Design of Energy Efficient and Dependable Health Monitoring Systems under Unreliable Nanometer Technologies

(Invited Paper)

Mohamed M. Sabry, Georgios Karakonstantis, David Atienza, Andreas Burg
Swiss Federal Institute of Technology (EPFL), Lausanne, VD 1015, Switzerland
{mohamed. sabry, georgios.karakonstantis, andreas.burg, david.atienza}@epfl.ch

ABSTRACT

In this paper we investigate the impact of potential hardware misbehavior induced by reliability issues and scaled voltages in wireless body sensor network (WBSN) nodes. Our study reveals the inherent resilience of popular algorithms in cardiac monitoring applications and argues that by exploiting the unique characteristics of such algorithms the energy efficiency and reliability of such systems can be significantly improved. This is achieved by developing a cross-layer design paradigm that utilizes low cost techniques at the hardware and software layers and by optimizing the synergy between them in order to provide intelligent trade-offs between energy, performance and quality. The main idea of the proposed approach is the selective application of costly robust techniques only to the most critical tasks identified at the application layer that are detrimental for obtaining sufficient output quality. Our results show that by ensuring the correct operation of only 37% of the total computations in an electrocardiogram (ECG) monitoring WBSN node we can achieve up to 70% power savings with only 9% degradation in ECG output quality.

Categories and Subject Descriptors

B.8.2 **[PERFORMANCE AND RELIABILITY]**: Performance Analysis and Design Aids.

General Terms

Algorithm, Design, Reliability.

Keywords

ECG Monitoring, Biomedical Applications, Error-Resilience, Memory Failures, Energy-Efficiency, Reliability, Hardware, Software Co-Design.

1. INTRODUCTION

The rapid aging of the world population and the increasing prevalence of unhealthy lifestyles are expected to deteriorate the personal health of the world population and increase substantially the costs for health monitoring and care. It is characteristic for instance that according to the World Health Organization, cardiovascular diseases are responsible for one third of the deaths worldwide [1]. Cardiovascular diseases are expected to soon require healthcare costs and medical management needs that are unsustainable for traditional healthcare delivery systems since they need close and potentially continuous medical supervision [10]. Wireless body sensor network (WBSN) technologies are poised to offer large-scale and cost-effective solutions to this problem. The use of wearable, miniaturized, and wireless sensors, able to continuously measure and report cardiac signals, can indeed provide the ubiquitous, long-term and even real-time monitoring required by the patients, as well as its integration with the patient's medical record and its coordination with nursing/medical support.

Technology scaling enables the integration of numerous transistors and complex functionality in a single chip, enabling the realization of such miniature WBSN systems, but as we move beyond the 65nm node, new nano-scale electronic devices exhibit a significant growth of reliability issues. Specifically, as transistors are getting smaller it becomes more difficult to fabricate them, which causes spatial and temporal variations in their characteristics that lead to unpredictable behavior. This situation is further aggravated by random-soft errors that occur due to external radiation and other environmental conditions [2, 3]. Such reliability issues threaten the correct functionality of circuits which is detrimental especially in WBSN nodes where any error may pose a threat to a person's life.

Traditional electronic system conception dictates designing WBSN devices with margins (up-scaling voltage or down-scaling frequency [3]) in order to tackle any unpredictability or adding extra hardware for detection and correction of any error [4]. However, such techniques limit the performance gains obtained by technology scaling and, in combination with the worst-case conditions can lead to large power overheads [4] contradicting with the other main challenge in WBSN nodes. Specifically, WBSN nodes need to perform complex computations within strict power budgets and limited energy resources. Indeed while voltage scaling may offer one the most efficient techniques for reducing the power consumption [2], unfortunately it makes circuits more sensitive to reliability issues in latest technology nodes. Specifically, low voltage does not only slow down circuits, but also increases delay variations and memory failures [3], further threatening the correct functionality of WBSN nodes. In addition, traditional techniques for low operation, such as dynamic voltage and frequency scaling (DVFS), present serious drawbacks for biomedical applications. In particular, those methods do not guarantee real-time operation in case of inappropriate DVFS settings, which can occur at run-time due to unexpected workloads. Hence, these techniques cannot be applied without including additional protection mechanisms against the aforementioned reliability issues.

In the last decade several techniques at the software or hardware layers have tried to address simultaneously the contradictory issues of reliability and power. However, such techniques can easily lead to many pipeline stalls in WBSN designs while processing the input biological signals while they try to correct all the detected errors [5, 6]. These corrections alter the power-performance profiles of the underlying hardware units and typically reduce the overall WBSN performance. Therefore, they cannot satisfy the real-time processing constraints of biomedical applications.

