Nano-Engineered Architectures for Ultra-Low Power Wireless Body Sensor Nodes

Rubén Braojos
David Atienza

École Polytechnique Fédérale
de Lausanne
CH-1005, Switzerland
{ruben.braojoslopez;
david.atienza} @epfl.ch

Mohamed M. Sabry Aly
Tony F. Wu
H.-S. Philip Wong
Subhasish Mitra

Stanford University,
CA- 94305, U.S.A.
{msabry; tonyfwu; hspwong;
subh}@stanford.edu

Giovanni Ansaloni

Università della Svizzera
Italiana
CH-6900, Switzerland
giovanni.ansaloni@usi.ch

ABSTRACT

Wireless body sensor nodes (WBSNs) are miniaturized devices that are able to acquire, process and transmit bio-signals (such as electrocardiograms, respiration or human-body kinetics). WBSNs face major design challenges due to extremely limited power budgets and very small form factors. We demonstrate, for the first time in the literature, the use of disruptive nanotechnologies to create new nano-engineered ultra-low power (ULP) WBSN architectures. Compared to state-of-the-art multi-core WBSN designs, our new architectures dramatically reduce power consumption by 5.42x and footprint by 5x, while fulfilling real-time processing requirements of bio-signal monitoring applications. Our WBSN architectures achieve these results by utilizing emerging non-volatile memory technologies (such as resistive RAM and spin-transfer torque RAM) and their ultra-dense and fine-grained three-dimensional integration with logic (such as monolithic three-dimensional integration naturally enabled by carbon nanotube field-effect transistors).

CCS Concepts

• **Hardware → Emerging architectures**.
• **Computer systems organization → Multicore architectures.**
• **Applied computing → Health informatics.**

1. INTRODUCTION

Ongoing demographic and lifestyle changes are increasing the prevalence of chronic disorders, which are now the major sources of death worldwide [WHO15]. These ailments require extensive monitoring, which is often uncomfortable for patients and represents major financial burden for healthcare providers. Wearable bio-signal monitoring devices record bio-signals of patients even outside the hospital environment and with little intervention from medical staff. Such devices can help lower healthcare costs, and also improve the quality of life of patients affected by chronic diseases.

Wearable bio-signal monitoring devices, also known as *Wireless Body Sensor Nodes (WBSNs)*, must autonomously acquire, record and wirelessly transmit bio-signals (such as electrocardiograms) over extended periods of time, while relying on small batteries or energy harvesters. Thus, power efficiency of the entire system, from acquisition to transmission, is essential for the ubiquitous use of WBSNs. A *naïve WBSN* transmits raw acquired bio-signals, and is not power-efficient [Zhang12]. *Smart WBSNs* overcome this limitation through on-node advanced Digital Signal Processing (DSP) (e.g., compression, feature extraction, and classification [Hao08]). Thus, the required transmission bandwidth over the energy-hungry wireless link is significantly reduced [Rincon11]. However, as a result, the DSP stage itself becomes important (Figure 1). To perform complex bio-signal processing within an ultra-low power envelope, embedded digital platforms must be carefully tailored to the specific domain and its workload characteristics. For instance, state-of-the-art electrocardiogram (ECG) compression and filtering algorithms [Mamaghanian11, Braojos14] experience extended idle periods (>90% of the inter-sample arrival time as shown in Figure 2). Typically, in commonly-used SRAM-based WBSN platforms, the power consumption during these idle periods can account for up to 86% of the overall power consumption (see Figure 1). This is primarily due to the low sampling frequency of the acquired signals. Due to the volatility of on-chip SRAMs, power-gating during these idle periods requires the backup of the full data memory and processor states to (off-chip) non-volatile memory.

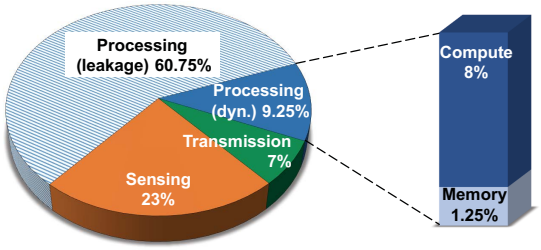

Figure 1. Power consumption breakdown for an SRAM-based WBSN executing a multi-channel bio-signal processing application. Values are computed based on [Braojos14], [Zhang12] and [TI-CC2540].

CODES/ISSS '16, October 01-07 2016, Pittsburgh, PA, USA

ACM 978-1-4503-4483-8/16/10…$15.00
DOI: http://dx.doi.org/10.1145/2968456.2968464

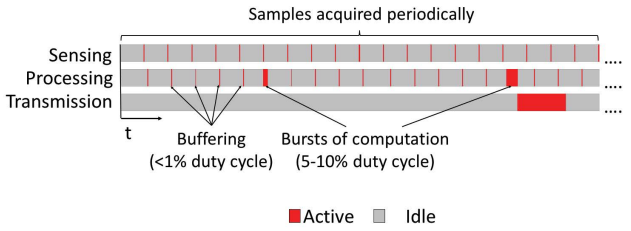

Figure 2. Activity profile of a typical WBSN.

Today's WBSNs typically use FLASH memory for non-volatile storage. FLASH memory stores the program and data memory contents when the system is switched off. At power-up, its contents are transferred to on-chip SRAM. However, such systems cannot support fine-grained power-gating over relatively short idle periods (as in Figure 2) to reduce the aforementioned leakage power for two reasons. First, strict real-time deadlines for this application domain can no longer be met due to very long write latencies (the time required to write a word into FLASH, ~120µs for small arrays [Mitani16, Nakashima15, Taito15]); i.e., the time needed to store the system state would exceed the inter-sample time. Second, the energy cost of shadowing the full data memory several hundreds of times per second can exceed the potential savings obtained from of power-gating.

To overcome the challenge, we present a new nano-engineered WBSN architecture that leverages the benefits of emerging nanotechnologies: low-voltage, non-volatile memory (NVM) structures (STTRAM [Kent15], RRAM [Wong15]) in conjunction with ultra-dense, fine-grained 3D integration (termed monolithic 3D [Wei13, Shulaker14]). To enable monolithic 3D integration, we use carbon nanotube field-effect transistors (CNFETs) for NVM access circuitry. Our main contributions in this paper are:

1) We present a new ultra-low power WBSN architecture which utilizes upcoming nanotechnology advances to obtain significant application-level power savings.

2) We present system management policy which allows low-overhead and fine-grained power gating of computation and storage elements to obtain power savings while meeting application-level real-time deadlines.

3) We present a detailed analysis of our nano-engineered architecture.

Figure 3 shows various architectures of WBSN integrated circuits (ICs) incorporating NVM that are analyzed in this paper. In Figure 3a, the NVM access transistors (as well as the transistors used for processor cores, etc.) are realized using conventional silicon CMOS transistors on the substrate as in typical (2D) silicon CMOS ICs. (The NVM memory elements reside on an upper metal layer and are connected to the access transistors on the bottom-most layer using conventional interconnects). Figure 3b shows a 3D architecture realized using through-silicon via (TSV) technology. In this case, the NVM together with its access transistors (NVM tier) is integrated on top of the processor cores and related circuitry (processing tier) using TSVs. The transistors are fabricated using conventional silicon CMOS technology. The architecture in Figure 3c is realized using monolithic 3D integration. After fabricating the processing tier using traditional silicon CMOS technology, CNFETs (used for access transistors of NVM) are fabricated directly on top to form the next tier of circuits. The NVM is then fabricated. In this architecture, inter-layer vias (ILVs) traditionally used for on-chip interconnects, are used to connect the tiers. Such monolithic 3D integration using RRAM has been experimentally demonstrated in [Shulaker 14].

