Energy vs. Reliability Trade-offs Exploration in Biomedical Ultra-Low Power Devices

Loris Duch[1], P. Garcia del Valle[1], Shrikanth Ganapathy[2], Andreas Burg[2] and David Atienza[1]
[1]Embedded Systems Laboratory (ESL) and [2]Telecommunications Circuits Laboratory (TCL)
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{loris.duch, pablo.garciadelvalle, shrikanth.ganapathy, andreas.burg, david.atienza}@epfl.ch

***Abstract*— State-of-the-art wearable devices such as embedded biomedical monitoring systems apply voltage scaling to lower as much as possible their energy consumption and achieve longer battery lifetimes. While embedded memories often rely on Error Correction Codes (ECC) for error protection, in this paper we explore how the characteristics of biomedical applications can be exploited to develop new techniques with lower power overhead. We then introduce the *Dynamic eRror compEnsation And Masking (DREAM)* technique, that provides partial memory protection with less area and power overheads than ECC. Different trade-offs between the error correction ability of the techniques and their energy consumption are examined to conclude that, when properly applied, DREAM consumes 21% less energy than a traditional ECC with Single Error Correction and Double Error Detection (SEC/DED) capabilities.**

I. Introduction and Related Work

The continuous shrinking of CMOS transistors enables the development of computing systems increasingly complex and including more computing capabilities within the same chip-size. Unfortunately, smaller transistor sizes have adverse consequences on the reliability and make the components, specially the memories, of such systems more prone to soft and hard faults [1]. On top of that, electronics design engineers are forced to deal with the high energy demand of memories. An effective technique to reduce it is to diminish the memory supply voltage, which translates into quadratic energy savings. However, as the supply approaches the threshold voltage of the transistors, permanent errors appear, which makes necessary to add error detection and correction mechanisms.

In general, the lower the voltage is, the more energy, delay and area overheads have to be introduced in hardware (HW), to guarantee error-free operation [2]. For this reason, system designers now focus on software (SW) robustness exploration. Recently, a HW/SW paradigm, called *significance-based computing* [3], has been proposed to selectively protect the execution of significant computations or data while allowing controlled errors to occur in other parts of the program.

In this paper, we evaluate the significance-based computing scheme in a complete set of applications for Ultra-Low Power (ULP) biomedical devices. We exploit their inherent characteristics and faults tolerance to make these algorithms more power efficient by protecting dynamically (and unequally) a near-threshold powered data memory. Compared to [4] and [5], our approach is, respectively, tailored for the targeted domain and does not require any modification of the memory at the transistor level, increasing its applicability. When compared to other HW-based Error Mitigation Techniques (EMTs) for reliability and power savings enhancement, like [6], our work avoids duplication of the memory or processing units, being suitable for area- and power-sensitive systems. Furthermore, for the first time, we model the effect of permanent errors on biosignal processing; those that appear in the memories of ULP devices due to the reduction of their operating voltages.

The main contributions of this paper are the following:

1) We study the significance of the data bits processed by different widely used biomedical applications. Additionally, we show the output quality degradation as a function of the permanent errors injected inside the memory word, by using a stuck-at fault model. This information is used to order by significance of criticality which data must be protected by any EMT.
2) We introduce the *Dynamic eRror compEnsation And Masking (DREAM)* technique, a new asymmetric EMT that consumes 21% less energy than traditional ECC SEC/DED. The correction ability, energy consumption and area overhead of DREAM are compared to the ECC SEC/DED.
3) We analyse the efficiency of different EMTs according to the supply voltage of the memory, to get the best trade-off between energy consumed and output quality in biomedical ULP devices.

The rest of the paper is organized as follows. Section II introduces a set of biosignals applications representative of this domain, that will be characterized in Section III to identify the critical data that should be protected, a necessary step before conceiving DREAM, presented in Section IV. Finally, the experiments in Sections V and VI, prove that our approach allows for effective power reduction, by using aggressive voltage scaling while keeping the output degradation to the allowed level for the different biomedical applications.