In addition, these aforementioned existing techniques are sub-optimal since they try to address the power-reliability issues only at a single layer of design abstraction: at the software or hardware layer. While such conventional design paradigms that try to correct all potential errors (subsequently resulting in similar overheads for all tasks in the system) seems like a reasonable choice for general purpose systems, due to the equal significance of all data/tasks, it might be possible in health monitoring systems to take advantage of the algorithmic fault tolerance and align it with the underlying

BODYNETS 2012, September 24-26, Oslo, Norway

DOI 10.4108/icst.bodynets.2012.249935

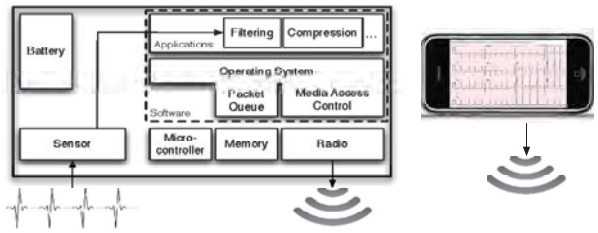


Figure 1. Schematic diagram of WBSN node and an example of the central coordinator

hardware capability and its reliability features. Interestingly, one of the main characteristics of biomedical systems is that they perform signal processing algorithms such as discrete wavelet transform (DWT) that inherently separate the processed information into significant and less-significant data, in terms of contribution to output quality/detection-capability.

Based on the previous observation, in this work we develop a new design paradigm for WBSN nodes that, as starting and key feature, classifies computations on biological signals based on their contribution to WBSN output quality. Then, our proposed approach gives priority to the reliable execution of the most significant WBSN tasks. Thus, the key idea of our approach is to improve the energy efficiency and limit the overheads of traditional approaches by protecting only the most significant tasks through the combination of low cost techniques at the software-and hardware layers. Moreover, by ensuring at design time that only less-significant computations may not be computed correctly under certain conditions (i.e., errors are acceptable due to scaled voltages or reliability issues), the proposed WBSN systems suffer from minimum quality degradation, thus having a reliable behavior for medical applications use. In this paper, we focus on the application of all the steps (from application down to hardware layers) of the proposed holistic design approach for WBSN nodes to an ECG monitoring system [9], where the issues arising from reliability and scaled voltages have never been addressed before. The contributions of this paper can be summarized as follows:

1) Develop a system-level fault simulation methodology for capturing the effects of any circuit misbehavior on the system performance. A generic hardware WBSN platform is used as the basis for our validation with biological signals. Our study focuses on the memory components of such WBSN platforms, which represent a very critical block in the system regarding reliability concerns in new nano-scale technologies.
2) Study the behavior of the ECG monitoring application to identify the impact of hardware defects and classify its computations according to degrees of significance. Our results prove that computations in WBSN applications can span from operation critical tasks, that need to execute reliably, to less-critical application specific tasks that do not affect substantially the overall WBSN output quality.
3) Explore low-overhead robust techniques for WBSN systems at multiple software and hardware levels, to provide intelligent trade-offs between energy, performance and quality for the targeted ECG monitoring application. The main idea of our method for the WBSN systems to limit the target system cost for error correction is the unequal protection of data; significant operations are ensured to be correct by using the appropriate software/hardware (SW/HW) methods and accepting the resulting overheads, while any errors in less-significant data do not activate any correction mechanism, thus do not incur any penalty (or system design cost). Therefore, our approach spends time and power whenever is necessary in order to obtain the right amount of performance and quality from the WBSN nodes under the given conditions.

Our results show that we can save up to 70% of the power consumed by a WBSN node for our ECG analysis case study by relying on protecting a small portion of the generated data, namely 37%. Interestingly, this only yields 9% of signal quality degradation, which is still acceptable for medical use [25].

The rest of the paper is organized as follows. Section 2 describes the fundamentals of WBSN nodes and the issues related to their operation in nano-meter nodes and scaled voltages. Section 3 briefly discusses the proposed approach and Section 4 presents the sensitivity analysis under hardware defects. Section 5 described the proposed approach and compares it with existing techniques after applying it to a DWT based ECG application. Finally, conclusions are drawn in Section 6.

2. NEXT GENERATION COMPUTING ARCHITECTURES IN SMART BIOMEDICAL SYSTEMS

As we discussed above, the WBSN nodes offer a smarter and smaller system architecture compared to traditional health care systems allowing the integration of more functionality in small portable devices. In this section we elaborate on the target next generation WBSN platforms by briefly summarizing their characteristics at the hardware and software layer and discussing the issues arising at the nano-meter technology nodes.

At the outset, the target WBSN systems include a group of nodes and a central coordinator. Figure 1 shows a schematic diagram of a typical WBSN node [17] along with an example of the central coordinator. Although the central coordinator has superior computing capabilities with respect to the node, recent work has shown that the coordinator can operate at real time on a typical smart phone [18]. In this paper we specifically focus on the node platform architecture due to their challenging operating conditions in terms of limited power budget and timing constraints.

The target platform nodes capture the required health monitor-related signals via ECG sensors, apply various filtering and compression transformations to these signals, and then transmit the processed data to the central coordinator. As shown in Figure 1, different node components exist in various architectural layers that collectively provide the node functionality. We further elucidate these layers in the following subsections.

2.1 Hardware Layer

The hardware layer includes all the physical components that are used in achieving the requested node functionality. In our target node architecture, these components are classified as *sensor, radio, memory,* and *microcontroller*.

The *sensor* component is responsible for acquiring the raw bio-signals from the physical environment. Then, the sensor applies the required preprocessing stages (quantization, down-sampling, etc.) to transform such signals into an appropriate format for further processing. For instance, Corventi's PiiX [13] is a typical example of health monitoring sensor.

The *radio* module [15] provides wireless connectivity between the node and the coordinator for streaming, continuous or demand-driven, of processed from the node to the coordinator.