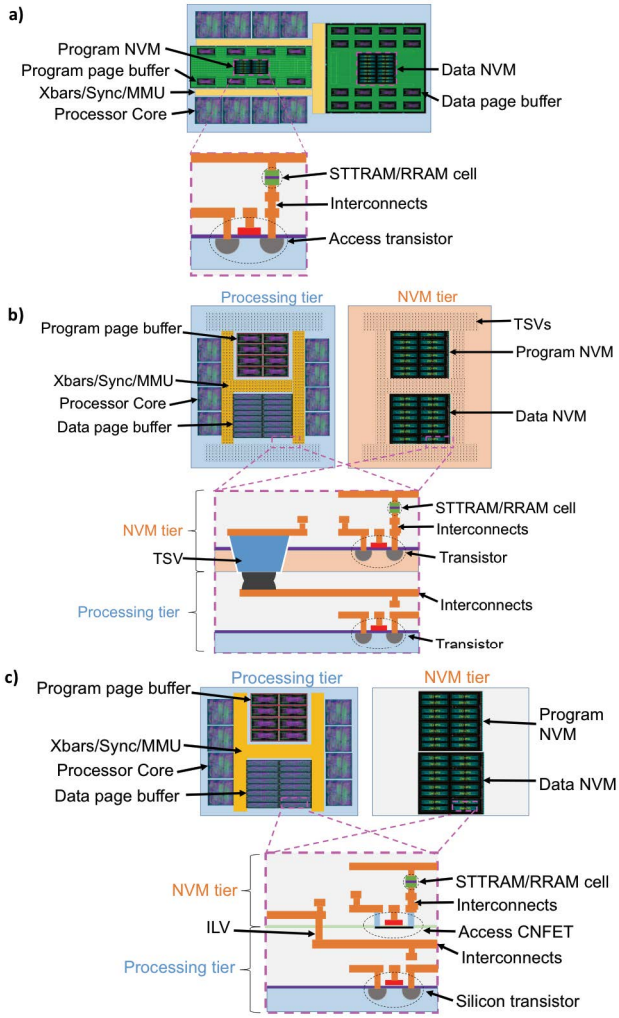

Figure 3. Overview of a) 2D architecture with NVM, b) 3D architecture with NVM using TSVs, and c) monolithic 3D architecture with NVM.

Our new WBSN architecture (following Figure 3c), enabled by NVM and monolithic 3D integration, achieves up to 5.42x power savings, while still meeting execution time constraints, versus an SRAM-based baseline system. Moreover, the footprint of such a nano-engineered system is reduced by 5x, compared to the 2D architecture of Figure 3a. A synergistic combination of several elements help us achieve such significant benefits. The NVM enables power gating of the system at idle times. Our architecture considers the nature of WBSN applications, and overcomes NVM limitations by building a 2-level memory hierarchy (latch-based level 1 and NVM-based level 2). The ultra-dense and fine-grained monolithic 3D integration enables efficient transfer (1-cycle transfer) between the levels of memory hierarchy for quick power gating. The required interconnection density is provided by monolithic 3D integration in an area-efficient manner vs. state-of-the-art TSVs [Shulaker15]. The footprint area benefits of our

approach greatly enhances implantability of such health monitoring systems [Bazaka12]. The footprint area benefit is even greater when compared to a traditional system with off-chip NVM connected using either a silicon interposer or through board-level integration.

The paper proceeds as follows: in Section 2, we present technological foundations used in this work. In Section 3, we present details of our WBSN architecture. In Section 4, we present simulation results. Section 5 presents an overview of related work. Section 6 concludes this paper.

Table 1. Comparison between different memory technologies for arrays ranging from 10KB-1MB. Latency and energy values are obtained from literature [Wong15].

		SRAM	STTRAM	RRAM	FLASH
Cell Size		Big ($120F^2$) [Kawasaki08]	**Small ($4\text{-}6F^2$)**		
Latency	Read	**Low (<10ns)**			High (>100ns)
	Write	**Low (<10ns)**	Medium (10's ns)		High (10's us)
Energy (pJ/bit)	Read	**Low (<2)**			High (>100)
	Write	**Low (<1)**	Medium (1-20)		High (>1,000)
	Leakage	High	**Low**		
Volatility		Yes	**No**		
Endurance		**High ($\gg 10^{15}$)**	Medium (10^{12}-10^{15})		Low (10^6)
Availability		**Yes**	Mostly experimental		**Yes**

2. TECHNOLOGY FOUNDATIONS

Ultra-low power bio-signal analysis platforms must satisfy multiple conflicting requirements. For example, complex applications must be executed within a sampling period of the system (inter-sample processing) with ultra-low power. Hence, such platforms must be highly flexible and process the data with an extremely low power budget. Hardware accelerators, although fast and power-efficient, are not necessarily flexible. In this section, we present technology foundations (non-volatile memories and 3D-integration) that enable ultra-low power and highly flexible WBSN architectures.

2.1 Non-Volatile Memories

Emerging non-volatile memories (NVM) such as Spin Transfer Torque RAM (STTRAM) and Resistive RAM (RRAM) satisfy the access latency and endurance requirements of the target domain [Kent15], as we show later in Section 4. Moreover, RRAM has been experimentally integrated in monolithic 3D fashion on top of both silicon CMOS as well as CNFETs [Shulaker14]. Our architecture, however, can also be used for other monolithic 3D-compatible low-voltage NVMs (e.g. Conductive Bridge RAM (CBRAM), and Phase Change RAM (PCRAM)) as well. Table 1 shows a (qualitative) comparison of various emerging memory technologies and existing volatile and non-volatile technologies. Here, energy is defined as the energy required to read or write a word (in an array including memory access circuits) divided by the number of bits in the word. Latency is defined as the amount of time required to read or write a word (in an array including memory access circuits).

Although cycling endurance (i.e., the number of times a memory cell can be written before it fails) of RRAM is significantly lower than in SRAM memories, architectural changes to the system design can still be made to achieve area and power benefits without any performance loss (discussed in Sections 4.3.1, 4.3.2 and 4.3.3). Although write latencies of STTRAM and RRAM are longer than SRAM, such long latencies are not critical for low-power systems. For example, the clock cycle of a low-power WBSN system can be 50ns (20MHz) which is longer than both the read and write latencies of STTRAM and RRAM (read 1-2ns, write 10-20ns) [Kent15]. Moreover, the low operating frequencies of such systems can be utilized to further lower the read and write energy of NVM memory technologies without device-level modifications. The required write current (or voltage) can be relaxed by increasing the write pulse width [Hosomi05, Koveshnikov12]. This relaxation is accompanied by a reduction of voltage (or current), due to the I-V characteristics of the access transistor. We follow this methodology in reducing the write energy of STTRAM and RRAM, in Section 4.1.1, based on the following relationships [Apalkov13, Park12].

STTRAM writing and reading current can be tuned to benefit from low operating frequencies, based on the relationship between writing current (I_c) and pulse width (t: time needed to change the magnetic material state) [Chun13]:

$$I_c = I_{co}\left[1 - \frac{1}{\Delta} ln\left(\frac{t}{\tau_o}\right)\right]$$

where I_{co} is the threshold write current (STTRAM-material dependent), Δ is the thermal stability factor (STTRAM-material dependent) and τ_o is the nominal switching time (~1ns). This relationship enables the tuning of write current by changing the pulse width, hence reducing the write energy of STTRAM cell. This tradeoff is used in our circuit-level characterization (Section 4.1.2). For example, for a Δ of 27, a threshold write current I_{co} of $170\mu A$, and a nominal switching time τ_o of 1ns [Park12], we can relax the write current to 150 μA if we have a pulsewidth of 25ns.

