Approximate Compressed Sensing: Ultra-Low Power Biosignal Processing via Aggressive Voltage Scaling on a Hybrid Memory Multi-core Processor

Daniele Bortolotti[†], Hossein Mamaghanian[‡], Andrea Bartolini[†], Maryam Ashouei[*]
Jan Stuijt[*], David Atienza[‡], Pierre Vandergheynst[‡] and Luca Benini[†]

[†]DEI - University of Bologna
Bologna, Italy
{daniele.bortolotti, a.bartolini, luca.benini}@unibo.it;

[‡]ESL, LTS2 - EPFL
Lausanne, Switzerland
{hossein.mamaghanian, david.atienza, pierre.vandergheynst}@epfl.ch;

[*]Holst Centre/imec
Eindhoven, The Netherlands
{maryam.ashouei, jan.stuijt}@imec-nl.nl

ABSTRACT

Technology scaling enables the design of low cost biosignal processing chips suited for emerging wireless body-area sensing applications. Energy consumption severely limits such applications and memories are becoming the energy bottleneck to achieve ultra-low-power operation. When aggressive voltage scaling is used, memory operation becomes unreliable due to the lack of sufficient Static Noise Margin. This paper introduces an approximate biosignal Compressed Sensing approach. We propose a digital architecture featuring a hybrid memory (6T-SRAM/SCMEM cells) designed to control perturbations on specific data structures. Combined with a statistically robust reconstruction algorithm, the system tolerates memory errors and achieves significant energy savings with low area overhead.

Categories and Subject Descriptors

C.1.4 [Mobile processors]

Keywords

Compressed Sensing; Approximate Computing; Ultra-Low Power; Hybrid Memory

1. INTRODUCTION

Emerging and future healthcare policies are fueling up an application driven shift toward long term monitoring of biosignals by means of embedded ultra-low power (ULP) devices. Modern human behavior-related diseases, such as cardiovascular pathologies, require accurate and non-stop medical supervision, which is unsustainable for the traditional healthcare system due to increasing costs and medical management needs [1]. Personal health monitoring systems are able to offer large-scale and cost-effective solutions to this problem.

Wearable health monitoring systems, enabled by Wireless Body Sensor Networks (WBSNs), face opposite requirements such as a continuously tighter power budget and an increasing demand of computation capabilities to pre-process locally the sensors information to reduce the amount of data to be transmitted as well as response time. To ensure minimal energy operation several aspects must be considered, combining optimizations of the signal processing aspects and of the technological layers of the ULP architecture. Recently the Compressed Sensing (CS) paradigm for signal acquisition and compression has proved to be effective in reducing energy consumption in embedded ECG monitors. Enabling a sub-Nyquist sampling rate for sparse signals, authors in [7] show $\approx$ 37% improved lifetime compared to state-of-the-art compression techniques. Motivated by the inherent parallel nature of medical grade ECG monitoring, where multi-channel signal analysis is often embarrassingly parallel, multi-core architectures demonstrated their efficiency compared to single-core solutions [10, 8]. In [10] is presented a multi-core architecture where individual leads are processed on different cores in parallel. Parallel processing enables more aggressive voltage-frequency scaling than single-core solutions, though at low workload requirements the single-core solution proved to be more efficient. Leakage power, mainly due to data and instruction memories, has a big impact and aggressive voltage scaling cannot be applied due to reliability issues of the memories.

Indeed, the failure probability of the conventional 6 Transistors (6T) SRAM cell increases considerably as the supply voltage is scaled down [11]. The usage of more reliable SRAM bit-cells, such as 8 Transistors (8T) or 10 Transistors (10T) cells, as well as standard cells memories (SCMEM) allows scaling to lower supply voltage, however, such solutions incur in large area penalties.

Approximate computing is an emerging paradigm that exploits intrinsic properties of multimedia and visual applications to tolerate errors to save energy allowing some final QoS degradation. [14] proposes an approximate full adder that reduces design complexity and power (up to 60%) while inducing a negligible QoS loss in JPEG and MPEG com-

ISLPED'14, August 11–13, 2014, La Jolla, CA, USA.

http://dx.doi.org/10.1145/2627369.2627629.

pression blocks. [12] uses a custom SRAM design with 6T memory for storing the LSB of each word and an 8T memory to store the MSB, applying such architecture to video decoder applications operating at low voltages. Both solutions are tailored to multimedia accelerators and are built based on the assumption that an error can be tolerated when occurring in the LSBs. As a matter of fact none of the state-of-the-art solution fully exploits the randomly distributed bit-flips errors that are typical in over-scaled SRAMs.

