Energy Dissipation Reduction of a Cardiac Event Detector in the sub-V$_t$ Domain By Architectural Folding

Blind, Blind, Blind,
Blind, Blind, and Blind

[1] Blind
[2] Blind
blind,blind
blind,blind,
blind,blind
Blind

Abstract. This manuscript presents the digital hardware realization of a wavelet based event detector for cardiac pacemaker applications. The architecture of the detector is partially folded to minimize hardware cost. An energy model is applied to evaluate the energy efficiency in the sub-threshold (sub-V$_T$) domain. The design is synthesized in 65 nm low leakage-high threshold CMOS technology, and it is shown that folding reduces the area cost by 30.6 %. Due to folding, energy efficiency of the circuit is increased by 14.4 % in the sub-V$_T$ regime, where the circuit dissipates 3.3 pJ per sample at V$_{DD}$=0.26 V.

Key words: Cardiac pacemaker, QRS detection, wavelet filterbank, folding, time-multiplexing, sub-threshold, energy model

1 Introduction

The application of implantable biomedical appliances has tremendously progressed during the last decades due to advances in CMOS technology scaling. The functionality of cardiac pacemakers has evolved from the steady-rate pacing in 1958, to programmable rate-responsive operation [5]. Traditionally, sensing, amplification and filtering of cardiac activity in the μV signal range is performed in the analog domain, before the signal is digitized [5, 11]. However, pacemaker functionality may be enhanced by performing signal processing in the digital domain, with the advantage of deploying more advanced algorithms.

The application of digital CMOS for cardiac event detection in favor of analog circuitry has previously been discarded because of the constraint on energy dissipation [5]. Technology scaling reduces dynamic power consumption due to smaller capacitive parasitics. However, disadvantageously leakage current has emerged as a major design constraint. Thus, leakage dissipation is seen as the

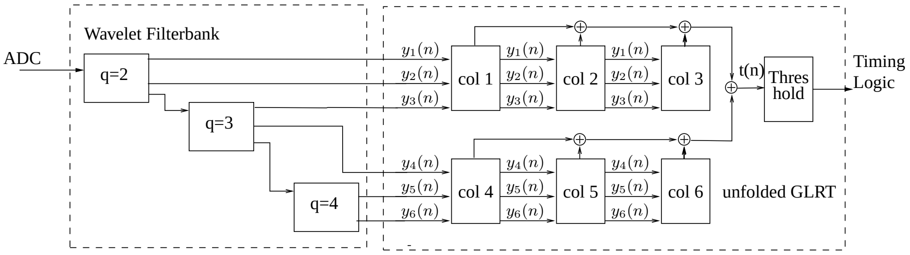

Fig. 1. Parallel architecture of the wavelet filterbank and GLRT.

major barrier, if targeting smaller technology nodes. However, if leakage is aggressively addressed, the overall energy dissipation may be competitive to analog circuitry.

An effective approach to minimize leakage is reduction of the gate count. This may be attained by hardware re-use, realized by time-multiplexing or folding. A folded architecture experiences an increased latency and needs to be triggered by a faster clock. However, gate count may be decreased if the overhead for extra registers and control logic does not eat up the gate reduction attained by folding. Moreover, in sub-threshold (sub-V$_T$) operation mode, power consumption is decreased significantly by aggressive supply voltage scaling.

Several successful implementations of digital circuits operating in the subthreshold regime are available in the literature [6, 10]. Circuits operating at these extreme low supply voltages work at much lower speeds, i. e., the FFT processor presented in [10] has a maximum clock frequency of 10 kHz with a power supply of 350 mV. Their extreme low power consumption results in excellent power delay product, making such circuits very interesting candidates for ultra-low power applications which do not have very high processing requirements.

The proposed architecture of a 3-scaled wavelet filterbank that feeds a generalized likelihood ratio test (GLRT), is optimized by folding the GLRT. The architecture is synthesized with 65 nm low leakage-high threshold (LL-HVT) CMOS technology. Energy efficiency is evaluated by deploying a SPICE-accurate energy model on the gate-level netlists [1]. These simulations require only a fraction of SPICE simulation time, and compute the supply voltage for the energy optimal operation point, maximum frequency, as well as dynamic and leakage dissipation.

In Sec. 2 the folding scheme of the cardiac event detector is presented. The energy model is presented in Sec. 3. In Sec. 4 the results of the energy dissipation reduction are discussed. Finally, conclusions are presented in Sec. 5.