Similarly, for RRAM, we can reduce the applied write voltage (V) to relax the pulse width [Ielmini11]:

$$\tau_{set} = \frac{\Delta\phi}{A} \exp\left(\frac{E_A - \alpha q V}{kT_0 + \frac{V^2}{8\rho k_{th}}}\right)$$

where $\Delta\phi$ is the change in conductive filament required for a sufficient change of resistance, A is the filament diameter, E_A is the activation energy required to set, ρ is the electrical resistivity of the conductive filament, and k_{th} is the thermal conductivity of the conductive filament (all these parameters are material-dependent and cannot be controlled at the circuit level). T_0 is the ambient temperature. However, we find that the pulse width required, τ_{set}, increases fast with decreasing applied voltage V. Thus, the write energy increases as the applied voltage is increased since $E_{write} \propto V^2\tau_{set}$. For example, using the parameters given in [Ielmini11], a 7ns pulse at 1.4V can be reduced to a 25ns pulse at 1.19V.

2.2 3D Integration

3D integration, whereby circuits are stacked vertically over one another, offers increased connectivity between various circuit components. 3D integration often relies on Through-Silicon Via (TSV) technology. However, TSVs can occupy significantly large area footprint [Xu13] (e.g., $6.25\mu m^2$ compared to $0.5\mu m^2$ area of a 2-input NAND gate standard cell for 28nm technology). Moreover, they require large keep-out-zones where no transistors may be placed. To achieve fine-grained and dense integration, we

rely on monolithic 3D integration, whereby each vertically-stacked tier of circuits is fabricated directly over previously fabricated tiers [Batude11, Wei13]. This technology uses inter-layer vias (ILVs), vias used for interconnects in conventional ICs, to connect circuits on various tiers. The significantly smaller via size ($0.0025\mu m^2$ compared to a TSV size of $6.25\mu m^2$ for 28nm technology) and absence of keep-out-zones, allows for dense vertical connectivity.

Monolithic 3D integration requires stacked subsequent tiers of circuits to be fabricated at low temperature (<400°C) to preserve the performance of the ones already finished. Both RRAM and STTRAM can be manufactured at the required low temperature [Wong07]. However, silicon CMOS requires high-temperature fabrication (temperature exceeding 1000°C) [Rotondaro02]. Carbon nanotube field-effect transistors (CNFETs) naturally overcome this temperature barrier since all fabrication steps can be accomplished below 200°C. Systems that monolithically integrate RRAM and CNFETs (on top of silicon transistors) have already been experimentally demonstrated [Shulaker14].

In our monolithic 3D WBSN architecture, CNFETs are monolithically integrated on top of traditional silicon CMOS logic to construct the NVM access circuitry (e.g., row decoders, selection transistors). CNFETs are only used in the upper tiers to demonstrate the benefits of NVM and their 3D dense integration. Processing cores can also be realized using CNFETs, which can provide further benefits.

3. NANO-ENGINEERED ARCHITECTURE

3.1 Overall Architecture

In this section, we compare our nano-engineered architecture (Figure 4b) with the SRAM-based one introduced in [Braojos14] (Figure 4a). [Braojos14] presents a state-of-the-art design, with all computing and storage elements residing on the same tier. It consists of eight 16-bit RISC cores. Each core contains a simple three-stage pipeline, implemented using ~12KGates. The system supports efficient synchronous SIMD execution on multiple cores [Dogan12a], managed by a hardware synchronizer unit, in addition to native MIMD mode. Processors interface to separate and multi-banked program (8 banks) and data (16 banks) memories through combinational mesh-of-trees crossbars [Rahimi11]. In [Braojos14], memory banks are realized using SRAMs. To fit the instructions and data of the host applications, the sizes are 64KB (4KB per bank) and 96KB (12 KB per bank) for data (DM) and instructions (PM) memories, respectively, corresponding to 85% of the chip area.

Our target nano-engineered 3D system in Figure 4b is partitioned into a lower processing tier and an upper NVM tier, where the latter hosts the main memory (non-volatile), its address decoding logic and the access transistors. In the target system, crossbars interface with small page buffers, which collectively act as a cache for the NVM main memory. They are implemented as latch-based memories, which are more compact (in area) for small arrays (<1kb) compared to SRAM-based memories [Andersson16]. Our array design also includes an additional read/write port (with data width equal to the entire memory size) to allow for reading and writing every bit of the entire array simultaneously. This additional read/write port is connected to the NVM main memory (residing on the upper tier) by high-density vertical interconnects, enabled by 3D integration. The number of words in each buffer influences the overall power efficiency (Section 4.3.4); the lowest power configuration corresponds to only 8 words for program page buffers (24 bits per word) and 8 words for data page buffers (16 bits per word). Such an approach allows single-cycle page transfers (detailed in Section 3.2) between the page buffers and the NVM. The NVM is partitioned into 64KB of data memory and 96KB of instruction memory (consistent with Figure 4a). Finally, a combinational Memory Management Unit (MMU) monitors the read and write requests, loading and evicting the pages into/from the page buffers according to a least-recently-used (LRU) policy (more sophisticated policies can be explored as part of future work). It also interfaces with the synchronizer, so that a) cores incurring a miss when attempting to access page buffers are stalled until the corresponding data/instruction is loaded, and b) the contents registers of processor cores, data loaded into data page buffers, are transferred to the NVM before entering deep sleep mode (see Section 3.2). Note that, program page buffers do not need to be stored as they are read-only and they will be re-populated on-demand when the system wakes up.

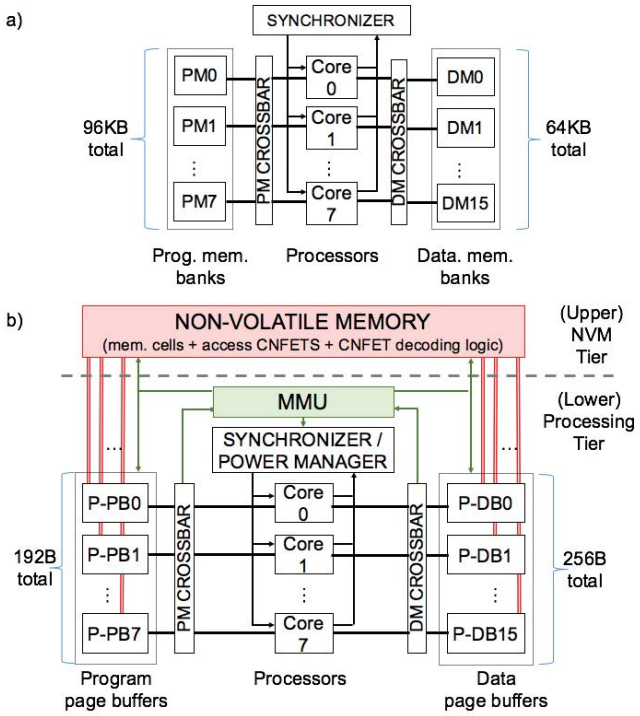

Figure 4: Block diagrams of the a) multi-core architecture in [Braojos14], featuring volatile SRAM and b) the target NVM-based platform.

