Unified V_{dd} - V_{th} optimization based DVFM controller for a Logic block

Kannan S A
akannan@cedt.iisc.ernet.in

Sreeram N S
nssreeram@cedt.iisc.ernet.in

Bharadwaj S Amrutur
amrutur@ece.iisc.ernet.in

Department of ECE, Indian Institute of Science, Bangalore, India - 560012

Abstract

In this paper analytical expressions for optimal V_{dd} and V_{th} to minimize energy for a given speed constraint are derived. These expressions are based on the EKV model for transistors and are valid in both strong inversion and sub threshold regions. The effect of gate leakage on the optimal V_{dd} and V_{th} is analyzed. A new gradient based algorithm for controlling V_{dd} and V_{th} based on delay and power monitoring results is proposed. A V_{dd}-V_{th} controller which uses the algorithm to dynamically control the supply and threshold voltage of a representative logic block (Sum of Absolute Difference computation of an MPEG Decoder) is designed. Simulation results using 65 nm predictive Technology models are given.

1. Introduction

Dynamically varying frequency of a logic block based on load requirements is effective in minimizing the energy consumptions. It has been shown that optimal supply and threshold voltages exist which minimizes the power consumed while meeting the speed constraint [1] [2] [3]. Most of the earlier attempts to derive the optimal V_{dd}, V_{th} focus either on the strong inversion region or the sub threshold region of operation. But in a generic case where the frequency specifications of a circuit has a wide dynamic range operations in both regions may be necessary. In this paper delay and power models based on the EKV model [4] for the transistors are developed from which we obtain the optimal V_{dd} and V_{th} solutions which are valid in both the regions.

In [5] delay and power monitoring schemes for dynamically controlling V_{dd}, V_{th} is proposed. The algorithm used in [5] varied V_{dd} based on the delay monitoring results and V_{th} based on power monitoring results. This is not satisfactory as the delay depends on V_{th} also and power has a V_{dd} dependency also. A control algorithm which varies V_{dd} and V_{th} in a combined fashion based on the delay and power will give a faster convergence and more accurate results. So a combined V_{dd}-V_{th} control algorithm based on the sensitivities of delay and power to V_{dd} and V_{th} is discussed in the paper. These sensitivities are obtained from the unified models developed.

The paper is organized as follows. Section 2 describes the analytical solution for optimal V_{dd}, V_{th} based on the EKV model. Section 3 describes the proposed V_{dd}-V_{th} control algorithm, delay and power monitoring circuits and the DVFM controller simulation results. Section 4 concludes the paper.

2 Unified V_{dd} - V_{th} optimization based on EKV model

2.1 Proposed EKV based current model

The V_{dd}-V_{th} optimization for a given target frequency has been done in [1] for above threshold operation and in [2] [3] for sub threshold operation. The transistor current models in these work were based on the classical alpha power law model. Fig.1 shows the comparison of the SPICE simulated I_d values and the alpha power law model fitted values. There are two issues in using this model for V_{dd}-V_{th} optimization.

(i) It uses two different equations to model the current. A power law model for strong inversion region and an exponential model for subthreshold region. So V_{dd}-V_{th} optimizations has to be done separately for these two regions.

(ii) The model does not fit well for regions close to V_{th}.

To overcome these difficulties we chose to use a transistor current model based on the EKV model. The proposed model is given by

$$I_d = I_0[ln(1 + e^{\frac{V_{gs}-V_{th}}{\alpha N_s}})]^{\alpha} \quad (1)$$

For high V_{gs} and low V_{gs} the proposed model reduces to the alpha power law model. But for V_{gs} values close to V_{th} it uses an interpolation function to predict the transistor current. This single equation closely follows the measured

1063-9667/08 $25.00 © 2008 IEEE
DOI 10.1109/VLSI.2008.69

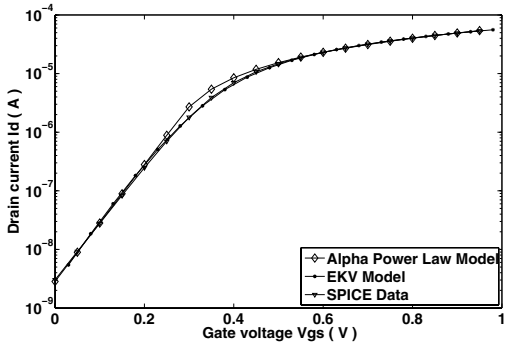


Figure 1. I_d - V_{gs} plot for 65nm transistor of w/l = 65nm/65nm at V_{ds} = 1V.