II. Biomedical Applications Case Studies

In this work we have selected five representative applications that are widely used in the field of Electrocardiogram (ECG) processing, either as standalone applications, or as the core components of more complex monitoring systems such as Wireless Body Sensor Node (WBSN) devices (c.f. Fig. 1). They have been tested using ECG traces from the MIT-BIH Arrhythmia database [7] with samples of 16-bits. The following paragraphs present each of these applications.

This work has been partially supported by the EC FP7 FET SCoRPiO project (grant no. 323872), the ONR-G (grant no. N62909-14-1-N072), and the BodyPoweredSenSE (grant no. 20NA21 143069) RTD project evaluated by the Swiss NSF and funded by Nano-Tera.ch with Swiss Confederation financing.

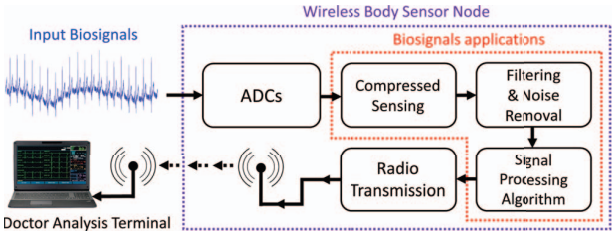


Fig. 1: Block Scheme of a Typical WBSN

1) Discrete Wavelet Transform (DWT): Used to analyze multi-lead ECG signals in commercial WBSNs [8]. In our implementation, the DWT takes as input a vector of ECG samples and performs on it several scales of low-pass and high-pass filtering.

2) Matrix Filtering: A signal processing algorithm that applies a given transformation (e.g.: low-pass or high-pass filtering) to a set of biosignal samples [9]. At the lower level, this operation consists of a series of matrix multiplication operations $[A] \times [B] = [C]$ repeated (iterations of the algorithm) until the quality of the result meets the desired level.

3) Compressed Sensing (CS): A signal acquisition/compression technique introduced as a new paradigm for energy-aware WBSNs. It helps to reduce airtime over energy-hungry wireless links. We have implemented our own version of the algorithm presented by the authors of [10], which takes as input a vector of ECG samples and applies a 50% lossy compression algorithm to convert it into a smaller one (half the size of the input vector).

4) Morphological Filtering: A special type of filtering algorithm developed to clean, thanks to different erosion and dilation steps, the raw ECG signals (often degraded due to factors such as the patients muscles activity or the system AC supply interferences) attending to the shape or morphology of certain expected features. They are widely used in the image- and biomedical signal-processing fields [8].

5) Wavelet Delineation: Typically used to perform an analysis of ECG signals to detect heartbeat fiducial points [8]. Our versions of such applications use the DWT algorithm presented in Section II-1 to generate, as output, the list of P, Q, R, S and T heartbeat points found.

III. Characterization of Biomedical Applications

In order to efficiently apply the significance-based computing paradigm, we first analyze and explore the nature of biomedical applications and conduct a significance analysis for each bit of information, with the goal to determine which data bits (from the input, intermediate and output buffers of the applications) in the main data memory must be protected first. To do so, we performed a quality analysis of the computation results under permanent error injection for each application. We use the Signal to Noise Ratio (SNR) metric, as defined in Formula 1, based on the Mean Square Error (MSE) metric, which is defined as the average of the squares of the difference between all the error-free $x_{theo}(i)$ (theoretical) and corrupted $x_{exp}(i)$ (experimental) output data.

$$SNR = 20 \times \log_{10} \frac{\sqrt{\frac{1}{n}\sum_{i=0}^{n-1} x_{theo}^2(i)}}{\sqrt{MSE}} \quad (1)$$

To generate output data corrupted by permanent errors, we successively set to "1" and "0" each bit located on the positions 0 to 15 of the 16-bits data buffers. Different ECG signals with different pathologies are used to produce each averaged point of Fig. 2. The gap between the SNR curve of the Matrix Filtering and the other curves stems from the fact that, when operating with matrices, each element of the resulting matrix depends on many elements (one full row and one full column) of the input matrices. As a consequence, a single error affects many positions in the output.

