdd of 0.75 Volts, which rapidly escalates as the supply level drops. Including different voltage supply levels confirming the reliable usage of the processor and its associated registers on one hand and the instruction and data memories on the other could be seen as a simple solution to tackle reliability issues at low voltages. However, such an approach requires multiple voltage regulators, as well as voltage converters to maintain uniform voltage level for logic. This solution is thereby not efficient for ultra-low power platforms [18]. More complex cell structures employing dedicated read and write paths (8T- or 10T- SRAMs) can reliably operate at a lower voltage supply, but at the cost of important area and energy overheads. In [18] and from the CACTI memory modelling tool [32], it is found that 8T cells occupy 30% more real-estate, and consume on average 25% more dynamic energy, than comparable 6T implementations at 40nm technology. An alternative path to ensure reliability is to detect and correct errors using redundant representations of the memory content, adding dedicated Error Correction Codes (ECCs) to transparently recover from bit-flips. Even in this case, area and power overheads have to be accounted for, due to the dedicated memory cells required to store the redundant information and the logic required to recover from errors.

Herein, we propose to minimize the above-mentioned overheads by imposing different reliability guarantees, depending on the criticality of the data. The IM content is highly susceptible towards errors, as a single bit-flip can lead to an unpredictable execution flow, or even to unrecoverable states (e.g.: if a jump to a random location is made). The impact of bit-flips in DM can instead be less pronounced. While control variables and address manipulations have indeed high criticality levels, a large portion of the DM in bio-signal analysis applications is employed to store windows of data containing inputs and outputs of the various stages of digital processing. Errors in these buffers do not lead to catastrophic failures, but can cause an inacceptable degradation of the output quality.

Crucially, the loss in end-to-end quality of service, or in other words the net performance of the system, deriving from random errors, is dependent on the statistical properties of the stored data. Herein, this characteristic is leveraged to guide the design of heterogeneous protection schemes, which provide correction, detection or only a best-effort guarantee on different memory sections, maximizing the quality of service for a target energy budget. In our approach, we distinguish between two important cases, addressing *sparse* and *non-sparse* buffers. In the latter case, the magnitude of each entry in a buffer array is randomly distributed. For these arrays, each entry equally contributes to the overall correctness of the computation, so each stored word must expose the same reliability level. Protection of non-sparse buffer must be therefore performed at the bit-level, ensuring the correctness of high-order bits, while possibly allowing a degree of inexactness for low-order ones.

This strategy, however, becomes sub-optimal when the memory content is mostly centred on an expected value, with only few words significantly deviating from it, which is often the case in WBSN applications [7]. This *sparsity* property allows the adoption of a word-based, instead of bit-based, protection scheme, in which error correction is employed for the small subset of data, which is not close to the expected value (and we term this subset *significant)*, while a much simpler error detection mechanism is used for the rest. In this way, correctness is ensured for significant words, while bit-flips in the non-significant parts are only partially countered, by adopting the expected value of the data upon the detection of an error.

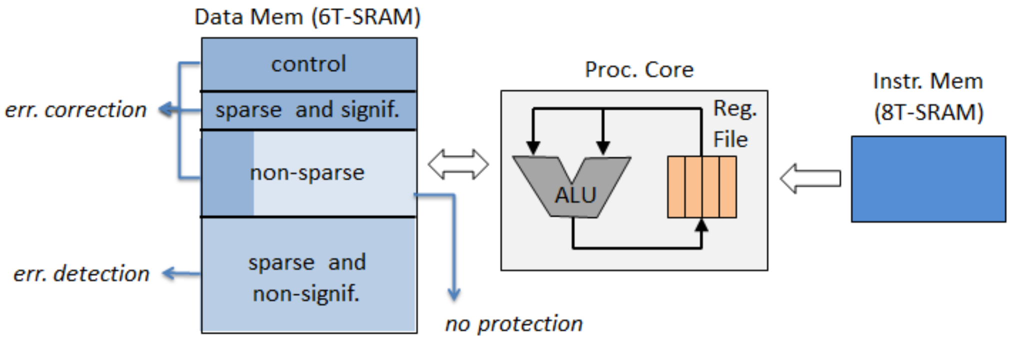

Fig 3. *The proposed heterogeneous protection scheme applies varying exactness guarantees depending on the data criticality.*

We apply the above-mentioned considerations to the design of the target inexact architecture, whose block scheme is provided in Fig. 3. To increase reliability at low supply levels, the instruction memory is realized with 8T-SRAM cells. As for the data memory, energy- and area- efficient 6T-SRAMs are employed, coupled with heterogeneous error detection/protection features. This arrangement results in a simple implementation of the IM (which is protected in its entirety), while allowing a fine-grained tuning of the DM reliability, dependent on the criticality of the stored values. Scalar and control variables (non-buffer data), as well as the highly significant portions of the buffer data, are fully protected with multi-bit error detection codes. Conversely, only error detection (implemented as a 1-bit parity code), but not correction, is employed for non-significant buffer data in sparse arrays. Finally, for the non-sparse arrays, errors in the most significant bits are corrected, while no detection or correction is performed for the least significant parts of each word.

3. Data-significance analysis of the Power Spectral Analysis system

3.1 Functionality of the PSA System

In this paper, we use as a case study the power spectral analysis of the HRV, which is a powerful tool for evaluating the autonomic control of the heart rate and identifying various health conditions [22]. Fig. 4 shows the block scheme of the system.