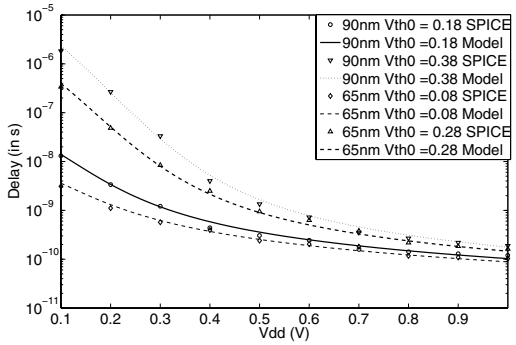


Figure 2. Comparisons of proposed model and SPICE data for 90nm and 65nm processes. Delays shown are that of a ring oscillator in the test circuit.

values in all regions with an maximum error of around 15%. Fig.1 shows the comparison between the spice simulated transistor currents and that given by (1).

2.2 Test circuit

To evaluate the validity of the proposed models a test circuit given in [8] is used. The circuit has a ring oscillator made of 11 NAND gates feeding chains of NAND gates. All the SPICE results given in this section have been obtained by the simulations of this test circuit.

2.3 Delay model

A simple delay model for a gate is given by

$$t_{gate} = \frac{KC_L V_{dd}\theta}{I_0[ln(1+e^{\frac{V_{dd}-V_{th0max}+\eta V_{dd}}{\alpha N_s}})]^{\alpha}} \quad (2)$$

where N_s is the subthreshold slope, V_{thmax} is the maximum V_{th} across temperature and process corners, C_L is the load of the gate and K is a delay fitting parameter. θ is a degradation factor to account for increase in delay due to finite rise/fall times. θ is a function of V_{dd}, V_{th}. Through simulations the degradation due to slope was found and modelled as a function V_{dd}, V_{th} as

$$\theta = \frac{1}{[ln(1+e^{\frac{V_{dd}-V_{th0max}+\eta V_{dd}}{\alpha N_s}})]^{\beta}} \quad (3)$$

β is a fitting parameter. Fig.2 shows the accuracy of the model in predicting the delay across a wide range of V_{dd} for 65nm(PTM) [9] and 90nm(UMC) processes. Maximum discrepancies of around 17% were seen across a wide range of V_{dd} and V_{th}.

2.4 Power model

Dynamic power is given by

$$P_{dyn} = aC_{Leff}V_{dd}^2 f \quad (4)$$

Where C_{Leff} is the effective switching capacitance and ais the activity factor. In this work the contribution of short circuit power dissipation has been neglected for simplicity as it is a small fraction of total power [1]. Leakage current in deep submicron devices is predominantly due to subthreshold leakage and gate leakage. In this optimization the gate leakage current is neglected. However in a subsequent section the validity of this assumption and its limitations are discussed. Fig.3 shows the comparison of leakage current of the test circuit using 65nm PTM process with the corresponding models.

$$I_{leak} = I_0 W_{eff} e^{\frac{-(V_{th0min}-\eta V_{dd})}{N_s}} \quad (5)$$

$$P_{leak} = I_0 W_{eff} e^{\frac{-(V_{th0min}-\eta V_{dd})}{N_s}} V_{dd} \quad (6)$$

$$V_{th0min} = V_{th0max} - \Delta V_{th0} \quad (7)$$

ΔV_{th0} is peak-to-peak V_{th0} variation across process corners and temperature. V_{th0min} is lowest V_{th0} in operation temperature and process variation range [1]. It is to be noted that delay is calculated at V_{th0max} and leakage power at V_{th0min} as in [1]. The total power is given by

$$P_t = P_{leak} + P_{dyn} \quad (8)$$

2.5 V_{dd} - V_{th} optimization

Using above models optimization for minimum energy is done. solving (2) for V_{thmin} in terms of V_{dd}

$$V_{th0min} = (1+\eta)V_{dd} - \alpha N_s ln(e^{\chi V_{dd}^{\frac{1}{\gamma}}} - 1) - \Delta V_{th0} \quad (9)$$

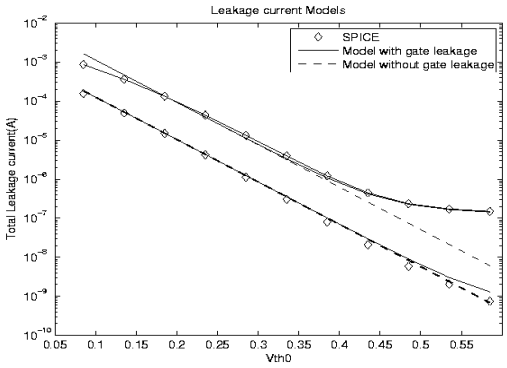


Figure 3. Leakage current of the test circuit. Comparison of SPICE results, equation(5) and equation(14).

