High-level Energy Estimation in the Sub-V_T Domain: Simulation and Measurement of a Cardiac Event Detector

Omer Can Akgun, *Member*, IEEE, Joachim Neves Rodrigues, *Member*, IEEE, Yusuf Leblebici, *Fellow*, IEEE and Viktor Öwall, *Member*, IEEE
Electrical and Information Technology, Lund University 22100 Lund, Sweden Email: {omercan.akgun, joachim.rodrigues,viktor.owall}@eit.lth.se
Microelectronic System Laboratory, EPFL 1015 Lausanne, Switzerland Email: yusuf.leblebici@epfl.ch

***Abstract*—This paper presents a flow that is suitable to estimate energy dissipation of digital standard-cell based designs which are determined to be operated in the sub-threshold regime. The flow is applicable on gate-level netlists, where back-annotated toggle information is used to find the minimum energy operation point, corresponding maximum clock frequency, as well as the dissipated energy per clock cycle. Simulation results, which are obtained during a fraction of SPICE simulation time, are validated by measurements on a wavelet based cardiac event detector that was fabricated in 65 nm low-leakage high-threshold technology. The mean of the absolute modeling error is calculated as 5.2 %, with a standard deviation of 6.6 % over the measurement points. The cardiac event detector dissipates 0.88 pJ/sample at a supply voltage of 320 mV.**

***Index Terms*—High-level energy estimation, sub-threshold, QRS detection, R-wave, cardiac pacemaker, energy model.**

I. Introduction

LOW energy design is a very crucial design constraint for biomedical implants. Significant reduction in energy dissipation is achieved by lowering the supply voltage [1]. This is possible by relaxing the constraints of classical strong-inversion operation of MOSFETs, and by accepting the notion that transistors are operated well below threshold, in the sub-threshold (weak-inversion) regime.

In sub-threshold (Sub-V_T) mode, the supply voltage may be scaled aggressively and thus power consumption is decreased by magnitudes. Sub-threshold operation of static CMOS logic has been analyzed using the EKV model in [2]. In this analysis, it is shown that static CMOS logic may be operated with a supply voltage as low as 50 mV at ambient temperature.

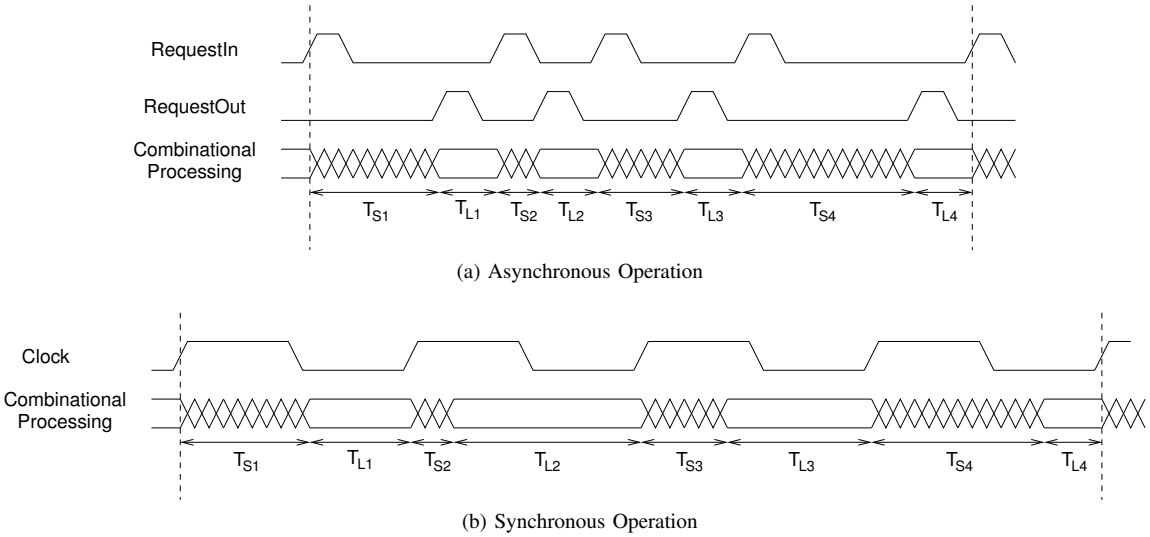
There are several successful implementations of digital circuits operating in the sub-threshold regime in the literature such as, an FFT processor that is operational down to 180 mV [3] and a sub-threshold SRAM which operates with a supply voltage of 160 mV [4]. Circuits operating at these extreme low supply voltages work at much lower speeds, as an example, the FFT processor presented in [3] works with a maximum clock frequency of 10 kHz with a power supply of 350 mV. Their extremely low power consumption results in excellent power delay product (PDP) values, making such circuits very interesting candidates for ultra-low power applications which do not have very high processing requirements; one such application being cardiac pacemakers.

In the sub-threshold regime, leakage current of the transistors are used for computation. The sub-threshold leakage current depends on the supply voltage exponentially, resulting in exponential increase in the circuit delay and lower leakage energy dissipation for lower supply voltages. Due to an exponential dependence of the leakage energy and quadratic dependence of the switching energy in the sub-threshold regime, sub-threshold operation has an energy-minimum operating voltage (EMV).

Several energy estimation models for sub-V_T operation has been published [2], [5], [6]. In [2] Vittoz investigated and proved the energy-minimum operation property of sub-threshold logic. In the developed model , an expression for the energy-minimum operating voltage was not derived and energy-minimum operating point was shown by numerically inverting the duty factor for minimum energy. In [5] occurrence of EMV was shown but the equation for EMV was solved by curve fitting. In [6] Calhoun solved the sub-threshold energy-minimum operating voltage analytically. In Calhoun's model the average switched capacitance and average leakage current were specified as parameters and extracted from SPICE level simulation results. However, previous models are only applicable on synchronous designs, and furthermore, none of the mentioned models is feasible for high-level design exploration.

The proposed energy model provides accurate results on sub-V_T design characteristics, without requiring computation and time intensive SPICE simulations, making the model usable for an early stage in the design flow. The model is applicable on gate-level, and thus, it is possible to characterize sub-threshold energy efficiency by architectural optimizations. The model is capable of modelling both asynchronous and synchronous designs.