The *microcontroller* executes the software applications of filtering and compression on the processed data, which is stored in the *memory*. Typical examples of utilized micro-controllers are based on RISC architectures as shown in various biomedical-oriented processing architectures [19, 20]. The on-board

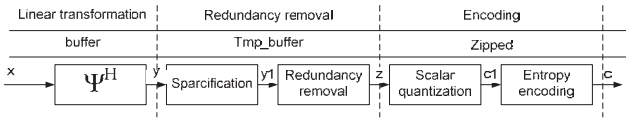


Figure 2. Schematic diagram of the DWT ECG monitoring application case study

microcontroller needs to carry out the required signal processing algorithms fast and reliably within strict power budgets.

2.2 Software Layer

The software layer includes both the operating system and the application layers for the desired functionality. On one side the operating system (OS) (e.g. FreeRTOS [16]) provides the common demands such as interlayer communication between hardware and applications, and power management.

On the other side, the applications include mainly filtering and compression tasks that try to exclude any insignificant or redundant information from the acquired raw data to transmit only what is necessary for achieving the required functionality. In our target WBSN, we focus on ECG monitoring. In particular, we utilize as case study a DWT-based ECG monitoring application, in which the following subtasks are executed sequentially: linear transformation, redundancy removal, and encoding [10].

Figure 2 shows a schematic diagram of the targeted ECG application, with the utilized memory segments for each stage. The application flow starts with the linear transformation stage, then with the redundancy removal for lower representation of the signals, then a final encoding scheme to compress the data even more to obtain an energy efficient data transmission. Each stage utilizes three distinct memory segments, namely *buffer, tmp_buffer, and Zipped*. Memory segment *buffer* is utilized in the main sensing algorithm (DWT), *temp_buffer* is used in redundancy removal, and *Zipped* is used in the final Huffman encoding [10].

2.3 Energy Efficiency and Error Resilience in Nanometer nodes

It is evident that the above mentioned memory segments represent a significant percentage of the overall system area and carry out critical operations. Therefore, their low power and reliable operation is very crucial for the correct and efficient execution of the overall application, as WBSN nodes need to ensure large battery lifetime in order to increase the system portability, as well as providing long-term bio-signals monitoring. However, in the design at nanometer technology nodes, reliability issues threaten the correct functionality of circuits and, especially, memories. We briefly discuss the sources and effects of such issues in the next paragraphs.

As geometries are being scaled down, fabrication processes become inaccurate and lead to spatial and temporal variations in transistor characteristics that can result in more that 30% performance degradation [2]. Furthermore, at these nano-scale technology nodes, the amount of charge required to upset a circuit node is reduced, raising the likelihood of having a large number of on-chip soft-errors [4]. Indeed studies have shown that such variations can cause typically up to 30% performance degradation, and error rates can reach up-to 10^{-6} and 10^{-4}/cycle [22]. Even more important is the fact that such variations can substantially increase the probability of memory failures, which increase almost exponentially with lowering of supply voltage [11]. These memory failures can therefore lead to a complete system failure.

While researchers have already tried to address these issues in embedded systems by attempting to detect and correct any error at the hardware or software layer, unfortunately such methods lead to large power and area overheads [11, 21], which are unacceptable for WBSN nodes. In addition, it is characteristic that, as the percentage of memory components increases, the overhead of such techniques makes them prohibitive, reducing their viability for real-life (cost-effective) WBSN designs. For instance, error correction coding (ECC) and novel bit-cell architectures (i.e., 8T) might tackle the high failure probability of traditional 6 transistor (6T) bit-cells (under reliability issues), but can lead to more than 50% power overheads [11].

Such additional costs are prohibitive for battery-based WBSN nodes, and a reliable operation of next generation ultra-low power health monitoring systems cannot be effectively achieved by focusing on a single, narrow area, of the system (i.e. hardware). Instead, it requires a holistic co-design of all system layers, namely, hardware, system software and applications layers. By identifying the most significant instructions and operations of a system and giving priority to its execution we could intelligently trade-off power and performance for slightly reduced output quality. The proposed significance driven approach is described in the following section.

3. PROPOSED SIGNIFICANCE DRIVEN APPROACH

As previously mentioned, all existing resilient techniques have tried to ensure the correctness of all the operations of an application either at the software or hardware layer irrespective of the nature/type of the tasks. However, in bio-signal processing applications this may not be required. In fact, it can be observed that not all computations are equally important in biomedical applications. For instance, during frequency analysis of bio-signals data (i.e., ECG) required by many biomedical applications, several frequencies might be of higher interest and might be enough to identify a potential issue/condition. The remaining frequencies might play consequently a less important role and be only useful in obtaining higher quality results and verifying any feature at a later processing stage. Such an attribute can be identified in many biomedical applications and in particular ECG monitoring that we are concerned in this paper. All of the required algorithms such as the wavelet transform are usually sampling and executing some desired data at very low frequencies in order to extract some features that indicate various conditions (i.e. heart rate, cardiac arrhythmia, etc.). It was shown for instance that 75% of DWT computations might be sufficient in order to reconstruct a very good quality signal that could be used for indicating various cardiac diseases [9].

