Ultralow-Power and Robust Embedded Memory for Bioimplantable Microsystems

Maryam S. Hashemian and Swarup Bhunia
Department of Electrical Engineering and Computer Science
Case Western Reserve University
Cleveland, OH, USA
Email: {mxh460 and skb21}@case.edu

***Abstract*—Bioimplantable microsystems, such as pacemaker and cochlear implant, interface with internal body parts to monitor and/or control their activity. These systems typically record biological signals; analyze them in real time; and then transmit them to outside world or take appropriate corrective action. They require ultralow-power miniaturized electronics for long-term reliable operation using on-board battery. Embedded memory used to temporarily store the recorded data, forms an integral and important part of these systems. In this paper, we explore the design space and propose an optimal design of embedded memory for implantable applications. First, we compare a conventional super-threshold implementation of memory with a sub-threshold design with respect to energy efficiency. Next, we propose a super-threshold static random access memory (SRAM) design operating at a frequency much higher than the sampling frequency. We show that it can achieve very low energy dissipation by taking advantage of extensive power gating. Moreover, compared to a sub-threshold memory, it provides significantly better area and higher robustness of operation, both of which are important requirements for implantable systems. As a case study, we consider a neural control system that records and analyzes neural spikes. Simulation results for 45nm CMOS process using pre-recorded neural data from sea-slug (Aplysia californica) show that the proposed design can lead to significant energy reduction, without compromising the robustness and performance, compared to its sub-threshold counterparts.**

I. Introduction

Miniaturized implantable systems provide an important interface to internal body parts for interpreting and engineering their activity [1]. Fig. 1(a) shows the interface of a common implantable system, namely a neural control system, with a micro-electrode array, analog signal conditioning, and transceiver electronics. The neural data recorded by the electrodes are conditioned using analog circuits and converted into digital signals, which go into the neural signal-processing block shown in Fig. 1(b). It analyzes the digitized data, compresses it, and extracts meaningful neural patterns. Finally, it sends out recorded signals to an outside receiver through a transmitter or provides stimulation in a closed loop framework [2]. With an increase in the number of the recording electrodes, the transmission bandwidth of the implanted telemetry device becomes insufficient and power-hungry. For example, neural recording from an array of 100 electrodes sampled at 25 KHz per channel with 10-bit precision yields an aggregate data of 25Mbps, which is beyond the state-of-the-art wireless telemetry. Therefore, it is extremely important to use on-chip electronics for data compression [3]. Furthermore, signal processing system is also responsible for recognition of meaningful patterns in order to trigger appropriate corrective actions through stimulation. To perform the data analysis from multiple recording electrodes, there must be an on-chip data storage to store the recorded signals and the detected events. An effective approach for pattern recognition is based on matching events detected from incoming signals with a "vocabulary" [3] of signatures. Such a vocabulary can be created dynamically by checking for new events and storing them in it. In general, implantable systems used in diverse applications would require several 100 kilo-bytes of memory for temporary data storage. The memory access, leakage power, and area largely affect the energy, reliability, and size of these systems. Hence, it is of critical importance to achieve unltralow power, robust, and area-efficient design of this memory.

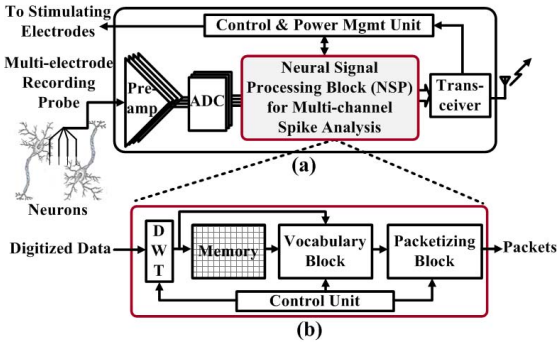

Fig. 1. (a) High-level functional block diagram for a typical neural interface system. (b) The digital-signal-processing block, which analyzes spike patterns on multichannel recorded data, constitutes a key part of the system [2].