In Sec. II the theory of the energy model is presented. The implementation flow of the energy model is presented in Sec. III. Section IV introduces the architecture of a cardiac event detector, which is used as a reference design to validate the energy model. Chip implementation details are given in Section V and model application is presented in Section VI. The energy model is validated by measurements in Sec. VII, and, finally, conclusions are presented in Sec. VIII.



(a) Asynchronous Operation

(b) Synchronous Operation

Figure 1. Timing diagram showing the same logic block operating in (a) asynchronous and (b) synchronous modes.

II. THEORY

This section presents the theory behind the energy estimation flow, applicable to both asynchronous and synchronous circuits. Furthermore, in order to be able to simplify mathematical operations and derivation of equations the following assumptions are made:

Assumption-1: The (asynchronous or synchronous) design operates with the highest possible throughput, i.e., new data is applied as soon as processing is finished (asynchronous), or with the next clock (synchronous) that operates at critical path speed. Thus, idle time of the hardware is minimized.

Assumption-2: Energy dissipation and processing delays of the circuit per computation are randomly distributed, guaranteed by processing a randomly distributed data set.

Assumption-1 guarantees that synchronous and asynchronous operations are compared fairly in terms of energy dissipation. Circuits run at the maximum speeds possible, and, hence leakage energy dissipation is minimized, while working at the energy-minimum operating voltage. The claim in Assumption-2 is necessary to simplify the statistical properties of the energy model. Moreover, as long as energy dissipation and processing delay of a circuit are randomly distributed with a mean, the model may be applied to any circuit that is operated with an arbitrary data set.

A. Energy Model Derivation

The energy model is developed by initial observations of an asynchronous design, and subsequently, the model is modified to enable numerical simulations of synchronous designs as well. In an asynchronous system, the operation of the system is dictated by the switching and delay properties, as well as the external request and acknowledge signals. A conceptual illustration of an asynchronous operation is shown in Figure 1a. The design is observed for an arbitrary time frame T, where four sets of input data are processed. The time intervals where the circuit is purely *leaking* (waiting for handshake completion) are denoted with T_{Li}, and the time interval where the gates are both *switching* (processing data) and *leaking* are denoted as T_{Si}. It is assumed that as soon as the *RequestOut* signal is lowered, *RequestIn* goes high (Assumption-1) and new data is processed. Thus, the purely leaking time intervals (T_{Li}s) are fixed, and due to the asynchronous protocol overhead.

Figure 1b shows the timing diagram of a synchronous design, which processes the same data as depicted in Figure 1a. The combinational processing times T_{Si} remain unchanged. However the interval T_{Li}, where the synchronous timed design is in idle mode, is longer compared to its asynchronous counterpart. Thus, the synchronous implementation will have a lower throughput and longer idles times, while operating at the same supply voltage. These observations will be considered while developing the energy model.

1) Asynchronous Designs: To calculate the energy dissipation of an arbitrarily long time frame T, the operation cycles of a design need to be monitored. During this time frame, it is assumed that the number of switchings is N. In any static digital CMOS the total energy dissipation is calculated as

$$E_T = E_{dynamic} + E_{leakage} + E_{short-circuit}, \qquad (1)$$

where $E_{dynamic}$ is the dynamic energy due to the switching of the capacitances, $E_{leakage}$ is the leakage energy during the time the gates are supplied by an energy source, and $E_{short-circuit}$ is the energy due to the short-circuit current flowing from the supply to the ground during switching. In this analysis the contribution of the short-circuit is neglected, as it is known to contribute only a small portion of the overall energy [2]. Dynamic energy dissipation in (1) during the ith time interval is specified as

$$E_{dynamic_i} = e_i C_{tot} V_{DD}^2, \qquad (2)$$

where e_i is a scaling parameter that defines the switching property of a design circuit for a specific input data transition, and C_{tot} is the maximum possible switched capacitance of the design. The switching energy scaling parameter e_i is in

the range $[0, 1]$. Without loss of generality e_i is expressed as a single value in a random process e (Assumption-2). Hence, it is possible to specify a mean μ_e, and thus (2) for N computations may be rewritten as

$$E_{dynamic} = N\mu_e C_{tot} V_{DD}^2. \tag{3}$$

In (2) and (3), the total capacitance C_{tot} may be normalized in terms of the 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. Furthermore, even while switching, the gates are leaking, and consequently leakage energy during the observation period T is defined as

$$E_{leak} = k_{leak} I_0 V_{DD} T, \tag{4}$$

where k_{leak} is the average leakage scaling factor over all gates, and I_0 represents the average leakage current of a single inverter. Total average leakage current of a design is calculated as $k_{leak} I_0$ from (4). The average leakage parameter k_{leak} is obtained from synthesis results by computing the mean of the leakage current for all combinations of input vectors applied to the logic gate inferred in the gate-level netlist, and normalizing the result to the average leakage current of a single inverter. Insertion of (3) and (4) in (1), results in

$$E_T = N\mu_e k_{cap} C_{inv} V_{DD}^2 + k_{leak} I_0 V_{DD} T, \tag{5}$$

which specifies the energy dissipation during a time interval T. The combinational processing time is specified as

$$T_S = \sum_{i=1}^{S} T_{s_i}, \tag{6}$$

and based on the switching/timing statistics of the design. T_{s_i} in Figure 1a is calculated as

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

where d_i in the range $[0, 1]$ is a scaling parameter that defines the delay properties during switching, k_{crit} defines the critical path delay per sample in terms of an inverter delay, and T_{sw_inv} is the delay of an inverter. Furthermore, by modelling d as a random process with the mean μ_d (Assumption-2), the total time spent during switching is computed as

$$T_S = N\mu_d k_{crit} T_{sw\_inv}. \tag{8}$$

During the observation frame T, N switchings and N handshakes take place, thus T is expressed as

$$\begin{aligned} T &= T_S + T_L \\ &= N\mu_d k_{crit} T_{sw\_inv} + N k_{oh} k_{crit} T_{sw\_inv}, \end{aligned} \tag{9}$$

where k_{oh} is a parameter that defines the overhead caused by the asynchronous communication in terms of critical path delay. The delay of an inverter working in the sub-threshold regime is given in [2] as

$$T_{sw\_inv} = \frac{C_{inv} V_{DD}}{I_0 e^{V_{DD}/(nU_t)}}, \tag{10}$$

where n is a process dependent term called slope factor typically in the range of 1.3-1.5 in modern CMOS technologies,I_0 the saturation off current, and U_t is the thermal voltage (26 mV at 300 K). Introduction of (10) into (9), gives