The above observations lead us to the proposed method

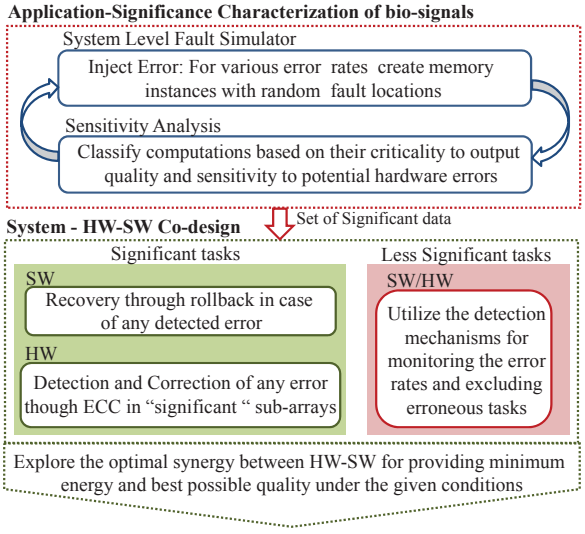


Figure 3: Proposed Significance-Driven Approach

according to which we try to distinguish the parts that are relevant and more critical for the output of a sensor node computation and provide a platform that can take advantage of this property to reduce energy and provide the right amount of reliability and quality. The overall design approach is shown in Figure 3, which we briefly describe in the following paragraphs.

3.1. System Level Simulator

Initially we have developed a system-level fault simulator to allow us to study the effect of potential hardware misbehavior at the system level. This simulator is based on the cycle accurate MPARM simulator [24], which was utilized and further developed to accommodate error injection, detection and mitigation in our previous work [23]. Note that we consider errors rates (N_f) of the order of 10^{-9} down to 10^{-6} /cycle which are typical rates reported in recent works [22]. We assume that at each time N_f defects are distributed across the memory using random fault-location maps. For a given algorithm, each computed bit is mapped to a specific memory cell in the memory array. If the mapped location of the 'bit' indicates a fault in the fault location map, the 'bit' is inverted to indicate a bit-error. Note that such an error injection method models both single event upsets (soft errors) as well as errors associated with operation of memories at minimum supply voltage (V_{CCmin}). While the random nature of bit-flips is evident in case of soft-errors, we assumed that the data stored are shuffled at every time by randomizing their address locations such that we can model V_{CCmin} associated errors (production defects e.g., due to variations at scaled voltages) also as random [22].

3.2. Phase 1: Characterization of Significance

Based on the system level fault simulator, in phase 1 of the approach a sensitivity analysis is being performed at the application layer in order to classify computations into significant and less-significant. At this phase, the metrics required for evaluating the computations are being defined. For instance, in case of ECG monitoring applications the Mean-Squared Error (MSE) or the percentage root-mean square difference (PRD) can be used for evaluating the output quality.

In this phase, the significant data are those parts that contribute significantly to the output quality while less-significant data are the ones that affect the output quality to a lesser extent. Note that at this step we do not only characterize the significance of computations of the algorithm but also the criticality of the operational tasks of the overall system. For instance, control tasks are more critical in maintaining the correct functionality of the overall system. Any error in these tasks may result in catastrophic failure of the overall system. This analysis indicates the percentage of computations that are sensitive to potential errors and need to be protected in order to obtain an acceptable output. At this step we also assign to groups of data a degree of significance which indicates the degree of priority that must be given to them by the system. For instance, the set of the most critical computations A_1 is assigned a significance of order 1, whereas the set A_2 is assigned a significance of order 2, etc. Indeed, the higher the significance of a group the higher the number of computations it contains ($A_1 < A_2 < A_N$). This proposed classification determines also both the quality and power tradeoffs that can be achieved; the less number of significant computations that are going to be protected, the less power overhead, but also less output quality. However, since the computations that are at least executed are very significant, the degradation of quality is ensured to be minimal by design.

3.3. Phase 2: Software-Hardware Significance Driven Co-Design

After classifying the computations we explore low cost techniques at the software and hardware layers that could support the proposed approach and provide reliable operation of the significant parts at minimum overhead. The main idea of our approach is to utilize the detection and correction mechanisms of such methods to correct any error only in significant data while utilize only their detection mechanisms in the less significant data.

3.3.1 Hardware Level Methods

In this step the significant components can be selected to be stored in, for instance, sub-arrays of the caches/buffers that include robustness mechanisms, such as upsized bit-cells or ECC. This comes at an increased cost of area. To circumvent this area overhead, we reduce the reliability constraint of other sub-arrays in cache where only less-significant tasks are being stored allowing the use of less robust and thus less costly mechanisms. In this way, we follow an asymmetric design where expensive mechanisms are utilized for the significant components and regular low power mechanisms are used for the less-critical tasks. By following such an approach the power and area overhead is limited as compared to conventional design paradigms, while reliable operation with acceptable quality is being obtained.

3.3.2 Software Level Methods

We advocate for the use of a similar approach to many other components of the system utilizing various mechanisms at the software level. In this context, the error correction mechanisms are activated only for recovering from any error in significant components, while the system continues to function normally in case of any error in less-significant tasks. Hence, the throughput penalty is limited since correction mechanisms only generate a throughput penalty due to extra cycles in case of significant tasks.