In this paper, we explore the design choices for embedded memory in implantable systems and propose an efficient memory design for these systems exploiting the nature of the recorded biological signals. As a case study, we consider a neural recording framework that monitors the neuronal action potentials or spikes. Outputs of multiple sensors go to a signal processing hardware, which performs spike detection, wavelet analysis based classification, and vocabulary construction. Finally, the recorded signals are encoded in terms of alphabets in the vocabulary and transmitted wirelessly to the outside world using the built-in telemetry device [3]. In the following

1063-9667 2013

DOI 10.1109/VLSID.2013.164

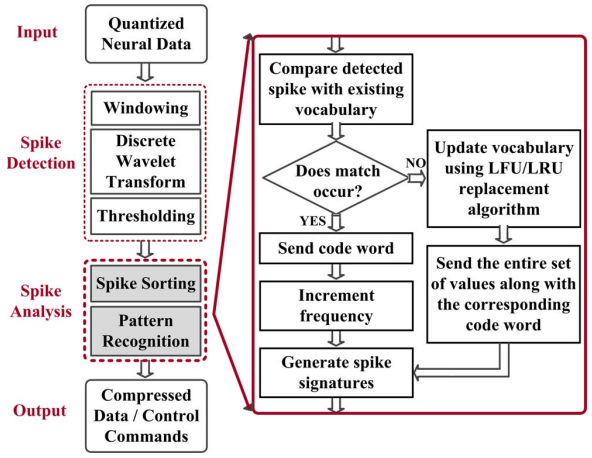

Fig. 2. Flow diagram of the vocabulary based neural signal processing algorithm.

sections, we compare the energy efficiency and reliability of super- and sub-threshold designs. We show that a super-threshold design with appropriate choice of supply voltage, device size, and threshold and also application of opportunistic power gating can dramatically reduce the power dissipation. In order to increase the opportunity for power gating in super-threshold design, we increase the operating frequency, than the one dictated by signal acquisition frequency. We observe that at increased frequency, we can achieve comparable energy behavior while maintaining the area and robustness advantages, compared to a sub-threshold design.

II. Background and Motivation

The steps of vocabulary-based neural data compression are shown in the flow diagram in Fig. 2. It takes the quantized recorded neural data as input and gives the compressed neural data as output. The encoded output is in form of packets containing some wavelet information about the detected spikes. At first the digital neural data is broken into fixed size overlapping windows. Next wavelet analysis is done for each window creating an 'approximation' and 'detail' coefficients. Then spikes are identified using a thresholding step and are matched with the existing set of spikes (referred as 'alphabets' [3]) in the database (referred as 'vocabulary' [3]). Finally, output packets are sent out reporting the location of the spike as well as the corresponding alphabet [3]. The vocabulary process contains two tasks: 1) matching the newly detected spike with an existing set of spikes in the vocabulary; and 2) updating the information in the vocabulary. In order to minimize the power requirement, the new spike is compared in parallel with all the old spikes stored in the vocabulary. The match is determined by simply comparing the magnitudes with equality checker hardware. If a match is found, then the memory address in which the spike is stored is sent out as the location of the spike.

To design the on-chip memory, the design parameters of interest are: area, power, and robustness. Area is important to ensure small form factor of the implant unit. Low power dissipation is important to enable long-term operation using the on-board battery and to prevent tissue damage due to heat generation. Further, we need to make it robust against run-time failures to ensure reliable long-term operation. Considering these design parameters, we propose a design choice for on-chip memory that achieves ultra low-power, area efficient, and robust implementation using nanoscale devices.

TABLE I
ALTERNATIVE MEMORY ARCHITECTURES FOR DATA STORAGE IN IMPLANTABLE SYSTEMS

	Array of FFs	Register File	Embedded SRAM
Merits	Very Fast Ease of Design	Very Fast	Fast Small Area Scalable
Demerits	Large Area Not Scalable	Power Hungry* Less Robust* More Design Effort**	More Design Effort**