These considerations motivate the idea of the present work: by using a hybrid memory architecture, combining classic 6T-SRAM with SCMEM cells, we are able to offer an architecture that can operate at low voltages with a heterogeneous memory map composed of an error-free portion and an error-prone one. With an accurate data allocation of the CS internal structures between the two different portions, we are able to reduce the size of the SCMEM portion leading to a significantly lower area-overhead and, on the other hand, tolerates error induced by bit-flips in the SRAM by an innovative CS reconstruction algorithm.

The main contributions of this work are the following:

- a novel hybrid memory architecture for ULP multi-core biosignal processors is proposed. The combination of 6T and SCMEM banks enables operating at low-voltage while preserving data-correctness for the most critical data structures.
- the novel Approximate Compress Sensing paradigm is presented. Based on a reconstruction algorithm the proposed Compressed Sensing framework is capable of tolerating random bit-flips errors in the 6T memory.
- the proposed architecture allows to trade-off signal reconstruction quality with voltage supply and this leads to a significant improvement in energy saving. When operating at 0.6V, the hybrid memory architecture proves to be 5x more energy efficient than a purely 6T architecture (@ 0.8V) counterpart with a reduced area overhead ($\approx 13\%$). At 0.7V our architecture saves 60% of power with same reconstruction performance of standard CS for single lead ECG. When compared to SCMEM-only design our architecture has comparable power savings but with almost 10% less area overhead.

The rest of the paper is organized as follows. In Section 2 the hybrid memory multi-core architecture is introduced. Section 3 discusses the CS algorithm, the memory errors in low-voltage operation and the reconstruction algorithm. Next, in Section 4 we describe the experimental setup and the results of the proposed architecture in terms of energy efficiency, reconstruction quality and area overhead. Finally, the conclusions of this work are presented in Section 5.

2. HYBRID MEMORY ARCHITECTURE

We consider for the digital sensor node a baseline architecture similar to several current multi-core architectures targeting biosignal processors [8, 10]. The considered architecture, presented in Figure 1, features 8 Processing Elements (PEs) each one with a private Instruction Memory (IM). The PEs do not have private data caches, therefore avoiding memory coherency overhead, while they all share a L1 multi-banked tightly coupled data memory (TCDM) acting as a shared data scratchpad memory. The TCDM has a

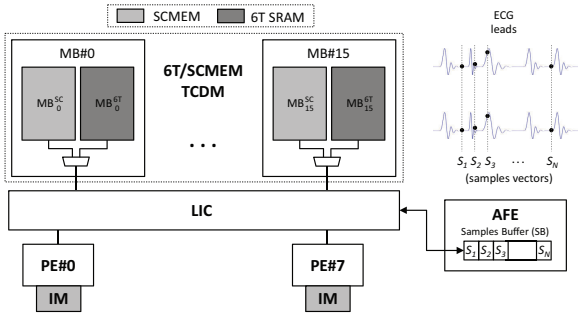

Figure 1: Multi-core architecture with hybrid 6T/SCMEM memory for Compressed Sensing

number of ports equal to the number of banks to have concurrent access to different memory locations. Intra-cluster communication is based on a low-latency high-bandwidth logarithmic interconnect (LIC) able to support single-cycle communication between PEs and memory banks (MBs). In case of multiple conflicting requests, for fair access to memory banks, a round-robin scheduler arbitrates the accesses. To ease the negative impact of banking conflicts we consider a banking factor of 2 (16 banks).

In the considered CS architecture, the input multi-channel signal is sampled by the analog front-end (AFE), with a sampling frequency according to the dynamics of the signal to analyze and the accuracy needed. The AFE is interfaced as a memory mapped buffer (SB in Figure 1) accessible through the LIC and can send interrupts to the PEs when the samples are ready for on the fly compression. Considering the limitations imposed by classic 6T-SRAM memory when operating aggressive voltage scaling and the characteristics of biomedical applications, we consider a hybrid memory architecture. By combining 6T and SCMEM-banks and a careful data allocation in the different memory portions, the system is capable of operating at ultra-low voltage and errors in the 6T portion are handled in the CS reconstruction algorithm. The 6T/SCMEM hybrid architecture is schematized in Figure 1 and it features a single voltage domain for the whole architecture, reducing area overheads and design complexity. The SCMEM portion of the TCDM offers reliable operation down to 400mV, while the 6T portion shows errors below 800mV as will be shown in Section 3.3.

3. ROBUST COMPRESSED SENSING

3.1 Compressed Sensing

Compressed sensing (CS), as an emerging tool has been investigated in many applications from low-power sensing and compression, radar and communication signal processing, high dimensional data analysis. The main idea behind CS is fairly simple and it assumes that given high dimensional data has a sparse representation which could be exploited to highly reduce the dimensionality of data.