$$T = N k_{crit} \frac{C_{inv} V_{DD}}{I_0 e^{V_{DD}/(nU_t)}} (\mu_d + k_{oh}), \tag{11}$$

which is the total observation time. Finally, insertion of (11) in (5), results in

$$\begin{aligned} E_{T-async} =& N C_{inv} V_{DD}^2 + \Big[\mu_e k_{cap} \\ &+ k_{crit} k_{leak} (\mu_d + k_{oh}) e^{-V_{DD}/(nU_t)}\Big], \end{aligned} \tag{12}$$

which is the total energy dissipation for N switchings. The average energy dissipation per clock cycle is calculated by setting $N = 1$ in (12). The optimal operating voltage for minimum energy operation is found by setting the derivative of (12) with respect to V_{DD} to zero. Thus, the energy-minimum operating voltage is given in (13) as

$$\begin{aligned} V_{opt-async} =& 2nU_t - \\ & nU_t W_{-1}\left[-\frac{2e^2 k_{cap}\mu_e}{k_{crit} k_{leak}(k_{oh} + \mu_d)}\right], \end{aligned} \tag{13}$$

where W_{-1} is the -1 branch of the LambertW function [7]. All the k-parameters in (12) and (13) are extracted from synthesis results, and the μ-parameters are obtained from toggle information generated by gate level simulations with back-annotated timing information.

2) Synchronous Designs: A similar modelling approach is developed for synchronous designs, which operate as depicted in Figure 1b. The model is developed according to Assumption-1, which constraints that the clock period is equal to the critical path, i.e., ($T = k_{crit} T_{sw\_inv}$); the energy per clock cycle is derived by modification of (11) and (12) as

$$\begin{aligned} E_T =& C_{inv} V_{DD}^2 \\ & \left[\mu_e k_{cap} + k_{crit} k_{leak} e^{-V_{DD}/(nU_t)}\right]. \end{aligned} \tag{14}$$

As in the asynchronous case, by taking the derivative of (14), we get the minimum energy operating point as

$$V_{opt-sync} = 2nU_t - nU_t W_{-1}\left[-\frac{2e^2 k_{cap}\mu_e}{k_{crit} k_{leak}}\right]. \tag{15}$$

So far, it was assumed that the design operates at the maximum frequency imposed by the operating voltage, hence operating with minimum leakage energy possible at that voltage. Usually, this is not the case in real world applications, where an operation frequency is dictated by external constraints. Thus, (14) cannot be used to calculate the energy dissipation of a circuit in such a scenario. Therefore, a model not constraining the leakage time by the maximum operating frequency needs to be developed. For externally constrained systems which work below the speed achievable at the energy-minimum operating point, (5) is modified to

$$E_T = \mu_e k_{cap} C_{inv} V_{DD}^2 + k_{leak} I_0 V_{DD} T_{CLK}, \tag{16}$$

where T_{CLK} is the period of the clock.

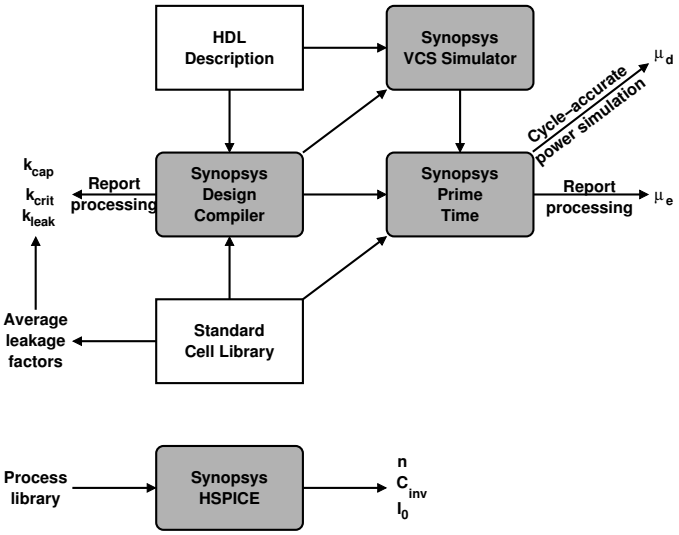


Figure 2. Model application flow emphasizing the tools used.

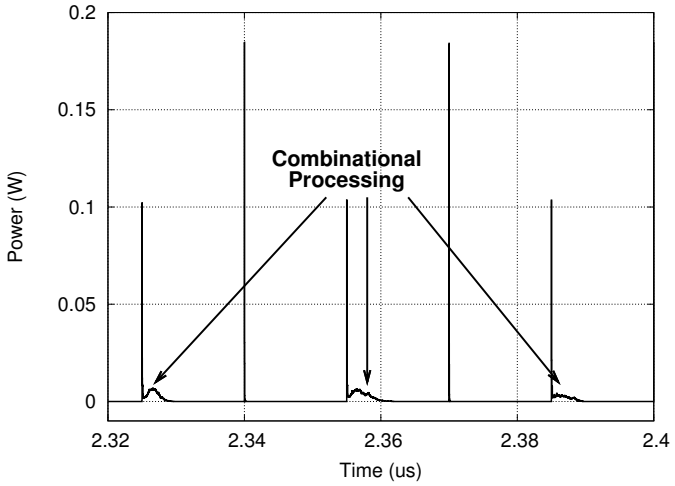


Figure 3. Synopsys PrimeTime Power cycle-accurate power waveform to emphasize the power consuming combinational operations. Combinational processing timing information is used for calculating the delay distribution of a mixed combinational-sequential circuit.

B. Maximum Operating Frequency Calculation

From the introduced model, the maximum operating frequency of a circuit may be easily calculated. The operating frequency of a synchronous design is defined as

$$f_{max} = \frac{1}{k_{crit} T_{sw\_inv}}, \tag{17}$$

and by introducing (10) into (17), the maximum operating frequency of a synchronous circuit is found as

$$f_{max} = \frac{I_0 e^{V_{DD}/(nU_t)}}{k_{crit} C_{inv} V_{DD}}. \tag{18}$$

For an asynchronous design the average throughput is a more important parameter, which can be obtained by the inclusion of the $\mu_d + k_{oh}$ factor as

$$f_{average} = \frac{1}{\mu_d + k_{oh}} \frac{I_0 e^{V_{DD}/(nU_t)}}{k_{crit} C_{inv} V_{DD}}. \tag{19}$$

When equations (18) and (19) are compared, it can be seen that as long as $\mu_d + k_{oh}$ is less than 1, the average throughput of an asynchronous implementation will be higher than that of a synchronous implementation.