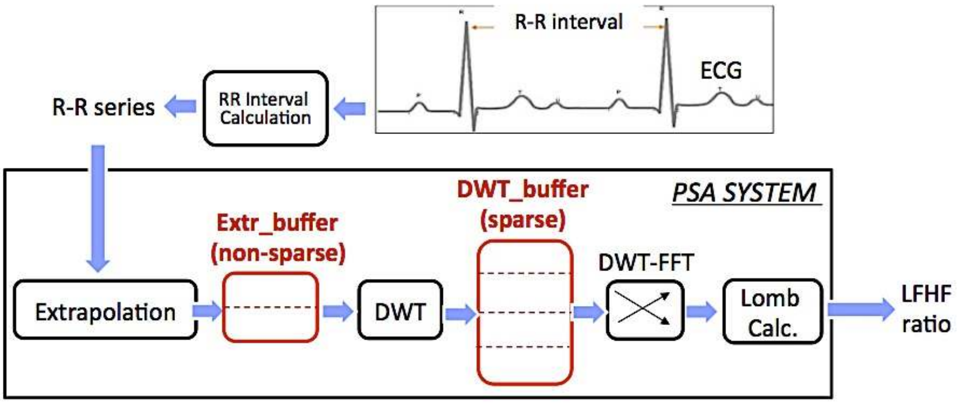

Fig 4. *Block scheme of the PSA application*

The PSA system is composed of 4 essential steps:

1) In the first step, the time differences between consecutive heartbeats (known as RR intervals) are extracted from ECG recordings of patients. The RR intervals are non-periodic signals, and thus require to be processed by dedicated algorithms, such has the Fast Lomb periodogram method [23].
2) In the next step, according to the Fast-Lomb method, the extracted RR intervals are extrapolated to a fixed size window (i.e. 512 samples). This procedure essentially converts the non-periodic signals into uniformly sampled ones.
3) Then the uniformly-sampled data are processed to estimate the specific trigonometric functions required by Fast-Lomb. Traditionally, such an estimation is performed by applying a Fast-Fourier Transform (FFT). Instead, and similarly to [23], in this paper we use a wavelet-based FFT (WFFT), which reduces substantially (up to 28% w.r.t the state-of-the-art) the complexity of the Fast-Lomb method [26] and tends to introduce *sparsity* in the bio-signals [30], especially in the heartbeats. In particular, the wavelet transform involved in the WFFT helps in revealing the sparse nature of bio-signals in the wavelet domain, exposing eventually the terms that are zero (or close to zero). Such close-to-zero terms and the following butterfly operations applied in the second stage of the WFFT can then be pruned, eventually reducing the computational complexity.
4) Finally, the Lomb calculator combines the output data obtained from WFFT, estimating the real-time power spectrum information. In clinical practice [23], the most used metric derived from PSA is the ratio between the power in low frequencies (LFP, defined as 0.04 – 0.15 Hz) and high

frequencies (HFP, 0.15 – 0.4 Hz), with *LFHF Ratio=LFP/HFP*. A deviation of the *LFHF Ratio* above or below normal values is indicative of various health issues [9].

3.2 Analysis of the PSA System

The application of our scheme requires the identification of the statistical characteristics of the target application for identifying blocks of data where it can be applied. To this end, we have analysed the PSA system and estimated the data distribution in the various stages, by performing several experiments with the ECG recordings. Fig. 5 focuses on the distribution of the data in the two memory buffers used in the system: the `Extr_buffer` is used to store the output of the extrapolation on the input data and the `DWT_buffer` is used to store the DWT outputs, as indicated in Fig. 4.

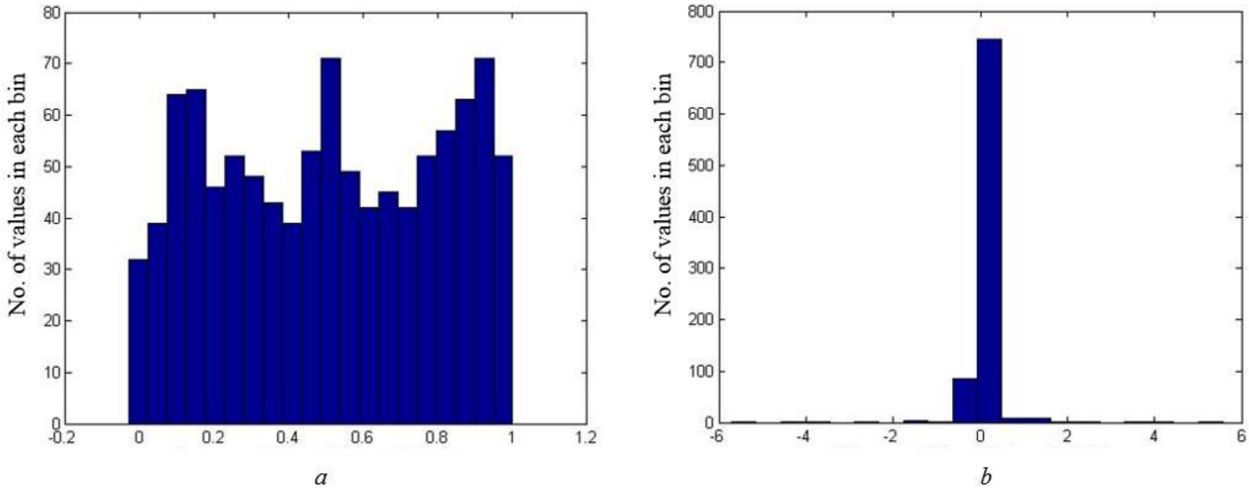

Fig 5. *Histogram of data values in `Extr_buffer` and `DWT_buffer` (normalized), distributed in 20 bins.*
a. `Extr_buffer` presents a non-sparse distribution
b. `DWT_buffer` presents a sparse distribution

We can observe in these two figures that the elements of the `DWT_buffer` (Fig. 5b) are mostly centred on zero, justified by their sparse nature, while the elements of the `Extr_buffer` (Fig. 5a) have a non-sparse distribution.