Let $\mathbf{x}$ be the real-valued N-dimensional signal vector ($\mathbf{x} \in \mathbb{R}^N$) that is sparse or has a sparse representation in some known dictionary $\mathbf{x} = \mathbf{\Psi}\boldsymbol{\alpha}$. By sparse we mean that $\boldsymbol{\alpha}$ has only few non-zero elements. If we collect a vector of linear measurement $\mathbf{y} \in \mathbb{R}^M$ by $\mathbf{y} = \mathbf{\Phi}\mathbf{x}$, it is possible to recover the original signal $\mathbf{x}$ form measurements vector by solving a convex optimization problem. In the CS context, $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ is called *sensing matrix* and preferably $M \ll N$, so that the size of the measurement vector is much smaller than the original vector $\mathbf{x}$. To guarantee the recovery, the sensing

matrix $\mathbf{\Phi}$ must obey the key *restricted isometry property* (RIP) [17]:

$$(1-\delta_S)\left\|\boldsymbol{\alpha}\right\|_2 \leq \left\|\mathbf{\Phi}\mathbf{\Psi}\boldsymbol{\alpha}\right\|_2 \leq (1+\delta_S)\left\|\boldsymbol{\alpha}\right\|_2 \qquad (1)$$

for all S-sparse vectors $\boldsymbol{\alpha}$ and $\|.\|_2$ denotes the 2-norm of the vector. δ_S is the isometry constant of matrix $\mathbf{\Phi}$, which must be not too close to one.

If RIP holds, then an approximate sparse signal reconstruction can be accomplished by solving the following convex optimization problem:

$$\min_{\tilde{\boldsymbol{\alpha}}\in\mathbb{R}^N} \left\|\tilde{\boldsymbol{\alpha}}\right\|_1 \qquad \text{s.t.} \qquad \left\|\mathbf{\Phi}\mathbf{\Psi}\tilde{\boldsymbol{\alpha}}-\mathbf{y}\right\|_2 \leq \sigma \qquad (2)$$

where σ bounds the amount of noise corrupting the data. Usually in CS context the ℓ_1 norm is used as a sparsity inducing norm and it is proven to reach the sparse solution, while $f(\boldsymbol{\alpha}) = \left\|\mathbf{\Phi}\mathbf{\Psi}\tilde{\boldsymbol{\alpha}}-\mathbf{y}\right\|_2$ is named data fidelity or data fitting function which is the least square estimate of the answer to the inverse problem (2).

3.2 Multi-lead ECG and joint Compression

ECG signals are known to be compressible in Discrete Wavelet Domain (DWT). By compressible we mean that even though they are not exactly sparse, it exists an S-sparse approximation which contains most of the information of the signal and the same principles can be applied to some extent [7].

For multi-lead signals we can write the same problem in matrix form. Let $\mathbf{X} \in \mathbb{R}^{N\times L} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L]$ be the real valued matrix of ECG signals where L is the total number of leads and each column corresponds to a single ECG lead. This matrix could be represented on the DWT domain by $\mathbf{X} = \mathbf{\Psi}\mathbf{A}$, where matrix $\mathbf{A}$ is the sparse coefficients matrix and $\mathbf{\Psi}$ is the DWT matrix. Then the CS recovery problem ((2)) could be solved in multi-lead case too. But in the case of multi-lead ECG compression, where there is a strong correlation between the sparsity structure among the leads, the sparse coefficients model should be refined to take it into account. In such a situation, where non-zero coefficients are naturally partitioned in subsets or groups, the best choice could be using a group-sparsity inducing term [22]. In a recent prior work [21], we proposed to replace the ℓ_1 norm with mixed ℓ_1/ℓ_2 norm. It behaves like an ℓ_1-norm on the vector $(\|\boldsymbol{\alpha}_i\|_2)_{i\in\mathcal{L}}$ in $\mathbb{R}^{|\mathcal{L}|}$, and therefore, induces group sparsity.

3.3 Low Voltage Memory Operation

The classic 6T SRAM is not able to reliably operate at lower supply voltage. One way to address the problem is to have different supply voltages for the memory and the logic sub-blocks [2]. This solution results in the overhead of generating multiple supply voltages and back-end complexity of having separate voltage domains, and the level-shifting overhead between the memory and logic voltage domains. More importantly, the solution does not allow for minimum energy operation due to higher operating voltage of the memory, and therefore not addressing the memory power wall. Another solution is the use of 8T (or even 9T, 10T) cells that were shown operating at lower voltages [3, 4, 5]. While foundries provide 8T SRAM bit cell, the cell is not characterized for low voltage operation. Furthermore, the commercial SRAM generators do not provide characterization points (e.g. timing and power information) for low voltage operation. This prevents proper timing closure at