3.2 Deep-sleep

When traditional volatile memories are considered [Dogan12a], the goal of the power manager is usually to minimize idle times by setting a clock frequency that allows processing in real-time (1MHz in [Braojos14]) including a marginal extra slack. However, thanks to emerging NVM, our architecture can operate at higher frequencies, thus, maximizing idle periods during which the platform is power-gated. In our case, for the target voltage (see Section 4.2.1), this frequency has been set to 20MHz for NVM-based architectures. The relatively low latency of STTRAMs and RRAMs (vs. the target operating frequency) and

the high density of connectivity due to monolithic 3D integration enable page replacements within a single clock cycle. Hence, our nano-engineered architecture supports frequent transitions between sleep and active modes, which would be difficult with traditional FLASH memory (because of its high latency).

Upon reaching an idle period, the entire architecture is power gated, waiting for new samples to be acquired. We term this state "deep-sleep." Before entering deep-sleep, a copy of the application state, namely the content of the data page buffers and the processor registers, is transferred to the NVM. At this point, the memories and the processing elements can be safely power-gated. At power up, each processor reloads the content of its own registers and execution can seamlessly resume. It is important to note that the time required to ramp up the voltage (few nanoseconds) is negligible at the considered clock frequency [Kim12].

The transition to and from the power-gated state is managed by monitoring the activity of each core (by the synchronizer) and the availability of input samples (notified by an external signal). Deep-sleep is entered when all cores have finished processing [Braojos14], while the active state is resumed when enough input samples are available.

4. SIMULATION

4.1 Explored WBSN architectures

• **2D_Baseline**: This architecture (Figure 4a) only employs on-chip SRAM and is fully implemented in 2D. The system features the synchronization mechanism described in Section 3.1 and is able to permanently power gate unused read-only program memory banks at boot time as described in [Braojos14].

• **2D_ACCESS_NVM**: This architecture (Figure 3a) integrates NVM subsystem with NVM access transistors on the same tier as other transistors.

• **3D_TSV_NVM**: In this architecture (Figure 3b), the NVM is placed on a separate tier than the processors, page buffers, synchronization circuits, MMU and crossbars. The different tiers are connected by TSVs placed at a 5 μm pitch with a keep-out zone of 5 μm, which corresponds to a 28nm technology node ([Jung14]). We found through physical design (Section 4.2.1) that TSVs still allow single-cycle access to the NVM.

• **3D_TARGET**: Our target architecture uses monolithic 3D integration (Figures 3c, 4b).

We do not consider 3D systems without on-chip NVM, as such strategy would not lead to a power-efficient implementation. Area-wise, it would also not be particularly appealing, as 85% of the area is devoted to memories in [Braojos14], which leads to a larger footprint than the studied NVM-based 3D alternatives.

4.2 Methodology

4.2.1 Circuit-level characterization

We performed full physical design (synthesis, place and route, parasitic extraction, and timing/power analysis) of the most power-hungry components (i.e., processor cores, crossbars, SRAM-based memories, and latch-based page buffers and their integration with NVM) using an industrial 28nm low-power, high-k metal gate, process design kit (PDK) (1.0V VDD) to extract area, power and performance characteristics. The synchronizer and MMU consume very low power compared to the other components and, thus, were only black-boxed for their area. The power values and latencies for the synchronizer and MMU were determined using the SystemC simulator described in Section 4.2.2. A full-chip floorplan with global routing was also performed to determine the final area of each architecture. Thus, detailed parasitics were taken into consideration. The clock frequency used in all NVM-based architectures was 20MHz while in 2D_Baseline it was 1 MHz as explained in Section 3.2. For power-gated components, high-threshold voltage power switch transistors were also inserted in the layout. The area overhead of these transistors is only 0.1% - 0.6% for the explored architectures. The system requires 30ns for the power-gating transistors to switch on and the voltage to reach target VDD. The resulting leakage power for all power-gated components (i.e. processor cores, page buffers, MMU, crossbars, synchronizers and NVM memory) combined is 1.2 nW.

For the considered NVM, we obtained device-level parameters from literature [Chun13, Koveshnikov12, Wong15]. We set the write pulse width of the NVM cells to 25ns to account for the additional overhead spent in memory access circuitry. Then, we calculate the required read and write current values based on the device-level equations in Section 2.1 to reduce the write energy from the nominal 1ns pulse. We then perform SPICE simulations by modeling the 1-transistor, 1-resistor, memory cell to deduce the voltage and transistor width required to provide the necessary write current or voltage using the 28nm PDK. These values, the transistor model, and parasitics are then linked with NVSim [Dong12] to estimate the corresponding parameters of the overall memory arrays (including memory interface circuits). Relaxing the pulse width to 25ns from 1ns provided a 25% decrease in total write energy for the entire memory.

Table 2. Parameters of various key components of the targeted WBSN[1] and the compared architectures

	Dyn. Energy (pJ/bit)		Leakage Power (uW)
Processing core	10.9 (pJ/operation)		41.37
8x12 KB PM SRAM bank	0.2 (rd)		3.53
16x4 KB DM SRAM bank	0.23 (rd)	0.27 (wr)	1.90
24B program page buffer	0.01 (rd)		6.81
16B data page buffer	0.01 (rd)	0.02 (wr)	4.65
96KB STTRAM PM	0.13 (rd)		1.05
64KB STTRAM DM	0.13 (rd)	1.3 (wr)	0.66
96KB RRAM PM	3.2 (rd)		3.46
64KB RRAM DM	3.3 (rd)	6.7 (wr)	2.31

Table 2 summarizes the power consumption of the main blocks of the target system, which uses 8 x 8-word (3 bytes per word) program page buffers (total 192 bits), 16 x 8-word (2 bytes per word) data page buffers (total 256 bits), 96 KB of program NVM and 64 KB data NVM. In addition, the 2D_Baseline architecture uses 8x12 KB program (PM) SRAM banks and 16x4 KB data (DM) SRAM banks.

4.2.2 Architecture-level framework

In order to speed up design space exploration without sacrificing accuracy, we developed a cycle-accurate SystemC simulator of the target platform defined in Section 3. The simulator embeds all the building components of the architecture (i.e. the processing

[1] These values have been obtained assuming 8-word program and data page buffers which represent the optimal configuration as it is later shown in section 4.3.1.

cores, the MMU, the page buffers the NVM and the rest of the logic). It reports detailed statistics about the run-time behavior of the architecture and all relevant events for the power estimation (e.g. power transitions, page transfers, core cycle counters, memory accesses, etc.). Power values obtained from post-place-and-route power analysis of various components (see details in Section 4.2.1) are afterwards used to annotate the simulator to compute the system-level power consumption [Dogan12b].

4.2.3 Examined applications

We used four different benchmarks widely utilized in the field of embedded electrocardiogram (ECG) processing [Rincon11, Braojos14, Mamaghanian11]. These applications process multiple ECG channels, which are bio-potential signals measured between pairs of specific locations of the body trunk that provide information about the electrical activity of the heart. Each of the employed benchmarks exploit different features of the proposed processing architecture and present diverse workload characteristics:

• **8L-CS**: Based on the algorithm in [Mamaghanian11], this benchmark efficiently compresses 8 ECG channels in parallel (one channel is processed per core) utilizing all the processors available in the platform. 8L-CS is extremely parallelized (i.e., fully SIMD), lacking any data-dependent branch. It has a moderate workload (requiring an average of more than 2,000 instructions per processed sample).