III. Model Implementation Flow

The application flow of the energy model derived in Sec. II is shown in Figure 2 and explained in this section with emphasis on the generation of the required parameters.

The flow requires a so called pre-characterization of the process and the cell libraries, which only needs to be done once for the used technology. First, the cell libraries are characterized to compute the average leakage factor of each cell, normalized to a chosen inverter implementation. This is achieved by processing the `.lib` file supplied by the foundry (or generated during full custom layout design), with custom developed scripts [8]. Second, the process library is characterized for the slope factor n, the inverter internal switched capacitance C_{int}, and saturation off current I_0. These parameters are not directly available in the `.lib` file. This is carried out by HSPICE simulations, where inverter characterization and single transistor testbenches are created.

Digital designs may be implemented by any hardware description language (HDL). Gate mapping is carried out by a synthesis tool, e.g., Synopsys Design Compiler. During synthesis, detailed reports of circuit properties are generated by the tool. For cycle accurate timing and power simulation, delay information of the design needs to be written in a standard delay format (`.sdf`) file. Thereafter, the synthesis reports are processed using custom developed scripts in combination with average leakage factors obtained during pre-characterization that are used to calculate k_{leak}. Finally, the values for k_{cap}, k_{crit} and k_{leak} are obtained, see Figure 2.

Input data dependent toggle information of the design is obtained by gate-level netlist simulations using back-annotated timing information from the `sdf` file. HDL simulators like Synopsys VCS or Modelsim may be used to create a *value change dump* (vcd) file. The gate-level netlist is subjected to Synopsys Prime Time to perform power simulation. Here the vcd information is used in combination with the .sdf file to carry out a cycle-accurate power simulation. By post processing the data, it is possible to extract the delay distribution and its average (μ_d), as well as the energy distribution and its mean (μ_e). A sample waveform from the Synopsys PrimeTime Power simulations is presented in Figure 3. Power spikes just before combinational processing occurs is due to the positive edge of the clock signal and spikes that are not followed by combinational processing are due to the negative edge of the clock signal. Thus, both the distributions and the average values of delay and energy characteristics of the circuit are obtained.

IV. Digital Hardware Implementation of a Cardiac Event Detector

As a reference design for a typical low power application, a cardiac event detector for pacemaker application is chosen. This section briefly presents the theory and architecture of

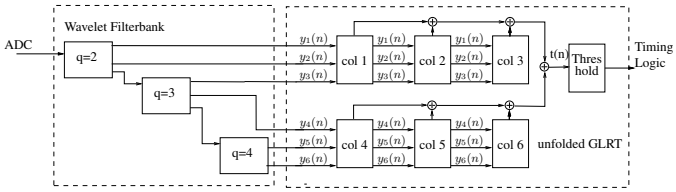


Figure 4. Block diagram of the wavelet filterbank and GLRT.

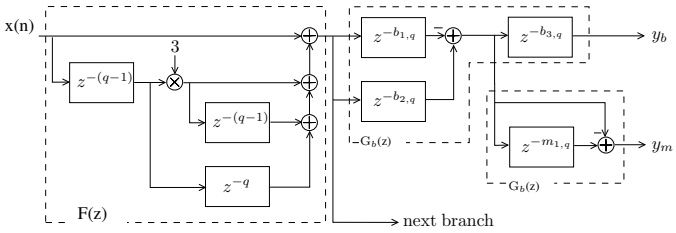


Figure 5. Data flow diagram of a wavelet branch using Mallat's algorithm, $(q = 2)$.

a 3-scaled wavelet filterbank, that scales and conditions the signal for hypothesis testing in a GLRT, see Figure 4. A more thorough description of the cardiac event detector may be found in [9].

A. Architecture

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 [10]. The implemented wavelet filterbank consist of three branches, $q = 2, 3, 4$, that scale and filter the signal $x(n)$, see Figure 4 and 5. The first biphasic branch realizes a straight-forward implementation as

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

and

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

Reusing $G_b(z)$ implements a monophasic filterbank using a single branch for one scale factor and realizes the output of the filterbank. In order to center the functions to the longest propagation delay in the third branch, it is necessary to introduce additional delays in $G_b(z)$, see Figure 5. The impulse responses of the filterbank are presented in Figure 6. The wavelet-based structure offers a high flexibility for various cardiac morphologies.

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

$$T(n) = \mathbf{x}^T(n)\mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}(n), \qquad (20)$$

where $\mathbf{H}$ holds the coefficients of the bi- and mono-phasic filter functions. Since $\mathbf{x}^T(n)\mathbf{H} = \mathbf{H}^T\mathbf{x}(n)$, the remaining part of (20) 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

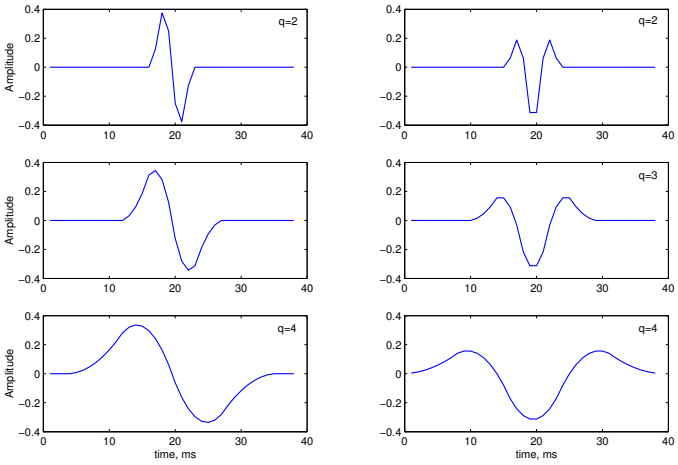


Figure 6. Impulse responses of the wavelet filterbank. The biphasic impulse responses $y_{b,k}(n)$ for $q = 2, 3, 4$ are displayed in the left panel and the monophasic impulse responses $y_{m,k}(n)$ in the right panel.