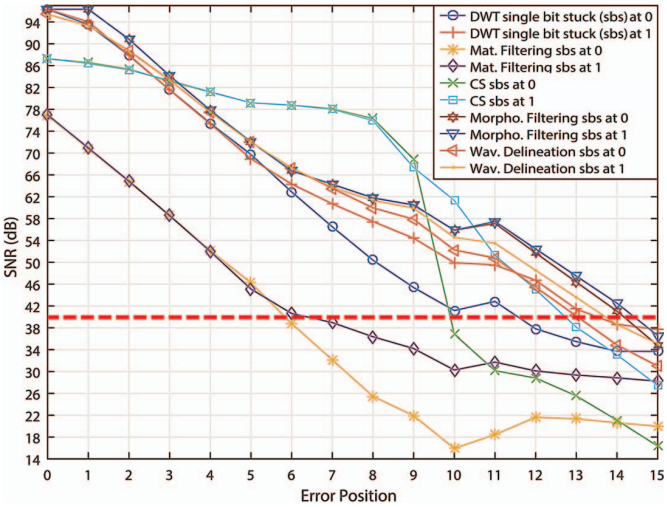


Fig. 2: SNR vs. Data Bit Positions of Injected Error

Fig. 2 outlines that, for the Matrix Filtering and CS applications, erroneous bits set to "1" on MSB positions have a smaller impact on the SNR than erroneous bits set to "0". This is due to the fact that most of the biosignal samples employed during the experiments are negative; thus, when a permanent error sets to "1" a bit on the MSB positions of a negative number, the effect is often hidden.

In addition, for all the applications in Fig. 2, we highlight the continuous decrease of the SNR as the erroneous bit is shifted towards the MSB positions, which demonstrates that errors on the MSBs have a stronger impact on the application results, whereas errors on the LSBs often have a small or negligible effect. This finding demonstrates the possibility of dealing with some degree of inexactness on the LSBs positions, as it is the case for input data samples acquired in real-life conditions, namely, from noisy analog sources (c.f. Section II-4).

Finally, some applications, such as the Heartbeat Classifier [9] (based on Wavelet Delineation + CS), produce statistical or qualitative results; After delineation, heartbeats are sorted out according to different classes of morphologies to detect patients' pathologies, and this task usually requires fine-tuning with human feedback to adjust the margin inherent to the classification algorithm. This classification is often performed visually by doctors and its precision depends on human interpretation with coarse-grained boundaries between classes. This process enables a relaxation of the traditional reliability requirements (i.e., 100% computational precision is not needed), which can be exploited by the CS application. In particular, the maximum required output SNR to get almost

a 100% reconstruction quality is only 35 dB in the case of multi-lead ECG [10] and 40 dB in the case of a single lead ECG [11]. Thus, as Fig. 2 shows, CS can tolerate errors on the bit positions from 0 to 10, for bits stuck-at-0; and from 0 to 12, for bits stuck-at-1.

IV. Proposed DREAM Technique

In embedded biomedical applications, most of the samples produced by the analog to digital converters (ADC) contain series of bits with the same value on the MSB positions; they do not need the full range of bits allowed by the system to encode the information extracted from the electric signal. Furthermore, as analysed in the previous section, errors occurring on the MSB positions have a stronger impact on the final result of each application.