• **3L-MF**: This benchmark performs morphological filtering over 3 ECG channels employing 3 cores of the platform. Unlike 8L-CS, 3L-MF runs partially in SIMD mode and partially in MIMD mode. This application exhibits numerous conditional blocks of code making the cores diverge during part of the execution and resume the parallel mode thank to the platform capability to recover synchronized execution [Dogan12a, Braojos14].

• **3L-MMD**: In addition to a 3-channel filtering stage similar to 3L-MF, two more advanced processing routines are executed in separated additional cores to first perform signal fusion, and then delimit the ECG characteristic waves [Rincon11]. Therefore, this application employs 5 of the available cores and exploits all the synchronization benefits of the architecture (described in [Braojos14]) to recover SIMD execution and efficiently manage producer-consumer relationships among cores.

• **RP-CLASS**: This application performs selective advanced multi-channel delineation (similar to 3L-MMD) triggered by the detection of an abnormality. Two cores are used to constantly monitor a single ECG channel by first filtering and second performing heartbeat classification. When an abnormal heartbeat is identified, 4 additional cores are employed to perform the multi-lead processing of the last 2 seconds of signal with a software scheme similar to the one of 3L-MMD. Therefore, this benchmark utilizes 6 cores in total. The structure of this application presents the most complex workload profile among the considered benchmarks, with both control and data dependencies among algorithmic phases running on the different cores.

4.3 Simulation Results

We report overall power, performance (expressed in terms of Instructions Per Cycle, IPC), and footprint area results in Table 3. The values reported are arithmetic averages over the applications, for the best page buffer configuration. All our architectures consume the least power with the following page buffer configuration: 8-words x 2 Bytes/word x 16 data page buffers, 8-words x 3 Bytes/word x 8 program page buffers, as shown in Section 4.3.5. Table 3 shows that 3D_TARGET, with STTRAM monolithically integrated on top of the computing units (designed using 28 nm technology), simultaneously achieves 4.4x power and 4x footprint benefits over 2D_Baseline, with only 2% reduction in IPC. Despite the small IPC drop (details in Section 4.3.2) 3D_TARGET continues to meet application-level real-time processing constraints.

Table 3. Power, performance, and footprint of various WBSN architectures, averaged for all target applications. Best values are highlighted in bold.

	Memory	Power (µW)	Performance (IPC)	Footprint (mm^2)
2D_Baseline	SRAM	543	**3.40**	0.374
2D_ACCESS_NVM	STTRAM	128	3.34	0.456
	RRAM	316	3.34	0.453
3D_TSV_NVM	STTRAM	124	3.34	0.232
	RRAM	287	3.34	0.232
3D_TARGET	STTRAM	**123**	3.34	**0.092**
	RRAM	285	3.34	**0.092**

4.3.1 Power

Figure 5 shows the power consumption for each of the studied benchmarks and its corresponding breakdown into three main parts: dynamic power for memory (power consumed in the page buffers described in Section 3.1 and non-volatile memory), dynamic power for computing (processing cores, synchronization logic, crossbars, memory management unit, and interconnect), and leakage power for the entire architecture (similar to the breakdown in Figure 1). We only show the results for 3D_TARGET since 3D_TSV consumes 1% more power (as shown in Table 3 as well). All power results for NVM configurations are for the case of STTRAM corresponding to lowest power (Table 3).

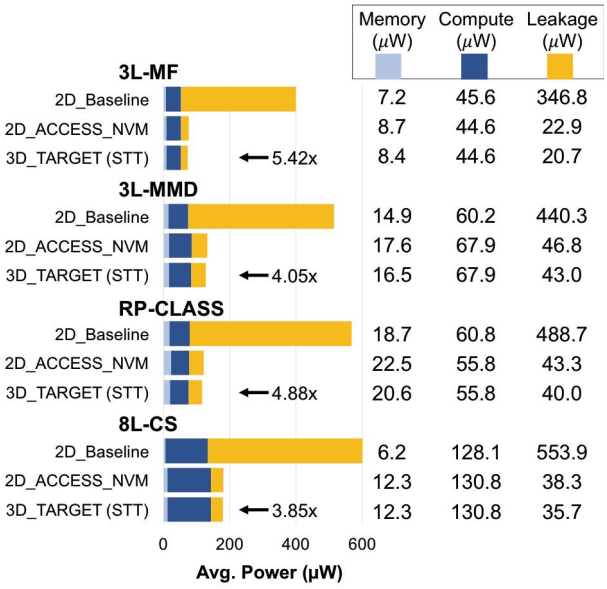

Figure 5. Power consumption of WBSN architectures.

The power consumption of our WBSN architectures improves (i.e., decreases) significantly with respect to 2D baseline, thanks to deep-sleep enabled by NVM. In particular, compared to 2D_Baseline, our 3D_TARGET architecture achieves up to 5.42x power savings (3D_TARGET).

In Figure 6, we show a detailed power breakdown of various components in 3D_TARGET. The figure shows further power improvement opportunities. For example, low-energy logic transistors (e.g., CNFETs) can be used to reduce the power of the processing cores.

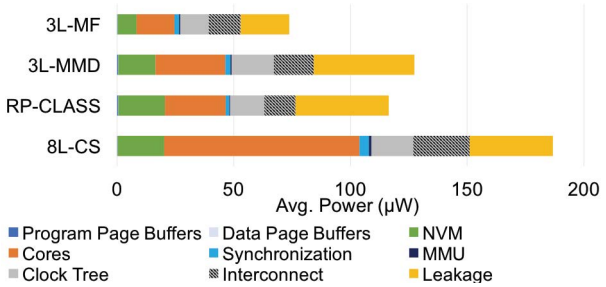

Figure 6. Power consumption breakdown for 3D_TARGET architecture.

4.3.2 Performance

We analyze the performance of our WBSN system by answering the following key questions: *1) does the new nano-engineered architecture meet real-time constraints? 2) does the introduced 2-level memory subsystem enhance run-time performance?; 3) does the page transfer policy required for deep-sleep introduce significant overhead?*

Table 4. Average execution time of the longest (software) pipeline stage for 2D_Baseline (per sample). For 500Hz input signal, execution time of each pipeline stage must be less than or equal to 2ms to meet real-time constraints (cf [Braojos14]).

	3L-MF	3L-MMD	RP-CLASS	8L-CS
2D_Baseline (ms)	1.58	1.65	1.80	1.99

Table 5. Average execution time for the 3D_TARGET architecture (per sample). For other considered architectures (2D_ACCESS_NVM and 3D_TSV) the time is similar. For 500Hz input signal, execution time of the complete system must be less than or equal to 2ms to meet real-time constraints (cf [Braojos14]).

	3L-MF	3L-MMD	RP-CLASS	8L-CS
3D_TARGET (ms)	0.08	0.15	0.13	0.11

We report application-level Instructions Per Cycle (IPC, Figure 7) and average execution time (Tables 4 and 5) to quantify performance. In the case of 2D_Baseline (running at 1 MHz), it uses the software pipelining technique described in [Braojos14], and Table 4 reports the execution time of the longest (software) pipeline stage for various applications. 3L-MMD and RPCLASS are composed of tasks organized in software pipelines (details in [Braojos14]). For 3D_TARGET (running at 20MHz), Table 5 reports the total execution time per sample and per application (i.e. full pipeline).

In both cases, for all the target applications in which input signals are sampled at 500 Hz, real-time constraints are met, since the average execution time per sample is less than 2 ms (per pipeline stage in the case of 2D_Baseline or per application in the case of 3D_TARGET). As these tables show, 3D_TARGET obtains a speed-up of the execution time per sample ranging from 11x to 19.75x, with respect to 2D_Baseline, allowing for longer idle periods.