2 Digital Hardware Implementation

This section presents the theory and architecture of a 3-scaled wavelet filterbank, supplemented with a GLRT, as presented in Fig. 1. Furthermore, event detection efficiency is discussed.

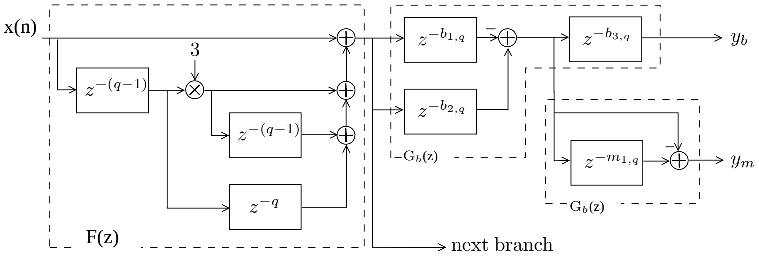

Fig. 2. Data flow diagram of the first wavelet filterbank branch using Mallat's algorithm, ($q = 2$).

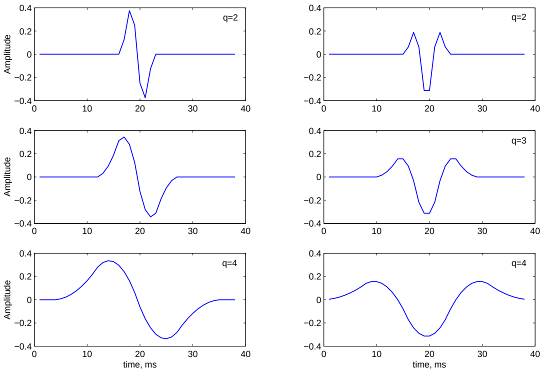

Fig. 3. Impulse responses of the wavelet filterbank. The biphasic impulse responses $y_{b,k}(n)$ for $q = 2, 3, 4$ are displayed in the left panel and the monophasic impulse responses $y_{m,k}(n)$ in the right panel.

2.1 Implementation of the R-wave detector

To achieve a power-efficient hardware mapping, short filters with integer values are chosen, i.e., first order difference, and the impulse response was chosen as a third order binomial function. A more detailed description of the wavelet filterbank and the GLRT is found in [2]. The implemented wavelet filterbank consist of three branches, $q = 2, 3, 4$, that scale and filter the signal $x(n)$ from the analog-to-digital converter, see Fig. 1 and 2. The first biphasic branch realizes a straight-forward implementation as

$$F(z) = 1 + 3z^{-(q-1)} + 3z^{-(2q-2)} + z^{-(2q-1)} \tag{1}$$

and

$$G_b(z) = -1 + z^{-q}. \tag{2}$$

Reusing $G_b(z)$ implements the monophasic filterbank using a single branch for one scale factor and realizes the output of the filterbank. However, in order to

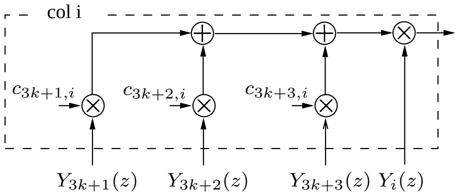

Fig. 4. Data flow diagram of a unfolded block in the GLRT.

center the functions to the longest propagation delay in the third branch, it is necessary to introduce additional delays in $G_b(z)$, see Fig. 2. The impulse responses of the filterbank are presented in Fig. 3. It can be observed that the wavelet-based structure offers a high flexibility for various cardiac morphologies.

The decision signal $T(n)$ is computed by the GLRT as

$$T(n) = \mathbf{x}^T(n)\mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}(n). \tag{3}$$

Since $\mathbf{x}^T(n)\mathbf{H} = \mathbf{H}^T\mathbf{x}(n)$, the remaining part of (3) to be implemented is the multiplication by $(\mathbf{H}^T\mathbf{H})^{-1}$, a matrix which is symmetric and sparse with half of its elements equal to zero,

$$(\mathbf{H}^T\mathbf{H})^{-1} = \begin{bmatrix} 4.3 & -2.8 & 0.7 & 0 & 0 & 0 \\ -2.8 & 4.5 & -1.8 & 0 & 0 & 0 \\ 0.7 & -1.8 & 1.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4.8 & -2.3 & 0.6 \\ 0 & 0 & 0 & -2.3 & 4.2 & -1.4 \\ 0 & 0 & 0 & 0.6 & -1.4 & 1.7 \end{bmatrix}. \tag{4}$$

The multiplication of $\mathbf{y}(n)$ with the first column of $(\mathbf{H}^T\mathbf{H})^{-1}$ and the first element of $\mathbf{H}^T\mathbf{x}(n)$ is carried out as depicted in Fig. 4, where $c_{i,i}$ are elements of $(\mathbf{H}^T\mathbf{H})^{-1}$ and $y_{3k+j}(n)$ the output of the filterbank.