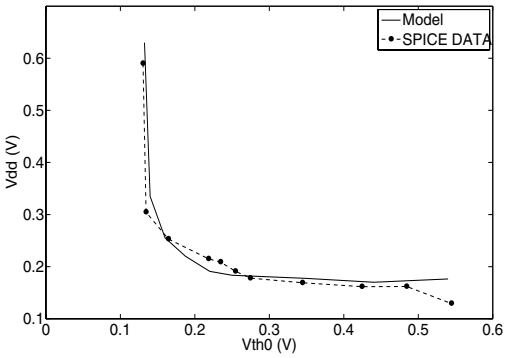


Figure 4. Optimal energy points for various frequencies on the V_{dd}-V_{th} plane for 65nm PTM process. Model and SPICE Data comparison.

Where $\gamma = \alpha + \beta$. Using (6), (9) leakage power is given by

$$P_{leak} = I_0 W_{eff} e^{\frac{-((1+\eta)V_{dd} - \Delta V_{th0})}{N_s}} (e^{\chi V_{dd}^{\frac{1}{\gamma}}} - 1)^{\alpha} V_{dd} \quad (10)$$

where

$$\chi = \left(\frac{K C_L f L_d}{I_0}\right)^{\frac{1}{\gamma}} \quad (11)$$

the optimum values are given by solving

$$\frac{dP_t}{dV_{dd}} = 2aC_{Leff} V_{dd} f +$$

$$P_{leak}\{\frac{1}{V_{dd}} - \frac{1+\eta}{N_s} + \frac{\alpha\chi V_{dd}^{\frac{1}{\gamma}-1} e^{\chi V_{dd}^{\frac{1}{\gamma}}}}{\gamma(e^{\chi V_{dd}^{\frac{1}{\gamma}}} - 1)}\} = 0 \quad (12)$$

V_{ddopt} is obtained by solving (12) for V_{dd}. Equation (12) is transcendental and is solved numerically. V_{thopt} obtained by substituting V_{ddopt} in (9).

Fig.4 compares the optimal V_{dd}/V_{th} values obtained using the models with those obtained from exhaustive SPICE simulation. The model and SPICE values are very close, further validating the models.

2.6 Discussion on the optimization results

(i) In the strong inversion region V_{thopt} remains almost constant with frequency whereas V_{dd} varies to achieve the necessary speed whereas in the subthreshold region the V_{ddopt} remains almost constant and V_{thopt} varies with frequency [3] [2] [1].

(ii) The optimal energy points occur where the leakage energy and the dynamic energy are comparable. [1] report that optimal energy point occurs when the ratio of dynamic to leakage power remains constant. This is true both in the strong inversion and the sub threshold regions but the optimality occurs for different ratios in these two regions. If a single ratio is used for the whole range of frequencies, the power level for some range of frequencies will be sub optimal. In fig.5 ratio of energy along the optimal curve to that along the curve Edyn/Eleak = constant is plotted for different delays. The sub optimality is within 10% as shown in fig.5.

2.7 Effect of gate leakage on the optimal curve

We model Gate leakage of a circuit empirically as

$$I_{gateleak} = I_{0gate} * V_{dd}^{k1} * e^{-k2 * V_{th0min}} \quad (13)$$

$$I_{leak} = I_{gateleak} + I_{subleak} \quad (14)$$

Where k1, k2 are fitting parameters.

Fig.3 shows the accuracy of the empirical fit for 65nm process. Gate leakage is strongly dependent on V_{dd} and increases with V_{dd}. However in the strong inversion region, optimal V_{th} is very low and hence subthreshold leakage dominates. In the sub-threshold regions V_{dd} is very low and hence gate leakage is again negligible. Hence gate leakage term affects the optimal curve very little. When the frequency is extremely low and optimal $V_{th}0$ is very high the subthreshold leakage falls even below the gate leakage. Only for such extremely low frequencies gate leakage tends to affect the optimal V_{dd} - V_{th} curve.

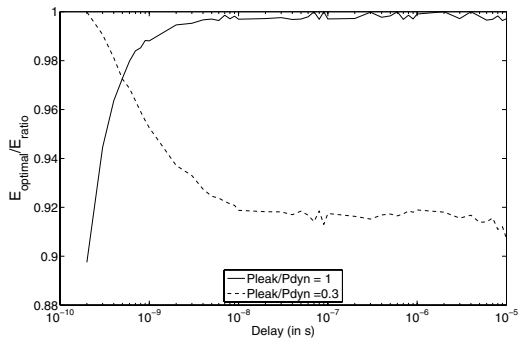


Figure 5. Suboptimality in choosing constant Eleak/Edyn ratio as the minimum energy criteria.