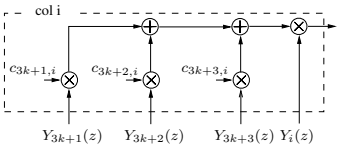


Figure 7. Data flow diagram of a direct-mapped block in the GLRT.

equal to zero,

$$(\mathbf{H}^T\mathbf{H})^{-1} = \begin{bmatrix} 4.3 & -2.8 & 0.7 & 0 & 0 & 0 \\ -2.8 & 4.5 & -1.8 & 0 & 0 & 0 \\ 0.7 & -1.8 & 1.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4.8 & -2.3 & 0.6 \\ 0 & 0 & 0 & -2.3 & 4.2 & -1.4 \\ 0 & 0 & 0 & 0.6 & -1.4 & 1.7 \end{bmatrix}. \qquad (21)$$

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 Figure 7, where $c_{3k+j,i}$ are elements of $(\mathbf{H}^T\mathbf{H})^{-1}$ and $y_{3k+j}(n)$ the output of the filterbank, with $k = 0, 1$ and $j = 1, 2, 3$.

The architecture of a wavelet scale and GLRT is mapped as illustrated in Figure 5 and 7, 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 Figure 7 represents the block referred to as *col i* in Figure 4, which needs to be replicated six times to realize the multiplication with the columns of the matrix $(H^T H)^{-1}$ in (21). To simplify the implementation, the matrix coefficients $c_{i,i} \cdots c_{i,i+2}$ are replaced with rounded integer values, which does not degrade detection performance. Thus, the multiplications are realized by *shift-add* instructions. Hence, the hardware 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 GLRT.

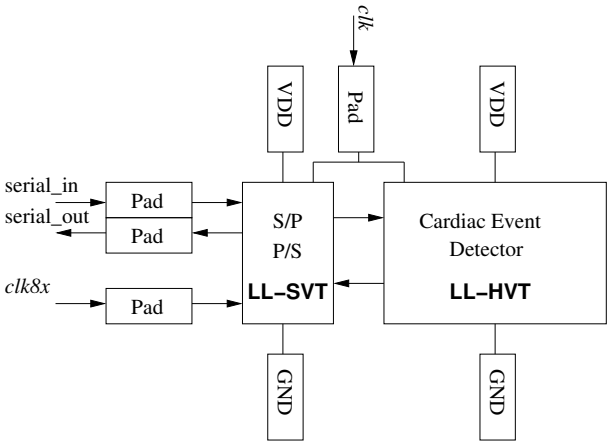


Figure 8. Cardiac event detector and peripherals.

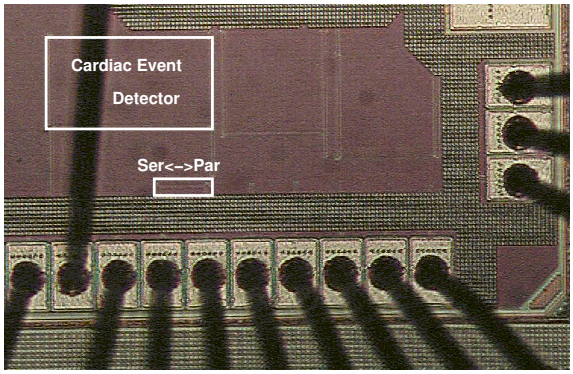


Figure 9. Chip micro-photograph. The area of the cardiac event detector is $19425\mu\text{m}^2$.

V. Test Chip Implementation

This section presents the details of the cardiac event detector test chip implementation for validating the energy model. Both synthesis and silicon implementation details are presented. The design is part of a multi project tape-out, where several different implementations are accommodated on the same pad-limited die.

The cardiac event detector was implemented with a 65 nm LL-HVT standard cell library, using constraints for minimum area and leakage. The gates are supplied by an independent power domain, where the power pads are isolated from any other power source, see Figure 8. The sequential logic is triggered by *clk*. Furthermore, timing is not a design constraint, and therefore, the clock is routed as an ordinary signal. The cardiac event detector consist of 727 registers and 4200 NAND2 equivalent gates occupying an area of $19425\mu\text{m}^2$.

In order to reduce silicon area and overcome pad limitation, both the data to/from the ASIC are serially supplied/sampled, see Figure 8. This is achieved by a module that receives serial input data and converts the bits to 8-bit words (S/P), and concurrently, the output of the ASIC is serialized (P/S). Two clocks, *clk* and *clk8x*, are connected to the module. By serializing the input and output data, the number of pads is reduced from 19 to 8. Moreover, the wordlength of the output is truncated to 8 bits which simplifies clocking, i.e., the module is triggered by a clock that is eight times faster than the the clock that triggers the ASIC. Furthermore, low-leakage standard-V_T (LL-SVT) cells are used in order to be able to drive the load of the pads and external measurement equipment. Thus, no level shifters were required. The serialization module is accommodated on an independent power domain, which allows accurate energy measurements on the cardiac event detector. The chip micrograph is shown in Figure 9, where the serialization module is labeled with `Ser<->Par`.

Table I
Design Characteristics in the Sub-V_T Domain.

Parameter	Value
k_{cap}	17820
k_{crit}	608
k_{leak}	13358
μ_{d}	0.47
μ_{e}	0.29

VI. Model Application and Process Variation Simulations

This section presents how the cardiac event detector is characterized in the sub-V_T domain by application of the energy model. Additionally, simulations on process variations and reliability are performed and results are presented.

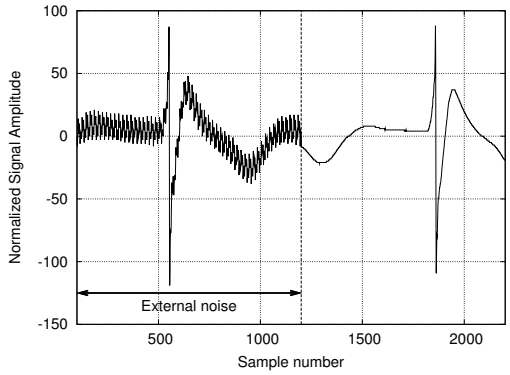
A. Model Application

The model is applied to the cardiac event detector presented in Sec. IV. The characteristics of the design are summarized in Table I, and computed according to the flow presented in Sec. III. All the model based calculations are made using the data in Table I.