3.3.3 Exploring the Optimum Synergy Between the Various Techniques for Maximum Energy Savings

After identifying the low cost techniques to use at the software and hardware level, in the next step of our methodology we explore the optimum synergy between them in order to achieve the minimum energy and best possible quality for each given operating condition (scaled voltage, variations, soft errors). For instance, in our previous work [23] we showed the effectiveness of combining a small hardware-protected buffer with fine-grained checkpoint and rollback mechanisms. In any case the detection mechanisms of reliability enhancement techniques, such as ECC, can be utilized to monitor the error rates and potentially take system level actions at the software level in case of large error rates.

Overall, the proposed approach applies the significance-driven design paradigm to all the layers of the design stack starting from the application and the characterization of tasks to the co-design of software and hardware in order to enable adaptation of application scalability to hardware capability which could vary depending on the operating conditions. Our approach not only limits the overhead of existing techniques, but also optimizes their synergy to achieve significant power savings by scaling the voltage. Furthermore, our approach tries to obtain the maximum possible quality with the minimum energy given any operating condition.

4. SIGNIFICANCE CHARACTERIZATION

As we discuss above, the first step in the proposed approach is the classification of computations into more and less-significant. To this end, in the next paragraphs, we investigate the error resilience limits of an ECG monitoring application by considering the effect of hardware failure mechanisms during system-level simulations. Our analysis focuses on DWT which is used for ECG data compression as discussed in section 2 and is one of the most popular kernels in WBSNs. Our analysis focuses on the memory components described in Section 2 that are very prone to errors. To

this end, based on the simulator of a typical biomedical processor [19], which we describe in Subsection 3.1, we inject various defect rates in the memory buffers required at various stages of the algorithm and evaluate their impact on the output quality.

4.1. Significance Quantification

In order to evaluate the impact of potential errors on the application we use the system level fault simulator described in section 3.1. As a metric we choose to use PRD which is defined as:

$$PRD_k = \frac{\|X - \overline{X_k}\|_2}{\|X\|_2} \quad (1)$$

where X is the error-free examined and $\overline{X_k}$ is the faulty signal due to injected fault pattern instance k.

It is important to mention that the values of PRD define the quality of the output bio-signal, as proposed in previous works [10, 25]. For instance, a signal with $PRD \leq 2\%$ is considered as "very good" signal for medical relevance analysis, and a signal with $2 \leq PRD \leq 9\%$ is considered as "good" signal. On the contrary, a signal with $PRD \geq 9\%$ does not have sufficient quality and real value for medical analysis. We assess the applications error sensitivity by running more than 1000 simulations, where each run is triggered by a different randomly-selected faulty bit location. Figure 4 shows an instance of the affected bit locations in the memories assumes in our simulations. At each faulty bit simulation, we are applying the DWT algorithm (c.f. Figure 2) to 50 traces of 1 minute ECG signal obtained from MIT-BIH arrhythmia Database [12].

4.2. Error Injection Impact

In the following paragraphs, we elaborate on the observed error calculated by (1) in each of the aforementioned cases. Figure 5 highlights the error impact at each memory segment. From this figure we observe that error varies significantly, particularly in both compression (buffer) and redundancy removal (tmp_buffer). Faults injected at the sensing memory segment (buffer) incur a large variation in PRD, such that signal quality is completely degraded, by high PRD values reaching up to 100%, when the error is in the odd bytes (most significant byte in each half word), while the PRD is within the "good" signal quality if the error is in the even bytes. This diverse behavior is related to identifying the DWT coefficients by their significance, as shown in Figure 6. The significant coefficients are characterized by their high values. If an

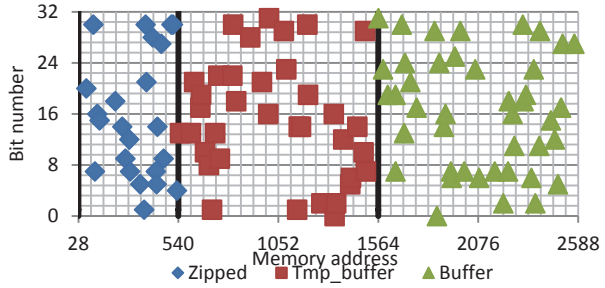


Figure 4: Error injection map applied to DWT application

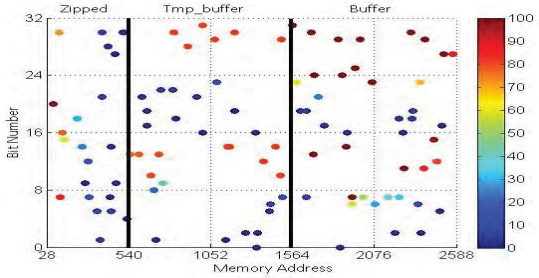


Figure 5. PRD of ECG signals, when DWT is applied, due to injected faults

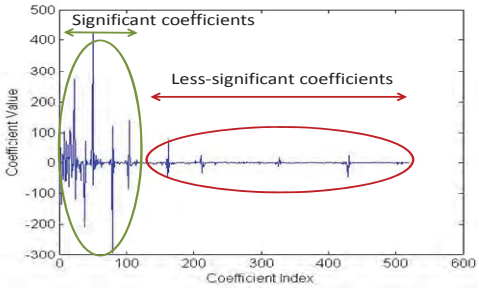


Figure 6. DWT output of a typical ECG signal

error occurs in the most significant bytes, the values are changed significantly from a low- to a high-valued coefficient. As a consequence, the output coefficient would resemble a different signal from the original one, that its quality cannot be determined.