The different data distribution patterns indicate that different protection approaches against memory faults can be applied for limiting the overhead. Intuitively, for taking advantage of the error resiliency of such an application we apply a scheme in which the most significant bits (MSBs) of every word in the `Extr_buffer` are protected with a state-of-the-art mechanism such as ECC, whereas the least significant bits (LSBs) are not protected against memory faults by any specific mechanism, as depicted in Fig. 6.

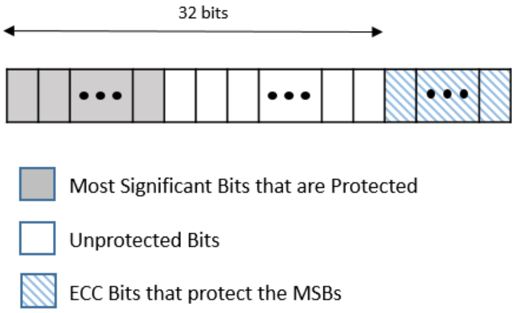
32 bits
Most Significant Bits that are Protected
Unprotected Bits
ECC Bits that protect the MSBs

***Fig 6.** A memory word in the* `Extr_buffer`

On the other hand, in case of the `DWT_buffer`, the distribution of the stored data allows us to apply a more elaborate protection scheme. In particular, rather than protecting groups of bits, here we can protect complete words, distinguishing between significant and less significant ones. In fact, in case of the less significant data, as most of the values are close to zero, it is possible to replace them with their expected value (zero) if an error occurs in a word. This ensures that the impact of such an error will not drastically affect the expected data, since a flipped bit within each of the close-to-zero data can alter completely the magnitude of the stored value. Error detection is supported by a single parity bit per word, resulting in a small overhead with respect to error correction.

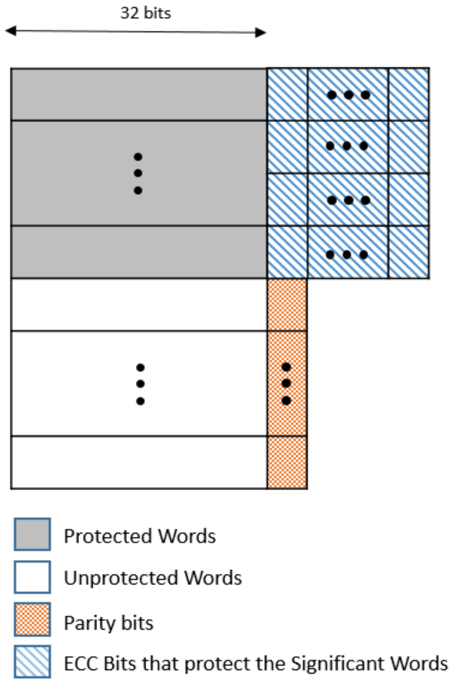

***Fig 7.** Block diagram of the* `DWT_buffer`

For the significant data, which has magnitudes much larger than zero, a more expensive error correction scheme must be applied for ensuring their correct storage. In the PSA application, such elements reside in the low-frequency outputs of the DWT. In this case, we considered SECDED (Single Error Correction, Double Error Detection) ECC (Fig. 7). Six ECC bits are required to protect a 32-bit word, as used in the data memories.

Note that the partition between significant and less significant words can be selected statically, i.e. independently from the particular window of inputs being processed. Identification of sparse and non-sparse data can be done by making an off-line analysis of the application, and thereby does not need any run-time overhead. In case of the Fast Lomb (FDWT), the data are distinguished based on the inherent properties of the DWT, resulting in the separation of the processed data into high and low frequencies. High frequencies (close to zero data) are termed as non- significant, whereas low frequencies are termed as significant.

4. Experimental setup

To evaluate the application of the heterogeneous scheme in the PSA application, we retrieved the input ECG signals from real-world recordings, available in the Physionet PAF prediction database [2]. This database includes 300 recordings, each of 30 minutes. Each recording consists of data acquired from two simultaneously operating ECG sensors, and the root mean squared (RMS) value of the data from the two sensors. We have considered time windows of 6 minutes, with an overlap of 5 minutes, for processing the input data. As a result we obtained 25 time windows for each of the 30-minute long recordings. The data from each window in each recording were independently processed by the application to retrieve their LFHF ratio.

To estimate the power consumption of the proposed system, we developed high-level models of its different components, including the processor, the data and the instruction memory, whose implementation is detailed as follows.

We considered a technology node of 40 nm as a realistic example for next-generation WBSN nodes (which are currently 65nm and 90nm), and an ambient temperature of 300 degrees Kelvin (27º C). To obtain the static energy from static power figures, we assumed a working frequency of 168 MHz and measured the run-time of the PSA application on all input windows on the target Cortex-M3 processor, which corresponds to 2.23 seconds. The dynamic energy consumption of the processor was calculated from the values reported in its datasheet [3]. Its leakage energy was derived by considering the leakage power values of the processor while in standby mode, with the clock inactive. Both static and leakage energy figures were scaled according to the target supply voltage.