* Compared to SRAM
** Compared to FFs

III. Design Space Exploration

A. Alternative Memory Architectures

Several alternative architectures can be used to implement memories: a) array of flip flops, b) register files, and c) embedded SRAMs. Table I, highlights the merits and demerits of each of these memory architectures in terms of area, scalability, power, performance, robustness, and price. Array of flip flops is not a good choice for an on-chip memory, since it is not area-efficient and hence not suitable for scaling to large storage size. Register file is not a good choice either as it is normally used when multiple read and/or write operations needs to be done at the same time. Considering the low recording rate and sampling frequency in neural interface systems, there is plenty of time, and there is no need to do multiple simultaneous reads and/or writes. Even single-port register file cannot be a good choice for on-chip memory as compared to embedded SRAM, since they use large-signal sensing which increases access power due to large voltage swing in bitlines, compared to SRAM arrays that typically use small-signal sensing. SRAMs are very area-efficient; easily scalable in size and typically consume less power than register files for the target size. Therefore, SRAMs can be the best choice of architecture for on-chip memories in implantable applications. The conventional super-threshold 6T cells are typically used for designing SRAM arrays. However, when minimum power consumption is the primary requirement and the operating frequency is low (in 100KHz), sub-threshold memory designs are very attractive. In sub-threshold design, in order to achieve ultralow-power operation, supply voltage is reduced to below the device threshold voltage. This significantly reduces power at the cost of increased delay. However, it typically suffers from huge area overhead and reduced reliability compared to a super-threshold design. In the following sections we will evaluate the sub-threshold and conventional super-threshold designs, and propose techniques to reduce power in super-threshold memory.

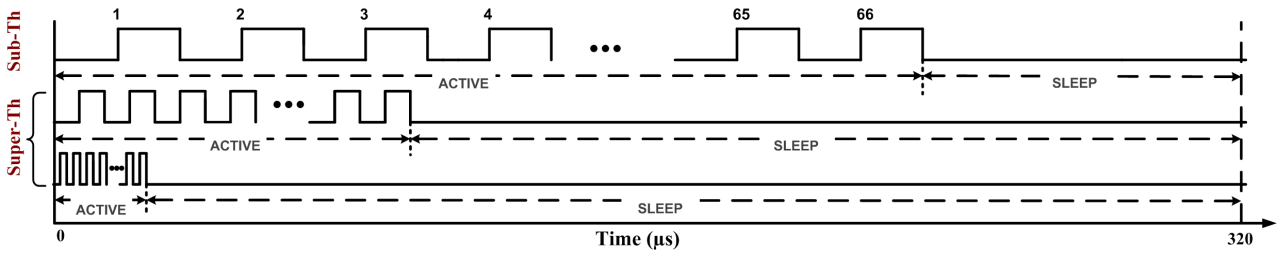

Fig. 3. Timing diagram comparing execution of 66 cycles of read, compare, and update in sub-threshold and super-threshold modes.

TABLE II
AREA, DELAY, ENERGY, AND NOISE-MARGIN FOR A 4X4 SRAM ARRAY IMPLEMENTED WITH 8T AND 10T SINGLE-ENDED CELLS

Cases	Area	Delay	Energy (zJ)			Noise Margin (mv)		
	μm^2	(ns)	dyn	leak	tot	Read	Write	Hold
8T	1.12	128.2	0.005	8.691	8.696	156	122	156
10T	1.44	200.4	0.02	6.576	6.596	156	122	156

B. Sub-threshold Memory

Conventional 6T SRAM do not function reliably in the sub-threshold regime because the ratio constraints for read stability and writability cannot be guaranteed. Therefore, several alternative configuration SRAMs are considered for sub-threshold operations [4] [5] [6] [7]. In this paper we have focused on single-ended 8T [7] and 10T [6]. With area and power as two important design parameters, we chose single-ended 8T and 10T because the former is comparatively smaller in area, and the latter reduces leakage power significantly due to stacking effect. The single-ended scheme in both cases is used for improved stability, and it comes at the cost of a read bitline, a read wordline, and two to four read transistors. 8T and 10T SRAM cells have very similar architecture. However, each one has some advantages over the other one. 8T is denser, while 10T can reduce leakage power more due to the two extra transistors in the read port. Although 8T and 10T can operate at a lower voltage than 6T, but they are not as dense as 6T.