Then, Figure 7 shows that the IPC of SIMD benchmarks increases up to 12% (for 8L-CS case). This improvement is enabled by its adopted memory subsystem. In fact, the 2D_Baseline architecture faces performance bottlenecks when SIMD execution experiences de-synchronization (when different processors execute different paths in a conditional branch). During these situations, cores often require different instructions or data from the same program/data memory bank. These accesses are then serialized, introducing performance penalties. The 2-level memory subsystem in 3D_TARGET minimizes such conflicts, because instructions and data are loaded into page buffers (D-PBs, P-PBs in Figure 4b), which are smaller than banks of 2D_Baseline (e.g., 24B PBs vs. 12KB PM bank) and serialization is reduced. Cores accessing different code regions will request disjoint pages that will be loaded by the MMU into different buffers (leading to conflict-free execution).

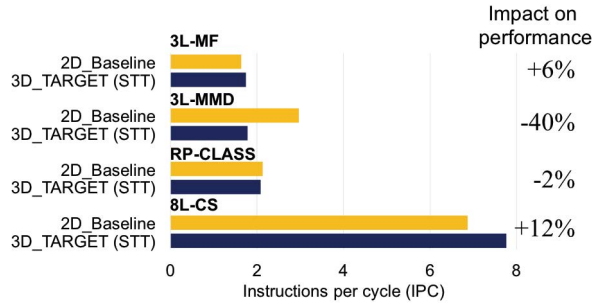

Figure 7. Instructions Per Cycle (IPC) for each application for 2D_Baseline and 3D_TARGET platforms. Other considered architectures (2D_ACCESS_NVM and 3D_TSV) have similar IPC behavior as 3D_TARGET. The theoretical maximum IPC for the target platform is 8.

However, we also observe degradation of IPC of -40% (3L-MMD) and -2% (RPCLASS). The reduction of IPC is due to the higher operating frequency of 3D_TARGET, which reduces the processing time of each task significantly. Therefore, on 3D_TARGET, the processing of each input sample (acquired every 2ms) by all cores is accomplished before the arrival of the subsequent one, reducing the possibility of pipelining code execution. The parallelism obtained by performing SIMD execution of code continues to be exploited by 3D_TARGET as well.

To analyze the impact of page transfer, we provide (in Table 6) a detailed breakdown of the transfer overhead, compared to processing time for 3D_TARGET (other NVM-based architectures show similar behavior). Table 6 shows that the total active time (both processing and data transfer) accounts for less than 9% of the inter-sample arrival time, for all the evaluated benchmarks allowing for long periods of deep-sleep. The time spent in page transfer is not dominant thanks to the low-latency 2-level memory subsystem (and also because our latch-based array

designs can transfer all contents to and from the NVM simultaneously).

Table 6. Runtime metrics of the analyzed benchmarks for 3D_TARGET. 2D_ACCESS_NVM and 3D_TSV have similar behaviors.

	3L-MF	3L-MMD	RP-CLASS	8L-CS
Platform Active time (%)	**4.67**	**8.16**	**7.04**	**5.49**
- Processing (%)	*4.41*	*7.81*	*6.63*	*5.36*
- Page transfer (%)	*0.26*	*0.35*	*0.41*	*0.13*
NVM → Page buffer (avg. MB/s)	51.41	79.34	137.36	44.60
Page buffer → NVM (avg. MB/s)	30.96	28.60	32.57	18.25

4.3.3 Area

As shown in Figure 8, 2D_ACCESS_NVM has the largest footprint (0.4568 mm^2). Having the access transistors for the NVM (STTRAM or RRAM) on the same tier as the processing units and the page buffers (as in 2D_ACCESS_NVM) creates significant routing congestion.

3D integration significantly reduces this congestion. For 3D_TSV, the space dedicated to TSVs is considerable, and the area of the NVM tier dominates. The monolithic 3D approach in 3D_TARGET architecture provides the most compact design. 3D_TARGET achieves 5x, 4x and 2.5x footprint area savings when compared to 2D_ACCESS_NVM, 2D_Baseline and 3D_TSV, respectively.

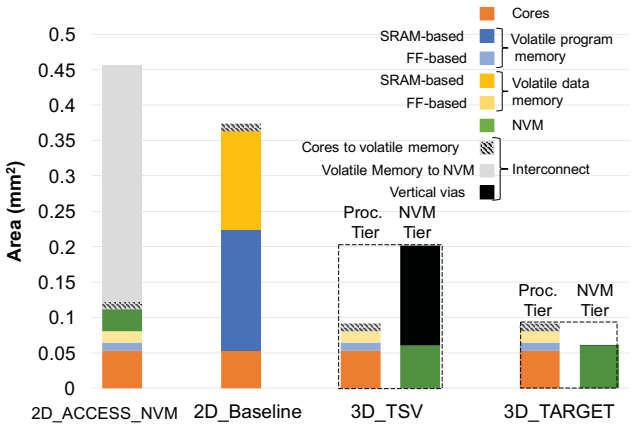

Figure 8. Area breakdown for the studied WBSN architectures

4.3.4 Architecture trade-off analysis: page buffers

We study the impact of program and data page buffers (Figure 4b) sizes on system power consumption. Figure 9 shows the power consumption of 3D_TARGET by sweeping both program and data page buffer (P-PB and D-PB respectively) sizes for the benchmark applications. Data and program buffer sizes are varied from 8 (minimum allowed by the designed MMU) to 512 words (for larger sizes we observed an increased consumption dominated by PB leakage power), with 2 bytes/word for data and 3 bytes/word for program buffers.

As observed in Figure 9, variants employing page buffers with fewer words consume less power. WBSN applications often exhibit inherent code locality through a series of compact (few instruction count) and iterative kernels [Braojos14] (loops to process a stream of input data), which manipulate small arrays of samples (usually circular buffers with low memory footprint). This type of code benefits from the latch-based memory buffers in our target architecture, as the code requires small storage for each kernel. With the system processing time clearly dominating the platform activity over the page transfer time (see Table 6), using a small-sized buffer reduces the overall power of the memory subsystem. Among various options, the best configuration uses 8-word banks for both instruction and data page buffers.

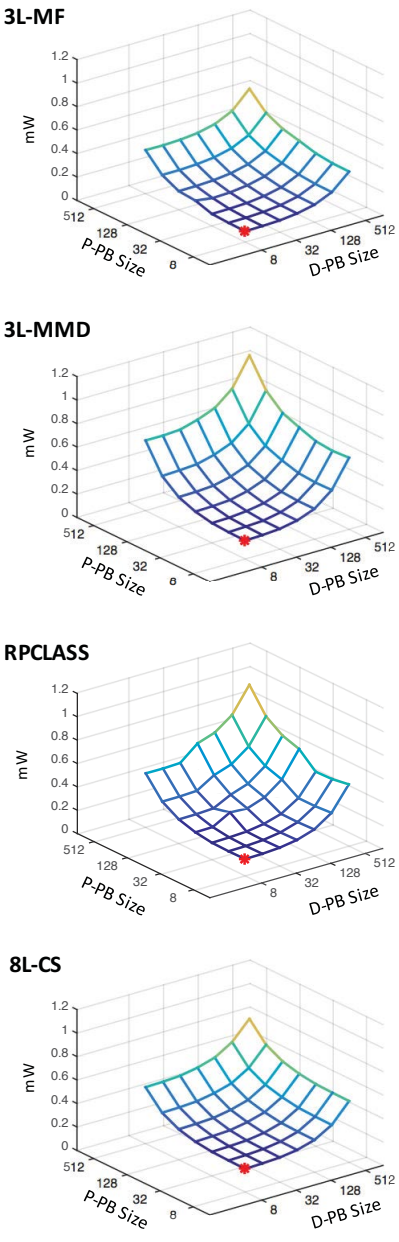

Figure 9. Average power consumption of 3D_TARGET employing different program page buffer (P-PB) and data page buffer (D-PB) sizes (in number of words). Local minima are marked with a red dot.