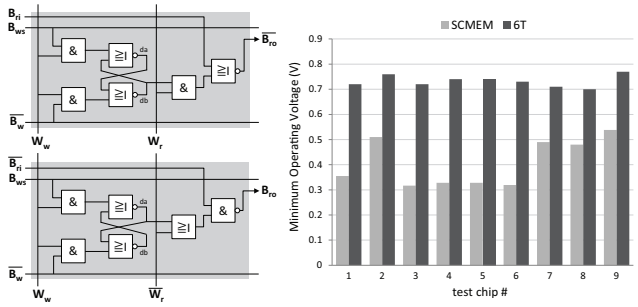

Figure 2: Schematic of SCMEM cell based on AOI/OAI gates (left) and Minimum operating voltage of different chips (right)

low voltage using commercial SRAMs. Custom design of such SRAMs presents a big design effort and can be limited to a few instance sizes. In this work, we use a standard cell-based memory (SCMEM) module. The approach has been proposed in the past [6] and similar to 8T or 10T SRAM, it suffers from an area overhead. We propose to use regular place & route (P&R) in the digital EDA flow for such memory to achieve significant area reduction.

A standard cell-based memory was designed and fabricated in a 40 nm CMOS technology. For comparison reason, also a commercial 6T memory was fabricated on the same chip. The SCMEM uses a cross-coupled pair of AND-OR-INV (AOI) as the storage element (Figure 2, left). The choice of the memory element, combined with the use of regular P&R using the CADENCE Encounter-SoC, results in more than 3x area saving compared to [6] that uses a latch as the storage element. Nine chips were measured and their corresponding minimum operating voltage point is shown in Figure 2 (right) for both the proposed SCMEM and the 6T memory. The results show that for the majority of the chips, the SCMEM operated correctly at voltages below 0.4V and on average it has 400mV lower minimum operating voltage point than the 6T memory. We also measured the minimum retention voltage and both SCMEM and 6T memory have similar retention voltage. The comparison of the memory with state of the art is shown in Table 1.

Error in memory: for the commercial memory, which fails at higher supply voltages, we also measured the errors in memory across the voltage range for nine chips. The errors was characterized by bit-flips randomly placed among all the bits [9]. The measured data was fitted to get the error probability as follows: $P_e = A(V_0 - V_{dd})^k$ where $V_0 = 0.85V$, $k = 6.14$. For the SCMEM, the same fitting equation is accurate when $V_0 = 0.55V$ [9].

3.4 Robust Compressed Sensing

Compressed sensing is known as a robust compression technique in case of noisy measurements vectors corrupted by i.i.d. Gaussian noise. The optimization still works pretty well as long as the amplitude of the noise is small. This is not true if the noise has a very harsh and coarse nature like the bit-flip errors which potentially could be very high in amplitude and not bounded.

Gaussian random distribution are a popular candidate for constructing the sensing matrix $\mathbf{\Phi}$. Such matrix is known to be measurement optimal. The lower bound for the number of measurements is proven to be $M = \Omega(S\log N)$. As far

Table 1: Comparison of different implementations of a 1k x 32b memory (TT corner, 1.1V, 25C)

Feature	Unit	6T commercial 40nm LP	Custom SRAM [3] 40nm LP	SCMEM [6] 65nm LP	SCMEM This work 40nm LP
Dynamic power	pJ	11.9	3.6	-	1.4
(reduced voltage)		-	-	0.93 @ 0.4V*2	0.19 @ 0.4V
Leakage	W	2.15	11*2	-	5.9
(reduced voltage)		-	-	$\geq$ 19.7 @ 0.25V*2	-
Area	mm^2	0.011	0.024*3	$\geq$ 0.186*3	0.058
Min Retention Voltage	V	0.29*1	-	0.25	0.32*1
Min R/W Voltage	V	0.7*1	-	0.35	0.33*1
Performance (nominal)	MHz	816	454 @ 1.2V	9.5 @ 0.65V	96*1
(reduced voltage)		-	-	0.1-0.2 @ 0.45V	0.4 @ 0.45V*1

*1measurements results *2scaled: $P_{dyn} \propto$ word length *3scaled: $A \propto$ total bits $\times(40nm/\lambda_{orig})^2$