In Table II, two 4x4 arrays based on 8T and 10T SRAM cells are compared in terms of area, performance, energy, and robustness at V_{DD} of 400 mV. It has to be mentioned that the delay measurement is based on a read operation, and the total energy is based on the dynamic energy and leakage energy dissipated during and after a read operation in a 100μs time interval. Both 8T and 10T SRAM cells are designed using high threshold transistors to reduce leakage [8]. They, however, suffer from reduced reliability. At very low supply voltages, they may experience functional failures due to increased delays caused by environmental variations. Knowing that area is a critical feature for implantable devices, for further simulations, we chose 8T as the nominal SRAM cell for sub-threshold design.

C. Super-threshold Memory

For super-threshold SRAM design we choose the conventional 6T SRAM cell. The cell is designed using nearly minimum-sized transistors to achieve high density. To reduce the leakage, we use high threshold transistors in the cell. The ratio between the pull down NMOS and the access transistor is sized to be grater than 1.2 to keep a proper noise margin during the read operation. Although the conventional 6T SRAM cell does not work reliably at an ultra-low power supply, but it is very dense, and provides better performance and robustness of operation. As demonstrated in the timing diagram in Fig. 3, the operations can be performed at very low voltage and frequency in sub-threshold mode. In the other hand, they can also be completed in a shorter time by operating at higher frequencies in super-threshold mode, which leaves longer idle time for more supply gating [2].

D. Power-Gating of Memory

Most of the subarrays in a large SRAM are not active at any given time. This inactive periods are even more observable in biomedical applications where the sampling rate is very low (in hundreds of KHz), and the system is idle most of the time. Supply-gating is a technique that can be used to decrease static power consumption. During a memory access, it turns on only the necessary subarrays and leaves the others in idle/sleep mode to minimize leakage power. To apply the gating technique, first we have to identify the idle subarrays in which gating can be applied. In the design under consideration the subarrays are corresponded to the rows of memory. Each row stores the information of one neural spike. When a new spike is detected, every single row in the memory is read and compared with the newly detected spike, and finally the SRAM gets updated if necessary. Therefore, at each time the row which is being read and compared is in active mode, and the rest of the rows are in sleep/idle mode and can be gated to save leakage power.

During sleep mode, we gate the idle rows using sleep transistors connected to each row, placed between GND and GNDV. A sleep transistor permits GNDV to rise by a control amount of voltage. The GNDV rise must be controlled so that it doesn't affect the cell's state. The sleep current is set to a level such that the rail-to-rail voltage across the transistors becomes equal to the minimum data retention voltage. When the row is about to be accessed, the sleep transistor corresponded to that row activates and connects the GNDV to GND. The transition from sleep mode to active mode takes some time and energy. In the design under consideration, the wake-up time and wake-up energy have 21% and 0.7% delay and energy overhead respectively.

TABLE III
TIMING PARAMETERS USED FOR DETERMINATION OF CLOCK CYCLE FOR A 'GATED' ARRAY

t_{read}(ns)	t_{write}(ns)	t_{comp}(ns)	t_{per}(ns)	t_{clock}(ns)
0.72	0.6	0.35	0.89	1.61

In sub-threshold region, the rail-to-rail voltage is already very low, and supply-gating cannot ensure reliable retention of data. Therefore, we only applied supply-gating to super-threshold design. Besides, gating can lead to unacceptable performance degradation for sub-threshold memory. We selected the optimal sub-threshold voltage to obtain acceptable memory performance in our case study. We need to preform 64 read and compare operations followed by a write operation for updating the memory. All the operations have to be completed within 320 μs, before the next spike is ready. We need 65 cycles to complete the read and compare operations in parallel. After adding one cycle for write operation, we have 66 cycles that need to be fit in the 320-μs time period. The length of the clock cycle (t_{clock}) can be represented as

$$t_{clock} = max(t_{read}, t_{comp}, t_{write}) + t_{per} \quad (1)$$

Where t_{read}, t_{comp}, and t_{write} are the time required to do a read, compare or a write operation, and t_{per} is the time required by the peripheral circuits like the decoder, the precharge unit, and the sense amplifier. Table III shows these timing parameters for a 'Gated' super-threshold design with a sleep transistor of 3000nm and V_{DD} of 1 V.