The instruction memory, which always needs to operate reliably, is implemented by using 8T SRAM cells. The energy consumption for loading the instructions into it from an NVM during start-up can be neglected, as the time spent for it (a few milliseconds) is in orders of magnitude smaller than the actual run-time of the application, which is typically from hours to days. Smart allocation policies have been proposed, which avoid full shadowing, and thus reduce the required on-chip SRAM sizes [29]. But this approach involves a careful integration with ultra-low power applications, as NVMs consume much higher energy per access with respect to SRAMs. Therefore they have not been considered in our proposed scheme. Novel NVMs have been presented which have much higher energy efficiency when compared to traditional flash memories [11]. Although in the future the static power consumption SRAMs will become more relevant, in the considered technology, NVMs still trail the energy efficiency of SRAMs, justifying the usage of the latter for implementing the instruction memory. The static power and dynamic energy figures of the IM were obtained from [32], adapting them to the target 40nm technology. To calculate the IM dynamic read energy, we considered a worst-case scenario in which an instruction is fetched every clock cycle. Conversely, for write energy, we assumed that the IM is written only once at the beginning of execution, as shown in Fig. 2.

The data memory, realized with 6T SRAMs, is itself divided in two sections. The first one comprises the part outside the `Extr_buffer` and `DWT_buffer`, which must also operate without errors irrespective of the supply voltage. It is therefore entirely protected by SECDED codes. The non-buffer data memory (DM_Rest) was modelled using CACTI [32] to retrieve the dynamic energy (read and write) per access, while the total number of accesses was estimated using software counters. The leakage power reported by CACTI was adopted to compute the leakage energy, considering the application run time.

The second data-memory section is composed of the data buffers (`Extr_buffer` and `DWT_buffer`), abbreviated as DM_Buff, and target of our approach of data-driven inexact scheme. For our experiments, we have considered a maximum of one error occurring in a memory word. In the case of the sparse `DWT_buffer`, we have employed 6 ECC bits for the protection of the most significant words, and one parity bit for detecting errors in the less significant words. To simulate this heterogeneous memory structure, two separate CACTI models were used as a starting point, either employing 6-bit ECC or 1-bit parity for the whole memory content. To derive the dynamic energy per read access of intermediate configurations, corresponding to the partial protection schemes, we employed the formula as described in (1).

$$Et = p * Ep + (1 - p) * Eu \tag{1}$$

where *Ep* is the read energy per access in the protected memory, and *Eu* is the read energy per access in the unprotected memory. Also, *p* is the percentage of considered significant words. *Et* is the net read energy per access in the heterogeneous memory. The write energy per access and the leakage power of the hybrid memory were also calculated in the same manner. For the memories having ECC protection, CACTI reports the total energy per access taking into account the cost of access only for the additional check bits. It also accounts for the leakage incurred due to these additional bits. In our work, we aim to enforce significant data based protection for the major part of the data memory, and discard ECC bits for that part by including a simple parity check to detect error. The energy dissipated per access to the memory in the logic associated to ECC bits is thus an overkill in this context, especially when operating at ultra-low voltages when it is very low compared to the total energy dissipated in ECC-protected memories, and hence is not reported.

To evaluate the impact of errors in unreliable memories, binary error masks were randomly generated for each buffer. We considered single bit-flip errors with probabilities of 0.07% and 0.22%, relative to the behaviour of a 6T SRAM cell working with supply voltages of 0.65V and 0.6V, respectively [21]. These masks were of the same size as the buffers corresponding to which they were created. A '1' in a mask position indicates a bit-flip error in the same position in the buffer, while a '0' indicates no error. The corresponding value in the buffer was accordingly modified by considering a bit-flip at the indicated error position. If that value belonged to the part of the memory that is protected, then the value remained unchanged. We have considered a maximum of one error per memory word as the probability of multiple bit-flips are extremely low, and even in the rare cases that they occur, they can be managed at the software level. In case more complex ECC protection is employed to correct multiple errors, it would result in higher area and energy overheads, thus underscoring the benefit of our approach of selectively protecting the significant part of the memory. The impact of these errors on the quality of the output of the PSA application, under the different protection schemes, was then measured by comparing the obtained LFHF ratio with respect to an error-free execution.

We compared our inexact architecture against two different baselines:

1) *High Vdd:* In the first case we considered a high supply voltage (1.1V), which does not impact the reliability of the system. All memories in this case were implemented as 6T SRAMs, whose energy values were computed by modelling them in CACTI.
2) *Low Vdd and total ECC protection:* In the second case we have considered exact operations at low supply voltage levels (0.65V). This requires the implementation of the instruction memory

with 8-transistor SRAMs, while all of the data memory (buffer and non-buffer) is completely protected by SECDED ECC codes.

5. Experiment Results

We evaluated our system in three parts. Firstly, we analysed the performance degradation of the PSA system in calculating the LFHF ratio, under the different configurations of the heterogeneous memory scheme. Next, we studied the energy savings achieved by using the proposed configurations. Finally, we reported the energy-performance trade-offs for the different protection schemes.

5.1 Analysis of Error Introduced

The results of the error simulations have been achieved by averaging individual results obtained by processing of data in each time window for each ECG recording. Figs. 8a and 8b show the percentage of error in the computation of the LFHF ratio by the PSA application, when compared to an error-free version of the same, under the different test-points of the proposed heterogeneous memory scheme at supply voltages of 0.65V and 0.6V, respectively.