Similarly, we observe the same behavior in the sparsification and redundancy removal memory segment (Tmp_buffer). In particular, high error values occur when the faults are in the most significant bytes of the set of half words. This is because flipping the most significant bit in these locations boosts the filtered out coefficients that in turn is translated to a significantly different signal when it is reconstructed.

Unlike the compression and redundancy removal stages, we observe less variation, with even lesser impact, of quantified error when the faults affect the Huffman encoding buffer (Zipped). As shown in Figure 5, the observed PRD does not exceed 9%, if the error is in the compressed least significant coefficients (c.f. Figure 6). However, if the fault is located in the compressed significant coefficients, the signal quality is heavily degraded. This observed behavior is related to the final compressed data stream. If the fault is placed outside the range of the compressed data, it is not sensed and does not affect the reconstructed signal. However, if the fault is found in the compressed data, particularly in the more-significant coefficients, the reconstructed signal is completely degraded.

4.3. Significance Analysis Evaluation

The above analysis indicates the critical memory segments that can lead to large quality degradation in term of PRD. In order to better understand the most significant computations of the particular application we need to further study its characteristics. To this end, we focus on Figure 6, which shows the DWT output of a typical ECG signal. We observe that there are some coefficients with much higher values compared to others. Such coefficients can indeed be classified as significant coefficients according to our methodology. Overall, we identify that in such an application we can classify significance in terms of bits (Figure 5) but also in terms of output computations/coefficients which seems to carry most of the ECG information. Such an observation can be also attributed to the DWT spectral analysis attribute that separates the data into high and low pass frequency content. Interestingly, the detection of many cardiac malfunctions requires the data within a specific frequency content (low pass) to have values above a pre-specified threshold. Any high pass frequency content usually augments slightly the output values which do not play a critical role if the value is already above the specific threshold.

Thus, we select the significant data as a percentage of the complete spectrum. In particular, we examine three significant cases that are shown in Table 1. Table 1 shows the PRD under the various cases. From the values in Table 1, only the output in case I has a "good" signal quality. The other two cases suffer from a significant increase in PRD, hence the signal quality is degraded. However, this degradation is based on the assumption that the data

identified as less significant is completely erroneous, which is not the case as we show later in Section 5.

Table 1. PRD in case of some sets of significant data

Case	percentage	PRD
I	50%	9%
II	25%	32%
III	12.5%	63%

5. EVALUATION OF PROPOSED WBSN SYSTEMS

Based on the classification of computations that we discussed above, in this section we describe the proposed method and compare its impact when combined with various error correction techniques. Moreover, we show the impact of significance driven computation on creating enough time slack that enable aggressive low-level power saving techniques, namely voltage scaling.

5.1. Experimental Setup

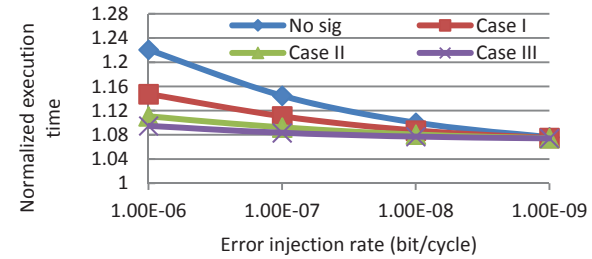
We conduct our experiments on the WBSN cycle-accurate processing node simulation platform mentioned in Subsection 3.1. In particular, we are simulating a single-core WBSN node that has the TAMARISC processing architecture [19, 20]. This processing unit is accompanied with a 64KB L1 SRAM that is split into instruction and data memories, 32KB each. We utilize the DWT encoding application developed in previous work [10], and cross-compile for the target architecture. We utilize the synthesized power consumption values of TAMARISC [19] and use CACTI [26] to estimate the power consumption of the memories, to assess the total power consumption of our proposal.

In our experiments, we measure the execution time and energy consumption overhead of the significance driven computation cases, when combined with various error mitigation techniques (c.f. Subsection 5.2). We normalize all of our results to the case of no error mitigation or significance driven application.

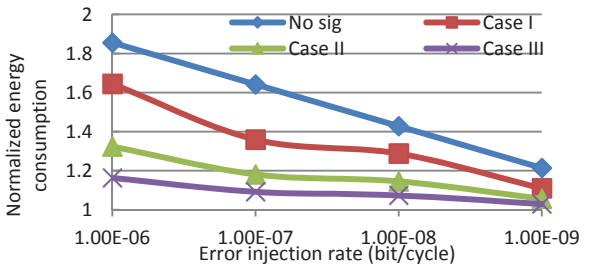
5.2. Explored Error Mitigation Techniques

In this paper, we examine various reliability mitigation techniques to study the impact of significance-driven computation on the data quality, execution time, and energy consumption. In particular we are using the following techniques:

- **Hardware (HW)**: the targeted memory is protected using multi-bit HW ECC circuitry.



(a). Execution time



(b). Energy Consumption

Figure 7. Observed overheads when HW error mitigation is used with significance-driven computation.