We can do all of the 66 operation cycles at 206 KHz frequency and scale down the supply voltage to 200 mV. Here, as a result of reducing the supply voltage the dynamic power is reduced, and delay is increased, leaving less time for supply-gating. In the next scenario, we can perform the operations at a faster rate of, say, 620 MHz with a supply voltage of 1 V, and then "gate" the idle rows for the remainder of the 320-μs period. In this case the leakage power is reduced at the cost of a slight increase in dynamic power. The timing diagram for the 66 cycles of the read-compare-write process in memory shown in Fig. 3, demonstrates the increased opportunity for power gating when operating at the super-threshold voltage at high frequency as opposed to the sub-threshold operation at ultra-low frequency.

As the frequency of operation increases, the active time decreases while the sleep time increases. On the other hand, the dynamic power increases but the leakage power does not change. Since the leakage power is only reduced during the sleep time, and the sleep time is increased, so more leakage energy can be reduced due to applying 'gating' technique for a longer time. Therefore, as the frequency of operation increases, the total energy of SRAM decreases.

IV. SIMULATION RESULTS

A. Simulation Setup

We consider a 64x80 SRAM array to explore the differences between the super-threshold and the sub-threshold designs in terms of energy, performance, area, and noise-margin. To determine the size of the array, we considered 64 neural samples which need to be stored, each sample has 8 coefficients, and each coefficient is 10-bit wide. To design the SRAMs, we use the conventional-6T and 8T SRAM cells for super-threshold and sub-threshold designs respectively.

TABLE IV
AREA, DELAY, ENERGY, AND NOISE-MARGIN FOR A 64X80 ARRAY REALIZED WITH THE CONVENTIONAL 6T SUPER-THRESHOLD AND THE SELECTED 8T SUB-THRESHOLD CELLS

Cases	Area	Delay	Energy (pJ)			Noise Margin (mv)		
	μm^2	(ns)	dyn	leak	tot	Read	Write	Hold
6T	256	0.37	40.48	69.61	110.09	209	420	383
8T	358.4	226	6.82	16.25	23.07	156	122	156

From the neural data recorded from the sea-slug (Aplysia californica), it can be observed that spikes usually appear every 100ms. We consider a window of 64 samples. Considering the sampling frequency of 10KHz, a new data sample for each channel arrives every 100μs. Therefore, for a single-channel recording system, we need to wait for 6.4ms (64 x 100μs) to have a full window of samples. However, if we consider a 20-channel recording system with time-multiplexed operation, data samples arrive every 5μs, and we only have to wait for 320μs (64 x 5μs). This is the time interval in which we have to fit the 66 cycles of read-compare-write operations. The simulations preformed for different V_{DD}s show that the read-delay dominates the write- and compare-delay, so it determines the length of the operation cycle. The comparator used in the simulations is an equality checker synthesized using the Synopsis design compiler and the OSU 45-nm standard cell library [9]. Simulations were performed in HSPICE for the 45-nm technology node [10]. For all the simulations the temperature is considered to be 40°C to simulate the body temperature for the implantable device.

B. Evaluation of Alternative Implementations

We compare the super-threshold and sub-threshold designs in terms of energy, performance, area, and noise-margin in Table IV. It should be noted that the nominal supply voltage for super-threshold and sub-threshold designs are considered as 1 V and 400 mV, respectively. As shown in the table, the leakage and dynamic energies are much smaller in sub-threshold design due to the lower supply voltage. While the energy decreases, delay and area increases significantly. We can observe that the super-threshold design is about 611X faster and 1.4X denser than its sub-threshold counterpart. Super-threshold design is also more robust than the sub-threshold design.

Supply-gating is applied to super-threshold design to save leakage energy. The gating is implemented through an NMOS sleep transistor which permits GND to rise by a controlled amount during the sleep mode. In Table V, we have compared the 'NotGated' and 'Gated' super-threshold SRAM arrays in terms of energy, performance, area, and noise-margin. The sleep transistor is considered to be 12X wider than the pull-down NMOS in the 6T SRAM cell. As it can be observed in Table V, supply gating has a huge effect on energy consumption. While energy is significantly reduced due to stacking

TABLE V
AREA, DELAY, ENERGY, AND NOISE-MARGIN VALUES FOR A 64x80 ARRAY REALIZED WITH THE 'NOTGATED' AND 'GATED' SUPER-THRESHOLD CELLS