Our obtained results show that selective protection of a small fraction of words (the significant ones) in the sparse buffers can still guarantee high-quality performance of the system, with respect to an error-free version. This shows the error-tolerance capabilities of WBSN applications. As an example, 1.3% relative error is incurred in the LFHF ratio by protecting 11 MSBs in the `Extr_buffer` and 15% significant words in the `DWT_buffer` (Fig. 8a). It is very low in comparison to the fact that 2 or 3 significant bits are needed to represent it, as reported in [9] and that it is dependent on multiple factors including, but not limited to, the age, race and gender of the patient [10]. This shows that bio-signal processing applications are tolerant to errors. However, the percentage of error in the LFHF ratio obtained by protecting only 4 MSBs in the `Extr_buffer` exceeded 20% in the case of full protection of the `DWT_buffer`. This high error-rate is not acceptable in bio-signal processing and thus we have excluded the condition of protecting only 4 MSBs in the `Extr_buffer` from the following sections of this paper.

In the case of 32 MSBs protected in the `Extr_buffer` and 15% of significant words protected in the `DWT_buffer`, the relative error is less than 1%. This figure is bound below 4% even when we consider the worst-case protection from our experimental setup (11 MSBs protected in the `Extr_buffer` and 5% of significant words protected in the `DWT_buffer`).

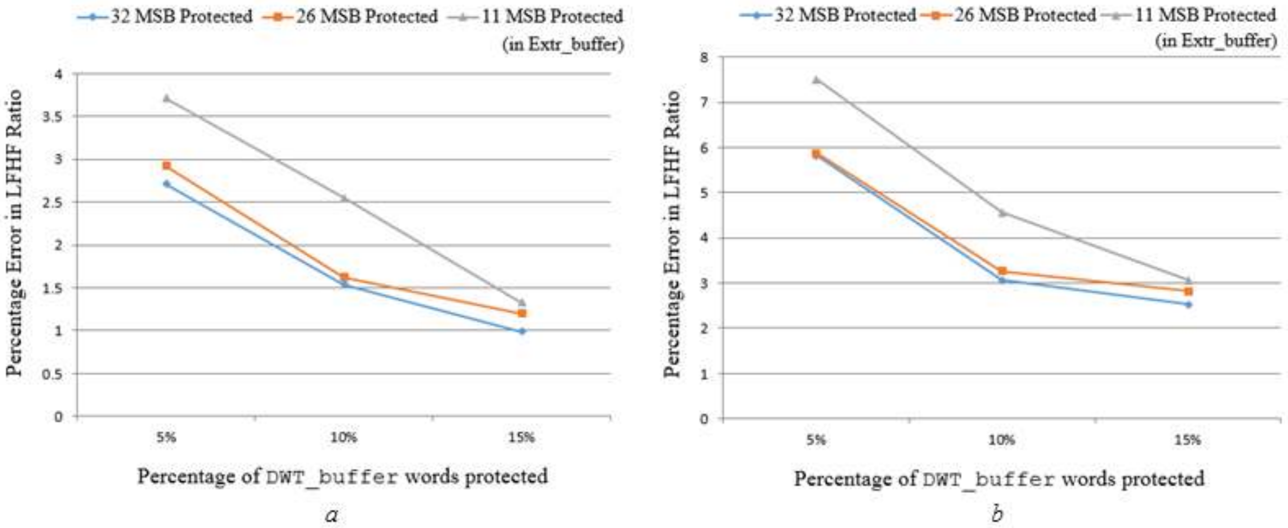

Fig 8. *Percentage of error in the calculation of the LFHF ratio under the different protection schemes*
a. At 0.65V supply
b. At 0.6V supply

The percentage error in the LFHF ratio for a supply voltage of 0.6V is shown in Fig. 8b. The same trends as in Fig. 8a are noticed also in this case, but with higher relative error with respect to operation at 0.65V supply voltage. This is due to the much higher number of bit-flip errors in the memories at 0.6V supply, when compared to 0.65V. Interestingly, even in this case, the error in the LFHF ratio, with respect to a fault-free execution, can be limited to 5% by allowing errors in the 21 LSBs of `Extr_buffer` and only checking (but not correcting) errors in 90% of `DWT_buffer`.

5.2 Analysis of Energy Consumption

Figs. 9 and 10 compare the energy consumption of the different system components, at the first and second baselines considered (high Vdd and low Vdd with complete protection, respectively). It can be seen that at low supply voltage, the system already shows substantial energy savings when compared to operation at high supply voltage. Moreover, at low Vdd, it can be observed that data memory accounts for a major part of the energy budget and that the targeted buffers account for most of the energy consumed by the data memory (Fig. 10). This justifies the application of the proposed memory protection scheme to save even more energy in these buffers, thereby further enabling energy benefits at low-voltage operating points.