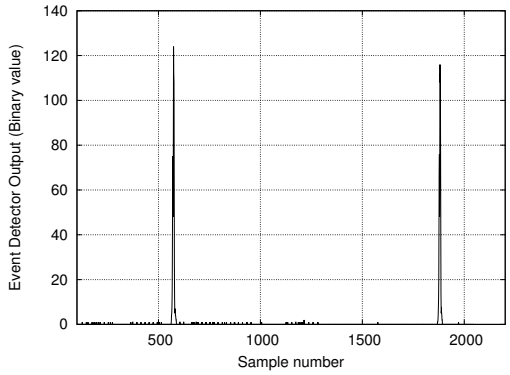
The signal that was supplied to the ASIC input is shown in Figure 10a, and is a typical electrogram that is distorted with noise for 1200 samples. Using a sequence where the input signal is partially distorted is supposed to represent an average use case. The signal in Figure 10b is the post-processed/re-constructed signal at the cardiac event detector output. Running the switch-level simulation by using real-world data, μ_{d} and μ_{e} values are extracted as explained in the previous sections. Both the model based analysis and measurements are made using the same data set.

B. Process Variation in 65 nm CMOS

Sub-threshold circuits are more susceptible to random process variations than their above-threshold counterparts. In some cases, these random process variations cause functional failure. In modern nanometer technologies random dopant fluctuation (RDF) and global process variations cause a shift from the nominal threshold voltage. Due to the fact that the sub-threshold drain current depends on the threshold voltage exponentially, any change in the threshold voltage dominates other process variation effects in the sub-threshold regime [11]. Functional failure of the static CMOS circuits due to random process variation, i.e., threshold voltage variation, may be investigated using the static noise margin values derived



(a) Electrocardiogram



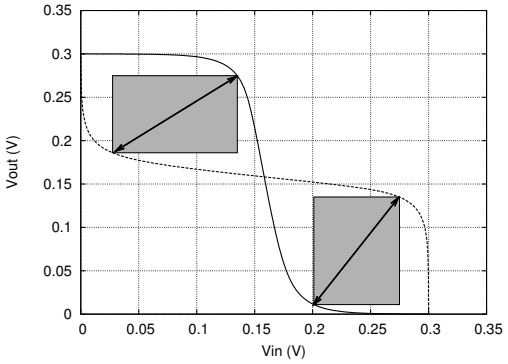
(b) ASIC output

Figure 10. (a) Electrocardiogram fed to ASIC, (b) measured ASIC output data.

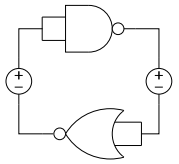
from the voltage transfer characteristic (VTC) curves of digital gates.

To investigate functional failure of digital gates, the used process is simulated following the methodology described in [12], [13], where the static noise margin (SNM) of the SRAM cells is calculated using butterfly plots. Butterfly plots for two gates are formed by superimposing the VTC of one gate over the mirrored VTC of the other. Superimposed VTC curves of an inverter are presented in Figure 11a. A case where the diagonals of the drawn rectangles are non-existent, i.e., at least one of the rectangles cannot be drawn, meaning at least one of the logic levels cannot be sensed by the gate, results in functional failure at the simulated voltage. This formation simulates whether a logic level can be regenerated in case these two successive gates are in a logic path one after another. A sample testbench which includes the static-noise sources is shown in Figure 11b. To simulate the scenario that different gates are connected in back-to-back fashion, testbenches which extract the SNM data automatically are setup according to the methodology in [12]. The SNM square diagonals are calculated by rotating the voltage axes by 45 degrees. In the rotated plane the diagonal of the SNM square for any voltage is calculated by subtracting one VTC curve from the other. If the calculated value is negative, the SNM is negative and the logic value cannot be regenerated.

The SNM failure rates of the gates are extracted from 10k-point Monte Carlo simulations. The simulations are run for supply voltage values which are varied between $0.1\,\mathrm{V}$ and



(a) SNM Curves



(b) SNM testbench

Figure 11. (a) Butterfly SNM curves for an subthreshold inverter with SNM diagonals emphasized and (b) SNM testbench for reliability simulations.

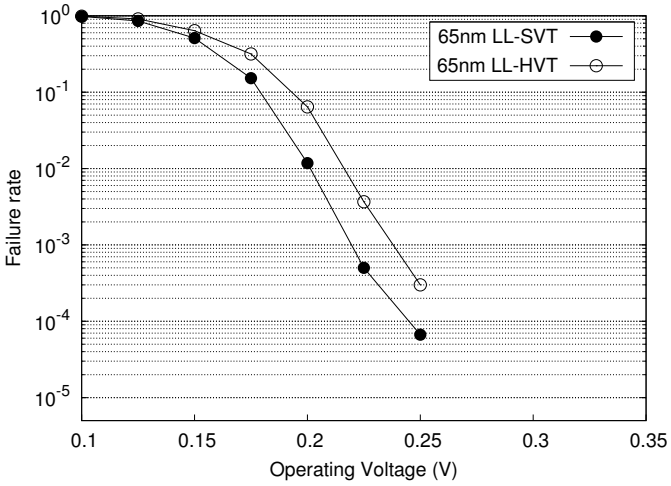


Figure 12. Simulated reliability for 65 nm CMOS.

$0.35\,\mathrm{V}$ with $25\,\mathrm{mV}$ steps. Simulated functional failure rates are presented in Figure 12. Supply voltage values which realize operation with less than 0.001 failure rate are taken as the minimum reliable operating voltage (ROV) in this analysis. It is shown that standard-threshold technology (SVT) has lower failure rates at all supply voltages. SVT technology has a lower failure rate because this technology has a lower threshold voltage and thus the ratio of variation of the threshold voltage of the transistors to the mean threshold voltage value is lower, resulting in less functional variation. Consequently, it is concluded that the design will malfunction if the supply voltage is below ROV.

VII. Model Validation

This section presents the measurement on the reference design, i.e., the direct-mapped cardiac event detector, that was

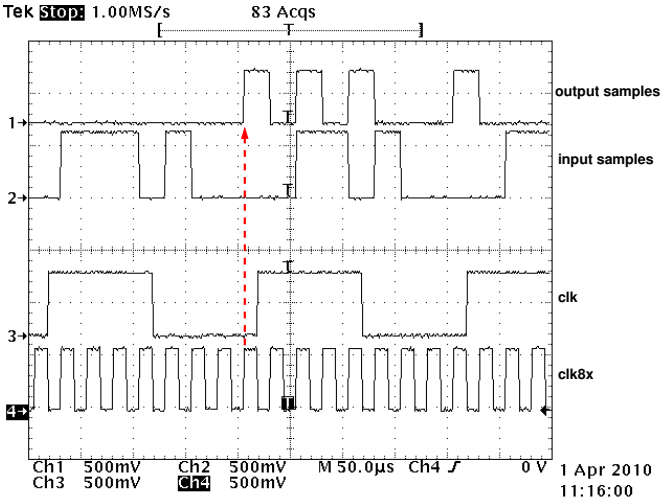


Figure 13. Oscilloscope measurements of the ASIC in- and output signals.