Cases	Area	Delay	Energy (pJ)			Noise Margin (mv)		
	μm^2	(ns)	dyn	leak	tot	Read	Write	Hold
NotGated Sup.th	256	0.37	40.48	69.61	110.09	209	420	383
Gated Sup.th	265.6	0.72	30.41	28.02	58.43	204	408	383

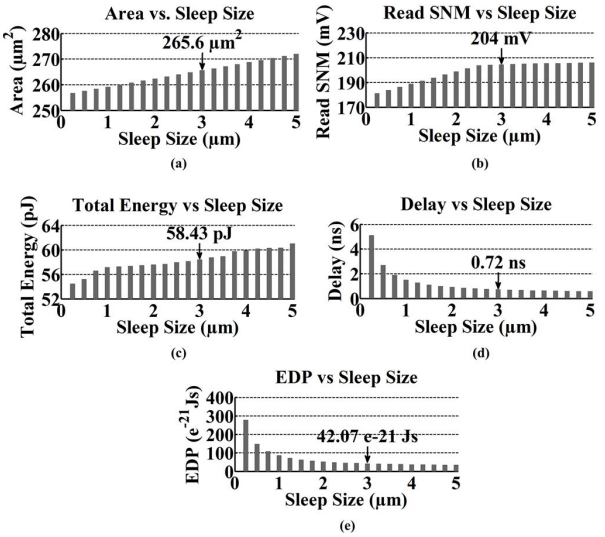

Fig. 4. (a) Area, (b) read noise-margin, (c) total energy, (d) delay, and (e) EDP of the 'Gated' Super-threshold SRAM with sleep transistor size scaling.

effect, small overhead can be observed in terms of area, delay, and noise-margin due to the imposed extra transistor per row. From the results shown in the table, it can be derived that supply-gating has significantly reduced the total energy by 47%, while it has increased the area and noise margin by only 3.6% and 2.4% respectively.

C. Case Study: Implantable Neural Interfaces

In this section, we investigate the proposed design solution for embedded memories in the context of a neural interface system. To implement supply-gating we consider only NMOS sleep transistor to minimize the area overhead. To account for the time required for the virtual ground node to discharge, the sleep to active transition is triggered during the negative phase of the clock cycle when the bitlines are getting charged, and the SRAM is still in sleep mode. The energy consumption overhead when the sleep transistor switches on/off was also taken into account when computing the total energy.

The effect of sleep transistor scaling on area, read static noise margin (SNM), energy consumption, delay, and energy delay product (EDP) is shown in Fig. 4. The size of sleep transistor is scaled from 0.25μm up to 5μm by steps of 0.25μm which is the size of the pulldown NMOS in the 6T SRAM cell. The maximum size of the sleep transistor is defined by the area and energy constraints. As it can be observed in Fig. 4(a), the

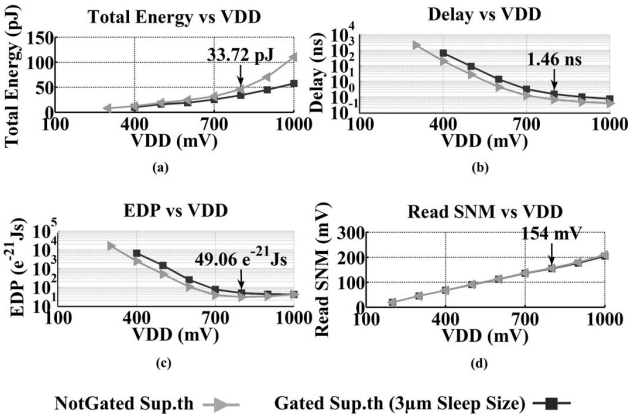

Fig. 5. (a) Total energy, (b) delay, (c) EDP, and (d) read noise-margin of the 'Gated' and 'NotGated' super-threshold SRAM with supply voltage scaling.