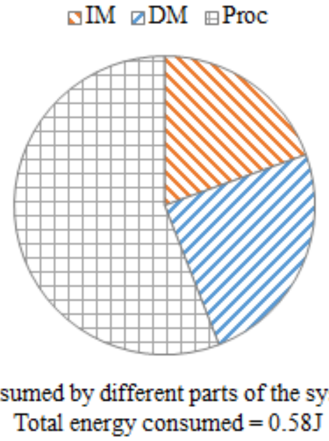

Fig 9. *Energy Consumptions at Baseline 1(high Vdd)*

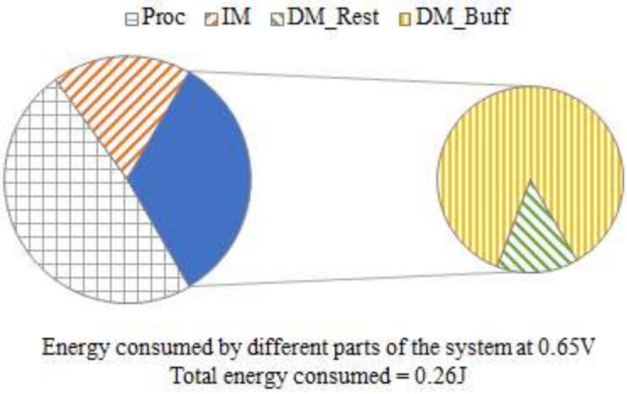

Fig 10. *Energy Consumptions at Baseline 2 (low Vdd)*

Fig. 11 shows the total energy consumption of the targeted memory buffers under the different protection schemes. We can observe from it that by using our proposed scheme with the condition where we protect 11 MSBs in the `Extr_buffer` and 10% significant words in the `DWT_buffer` (corresponding to a relatively low error of 2.6% in the LFHF ratio as in Fig. 8a), we were able to save about 18% of the energy in the buffers compared to the second baseline.

It can be further observed from Fig. 11 that by using the proposed heterogeneous memory protection scheme we could achieve almost 20% of savings in energy in the targeted buffers in the most energy-efficient case of protection considered (11 MSBs in `Extr_buffer` and 5% of significant words in `DWT_buffer`), over a scheme which involves protecting the buffers completely with SECDED codes.

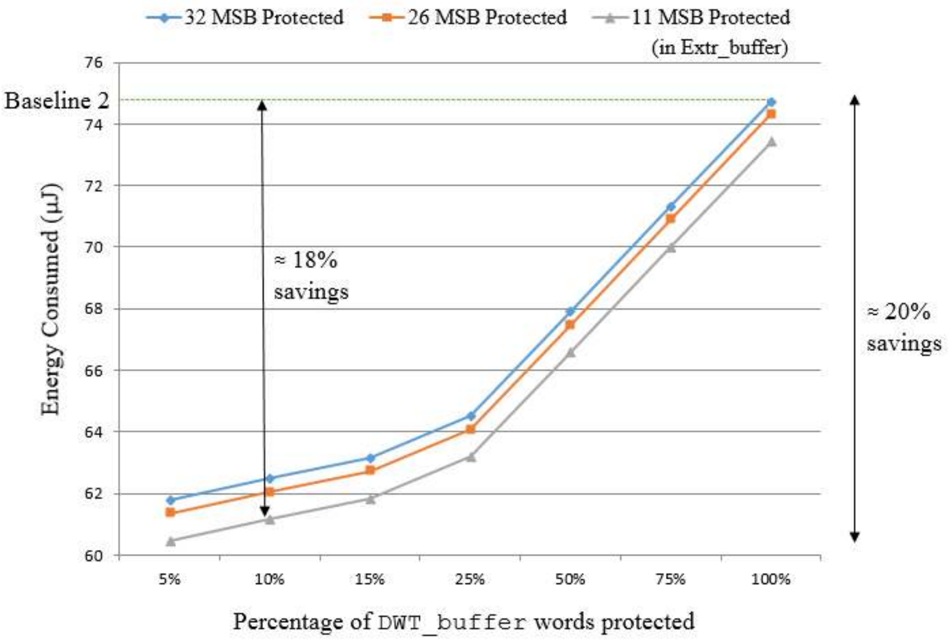

Fig 11. *Total energy consumption by the targeted memory buffers at 0.65V supply under different memory protection schemes for an execution time* $\approx$ *2.23s*

5.3 Energy / Performance Trade-off Analysis

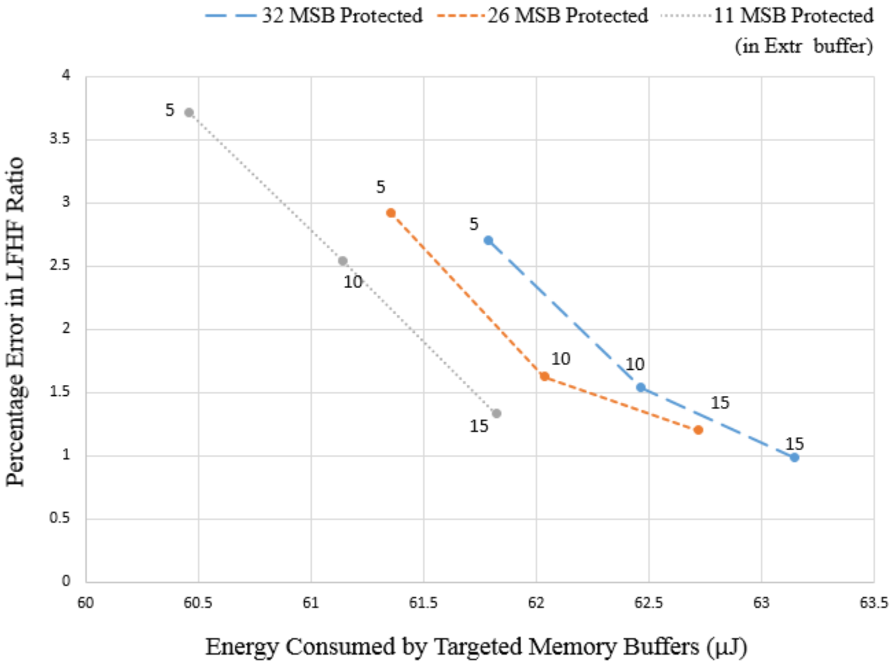

Fig 12. *Percentage error in LFHF ratio and the corresponding energy consumption under different memory protection schemes at 0.65V. Numbers beside the points represent the % of* `DWT_buffer` *protected*

Combining the results obtained in the previous sections, we can investigate the trade-offs between relative error in LFHF ratio and the energy consumed. This enables the selection of the optimum memory architecture considering both energy consumption and the performance degradation. The total energy consumed by the targeted buffers under the different protection schemes are plotted against the percentage error in the computation of the LFHF ratio at 0.65V, as shown in Fig. 12.