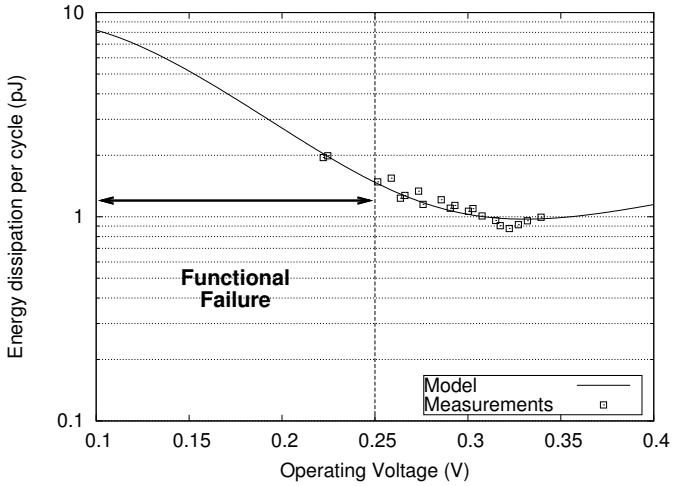


Figure 14. Measured and simulated data of the reference design.

fabricated in 65 nm LL-HVT technology. The measured data is used to evaluate the accuracy of the energy model for energy dissipation and operation speed at different temperatures.

A. Measurement Setup

The measurements are carried out by sweeping both the supply voltage and clock frequency of the cardiac event detector. The former is supplied by a programmable voltage source, and the latter is generated by a XILINX Spartan-3 FPGA. The output of the circuit is monitored by a logic analyzer. Furthermore, the current drawn by the ASIC is measured with an integrator IC, which is accommodated on a custom made printed circuit board (PCB) that supplies the ASIC core. The partially in-house developed energy acquisition equipment is applicable to slowly clocked ($<=$ 10 kHz) sub-threshold ASICs while measuring the ASICs in a sweeping fashion, i. e., frequency-supply voltage grid. However, this limitation in clock frequency is sufficient for the validation of the energy model.

The energy is measured on ASIC samples, which were available in an earlier tapeout [14], [15], where the ASIC cores could be triggered by a synchronous clock or in a self-timed manner depending on the implementation. The measured cores are identical, however, the ASIC in [14] has a sensing transistor in the supply rails. Speed degradation of 40% is due to this transistor and is taken into consideration in our analysis. Energy dissipation per sample is measured by sweeping V_{DD} from 220 to 350 mV, in steps of 10 mV, while clk is increased from 1000 to 10000 Hz. The frequency step size is 1 kHz up to 10 kHz. The supplied clock signals, as well as a sequence of input and output samples are presented in Figure 13. It may be observed that the 8th rising clock edge of *clk8x* occurs before the rising edge of *clk* (dashed arrow). This guarantees that 8 bits are stored in the registers before the 1 kHz clock submits the input sample to the design. The output samples are bitwise fed with *clk8x* by the serialization module, as indicated by the dashed arrow.

The supplied signals from the pattern generator have an amplitude of 550 mV, see Figure 13. These low voltage levels are obtained by external level shifters between the pattern generator pods and the PCB. The amplitude of the output samples is kept at the same level, as the independent power domain of the serialization module is connected to a 550 mV voltage source. Consequently, the captured samples have clear and sharp pulses. The output samples are captured and saved by a logic analyzer, and afterwards, correctness of the signal is approved by post-processing of the output data.

B. Sub-V_T Energy and Failure Rate Measurements

The measured values as well as the data obtained by the energy model are plotted in Figure 14. Simulation data is represented by the solid curve, and measured data is indicated by squares. It is shown that the measured data is in the near vicinity of the simulated data. The mean of the absolute modeling error is calculated as 5.2 %, with a standard deviation of 6.6 %.

At an EMV of 320 mV the cardiac event detector dissipates as little as 0.88 pJ/per sample, operating at a clock speed of 20 kHz. It should be noted that the measurements in Figure 14 are made in a point-by-point fashion, i. e., frequency-voltage pairs generated from the model are used for the setup and energy dissipation is measured.

Furthermore, to compare the maximum operating speed of the implemented circuit to the model, the ASIC output samples are monitored for their correctness. This analysis is carried out for different die temperatures, i,e., 0°, 27°, and 37°. For these measurements, V_{DD} is increased in steps of 10 mV from 250 to 350 mV, while sweeping f from 1 to 10 kHz.

The corresponding failure rate measurement plots are presented in Figure VII-B. Functional correctness of the ASIC is indicated by the white area, whereas the number of wrong samples is indicated by the shaded area. The highest measured failure rate is 1200, which represents the number of samples which were wrongly computed. The simulated clock frequency from the model (18), which is constrained by the propagation

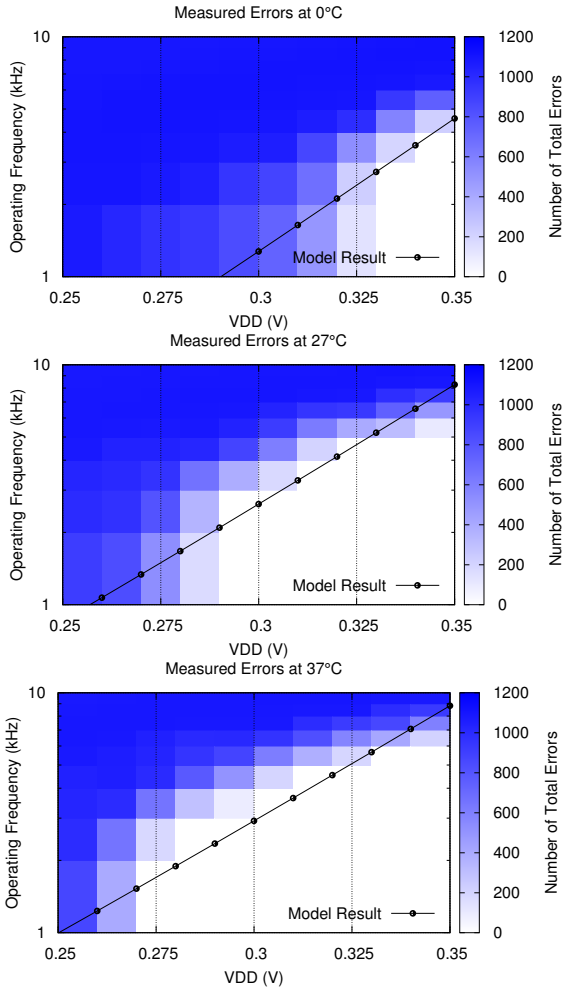


Figure 15. Error measurements for different temperatures with the operation frequency calculation from the model plotted over.