area has an increasing trend as the sleep transistor is scaled up. The area overhead due to sleep transistors ranges from 0.3% to 6% as compared to the 'NotGated' Super-threshold design. Fig. 4(b) shows the effect of sleep transistor scaling on read-SNM. As the size of the sleep transistor increases, the read-SNM gets bigger and bigger and saturates at the end. It has to be mentioned that since sleep transistor has more impact on read-SNM rather than write-SNM or even hold-SNM, we only considered the read-SNM as the deciding parameter to choose the best sleep size. The effect of sleep transistor scaling on total energy is plotted in Fig. 4(c). The increasing trend shows that the smaller the size of the sleep transistor the less the total energy. Fig. 4(d) shows a decreasing trend in delay as the sleep transistor is scaled. Since we have a long time interval to perform the vocabulary process, the operations can be performed at slow rate without violating the timing requirements. Hence we could choose a small sleep transistor to save energy at the cost of delay overhead. However, it should be noted that less delay overhead gives the opportunity of performing the operations at a faster rate and leaving more time for supply gating. EDP is widely used as a metric to determine the effectiveness of voltage scaling. In Fig. 4(e), we have plotted the effect of sleep transistor scaling on EDP. We observe that EDP has a decreasing trend as the sleep transistor is scaled. Evaluating all the plots, there is a point at which area has only 3% overhead, read-SNM is almost at its maximum value, and EDP is at one of its lowest values. This size point is 3000nm. At this point, energy is reduced up to 31%, and delay is increased by less than 100% as compared to the 'NotGated' super-threshold design.

Considering 3000nm as the nominal width of the sleep transistor, we try to reduce the total energy by scaling the supply-voltage down [11]. Fig. 5 shows the effect of supply-voltage scaling on energy, delay, EDP, and read-SNM. Fig. 5(a) and 5(b), show that while energy decreases steadily with voltage scaling, there is an overhead in terms of delay. Considering the EDP plot in Fig. 5(c), there is an optimal voltage point at which the EDP is minimized. This voltage point is 0.8 V, which is used in our simulations as the optimal super-threshold

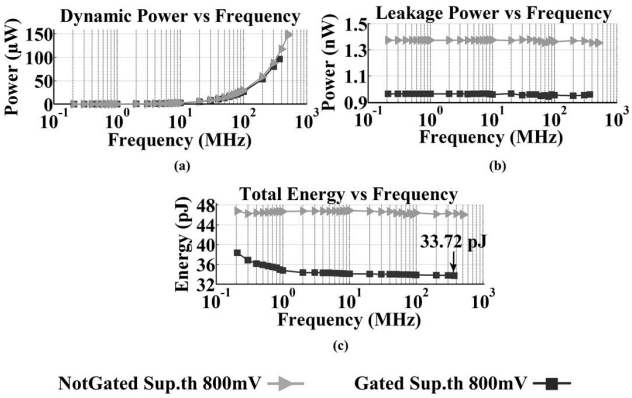

Fig. 6. (a) Dynamic power, (b) leakage power, and (c) total energy of the 'Gated' and 'NotGated' super-threshold SRAM with frequency scaling.

supply-voltage. Fig. 5(d), shows a linear decrease in read-SNM as the supply voltage is scaled. The read-SNM at the selected nominal supply voltage is 154 mV which is still high enough to assure the robustness of operations. It should be noted that the minimum energy point [see Fig. 5(a)] has a much lower supply voltage than the point with the minimum EDP (see Fig. 5(c)). Taking 0.8 V as the nominal super-threshold supply-voltage, we increased the idle time interval for supply gating by increasing the operating frequency.

Fig. 6, shows the effect of frequency scaling on dynamic power, leakage power, and total energy. The dynamic power increases with frequency, whereas the leakage power is independent of frequency. The leakage power for a 'Gated' design is also much lower than for a 'NotGated' design. These trends can be observed in Fig. 6(a) and 6(b). However, as frequency increases, the active time decreases while the idle time increases. Since the leakage power is reduced only during the idle period, the total energy decreases considerably for a 'Gated' design at high frequencies, as it is shown in Fig. 6(c). It can also be observed that the total energy is independent of frequency for a 'NotGated' design as no gating technique is applied to reduce the leakage energy. As it is shown in Fig. 6(c), the lowest total energy for the proposed design is 33.72pJ which is achieved at the highest frequency of operation 370MHz.