2.2 Unfolded Architecture

The unfolded architecture of a wavelet scale and GLRT is mapped as illustrated in Fig. 2 and 4, respectively. Three elements of the wavelet scale are cascaded to realize the scaling factors $q = [2, 3, 4]$ of the wavelet filterbank. The schematic in Fig. 4 represents the block referred to as *col i* in Fig. 1, which needs to be replicated six times to realize the multiplication with the columns of the matrix $(H^TH)^{-1}$ in (4). To simplify the implementation the matrix coefficients $c_{i,i} \cdots c_{i,i+2}$ are replaced with rounded integer values, which did not degrade performance. Thus, the multiplications are realized by *shift-add* instructions. Hence, the unfolded realization of the GLRT requires six generic multipliers and 17 adders. Furthermore, the architecture is optimized by register minimization, numerical strength reduction, and internal word-length optimization, which, in turn, results in narrower adders and multipliers in the following GLRT.

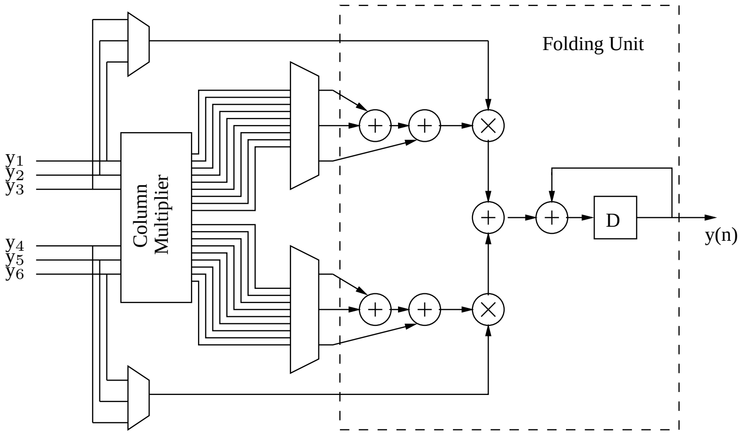

Fig. 5. Folded by three architecture of the GLRT.

2.3 Architectural Folding

Initially, both the wavelet filterbank and the GLRT, see Fig. 1, were folded. However, for the wavelet filterbank it turned out that the controller and register overhead were larger than savings achieved by reusing the adders. Consequently, only the GLRT is mapped as a by three and six folded architecture, i.e., the unfolded GLRT in Fig. 1 is replaced by the architecture in Fig. 5. In Fig. 5 folding by three is illustrated. The output $y_1 \cdots y_6$ of the wavelet filterbank is subjected to a block called *Column Multiplier* (CM). This block realizes concurrently the multiplications by $c_{i,i} \cdots c_{i,i+2}$, and holds the products for several clock cycles until processed by the *folding unit*. Folding of CM would lead to an area overhead since the coefficients are integer values. The de-multiplexers receive a control signal from a controller and switch the products to the adders, and switches $y_1 \cdots y_6$, correspondingly. The arrangement in Fig. 5 realizes the unfolded structure presented in Fig. 1 and 4. The HW cost of the folded architectures are listed with the unfolded realization in Table 1. The numbers show clearly the gain in area, i.e., the area cost for GLRT in PF3 and PF6 is reduced by 42 % and 49 % respectively. To maintain throughput the folded GLRT needs to be clocked three or six times higher than the wavelet filterbank.

Table 1. HW cost of a by three (PF3) and six (PF6) folded GLRT.

	Add	Mult	GLRT Area [μm^2]
Unfolded	35	6	6793
PF3	25	2	3912
PF6	21	1	3436

2.4 Detector Performance

The detector implemented in this study qualifies for pacemaker applications with reliable detection performance in noisy environments, validated on cardiograms digitally recorded during pacemaker implantation [8]. Detection performance is measured by computing the probability of true detections (PD) and false alarms (PFA). A true detection is defined as an event occurring within 50 ms of the annotation, whereas events outside this interval are declared as false alarms. All signals in the electrogram database (3200 events), are fed to the detector, and the detected events are classified as PD and PFA. It is found that the detector has a PD of 0.997 and PFA of < 0.001, which is rated as reliable performance.