- **Software (SW)**: the data is corrected whenever an error is occurred using rollback technique.
- **Hybrid (Hyb)** [23]: the targeted data is divided into chunks, where each chunk is placed in a protected buffer, one at a time to be used whenever an error occurs for successful recovery.

These techniques have been used in for reliability mitigation with various error injection rates (10^{-9} to 10^{-6} bit per cycle).

5.3. Significance Driven Computation Impact

We start by showing the impact of Significance driven computation on energy efficiency and time overhead when it is combined with HW error correction circuitry. Figure 7 shows the execution time and energy consumption overheads of all the considered significant cases (c.f. Table 1), as well as the case with 100% significance, referred to by "no_sig". From Figure 7 we observe that at low error rates, all of the significance-driven cases behave equally. This is mainly due to the extreme low error rates that do not trigger any correction.

Moreover, we observe that significance-driven computation achieves significant energy and time savings. Here, by significance we mean the protection of only the most significant data by any of the above techniques. In particular, in case of HW assisted reliability we assume that an 8-bit ECC circuitry is applied only in the memory array segment that stores the significant data.. Such savings can be as high as 35%, which is found at error rate 1E-6. However, it is important to mention that such savings are accompanied by a significant degradation of the quality of output data, as shown in Table 1. Finally, we observe a different behavior of the quality of output data than the ones in Table 1.

While the values in Table 1 show the worst cases, we have not experienced such cases in our simulations. The values in Table 1 correspond on total data deterioration of the less-significant data. However, we experience a worst case of 41% deterioration of the DWT coefficients at high error rates. Thus, in addition to the guaranteed error-free significant data, an adequate amount on the non-significant data is error free, hence usable. In our experiments, we observe the PRD of the output data, at various error percentage values lower than the worst case (41%), in Figure 8.

Figure 8 shows that we can achieve good signal in any case, according to the expected error rate. If the error percentage is below 10%, then we can choose Case III due to its superior power savings with respect to the other cases. However, if we want to guarantee a good signal quality regardless the errors in the less significant coefficients, Case I can be the best selection. Choosing Case I guarantees a good signal, but with minor energy savings with respect to the default case. Another observation is that Case I provides good signal quality with PRD values less than 9%, which may not be the optimum case for energy savings. Thus, we study the possibility of having a new significance case such that the maximum observed PRD is exactly equal to 9%. We find that by

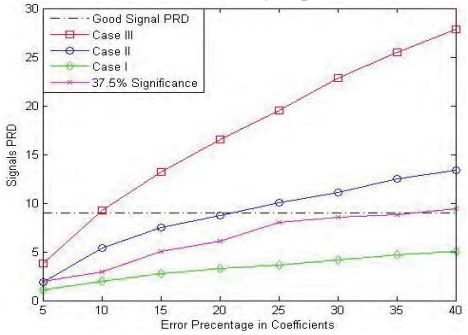
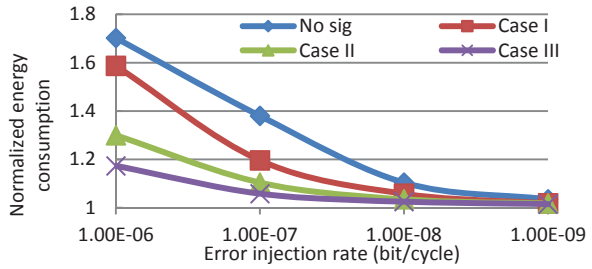
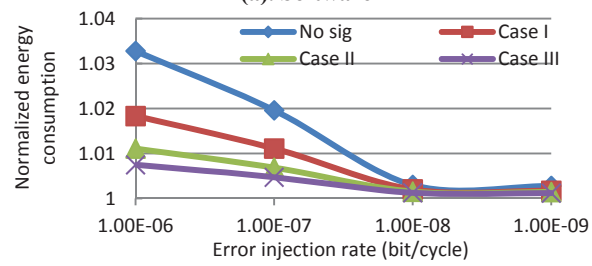


Figure 8. Observed PRD values for actual signals with the observed error percentage



(a). Software



(b). Hybrid

Figure 9. Energy consumption observed for DWT.

selecting 37.5% as the percentage of most significant data, we achieve the constraint of having a maximum PRD=9%. Hence, we utilize this new case in monitoring the energy and time savings and we observe a 20% energy savings with this new case.

5.4. Error Mitigation Techniques Impact

With the significant energy reduction of significance-driven computation including HW error mitigation, we explore the impact of the other error mitigation techniques (c.f. Subsection 5.2.). We show the energy consumption overheads of both SW and hybrid mitigation techniques in Figure 9. We observe that the hybrid approach [23] is superior to both HW and SW mitigation techniques. While Software technique rollbacks to the beginning of the application for a certain ECG stream in case of an error, the hybrid approach rollbacks to a much advanced state such that the re-executed portion is minimal. Moreover, while the hardware approach applies ECC to the total memory storage, which consumes substantial energy and time overheads (due to codeword generation and checking), the hybrid approach uses ECC on the protected data chunk, which is smaller than the actual memory segment used. In the DWT, the data chunk size is 15% of the total data size used to process a single ECG stream. With respect to HW mitigation at 37% significance, the hybrid mitigation technique achieves an additional 31% time and 38% energy savings. We have observed a similar trend in time overhead and time savings, but we do not show the results for brevity.