It shows that protection schemes with lesser energy consumption result in higher performance degradation. As an example for selecting the optimum protection scheme, for a maximum tolerable error of 3%, a solution with 11 MSBs protected in the `Extr_buffer` and just 10% of the most significant words in the `DWT_buffer` is the best one in terms of energy efficiency. On the other hand, for an energy budget of 61.5 μJ, the smallest percentage error can be achieved by protecting just 11 MSBs in the `Extr_buffer` and 10% of the `DWT_buffer`.

6. Conclusions

In this study we have introduced a novel heterogeneous memory architecture to increase the power efficiency of WBSNs by selectively protecting data with high criticality. Our experiments show that, by guaranteeing different amount of reliability in the bits and words of varying *significance*, the energy required by the considered PSA bio-signal processing application can be reduced beyond the levels attainable by voltage/frequency scaling alone, with a minimal degradation in the quality of service.

The results of our experiments have shown that by applying the resulting heterogeneous protection scheme, we were able to reduce approximately 20% of the energy budget of the data memory used in the intermediate data buffers in prospective real-life wearable ECG monitoring systems. Moreover, our approach is able to tolerate the high error rates incurred at ultra-low voltage supply levels. This supports scaling to ultra-low operating voltages, which itself has substantial energy benefits compared to high voltage operation.

Our framework is applicable in various health-monitoring applications beyond the PSA, as they share similar characteristics of acquiring noisy inputs, proving a statistical or qualitative output, and consisting of intermediate buffers, which show sparse data distribution.

7. Acknowledgements

This work has been partially supported by the ONR-G grant no. N62909-14-1-N072 and the E4Bio RTD project (no. 200021_159853) evaluated by the Swiss NSF.

8. References

8.2 Websites

[1] MEP Heart Group: “Cardiovascular diseases facts and figures”. Available online: http://www.mepheartgroup.eu/index.php/facts-a-figures

[2] PhysioBank Database. Available online: www.physionet.org/physiobank/

[3] The Cortex M3 Processor. Available online: http://www.arm.com/products/processors/cortex-m/cortex-m3.php

8.3 Journal articles

[4] Massé, Fabien, Martien Van Bussel, Aline Serteyn, Johan Arends, and Julien Penders. "Miniaturized wireless ECG monitor for real-time detection of epileptic seizures." ACM Transactions on Embedded Computing Systems (TECS) 12, no. 4 (2013): 102.

[5] Weckx, Pieter, Ben Kaczer, María Toledano-Luque, Praveen Raghavan, Jacopo Franco, Philippe J. Roussel, Guido Groeseneken, and Francky Catthoor. "Implications of BTI- induced time-dependent statistics on yield estimation of digital circuits." Electron Devices, IEEE Transactions on 61, no. 3 (2014): 666-673.

[6] Verma, Naveen, and Anantha P. Chandrakasan. "A 256 kb 65 nm 8T subthreshold SRAM employing sense-amplifier redundancy." Solid-State Circuits, IEEE Journal of 43.1 (2008): 141-149.

[7] Rajoub, Bashar. "An efficient coding algorithm for the compression of ECG signals using the wavelet transform." Biomedical Engineering, IEEE Transactions on 49.4 (2002): 355- 362.

[8] Sanchez-Macian A., Reviriego P., Maestro, J.A. “Hamming SEC-DAED and Extended Hamming SEC-DED-TAED Codes Through Selective Shortening and Bit Placement.” Device and Materials Reliability, IEEE Transactions on Volume: 14, Issue: 1 (2013): 574-576.

[9] R.J. Winchell and D.B. Hoyt, “Spectral Analysis of Heart Rate Variability in the ICU: A Measure of Autonomic Function”, Journal of Surgical Research, 1996.

[10] D. Liao, R.W. Barnes, L.E. Chambless, R.R. Simpson, Jr, P. Sorlie, “Age, Race, and Sex Differences in Autonomic Cardiac Function Measured by Spectral Analysis of Heart Rate Variability – The ARIC Study”, American Journal of Cardiology, 1995.

[11] X. Dong, C. Xu, Y. Xie, N.P. Jouppi, “NVSim: A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory”, IEEE CAD, 2012.

[12] Y. Hao, R. Foster. “Wireless body sensor networks for health-monitoring applications”, Physiological Measurement, 2008

8.4 Conference Papers

[13] **Braojos, Rubén, Giovanni Ansaloni, and David Atienza. "A methodology for embedded classification of heartbeats using random projections." In Design, Automation & Test in Europe Conference & Exhibition (DATE), 2013, pp. 899-904. IEEE, 2013.**

[14] Ganapathy, Shrikanth, Georgios Karakonstantis, Adam Shmuel Teman, and Andreas Peter Burg. "Mitigating the Impact of Faults in Unreliable Memories for Error-Resilient Applications." In Proceedings of the Design Automation Conference, 2015.