4.3.5 RRAM/STTRAM trade-off analysis

We present the impact of STTRAM and RRAM technologies on power, footprint, and reliability (in terms of endurance) of WBSN applications executed on our target architecture. We do not analyze the impact on runtime, since both memory technologies can achieve a single cycle access (read or write) for the targeted 20MHz operating frequency.

STTRAM arrays provide 4x lower power compared to RRAM arrays of the same size. However, by relaxing the write pulse width (Section 2.1), the array-level power benefits of STTRAM vs. RRAM increase to 5x (Table 2). At the application level, our results indicate that the STTRAM-based architecture consumes 2x less power than the RRAM-based architecture.

Footprint-wise, an RRAM cell is 33% smaller than an STTRAM cell (Table 1). However, for the particular NVM capacity (64 KB) of our architecture, we find that 70% of the memory area is used by access circuitry. Thus, at the system level, the footprint of the STTRAM-based architecture is comparable to the RRAM-based architecture.

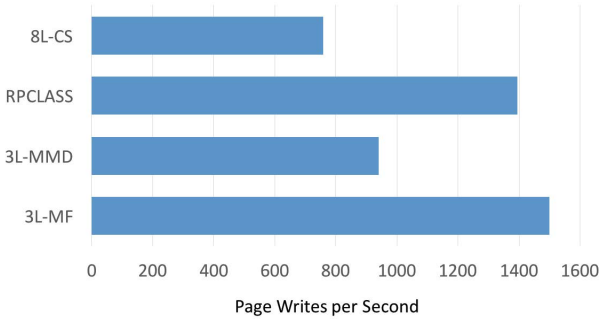

Figure 10. Number of page writes per second to NVM (STTRAM or RRAM) for the studied benchmarks. The values correspond to the page with the maximum number of writes

STTRAM has a better write endurance (~10^{15} writes) than RRAM (up to 10^{12}), which makes STTRAM favorable. However, due to relatively few writes in the targeted applications, high endurance may not be needed. Figure 10 shows the number of page write operations to the NVM per second for each application, indicating a maximum rate of 1500 writes/second of the same page. At this maximum rate, RRAM-based system is able to achieve a lifetime of >20 years, even with a write endurance of 10^{12}.

5. PRIOR WORK

The applicability of WBSNs has been investigated in a variety of scenarios [Hao08], including the automated analysis of ECGs [Rincon11], the estimation of the respiration rhythm [Berset12] and the detection of epileptic seizures [Masse13]. Recently, dedicated architectures have been published to support these workloads at ultra-low power levels. The authors of [Kwong11] and [Sridhara11] advocate the use of custom accelerators (such as FFT and Cordic engines) to efficiently support commonly-used kernels. This approach has limited flexibility, as it assumes the knowledge, at design time, of the computationally-intensive segments of applications. A different solution, illustrated in [Seok08], is to aggressively scale the supply voltage to decrease both static and dynamic power. Multicore architectures [Braojos14, He10] are good candidates for this strategy; they can distribute workload over a plurality of computing elements, with each of them operating at a low frequency (e.g., in a near-threshold regime). However, traditional SRAMs create a lower bound on the operating voltage, dictated by the minimum level at which data can be reliably stored and accessed [Bortolotti14].

[Bortollotti15] is another recent effort considering low-power NVMs integrated in WBSNs. The authors employ NVMs to implement temporary buffers for input samples. Other related efforts include [Ransford11] and [Jayakumar14], which utilize NVMs to create checkpoints in transiently-powered systems such as RFID implantable devices. The granularity of such checkpoints is much coarser with respect to the one considered in this paper.

6. CONCLUSION

The ever-growing demand for convenient wearable health monitoring devices is a major drive to improving power consumption and footprint of WBSN platforms. As demonstrated in this paper, new nano-engineered WBSN architectures are key to achieving major power and footprint area benefits. Such architectures utilize nanotechnology advances, such as, (a) emerging non-volatile memories and (b) highly dense and fine-grained three-dimensional integration (e.g., monolithic three-dimensional integration naturally enabled by carbon nanotube field-effect transistors). The overall application-level benefits of our new nano-engineered WBSN architectures, compared to state-of-the-art, are dramatic: up to 5.42x power and 5x footprint area improvements, while meeting real-time processing requirements of essential health monitoring applications.

Future research directions include: (1) New nano-engineered architectures for WBSNs targeting a wider variety of health-monitoring applications including electroencephalogram (EEG); (2) The use of emerging nanotechnologies (e.g., CNFETs) for efficient processing of bio-signals (e.g., using CNFETs for processor cores as well); (3) Real-time embedded data fusion from a large number of input channels (e.g., for EEG) or embedded ultra-low-power multi-modal bio-signal analysis (e.g., interpreting ECG, EMG or accelerometer data to derive global knowledge of the state of a person) enabled by new nano-engineered WBSN architectures.

7. ACKNOWLEDGMENT

This work has been partially supported by the BodyPoweredSenSE (no. 20NA21_143069) and E4Bio (no. 200021_159853) RTD projects evaluated by the Swiss NSF. It is also supported in part by DARPA, National Science Foundation, STARNet SONIC (one of the six SRC STARnet Centers sponsored by MARCO and DARPA), Swiss NSF Early Postdoc. Mobility Fellowship (no. 151965) for M. S. Aly, and the Stanford SystemX Alliance.

8. REFERENCES

[Andersson16] O. Andersson et al., "Ultra Low Voltage Synthesizable Memories: A Trade-Off Discussion in 65 nm CMOS," in IEEE TCS, no.99, pp.1-12, 2016.

[Apalkov13] D. Apalkov et al, "Spin-transfer torque magnetic random access memory (STT-MRAM)," ACM JETC, 2013.

[Batude11] P. Batude et al, "Advances, Challenges and Opportunities in 3D CMOS Sequential Integration," In Proceedings of IEDM, 2011.

[Bazaka12] K. Bazaka et al, “Implantable devices: issues and challenges.” MDPI Electronics 2.1 (2012): 1-34.

[Berset12] T. Berset *et al.*, “Robust Heart Rhythm Calculation and Respiration Rate Estimation in Ambulatory ECG Monitoring,” *BHI*, pp. 400–403, 2012.

[Bortolotti14] D. Bortolotti et al., *“*Hybrid Memory Architecture for Voltage Scaling in Ultra-low Power Multi-core Biomedical Processors*”*, In Proc. DATE, 2014.

[Bortolotti15] D. Bortolloti et al., “Long-Term ECG Monitoring with Zeroing Compressed Sensing Approach”, In Proc. NORCAS, 2015

[Braojos14] R. Braojos et al. “Hardware/software approach for code synchronization in low-power multi-core sensor nodes” In Proc. DATE, pp. 1-6, 24-28, 2014.

[Dogan12a] A. Dogan et al. “Multi-Core Architecture Design for Ultra-Low-Power Wearable Health Monitoring Systems” In Proc. DATE, pp. 988-993, 2012.