In Table VI, we compared the proposed 'Gated' super-threshold design with the 'NotGated' super-threshold and sub-threshold designs in terms of area, delay, total energy, and read-SNM. Using the proposed technique we reduced the total SRAM energy of the super-threshold design by 27% without a huge impact on area and robustness. The area and read-SNM for the proposed design are also better than the sub-threshold counterpart by 35% and 12%, respectively.

V. Conclusion & Future Work

We have presented design space exploration for implementing on-chip data storage in bioimplantable systems. Although sub-threshold design appears to be the natural choice for hardware implementation of the ultralow-power SRAM, we have shown that a well-optimized super-threshold design, which exploits the nature of the recorded signal, can achieve comparable energy efficiency while maintaining higher robustness and smaller area. This is possible by judiciously employing extensive supply gating in the design that leverages the speed gap between fast super-threshold operation and slow signal-acquisition frequency. We have shown that optimal choice of device size, operating voltage, and frequency can greatly minimize the power dissipation in the super-threshold design. Future work will involve application of the design approach to other implantable systems and hardware validation through test chip fabrication and measurement.

TABLE VI
Comparison of total energy, area, read-SNM, and delay among the 'NotGated' super-threshold, the proposed 'Gated' super-threshold, and sub-threshold SRAM designs

Cases	Area μm^2	Delay (ns)	Total Energy (pJ)	Read SNM (mV)
Sup.th NotGated 800 mV	256	0.65	46.01	156
Sup.th Gated 800 mV	265	1.45	33.72	154
Sub.th NotGated 400 mV	358	226	23.07	138

Acknowledgment

This work is supported in part by National Science Foundation (NSF) grant #CCF-0964514.

References

[1] K. D. Wise, D. J. Anderson, J. F. Hetke, D. R. Kipke, and K. Najafi, "Wireless implanatble microsystems: High-density electronic interfaces to the nervous system," *Proc. of the IEEE*, vol. 92, No. 1, pp. 76-97, 2004.

[2] S. Narasimhan, H. J. Chiel, and S. Bhunia, "Ultra-Low-Power and Robust Digital-Signal-Processing Hardware for Implantable Neural Interface Microsystems," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 5, No. 2, pp. 169-178, 2011.

[3] S. Narasimhan, Y. Zhou, H. J. Chiel, and S. Bhunia, "Low-Power VLSI Architecture for Neural Data Compression Using Vocabulary-based Approach," *IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2007.

[4] I. J. Chang, J. J. Kim, S. P. Park, and K. Roy, "A 32 kb 10T Sub-Threshold SRAM Array with Bit-Interleaving and Differential Read Scheme in 90nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 44, pp. 650-658, 2009.

[5] Tea-Hyoung Kim, J. Liu, and C. H. Kim, "An 8T subthreshold SRAM Cell Utilizing Reverse Short Channel Effect for Write Margin and Read Performance Improvement," *Custom Integrated Circuits Conference*, pp. 241-244, 2007.

[6] B. H. Calhoun and A. P. Chandrakasan, "A 256-kb 65-nm Sub-threshold SRAM Design for Ultra-Low-Voltage Operation," *IEEE Journal of Solid-State Circuits*, vol. 42, No. 3, pp. 680-688, 2007.

[7] N. Verma and A. P. Chandrakasan "A 256 kb 65 nm 8T Subthreshold SRAM Employing Sense-Amplifier Redundancy," *IEEE Journal of Solid-State Circuits*, vol. 43, No. 1, pp. 141-149, 2008.

[8] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechansim and leakage reduction techniques in deepsubmicrometer CMOS circuits," *Proc. of the IEEE*, vol. 91, No. 2, pp. 305-327, Feb. 2003.

[9] J. S. et al, "FreePDK: An Open-Source Variation-Aware Design Kit," *IEEE Intl. Conference on Microelectronic Systems Education*, pp. 173-174, 2007.

[10] Predictive technology model. [Online]. Available: *http://www.eas.asu.edu/ptm/*.

[11] R. Gonzalez, B. M. Gordon, and M. A. Horowitz "Supply and threshold voltage scaling for low power CMOS," *IEEE Journal of Solid-State Circuits*, vol. 32, No. 8, pp. 1210-1216, 1997.