3 Sub-V$_T$ Energy Model

The energy dissipation model presented in this section is comparable to other sub-V_T energy dissipation models [9, 12, 3]. In [9] Vittoz investigated and proved the energy-minimum operation property of sub-V_T logic. In this model, an expression for energy minimum operating voltage (EMV) is not derived, but determined by numerically inverting the duty factor for minimum energy. In [12] occurrence of the EMV is shown, but the corresponding equation is solved by curve fitting. In [3] sub-V_T EMV is solved analytically, where the model average switched capacitance and leakage current parameters are extracted from SPICE level simulation results.

As shown in this section, the model employed in this study uses parameters derived from high level simulations. The proposed model delivers SPICE-accurate data, but requires only a fraction of SPICE simulation runtime to obtain the internal energy dissipation of a single inverter equivalent capacitance value, which is not directly available in the synthesis library.

The total energy dissipation of static CMOS digital circuits is given by the following well-known equation:

$$E_{total} = \underbrace{\alpha C_{tot} {V_{DD}}^2}_{E_{dyn}} + \underbrace{I_{leak} V_{DD} T_{clk}}_{E_{leak}} + \underbrace{I_{peak} t_{sc} V_{DD}}_{E_{sc}}, \tag{5}$$

where E_{dyn} and E_{leak} are the average switching and leakage energy dissipated during a clock cycle T_{clk}, respectively. The contribution of the short circuit energy (E_{sc}) in the sub-V_T regime is neglected, as it is known to contribute only a small share of the overall energy dissipation [9]. In (5), E_{dyn} during one clock period is specified by the switching activity factor (α), and the maximum possible switched capacitance of the circuit (C_{tot}). The total capacitance C_{tot} is normalized in terms of total inverter capacitance using a capacitance scaling factor k_{cap} as $C_{tot} = k_{cap} C_{inv}$, where C_{inv} is the switched capacitance of an inverter. To calculate k_{cap}, the total capacitance obtained by the synthesis is normalized by the gate capacitance value of an inverter from the synthesis library. The leakage energy dissipation during a clock period T_{clk} is defined as

$$E_{leak} = k_{leak} I_0 V_{DD} T_{clk}, \tag{6}$$

where k_{leak} and I_0 are the average leakage scaling factor of the circuit and the average leakage current of a single inverter, respectively. The value for k_{leak} is obtained from the synthesis results by summing the individual average leakage currents of the gates, where the average leakage current is the mean of the leakage current for all the combinations of input vectors applied to the logic gate, and normalizing the result to the average leakage current of a single inverter.

The critical path that constraints the maximum clock frequency is specified as

$$T_{clk} = k_{crit} T_{sw\_inv}, \tag{7}$$

where k_{crit} is a coefficient that defines the critical path delay of the circuit in terms of the inverter delay T_{sw_inv}. The parameter k_{crit} is calculated by dividing the critical path from the synthesis results by the average delay of the inverter while operating at nominal supply.

The delay of an inverter operating in the sub-V_T regime is given in [9] as

$$T_{sw\_inv} = \frac{C_{inv} V_{DD}}{I_0 e^{V_{DD}/(nU_t)}}. \tag{8}$$

By introducing (8) into (7), the clock period is specified as

$$T_{clk} = k_{crit} \frac{C_{inv} V_{DD}}{I_0 e^{V_{DD}/(nU_t)}}, \tag{9}$$

and by combining (5), (6) and (9), the final total energy dissipation while working at the maximum clock frequency is specified as

$$E_T = C_{inv} V_{DD}^2 \left[\alpha k_{cap} + k_{crit} k_{leak} e^{-V_{DD}/(nU_t)} \right]. \tag{10}$$

EMV is found by taking the derivative of (10) with respect to V_{DD}. Thus, EMV is specified as

$$V_{opt-sync} = 2nU_t - nU_t W_{-1} \left[-\frac{2e^2 \alpha k_{cap}}{k_{crit} k_{leak}} \right], \tag{11}$$

where W_{-1} is the -1 branch of the LambertW function [4]. This result for the EMV confirms the result presented in [3]. All k-parameters in (10) and (11) are obtained from synthesis results, and the switching activity factor α is calculated after running gate level simulations where toggle information is generated from real data [8]. Hence, the total simulation time to characterize sub-V_T performance is highly reduced compared to the SPICE-level simulations. The only parameters that need to be extracted from SPICE simulations for numerical calculations are the slope factor n and the switched inverter capacitance C_{inv}. The accuracy of the model is checked with respect to the SPICE level simulations using the ISCAS85 benchmark circuits and the first quantile of error of the energy dissipation model is found to be below 6% with a mean error of 0.61% in the sub-V$_T$ regime.