[Dogan12b] A. Dogan, et al. "Low-power processor architecture exploration for online biomedical signal analysis," in IET Circuits, Devices & Systems, vol. 6, no. 5, pp. 279-286, Sept. 2012.

[Chun13] K. C. Chun et al., “A Scaling Roadmap and Performance Evaluation of In-Plane and Perpendicular MTJ Based STT-MRAMs for High-Density Cache Memory”, in IEEE JSSC, vol. 48(2), pp. 598-610, 2013

[Hao08] Y. Hao et al., “Wireless Body Sensor Networks for Health-Monitoring Applications,” Physiological Measur., vol. 29, no. 11, p. R27, 2008.

[He10] Y. He et al., “Xetal-Pro: an Ultra-Low Energy and High Throughput SIMD Processor,” DAC, pp. 543–548, 2010.

[Hosomi05] M. Hosomi et al., "A novel nonvolatile memory with spin torque transfer magnetization switching: spin-ram," IEEE IEDM, Washington, DC, 2005.

[Ielmini11] D. Ielmini, “Modeling the Universal Set/Reset Characteristics of Bipolar RRAM by Field- and Temperature-Driven Filament Growth”, in IEEE TED, Vol. 58(12), 2011.

[Jayakumar14] H. Jayakumar et al. “QickRecall: A Low Overhead HW/SW Approach for enabling Computations across Power Cycles in Transiently-Powered Computers”, In Proc. VLSID and ICES, 2014.

[Jung14] M. Jung et al., “TSV Stress-Aware Full-Chip Mechanical Reliability Analysis and Optimization for 3D IC”, In Communications of the ACM, Vol. 57(1), pp. 107-115, 2014

[Kawasaki08] H. Kawasaki et al., "Demonstration of highly scaled FinFET SRAM cells with high-κ/metal gate and investigation of characteristic variability for the 32 nm node and beyond," IEEE International Electron Devices Meeting, San Francisco, CA, 2008.

[Kent15] A. D. Kent and D. C. Worledge, A new spin on Magnetic Memories. Nature Nanotechnology, vol. 10, pp. 187-191, 2015.

[Kim12] W. Kim, et al., “A fully-integrated 3-level dc- dc converter for nanosecond-scale dvfs,” Solid-State Circuits, IEEE Journal of, vol. 47, no. 1, pp. 206–219, 2012.

[Koveshnikov12] S. Koveshnikov et al., "Real-time study of switching kinetics in integrated 1T/ HfOx 1R RRAM: Intrinsic tunability of set/reset voltage and trade-off with switching time," IEDM, San Francisco, CA, 2012.

[Kwong11] J. Kwong *et al.*, “An Energy-Efficient Biomedical Signal Processing Platform,” *Solid-State Circuits*, vol. 46, no. 7, pp. 1742–1753, 2011.

[Mamaghanian11] H. Mamaghanian et al. “Compressed Sensing for Real-Time Energy-Efficient ECG Compression on Wireless Body Sensor Nodes”, In IEEE Trans. on Biomedical Engineering vol 58, no.9, pp.2456-2466, 2011

[Masse13] F. Massé *et al.*, “Miniaturized Wireless ECG Monitor for Real-Time Detection of Epileptic Seizures,” *ACM TECS*, vol. 12, no. 4, pp. 102:1– 102:21, 2013.

[Nakashima15] M. Nakashima, “High Performance and Highly Reliable SSD -Proposal of the Fastest Storage with B4-Flash”, in Flash Memory Summit 2015

[Mitani16] H. Mitani et al., “A 90nm Embedded 1T-MONOS Flash Macro for Automotive Applications with 0.07mJ/8kB Rewrite Energy and Endurance Over 100M Cycles Under Tj of 175°C”, in ISSCC 2016

[Nigam11] A. Nigam et al. Delivering on the Promise of Universal Memory for Spin-Transfer Torque RAM (STT-RAM). In Proc. ISLPED, pp. 121-126, 2011.

[Park12] J.-H. Park et al., “Enhancement of data retention and write current scaling for sub-20nm STT-MRAM by utilizing dual interfaces for perpendicular magnetic anisotropy,” In Proceedings of VLSIT, 2012.

[Rahimi11] A. Rahimi et al., “A fully-synthesizable single-cycle interconnection network for Shared-L1 processor clusters,” In Proceedings of DATE, 2011.

[Ransford11] B. Ransford et al., “Mementos: System Support for Long-Running Computations on RFID-Scale Devices,” In Proc. ASPLOS, 2011.

[Rincon11] F. Rincon *et al.*, “Development and Evaluation of Multilead Wavelet-Based ECG Delineation Algorithms for Embedded Wireless Sensor Nodes,” *Info. Tech. in Biomedicine*, vol.15, no.6, pp. 854–863, 2011.

[Rotondaro02] A. Rotondaro et al., “Advanced CMOS transistors with a novel HfSiON gate dielectric,” VLSI Tech, pp. 148-189, 2002.

[Sampaio14] F. Sampaio et al., “Energy-Efficient Architecture for Advanced Video Memory”, In Proc. ICCAD, 2014

[Seok08] M. Seok *et al.*, “The Phoenix Processor: A 30pW Platform for Sensor Applications,” *VLSI Circuits*, pp. 188–189, 2008.

[Shulaker14] M. M. Shulaker et al., “Monolithic 3D integration of logic and memory: carbon nanotube FETs, resistive RAM, and silicon FETs,” IEDM, 2014.

[Shulaker15] M. Shulaker et al. Monolithic 3D Integration: A path from Concept to Reality. In Proceedings of DATE, 2015.

[Sridhara11] S. Sridhara *et al.*,“Microwatt Embedded Processor Platform for Medical System-on-Chip Applications,” *Solid-State Circuits*, vol. 46, no. 4, pp. 721–730, 2011.

[Taito15] Y. Taito et al., "7.3 A 28nm embedded SG-MONOS flash macro for automotive achieving 200MHz read operation and 2.0MB/S write throughput at Ti, of 170°C," In Proc. ISSCC, San Francisco, CA, 2015.

[TI-CC2540] Online http://www.ti.com/lit/ds/symlink/cc2540.pdf

[Wei13] H. Wei et al., “Monolithic Three-Dimensional Integration of Carbon Nanotube FET Complementary Logic Circuits,” IEDM, pp. 511-514, 2013.

[WHO15] World Health Organization., “Cardiovascular diseases,” 2015. [Online]. Available: www.who.int/topics/cardiovascular_diseases/en/

[Wong07] S. Wong et al., “Monolithic 3D Integrated Circuits”, VLSI-TSA, pp. 1-4, 2007.

[Wong15] H.-S. P. Wong, C. Ahn, J. Cao, H.-Y. Chen, S. W. Fong, Z. Jiang, C. Neumann, S. Qin, J. Sohn, Y. Wu, S. Yu,and X. Zheng, “Stanford Memory Trends,” https://nano.stanford.edu/stanford-memory-trends, accessed November 20, 2015.

[Xu13] Z. Xu and J.-Q. Lu, “Through-silicon-via Fabrication Technologies, Passive Extraction, and Electrical Modeling for 3-D Integration/ Packaging,” IEEE TSM, vol. 26(1), pp. 23-34, 2013.

[Zhang12] F. Zhang *et al.*, “Design of ultra-low power biopotential amplifiers for biosignal acquisition applications.” *Biomedical Circuits and Systems*, vol 6, no. 4, pp 344-355, 2012.