Based on these two observations, we propose in this paper the *Dynamic eRror compEnsation And Masking (DREAM)* technique to dynamically preserve the value of the constant MSBs of each memory word. DREAM relies on bit masks to keep track of the series of MSBs with the same value in each sample. These bits must be kept constant all along the storage of the sample into the faulty memory. To do so, similarly to ECC, some extra logic is required to determine and apply dynamically the mask on data-words. Also, some extra memory is needed to store the mask identifier (mask ID) and sign bit of each data-word. This memory overhead per data-word can be determined using Formula 2:

$$\begin{aligned} Extra\ Bits/Word &= Sign\ Bit + Mask\ ID\ Size \\ &= 1 + \log_2(Data\ Size)\ bits \end{aligned} \quad (2)$$

DREAM is able to correct multiple errors located in the series of MSBs highlighted by the mask. In fact, an additional data bit is always protected because the most-significant bit of the data part not covered by the mask is always set to the inverted value of the data sign bit; therefore, the position of this bit in the data-word can be specified by the mask ID and then inverted by a simple NOT gate (*Set one bit* block in Fig. 3). Also, it must be noted that, the smaller the data encoded inside the data-word is, the bigger the number of MSBs set to the same value; which means our EMT offers great error correction capabilities in the scenario of biomedical applications, since they typically manipulate signals with big dynamic ranges and a distribution of the values centred around zero. In the following, we detail how DREAM operates.

A. Write Operating Mode

During write access to the data memory, both sample storage and mask ID determination are done in parallel. Each "error-free" sample produced by the ADC, for instance, is stored into the error-prone data memory and, at the same time, passes through a logical block used to determine the sign and the number of MSBs set at the same value in the data-word. Subsequently, the bit count (number of MSBs) is used to determine a mask ID, that will be concatenated with the sign bit and stored in an independent error-free memory running at a high supply voltage level to prevent the occurrence of permanent errors induced by voltage scaling, contrary to the big data memory that runs below nominal supply voltage to consume less.

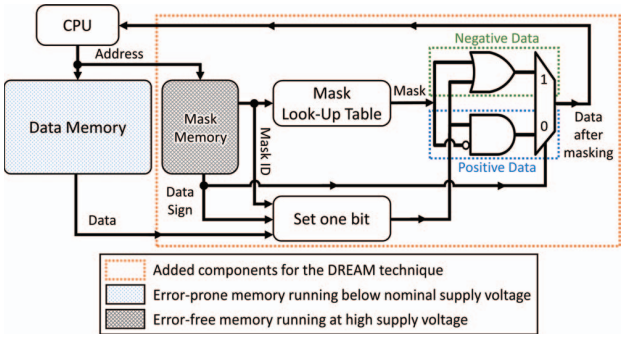


Fig. 3: DREAM technique - Read mode diagram

B. Read Operating Mode

In read mode, see Fig. 3, the data stored in the memory (possibly corrupted) and the mask ID concatenated with the data sign bit, are fetched from the memories. Secondly, the mask ID is converted into a full mask thanks to a lookup table. Then, two logical operations (AND and OR) are performed with the mask and the corrupted data and, finally, a 2 to 1 multiplexer controlled by the data sign bit selects the corrected data.