5.5. Low-level Aggressive Power Savings

As we mention in section 2, one of the most effective techniques for saving power is voltage scaling. The proposed approach can be utilized to achieve further power savings by utilizing the slack obtained in various cases for reducing the supply voltage. For instance in case of 37% significance we can achieve a slack in throughput up to 25%. This we can use it in order to scale voltage by up to 31% and obtain a total of 38% power savings compared to a reference system. Thus, if combined with hybrid mitigation technique, we can achieve up to 70% energy savings.

6. CONCLUSION

In this paper, we investigated the impact of potential hardware errors on biomedical systems and in particular ECG systems. By focusing on a popular application, the DWT, we show that various parts of the systems have different sensitivity to errors. Based on such an observation some computations are classified as more significant and absolutely critical for obtaining an acceptable output quality. In addition, by comparing various methods we propose a highly energy efficient method that can ensure reliable operation at minimum energy. Our results indicated that we can achieve 70% power savings with minimum quality degradation.

7. ACKNOWLEDGEMENTS

This work was supported in part by a PhD research grant for ESL-EPFL funded by IMEC and was supported by Swiss National Science Foundation under the project number PP002-119052.

8. REFERENCES

[1] World Health Organization. (2009). Cardiovascular diseases, [Online]. http://www.who.int/topics/cardiovascular diseases/

[2] J. M. Rabaey, "Low Power Design Essentials", Springer, 2009.

[3] S. Borkar et. al., "Designing reliable systems from unreliable components: The challenges of transistor variability and degradation," *IEEE Micro*, 2005, pp. 10-16.

[4] S. Bhunia, et al, "Low-Power Variation-Tolerant Design in Nanometer Silicon," Springer 2011.

[5] Dan Ernst, et al., "Razor: Circuit-Level Correction of Timing Errors for Low- Power Operation," *IEEE Micro*, 2004, pp.10-20.

[6] A. B. Kahng, et al., "Designing a Processor from the Ground Up to Allow Voltage/Reliability," *HPCA*, 2010.

[7] B. Shim, et al., "Reliable Low-Power Digital Signal Processing via Reduced Precision Redundancy," *IEEE TVLSI,* 2004, pp. 497-510.

[8] G. Karakonstantis, et al., "System Level DSP Synthesis Using Voltage Overscaling, Unequal Error Protection & Adaptive Quality Tuning," IEEE SiPS, 2009.

[9] F. Rincon, et al., "Development and Evaluation of Multi-Lead Wavelet-Based ECG Delineation Algorithms for Embedded Wireless Sensor Nodes", *IEEE TITB*, 2011.

[10] H. Mamaghanian et al., "Compressed Sensing for Real-Time Energy-Efficient ECG Compression on Wireless Body Sensor Nodes", IEEE TBME. 85(9), Sep. 2011.

[11] C. Wilkerson, et al., "Trading off Cache Capacity for Reliability to Enable Low Voltage Operation," IEEE ISCA, 2008.

[12] MIT-BIH arrhythmia database. (2005). [Online]. http://www.physionet.org/physiobank/database/mitdb/

[13] Corventis, 2009. [Online]. http://www.corventis.com/AP/nuvant.asp

[14] R. F. Yazicioglu et al., "Ultra-low-power wearable biopotential sensor nodes," in *Proc. IEEE EMBC*, Sep. 2009.

[15] C. Enz, N. Scolari, and U. Yodprasit, "Ultra low-power radio design for wireless sensor networks", in *Proc. IEEE RFIT,* Dec. 2005.

[16] "The FreeRTOS Project," http://www.freertos.org

[17] I. Beretta et al., "Model-Based Design for Wireless Body Sensor Network Nodes", in *Proc. IEEE LATW,* Jun. 2012.

[18] K. Kanoun et al., "A Real-Time Compressed Sensing-Based Personal Electrocardiogram Monitoring System". in *Proc. DATE,* Mar. 2011.

[19] J. H. F Constantin et al., "An Ultra-Low-Power Application-Specific Processor for Compressed Sensing". in *Proc. IEEE VLSI-SoC,* 2012.

[20] A. Y. Dogan et al. "Multi-Core Architecture Design for Ultra-Low-Power Wearable Health Monitoring Systems". in *Proc. DATE*, 2012.

[21] J. Kim et al. Multi-bit Error Tolerant Caches Using Two-Dimensional Error Coding. In MICRO-40, 2008.

[22] L. Leem et al. ERSA: Error Resilient System Architecture For Probabilistic Applications in *Proc. DATE,* 2010.

[23] M. M. Sabry et al. "A Hybrid HW-SW Approach for Intermittent Error Mitigation in Streaming-Based Embedded Systems". in *Proc. DATE,* 2012.

[24] L. Benini et al "MPARM: Exploring the Multi-Processor SoC Design Space with SystemC" Journal of VLSI processing systems, 2005.

[25] Y. Zigel et al."The weighted diagnostic distortion (WDD) measure for ECG signal compression," IEEE TBME., 47(11), Nov. 2000.

[26] CACTI: an Integrated Access Time, Cycle Time, Area, Leakage, and Dynamic Power Model for Cache Architectures. http://www.cs.utah.edu/ rajeev/cacti6.