delay of the critical path is plotted as a solid line. To be able to calculate the maximum operating speed for different operating temperatures, I_0 in (18) needs to be extracted for the required temperature values and U_t recalculated. The highest failure is measured when the ASIC operates at 0°, where ROV is above 325 mV. With an higher temperature, ROV is reduced to 290 and 270 mV at 27°, and 37°, respectively. From this observation it becomes evident that the the failure rate decreases while increasing the die temperature. This is expected as the leakage current of a CMOS gate increases with increasing temperature [16], resulting in faster operation according to (18). Moreover, it is shown that simulated maximum frequency (imposed by V_{DD}) matches well with the frequency where the ASIC starts to malfunction. Thus, by these observations, it is shown that the energy model provides reliable simulation data for different ASIC operation temperatures.

As in the maximum operating frequency, energy dissipation of a circuit also varies with changing temperature according to (14). Similar to the maximum frequency calculation, the value of the thermal voltage U_t in equation (10) needs to be updated to be able to retrieve data by simulations. The accuracy of the energy model for different operating temperatures is evaluated

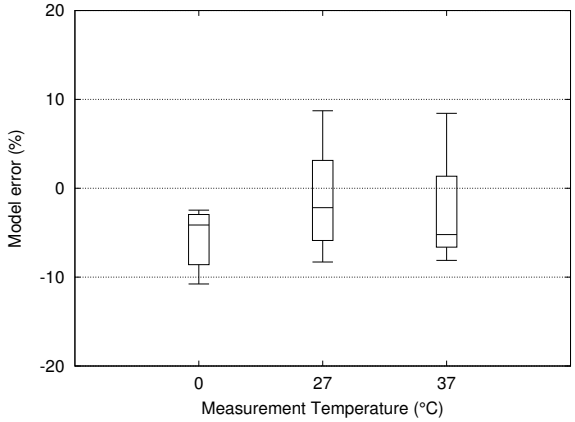


Figure 16. Energy model accuracy for different operating temperatures. For each temperature error plot, energy values and measurements on the energy curve from equation (14) are used.

with respect to measurements and the resulting error box-and-whisker plot is presented in Figure 16, where the modeling error is lower than 10% at different operating temperatures.

VIII. Conclusions

A high level energy estimation flow for ASICs that operate in the sub-threshold domain is proposed. All parameters are directly obtained by applying a traditional synthesis and power estimation flow, and by post-processing data that was generated during synthesis. The model is suitable for design and technology space exploration and is orders of magnitude faster than SPICE simulations. An ASIC in 65 nm LL-HVT technology for cardiac pacemaker application was fabricated. The accuracy of the model is validated by ASIC measurements under various operating conditions, i.e., V_{DD}, f_{clk}, and temperature. It is shown that the proposed model is able to accurately simulate the energy dissipation and maximum operating frequency of the implemented design.

References

[1] A. Chandrakasan, S. Sheng, and R. Brodersen, "Low-power CMOS digital design," *Solid-State Circuits, IEEE Journal of*, vol. 27, no. 4, pp. 473–484, 1992.

[2] E. Vittoz, *Low-Power Electronics Design*. CRC Press LLC, 2004, ch. 16.

[3] A. Wang and A. Chandrakasan, "A 180-mV subthreshold FFT processor using a minimum energy design methodology," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 1, pp. 310–319, 2005.

[4] J. P. Kulkarni, K. Kim, and K. Roy, "A 160 mv robust schmitt trigger based subthreshold SRAM," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 10, pp. 2303–2313, Oct. 2007.

[5] 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*. ACM New York, NY, USA, 2004, pp. 868–873.

[6] B. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and sizing for minimum energy operation in subthreshold circuits," *Solid-State Circuits, IEEE Journal of*, vol. 40, no. 9, pp. 1778–1786, 2005.

[7] R. Corless, G. Gonnet, D. Hare, D. Jeffrey, and D. Knuth, "On the LambertW function," *Advances in Computational Mathematics*, vol. 5, no. 1, pp. 329–359, 1996.

[8] A. Robbins, *Effective awk programming*. O'Reilly & Associates, Inc. Sebastopol, CA, USA, 2001.

[9] 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,*, vol. 52, no. 12, pp. 2686–2698, Dec. 2005.

[10] M. Åström, S. Olmos, and L. Sörnmo, "Wavelet-based event detection in implantable cardiac rhythm management devices," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, March 2006.

[11] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, "Analysis and mitigation of variability in subthreshold design," in *Proceedings of the 2005 international symposium on Low power electronics and design*. ACM New York, NY, USA, 2005, pp. 20–25.

[12] E. Seevinck, F. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *IEEE Journal of Solid-State Circuits*, vol. 22, no. 5, pp. 748–754, 1987.

[13] 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.

[14] O. C. Akgun, J. Rodrigues, and J. Sparsø, "Minimum-energy subthreshold self-timed circuits: Design methodology and a case study," in *Proceedings of 16th IEEE International Symposium on Asynchronous Circuits and Systems (ASYNC) 2010*, 2010, pp. 41–51.

[15] J. Rodrigues, O. C. Akgun, and V. Owall, "A $<$1 nJ Sub-VT Cardiac Event Detector in 65 nm LL-HVT CMOS," in *Proceedings of 18th IEEE/IFIP International Conference on VLSI and System-on-Chip (VLSI-SOC), in press*.

[16] C. Enz, F. Krummenacher, and E. Vittoz, "An analytical mos transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications," *Analog Integrated Circuits and Signal Processing*, vol. 8, pp. 83–114, July 1995.