as a sparsity S considered, Gaussian matrices are not space optimal and need a huge storage space and nor time optimal since encoding and reconstruction complexity is in the order of $\mathrm{O}(MN)$, which makes them not practical for limited resources real-time digital nodes. As an alternative, a sparse binary matrix is used as our sensing matrix $\mathbf{\Phi}$, where each column contains only d non-zero elements equal to 1, ($d \ll M \ll N$) with required space and time complexity of $\mathrm{O}(dN)$ [7]. For such a sensing matrix, the RIP property of (1) is not valid, however, it satisfies a different form of this property, with slightly more required measurements to guarantee the reconstruction [18]. This sensing matrix choice could be very efficient in terms of storage in TCDM, since we only need to store the indexes of non-zero elements that requires $\mathrm{O}(dN)$ space instead of $\mathrm{O}(MN)$, where $d \ll M$. When storing this matrix in an error-prone memory (6T), it means that the positions of the non-zero elements would change. As a consequence the corrupted sensing matrix $\mathbf{\Phi_c}$ multiplied with the original signal x are represented as $\mathbf{\Phi_c} = (\mathbf{\Phi} + \mathbf{E})$, where $\mathbf{E}$ is the error matrix. $\mathbf{E}$ has only ± 1 entries, $+1$ where an entry is added and -1 where is missed. As long as the number of errors is small and $\mathbf{E}$ is sufficiently sparse, the reconstruction problem (2) can be rewritten with an additional term for minimization and recover the original signal by solving the following optimization problem:

$$\min_{\tilde{\mathbf{A}},\tilde{\mathbf{E}}} \left\|\tilde{\mathbf{A}}\right\|_{1,2} + \lambda \left\|\tilde{\mathbf{E}}\right\|_1 \quad \text{s.t.:} \left\|(\mathbf{\Phi} + \tilde{\mathbf{E}})\mathbf{\Psi}\tilde{\boldsymbol{A}} - \mathbf{Y}\right\|_2 = 0 \quad (3)$$

where $\|.\|_{1,2}$ denotes the joint ℓ_1/ℓ_2 norm of a matrix. Due to presence of multiplicative noise, the problem (3) in general is not convex and our main goal is to design the recovery algorithm to reach at least the local minimum, and hopefully the global one, if the amount of corrupting noise is limited.

3.5 Reconstruction Algorithm

The formulation of the problem in (2) shows that due to presence of term $\mathbf{EA}$ in constraint, the problem is not convex, but if one of each is given then the problem is very like to the normal Lasso problem and could be treated as a convex optimization problem. If $\mathbf{E}$ is given the problem is the normal Joint CS problem while if $\mathbf{A}$ given then it is similar to the Lasso problem. Similar to the S-TLS problem [24], this suggest to have an iterative decent algorithm yielding successive estimates of the $\mathbf{E}$ and $\mathbf{x}$ with $\mathbf{x}$ and $\mathbf{E}$ fixed respectively. Then the recovery algorithm is two folded and includes two convex optimization problems. First, when $\mathbf{E}$ is considered to be fixed we solve:

$$\tilde{\mathbf{A}}^k = \arg\min_A \left\|(\mathbf{\Phi} + \tilde{\mathbf{E}}^{\mathbf{k}})\mathbf{\Psi}\mathbf{A} - \mathbf{Y}\right\|_2^2 + \lambda_1 \|\mathbf{A}\|_{1,2} \quad (4)$$

where k is the iteration number and in the next iteration when the estimate of $\mathbf{A}$ is available, we fix $\mathbf{A}$ and try to solve the problem for $\mathbf{E}$:

$$\tilde{\mathbf{E}}^{k+1} = \arg\min_E \left\|\mathbf{E}\mathbf{\Psi}\mathbf{A}^k - (\mathbf{Y} - \mathbf{\Phi}\mathbf{\Psi}\mathbf{A}^{\mathbf{k}})\right\|_2^2 + \lambda_2 \|\mathbf{E}\|_1 \quad (5)$$

To solve these set of equations we use the proximal gradient methods which are computationally not hard for the type of problems (4) and (5). More specifically, here we are using an accelerated version of the proximal algorithms proposed by [19, 16] which are proven to have faster convergence. Algorithm 1 shows the pseudo code of the proposed algorithm.

Data: $Y, \Phi, \lambda_1, \lambda_2, L_1, L_2$
initialization;
while *not converged* **do**
  $G^A = Y^A + \frac{2}{L_1}\big(\Psi(\Phi+\hat{E})^T(Y-(\Phi+\hat{E})\Psi^T Y^A)\big)$;
  $\hat{A} = \mathcal{P}_{L12}(G^A, \lambda_1/L_1)$;
  $t = \frac{1+\sqrt{1+4*told^2}}{2}$;
  $Y^A = \hat{A} + \frac{told-1}{t}(\hat{A} - \hat{Aold})$;
  $G^E = Y^E + \frac{2}{L_2}\big(Y-(\Phi+Y^E)(\Psi^T\hat{A})\big)(\Psi^T\hat{A})^T(I+\hat{A}\hat{A}^T)^{-1}$;
  $\hat{E} = \mathcal{P}_{L_1}(G^E, \lambda_2/L_2)$;
  $Y^E = \hat{E} + \frac{told-1}{t}(\hat{E} - \hat{E}_{\text{old}})$;
  $\hat{E}_{\text{old}} = \hat{E}$, $\hat{A}_{\text{old}} = \hat{A}$, $t_{\text{old}} = t$;
end

Algorithm 1: Reconstruction algorithm

The proof is based on the basic convergence of the proximal gradient decent algorithms, the first term of problem (4) and (5) are differentiable and the non differentiable term (ℓ_1 and $\ell_{1,2}$) are separable in the entries of A and E. The convergence to a local minimum is also guaranteed since it always iterates towards a reduced cost for the problem (3). Moreover, simulated tests also demonstrate the convergence of the algorithms and in case of less corrupted data the local optimum is very close to the global optimum.