[15] Du, Zidong, Avinash Lingamneni, Yunji Chen, Krishna Palem, Olivier Temam, and Chengyong Wu. "Leveraging the error resilience of machine-learning applications for designing highly energy efficient accelerators." In Design Automation Conference (ASP-DAC), 2014 19th Asia and South Pacific, pp. 201-206. IEEE, 2014.

[16] Milosevic, Jelena, Andreas Dittrich, Alberto Ferrante, Miroslaw Malek, Camilo Rojas Quiros, Rubén Braojos, Giovanni Ansaloni, and David Atienza. "Risk Assessment of Atrial Fibrillation: a Failure Prediction Approach." In Computing in Cardiology Conference (CinC), 2014, pp. 801-804. IEEE, 2014.

[17] Chou, C. C., Tseng, S. Y., Chua, E., Lee, Y. C., Fang, W. C. and Huang, H. C., "Advanced ECG processor with HRV analysis for real-time portable health monitoring." In Consumer Electronics- Berlin (ICCE-Berlin), pp. 172-175. September 2011.

[18] Bortolotti, Daniele, Andrea Bartolini, Christian Weis, Davide Rossi, and Luca Benini. "Hybrid memory architecture for voltage scaling in ultra-low power multi-core biomedical processors." In Design, Automation and Test in Europe Conference and Exhibition (DATE), 2014, pp. 1-6. IEEE, 2014.

[19] Wang, Alice, and Anantha Chandrakasan. "A 180mV FFT processor using subthreshold circuit techniques." In Solid-State Circuits Conference, vol. 1, pp. 292-529. 2004.

[20] Ashouei, Maryam, Jos Hulzink, Mario Konijnenburg, Jun Zhou, Filipa Duarte, Arjan Breeschoten, Jos Huisken et al. "A voltage-scalable biomedical signal processor running ECG using 13pJ/cycle at 1MHz and 0.4 V." In Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International, pp. 332-334. IEEE, 2011.

[21] Bortolotti, Daniele, Hossein Mamaghanian, Andrea Bartolini, Maryam Ashouei, Jan Stuijt, David Atienza Alonso, Pierre Vandergheynst, and Luca Benini. "Approximate compressed sensing: ultra-low power biosignal processing via aggressive voltage scaling on a hybrid memory multi-core processor." In Proc. of 2014 IEEE International Symposium on Low Power Electronics and Design (ISLPED 2014), vol. 1, no. EPFL-CONF-200128, pp. 40-45. IEEE/ACM Press, 2014.

[22] Akselrod, Solange, David Gordon, F. Andrew Ubel, Daniel C. Shannon, A. C. Berger, and Richard J. Cohen. "Power spectrum analysis of heart rate fluctuation: a quantitative probe of beat-to-beat cardiovascular control." science 213, no. 4504 (1981): 220-222.

[23] Karakonstantis, Georgios, Alamelu Sankaranarayanan, Mohamed M. Sabry, David Atienza, and Andreas Burg. "A quality-scalable and energy-efficient approach for spectral analysis of heart rate variability." In Design, Automation and Test in Europe Conference and Exhibition (DATE), 2014, pp. 1-6. IEEE, 2014.

[24] S. S. Basu, P. Garcia del Valle, G. Ansaloni, G. Karakonstantis and D. Atienza Alonso. "Heterogeneous Error-Resilient Scheme for Spectral Analysis in Ultra-Low Power Wearable Electrocardiogram Devices." In IEEE Annual Symposium on VLSI, 2015.

[25] H. Mamaghanian, N. Khaled, D. Atienza Alonso and P. Vandergheynst. “Compressed Sensing for Real-Time Energy-Efficient ECG Compression on Wireless Body Sensor Nodes.” In IEEE Transactions on Biomedical Engineering Bme, vol. 58, num. 9, pp. 2456-2466, 2011.

[26] Karakonstantis, Georgios, Aviinaash Sankaranarayanan, Andreas Burg. “Low Complexity Spectral Analysis of Heart-Rate-Variability through a Wavelet based FFT”. In Computing in Cardiology Conference (CinC), 2012, pp. 285-288. September, 2012.

[27] Stefano Di Carlo, Alessandro Savino, Alberto Scionti, Paolo Prinetto, “Influence of Parasitic Capacitance Variations on 65 nm and 32 nm Predictive Technology Model SRAM Core-Cells”. IEEE 17th Asian Test Symposium (ATS), November, 2008.

[28] R. Braojos, A. Dogan, I. Beretta, G. Ansaloni and D. Atienza, “Hardware/software approach for code synchronization in low-power multi-core sensor nodes”, DATE 2014.

[29] L. Zuolo, G. Morandi, C. Zambelli, P. Olivo, D. Bertozzi, “System Interconnect Extensions For Fully Transparent Demand Paging In Low-Cost MMU-less Embedded Systems”, International Symposium in System on Chip, 2013.

[30] N. Boichat, D. Atienza, N. Khaled, “Wavelet-based ECG delineation on a wearable embedded sensor platform”, BSN, 2009.

8.5 Book

[31] Sörnmo, Leif, and Pablo Laguna. Bioelectrical signal processing in cardiac and neurological applications. Academic Press, 2005.

8.6 Report

[32] Muralimanohar, Naveen, Rajeev Balasubramonian, and Norman P. Jouppi. "CACTI 6.0: A tool to model large caches." HP Laboratories (2009): 22-31.