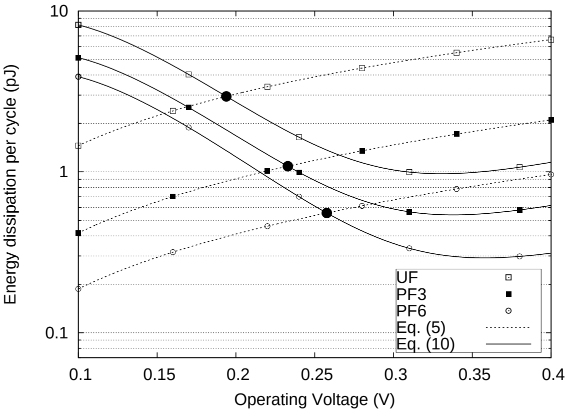

Fig. 6. Sub-threshold energy dissipation curves of different architectures.

4 Sub-V$_T$ Operation Mode

This section presents the energy dissipation results of the unfolded (UF), and by three (PF3) and six (PF6) folded architectures. Static noise margin (SNM) failure rates are taken into consideration to find an optimal operation point. Table 2 shows the circuit parameters of the synthesized architectures. By employing a higher folding factor the total gate count of the circuit is reduced. This results in lower leakage energy dissipation for the same amount of operation time. The data is fed to the cardiac event detector at a speed of 1 kHz, and in order to maintain throughput, the GLRT operation frequency in UF, PF3 and PF6 architectures needs to be 1, 3 and 6 kHz, respectively. Fig. 6 shows the sub-V_T energy dissipation curves for one clock cycle. The continuous lines show the energy dissipation while working at the speed of the critical path, i. e., minimum leakage time, and the dashed lines show the dissipation while working with a fixed clock. The circuits need to be operated at least at a V$_{DD}$ value that meets the requirement on the maximum clock frequency, i.e., 3 and 6 kHz, indicated by the black dots, which are lower than the EMV values. If V$_{DD}$ is raised higher than indicated by the black dots while working at an externally set speed, E_{total} from (5) will increase (dashed lines). The higher achievable clock frequency at EMV due to a higher V$_{DD}$, hence lower leakage time, can not be utilized since the clock speed constraint is external. Thus, if there is an external speed constraint, then, working at a voltage value higher than the value that sat-

Table 2. Composition properties of the synthesized circuits.

Architecture	k_{cap}	k_{leak}	k_{crit}	EMV (V)
Unfolded	17820	13358	608.051	0.33
PF3	12550	10991	463.259	0.34
PF6	10303	9794	396.805	0.36

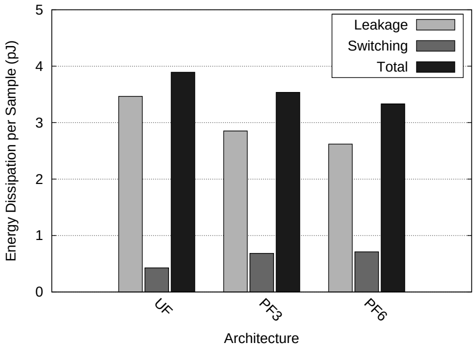

Fig. 7. Energy dissipation components of different architectures.

isfies the speed requirement results in energy overhead. The only way to operate at the EMV with external speed requirements when the speed requirement is lower than the operating speed and reduce the energy dissipation to a minimum is to employ power-shutdown, which is not investigated in this paper. It should be noted that although power-shutdown and working at EMV will reduce the energy dissipation, it will introduce energy overhead and a more complicated design process.

Theoretically, the supply voltage of sub-V_T circuits can be reduced down to 50 mV [9], in practice at such voltage values functional failures occur due to the process variations. Thus, circuits need to be checked for failure rates while operating at extremely low voltages. In this paper the static noise margin (SNM) failure rates of the gates are extracted from 5k-point Monte Carlo analysis, which follows the methodology in [7]. It is found that the supply voltage value which realizes operation with less than 0.001 failure rate for a 65 nm process is $0.25V$ and this value is taken as the minimum reliable operating voltage (ROV). This results in UF and PF3 operating voltages rising to $0.25V$, causing energy dissipation overhead. The PF6 architecture still operates at $0.26V$ as in Fig. 6 to satisfy the speed requirement. From now on, the mentioned supply voltages will be taken as the operating voltages of different architectures.