4. EVALUATION

4.1 Hybrid Memory Partitioning

To implement a multi-lead Compressed Sensing (CS) we have considered a window size of $N = 512$ samples for a fixed Compression Ratio (CR) of 50% and a sampling frequency of 512 Hz. The sensing matrix is constructed off-line and stored at boot time in the error-prone 6T memory portion of the TCDM and shared among all the channels. Since only the indexes of non-zero elements (d) are stored, the required space is equal to $d \cdot N$ (where $d = 16$) and the entries can be represented with a single byte. Considering our multi-core architecture, each processing element works on a separate

input data-sets associated to different leads performing on the fly compression. An interrupt generated by the AFE triggers execution on the new sample, thus not requiring to store the input vectors in TCDM. The memory footprint of the CS algorithm consists of 348B for instructions and 16KB for data. The data section comprises the sensing matrix in form of a LUT (8192B), i.e. a vector of random coefficients for the CS projections, and the output buffers for the 16 leads (8KB).

Such CS algorithm analysis was used at design time to choose the appropriate memory cuts and allocate the data. The total TCDM size is assumed to be 18KB (for the data section and 256B of stack per-core), while an instruction memory of 512B (private, per-core) is chosen. The size of the sensing matrix defines the 6T memory portion: 8192B split in 16 banks leading to 512B per-bank. The remaining portion of the TCDM (10KB) is split in 16 SCMEM banks leading to 640B per-bank. Address interleaving performed by the logarithmic interconnect (Section 2) allows a contiguous logical memory map and static allocation in the two portions is easily done by means of linker script sections and compiler variable attributes.

4.2 Area Overhead (iso-size)

To evaluate the area overhead of our solution, in an *iso-size* comparison, we quantified the overhead introduced by the SCMEM memory portion in the hybrid architecture compared to a system where all memory instances are 6T (*6T-only*) and another where all instances are SCMEM (*SCMEM-only*). For the SCMEM memory cuts we considered the numbers presented in Table 1, while for the 6T memory numbers are taken from a low power 40nm technology library. Area figures for the processing elements (PEs) and the logarithmic interconnect are scaled from a 28nm RTL design. Results presented in Table 2 show the evaluation of how of each element impacts on total area.

Table 2: Area comparison (6T/SCMEM, 6T-only, SCMEM-only). Numbers presented in μm^2.

ELEMENT	6T/SCMEM	6T-ONLY	SCMEM-ONLY
PEs	323439	323439	323439
IM	132819	97960	132819
SCM TCDM	332048	-	597686
6T TCDM	195920	431968	-
TOT TCDM	527968	431968	597686
LIC 8x16	88420	88420	88420
TOTAL	1072646	941787	1142364

The extra-circuitry required for the hybrid memory consists of a basic decoder on the address line and a multiplexer on the output line. Memory area is dominated by the memory matrix itself and not by the decoder/muxes making this contribution negligible. The area overhead of the hybrid memory with respect to a 6T-only architecture is $\approx$ 13%. On the other hand, the SCMEM-only architecture incur in a higher area penalty, the overhead on the overall system would be $\approx$ 21%.

4.3 Energy Efficiency

The proposed architecture has been modeled and integrated in a SystemC-based cycle-accurate virtual platform [15] with back-annotated power numbers for the memory subsystem (Table 1) and the rest of the logic (LIC, PEs)

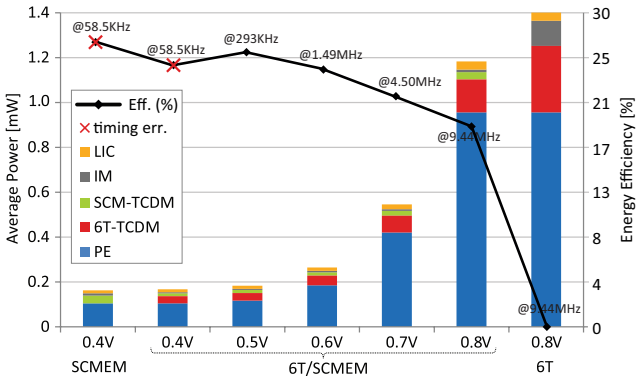

Figure 3: Average power at different operating points for architectures (6T/SCMEM, 6T-only, SCMEM-only) and Energy Efficiency (T=25°C)

extracted from a RTL-equivalent architecture with a customized OpenRISC core for minimum energy. Considering the CS application described in Section 4.1, the virtual platform shows a maximum error in timing accuracy below 6% with respect to RTL simulation. The architecture was configured with 8 cores, an 8x16 logarithmic interconnect and 6T/SCMEM portions as determined in Section 4.1. The CS algorithm, with 16 leads and 8 cores performing compression on 512 samples per window (1 sec), executes in $\approx$ 104 Kcycles.