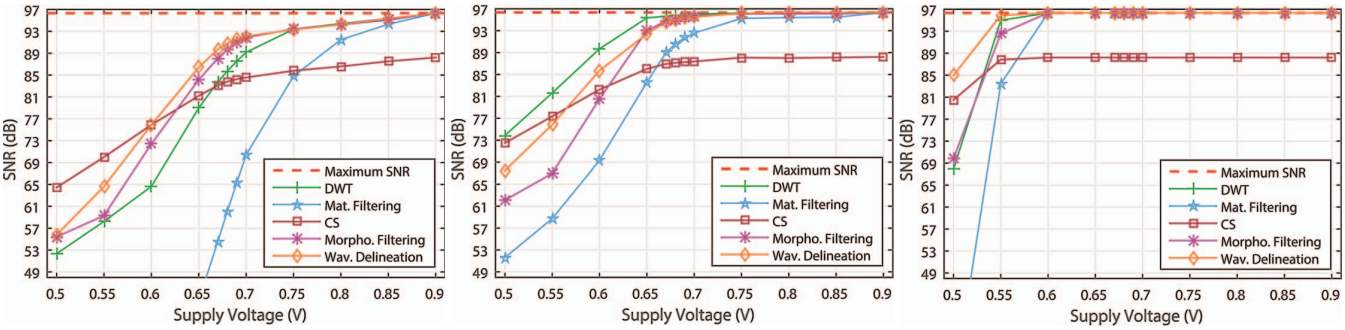
V. Experimental Setup

To evaluate the proposed approach using the DREAM technique, in terms of correction capabilities and power consumption, we model the architecture of the biomedical computing device INYU [12] by extending VirtualSOC [13], an existing multi-processor cycle-accurate architectural simulator. It can instantiate up to 16 ARM V6 cores with local and shared memories, accessed at a clock frequency of 200 MHz.

The biomedical applications work with 16-bit data-words stored into a shared memory of 32 kB, divided into 16 banks accessible by the cores through a crossbar. In order to compare EMTs, the memory has been enhanced to fully support DREAM and the well-known ECC SEC/DED [14], and instrumented for permanent error injection. To protect a 16-bit data-word, the memory overhead of these two EMTs are $1+log_2(16) = 5$ extra-bits for the DREAM technique and $2 + log_2(16) = 6$ extra-bits for the ECC SEC/DED.

Data corruption is caused by permanent errors that occur at random positions and set the affected memory bits to "1" or "0". Therefore, to provide fair comparisons, all the EMTs are tested reusing the same set of error locations/mappings. The amount of permanent errors or stuck-at faults injected depends on the Bit Error Rate (BER) [2], obtained profiling the memory for each voltage level for the selected technology node (32 nm) with low-power memory cells.

In order to get a representative sample, we ran 200 simulations per voltage level. We assume a different random fault-location map for every run, which can be generated even in the presence of stuck-at faults by adding a small logic to randomize the mapping between logical and physical addresses and bit locations. Then, an average of the 200 SNRs in dB is performed for every point. The dynamic and static energy consumption of the memory (Data + Mask memory) has been determined using CACTI 6.5 [15] and the synthesis power reports from Synopsys® Design Compiler for both the encoder and decoder of the different error correction mechanisms, assuming an operating temperature of 343 K.



(a) No protection (b) DREAM (c) ECC SEC/DED

Fig. 4: SNR Evolution versus Supply Voltage for each Application

VI. Experimental Results

A. Output Degradation Analysis

Fig. 4 illustrates the output SNR as a function of the memory supply voltage for all the applications and for different error protections. In all the graphs, as the data memory supply voltage decreases from 0.9 to 0.5V, the Bit Error Rate (BER) increases and the SNR decreases. A dashed line points out the maximum SNR value obtained with a data-word size of 16 bits when no errors affect the output. For the CS application, however, the maximum SNR is around 85 dB, because CS is, by construction, a lossy data compression algorithm which deteriorates the data even in the case of an error-free execution.

Fig. 4.b and Fig. 4.c, show that ECC SEC/DED offers a slightly better protection overall than DREAM in the range 0.55 to 0.65V. Below 0.55V (with multiple errors in the same data word) ECC SEC/DED underperforms, as it will detect but not correct the errors as DREAM does. However, in order to perform a complete and fair comparison between these two techniques, we need also to evaluate their energy consumption.

B. Energy Consumption Analysis

When ECC SEC/DED is used to protect the under-powered data memory, our experimental results show that the system consumes approximately 55% more energy for each voltage compared to the case with no error protection. With DREAM, the overall energy overhead is only 34%, reducing by 21% the overhead of ECC. This decrease is due to the reduction on the number of protection bits (cf. Section V), as well as the smaller design of the encoder and decoder used by DREAM (ECC requires 28% of area overhead for the encoder and 120% for the decoder, compared to those of DREAM).