To sustain throughput in a folded architecture, the computation of one sample requires 3 and 6 clock cycles for PF3 and PF6, respectively. Therefore, the switching energy per cycle for the folded architectures should be multiplied by their respective folding factors to obtain the switching energy per sample. Moreover, since the idle part of the circuitry leaks during the calculation, the total leakage time of all the architectures is the same and is 1 ms per sample. Thus, it is necessary to multiply leakage energy per clock cycle by the applied folding factor. Since the throughput is an external speed constraint, all the architectures process the data at the same amount of time. Gate count reduction minimizes leakage energy, and hence the average leakage scaling factor (k_{leak}) of the circuit. Fig. 7 shows the energy dissipation components of the designed architectures per sample. Since all circuits need to be supplied with a voltage lower than EMV, they will operate in the leakage dominated region. From Fig. 7, it is seen that by increasing the folding factor, the switching energy increases. This is due to the increase in the complexity of the control circuit. However, although the switching energy increases, it is offset by the reduction in the leakage energy, reducing the overall energy dissipation. By going from the UF architecture to the PF6 architecture, the overall energy dissipation per sample point is reduced by 14.4%.

5 Conclusions

This manuscript presents architectural folding of a wavelet based cardiac event detector. It is shown that the total area in the most optimized architecture is reduced by 31 %, which results in corresponding leakage reduction. Thereby, energy dissipation is reduced by 14.4 %. The switching energy due to controller and register overhead increases, but the total leakage reduction offsets this increase in energy dissipation. The operating voltage, which satisfies both speed and failure rate requirement, is determined as 0.26 V, where the circuit dissipates 3.3 pJ per sample.

References

1. O. Akgun and Y. Leblebici. Energy Efficiency Comparison of Asynchronous and Synchronous Circuits Operating in the Sub-Threshold Regime. *J. Low Power Electronics*, 3(3):320–336, 2008.
2. M. Åström, S. Olmos, and L. Sörnmo. Wavelet-based event detection in implantable cardiac rhythm management devices. *IEEE Trans. Biomed. Eng.*, 53(3), March 2006.
3. B. Calhoun, A. Wang, and A. Chandrakasan. Modeling and sizing for minimum energy operation in subthreshold circuits. *Solid-State Circuits, IEEE Journal of*, 40(9):1778–1786, 2005.
4. R. Corless, G. Gonnet, D. Hare, D. Jeffrey, and D. Knuth. On the LambertW function. *Advances in Computational Mathematics*, 5(1):329–359, 1996.
5. S. Haddad, R. Houben, and W. Serdijin. The evolution of pacemakers. *Engineering in Medicine and Biology Magazine, IEEE*, 25(3):38–48, May-June 2006.
6. J. P. Kulkarni, K. Kim, and K. Roy. A 160 mV robust schmitt trigger based subthreshold SRAM. *IEEE Journal of Solid-State Circuits*, 42(10):2303–2313, Oct. 2007.
7. J. Kwong and A. Chandrakasan. Variation-driven device sizing for minimum energy sub-threshold circuits. In *Proc. of the 2006 International symposium on Low power electronics and design.* ACM New York, NY, USA, 2006.
8. J. Rodrigues, L. Olsson, T. Sörnmo, and V. Öwall. Digital implementation of a wavelet-based event detector for cardiac pacemakers. *IEEE Transactions on Circuits and Systems I: Regular Papers,*, 52(12):2686–2698, Dec. 2005.
9. E. Vittoz. *Low-Power Electronics Design*, chapter 16. CRC Press LLC, 2004.
10. A. Wang and A. Chandrakasan. A 180-mV subthreshold FFT processor using a minimum energy design methodology. *IEEE Journal of Solid-State Circuits*, 40(1):310–319, 2005.

11. L. Wong, S. Hossain, A. Ta, J. Edvinsson, D. Rivas, and H. Naas. A very low-power cmos mixed-signal ic for implantable pacemaker applications. *IEEE Journal of Solid-State Circuits*, 39(12):2446–2456, Dec. 2004.
12. B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner. Theoretical and practical limits of dynamic voltage scaling. In *Proceedings of the 41st Annual Conference on Design Automation*, pages 868–873. ACM New York, NY, USA, 2004.