Figure 4 shows the average power consumption, and its breakdown, during CS execution for the proposed architecture and the 6T-only and SCMEM-only reference designs. In addition the figure shows the energy efficiency (computed in a 512 samples window) at the different design corners for the different cases. The plot clearly shows that the PEs account for the majority of the power consumption and this is mainly due to the small memory size, while the IM and LIC power consumption is negligible. We can notice that the proposed hybrid architecture has a significant power saving (13%) w.r.t. 6T-only at the same voltage (0.8V) and by scaling the voltage supply the proposed architecture gains extra power saving. At 0.6V we save the 81% of power but scaling further ($<$ 0.5V) the system fails to compress the 512 samples within a window. Moreover, when compared to SCMEM-only design our architecture has similar power performance. It results that our hybrid architecture can effectively trade-off the 6T-only and SCMEM-only designs. At 0.6V our architecture has similar power saving and reconstructed quality of the SCMEM-only but with significantly less area overhead.

4.4 Reconstruction Quality

To characterize the error in the 6T memory, 10 trials of separate read/write sequences were performed varying the voltage. Measurements results are shown in Table 3 in the voltage range 400mV-750mV. The probability of er-

Table 3: Probability of bit-flip errors in the 6T memory at different voltages

Voltage [V]	P(bit-flip)	Voltage [V]	P(bit-flip)
0.40	0.0707	0.60	0.0022
0.45	0.0356	0.65	0.0007
0.50	0.0162	0.70	0.0001
0.55	0.0065	0.75	1.3e-5

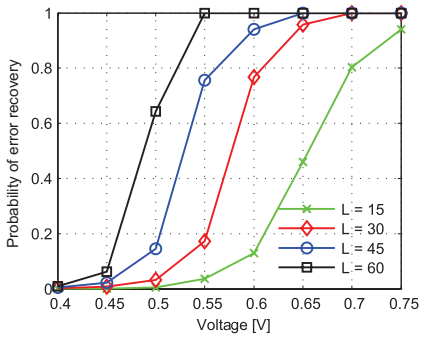

Figure 4: Probability of error recovery in reconstruction for different number of leads and voltages

rors in the whole sensing matrix then can be represented as $P = 1-(1-P_b)^{nb}$, where P_b is the probability of the error in each bit (Table 3) and nb represents the number of bits. For very low voltage values the number of errors can be significantly high, situation where the classic CS reconstruction algorithm fails. Figure 4 shows the results of the Reconstruction Algorithm for a simulated data averaged over 100 tests. It is clear that as the number of leads increases, the error recovery improves and more bit-flips in the sensing matrix can be recovered. This feature has major potential benefits for processing biosignals that require a larger number of leads (i.e. EEG).

To validate the performance of the proposed robust compression scheme, we use the PTB Diagnostic ECG Database, available online [20]. The database contains 549 records of 15-lead ECG from 290 subjects. Signals are sampled at 1 KHz with 16-bit resolution. Here we have down-sampled the signals to 512 Hz to be consistent with our system requirements. To quantify the compression performance while assessing the diagnostic quality of the compressed records, we consider the Signal to Noise Ratio (SNR) defined as $SNR = 20\log_{10}\|\mathbf{x}\|_2/\|\mathbf{x}-\tilde{\mathbf{x}}\|_2$. Figure 5 shows the average SNR for the Robust CS over different memory voltages. The results are all for 50% compression ration (i.e. $M = N/2$).

To clearly demonstrate the quality of the reconstructed signal, two windows of 1 second for 0.6V and 0.7V are shown in Figure 5 (right). The results show that even for low voltage (0.6V) the algorithm is still able to converge to the solution while the normal CS would fail.

5. CONCLUSIONS

In this work we present a 6T/SCMEM hybrid memory multi-core architecture for biosignal processing. Classic 6T memories face reliability issues when reducing supply voltage to threshold. By partitioning the compressed sensing data structures in the hybrid memory, combined with a novel reconstruction algorithm, we can tolerate bit-flips in 6T memory trading-off reconstruction quality for energy savings. Our solution offers significant improvements in power ($\approx$ 60%, -2.5dB @ 0.7V) with a low ($\approx$ 13%) area overhead.