C. Trading-off Result Quality for Energy Consumption

Combining the two aforementioned techniques and triggering, selectively, one or the other, according to the memory supply voltage and level of protection required, we can further reduce the energy budget. For example, in a real scenario where the DWT application can run with an output degradation tolerance of -1 dB, based on Fig. 4, three ranges of voltages could be used (e.g.: [0.9 ; 0.85], [0.85 ; 0.65] and [0.65 ; 0.55] Volts), to save up to 12.7% with no protection, 30.6% with DREAM and 39.5% with ECC SEC/DEC, compared to a system running at the nominal supply voltage (0.9V) with no protection. For voltages <0.55V, EMTs for multiple errors correction must be used to guarantee a reliable medical output.

VII. Conclusion

By studying several biomedical applications for embedded health monitoring systems, we have crafted the DREAM technique, an EMT which provides power savings by protecting dynamically and unequally an under-powered data memory in a new way compared to regular error protection schemes. It reduces approximately by 21% the energy budget consumed by traditional ECC. Moreover, experimental results show that different EMTs must be used, depending on the voltage, in order to find a trade-off between the error correction capability required and the energy budget. By demonstrating the possibility to produce acceptable results at near-threshold voltages with reduced energy consumption, we have paved the way for the development of promising ultra-low power wearable biosignal analysis devices.

References

[1] V. Chandra et al., "Impact of technology and voltage scaling on the soft error susceptibility in nanoscale CMOS," in *Proc. DFTVS*, Oct 2008, pp. 114–122.

[2] S. Ganapathy et al., "Variability-aware design space exploration of embedded memories," in *Proc. IEEE Israel*, Dec 2014, pp. 1–5.

[3] D. Nikolopoulos et al., "Energy efficiency through significance-based computing," *Computer*, vol. 47, no. 7, pp. 82–85, July 2014.

[4] W. Zheng et al., "Processor design with asymmetric reliability," in *Proc. ISVLSI*, July 2014, pp. 565–570.

[5] G. Karakonstantis et al., "Logic and memory design based on unequal error protection for voltage-scalable, robust and adaptive DSP systems," in *Signal Processing Systems*, vol. 68, no. 3, 2012, p. 415431.

[6] A. Ejlali et al., "A standby-sparing technique with low energy-overhead for fault-tolerant hard real-time systems," in *Proc. CODES+ISSS '09*. New York, NY, USA: ACM, 2009, pp. 193–202.

[7] "PhysioBank," http://www.physionet.org/physiobank/.

[8] F. Rincon et al., "Development and evaluation of multilead wavelet-based ECG delineation algorithms for embedded wireless sensor nodes," *IEEE T-ITB*, vol. 15, no. 6, pp. 854–863, Nov 2011.

[9] R. Braojos et al., "Early Classification of Pathological Heartbeats on Wireless Body Sensor Nodes," *Sensors*, vol. 14, no. 12, pp. 22 532–22 551, 2014.

[10] H. Mamaghanian et al., "Power-efficient joint compressed sensing of multi-lead ECG signals," in *Proc. ICASSP*, May 2014, pp. 4409–4412.

[11] H. Mamaghanian and N. Khaled et al., "Compressed sensing for real-time energy-efficient ECG compression on wireless body sensor nodes," *IEEE T-BME*, vol. 58, no. 9, pp. 2456–2466, Sept 2011.

[12] "INYU - The Inner You," http://www.smartcardia.com/inyu/.

[13] D. Bortolotti et al., "Virtualsoc: A full-system simulation environment for massively parallel heterogeneous system-on-chip," in *Proc. IPDPSW)*, May 2013, pp. 2182–2187.

[14] D. Rossi et al., "Error correcting code analysis for cache memory high reliability and performance," in *Proc. DATE'11*, March 2011, pp. 1–6.

[15] "CACTI 6.x," http://www.hpl.hp.com/research/cacti/.