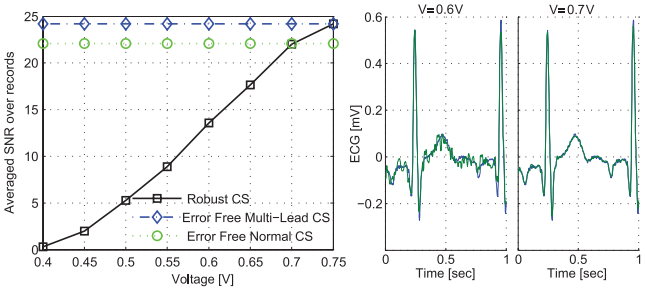

Figure 5: Average SNR for robust CS at different voltages (left). Reconstructed signal quality at 0.6V and 0.7V (right).

Acknowledgments

Work supported by the EU project PHIDIAS (g.a. 318013) and ONR-G grant no. N62909-14-1-N072.

6. REFERENCES

[1] http://www.who.int/mediacentre/factsheets/fs317/en.
[2] Ashouei, M. et al., *"A voltage-scalable biomedical signal processor running ECG using 13pJ/cycle at 1MHz and 0.4 V"*, ISSCC, 2011.
[3] Rooseleer, B. and Wim D., *"A 40 nm, 454MHz 114 fJ/bit area-efficient SRAM memory with integrated charge pump"*, ESSCIRC, 2013.
[4] Sharma, V. et al. *"8T SRAM with mimicked negative bit-lines and charge limited sequential sense amplifier for wireless sensor nodes"*, ESSCIRC, 2011.
[5] Verma, N., and A. 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.
[6] Andersson, O. et al., *"Dual-VT 4kb sub-VT memories with < 1pW/bit leakage in 65 nm CMOS"*, ESSCIRC, 2013.
[7] Mamaghanian, H. et al., *"Compressed sensing for real-time energy-efficient ECG compression on wireless body sensor nodes"*, IEEE Transactions Biomedical Engineering, vol. 58, no.9 pp. 2456–2466, 2011.
[8] Dreslinkski, R. G., et al., *"An energy efficient parallel architecture using near threshold operation"*, PACT, 2007.
[9] Gemmeke, T. et al., *"Resolving the Memory Bottleneck for Single Supply Near-Threshold Computing"*, DATE, 2014.
[10] Dogan A.Y. et al., *"Multi-core architecture design for ultra-low-power wearable health monitoring systems"*, DATE, 2012.
[11] Calhoun, B. H. et al., *"Analyzing static noise margin for sub-threshold SRAM in 65nm CMOS"*, ESSCIRC, 2005.
[12] Chang I.J. et al., *"A Priority-Based 6T/8T Hybrid SRAM Architecture for Aggressive Voltage Scaling in Video Applications"*, IEEE transactions on circuits and systems for video technology, vol. 21, no. 2, Feb 2011.
[13] Bortolotti D. et al., *"Hybrid memory architecture for voltage scaling in ultra-low power multi-core biomedical processors"*, DATE, 2014.
[14] Gupta, V. et al., *"IMPACT: imprecise adders for low-power approximate computing"*, ISLPED, 2011.
[15] Bortolotti D. et al.,*"VirtualSoC: a Full-System SimulationEnvironment for Massively Parallel Heterogeneous System-on-Chip"*, IPDPWS, 2013.
[16] Beck A. and Teboulle M., "*Fast iterative shrinkage-thresholding algorithm with application to wavelet-based image deblurring*", ICASSP, 2009.
[17] Candes E. et al., *"Stable signal recovery from incomplete and inaccurate measurements"*, Communications on Pure and Applied Mathematics, 59:pages 1207–1223, 2006.
[18] Chandar V., *"A negative result concerning explicit matrices with the restricted isometry property"*, Tech. report, 2008.
[19] Nesterov, Y. "*A method of solving a convex programming problem with convergence rate O(1/k2)*", Soviet Mathematics Doklady. Vol. 27. No. 2. 1983.
[20] Goldberger A. L. et al., *"Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals"*, Circulation,101(23):pp. 215–220, 2000.
[21] Mamaghanian H. et al.,*"Power-efficient joint compressed sensing of multi-lead ecg signals"*, ICASSP, 2014.
[22] Kowalski M. et al., *"Sparsity and persistence: mixed norms provide simple signal models with dependent coefficients"*, Signal, Image and Video Processing, 3(3):pages 251–264, 2009.
[23] Lin Z. et al., *"The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices"*, arXiv preprint arXiv:1009.5055, 2010.
[24] Zhu, H. et al., *"Sparsity-cognizant total least-squares for perturbed compressive sampling"*, Signal Processing, IEEE Transactions on 59, no 5 (2011): pp. 2002–2016.