3 DVFM controller

A DVFM controller based on delay and power measurement results is shown in fig.6(a). It adjusts the supply and body bias to achieve minimum power for a frequency target. The control algorithm uses the measured values of delay and power to adjust V_{dd} and body bias and is described next.

3.1 Gradient based V_{dd} - V_{th} control algorithm

In [5] V_{dd} is controlled based on the results of the delay measurement and V_{th} is varied based on the results of power measurements. V_{dd} and V_{th} affects both delay and power and hence a combined control of V_{dd} and V_{th} based on the power and delay measurements will be more effective. We propose a new control algorithm below.

The difference between targeted delay(Dtarget) and the actual circuit delay(Dactual) is obtained. Also the ratio of Leakage Energy to Dynamic Energy(Eleakratio) is obtained and its difference from the optimal Energy ratio (Eratioptimal) is computed.

$$delD = Dtarget - Dactual \quad (15)$$

$$delEratio = Eratiotarget - Eratioactual \quad (16)$$

The Sensitivity matrix is defined as

$$S = [\ \frac{\partial D}{\partial V_{dd}}\ \ \frac{\partial D}{\partial V_{th}}\ \ \frac{\partial Eratio}{\partial V_{dd}}\ \ \frac{\partial Eratio}{\partial V_{th}}] \quad (17)$$

We can write

$$[delD\ delEratio]^T = S * [delvdd\ delvth]^T \quad (18)$$

Then the direction in which we should move in the V_{dd}-V_{th} space to reach the optimal point is computed by inverting the above equation.

$$[delvdd\ delvth]^T = R * [delD\ delEratio]^T \quad (19)$$

where R is related to the inverse of the sensitivity matrix S.

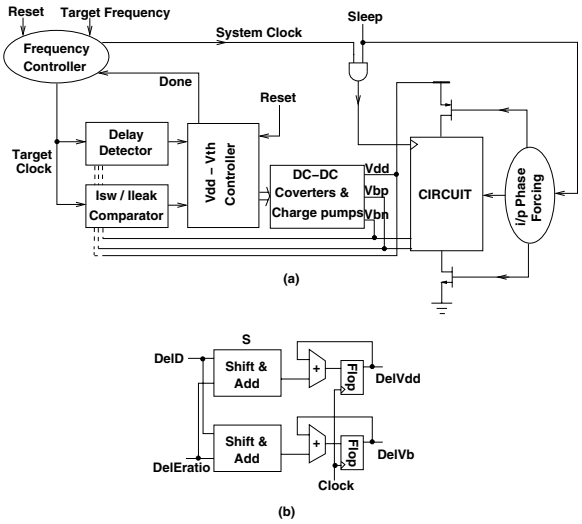


Figure 6. (a)Block Diagram of DVFM controller. (b)Proposed V_{dd} - V_{th} controller.

The matrix multiplication in (19) is implemented in a shift and add fashion as shown in fig.6(b). Choosing the elements of R is critical as they control the dynamics and stability of the loop. The following heuristic steps are developed to avoid oscillations in the proposed algorithm.

(i) Given a range of frequencies the optimal V_{dd} - V_{th} points are estimated using the delay and power models given in equations (2) and (8). The range of V_{dd} and V_{th} is thus known.

(ii) The sensitivity matrix S is evaluated using (17) at each point in this operational range. Then S^{-1} is computed. R(i,j) is then computed as minS^{-1}(i,j) over the operational range.

The performance of the proposed algorithm was evaluated using MATLAB simulations. The operation of the algorithm to control the V_{dd} and V_{th} of test circuit for a frequency range on 100M - 1G was simulated. The initial values of V_{dd} and V_{th} was chosen to be the optimal V_{dd} - V_{th} for a frequency of 2GHz. A step decrease in frequency is then given and the number of iterations to converge was found for the proposed algorithm and the algorithm in [5]. The proposed algorithm is faster by more than 50% than the previous work as shown in fig.7. In a DVFM

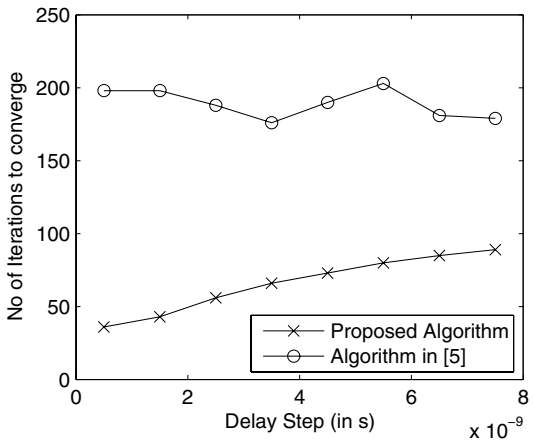


Figure 7. Performance comparison of the proposed algorithm and algorithm in [5].

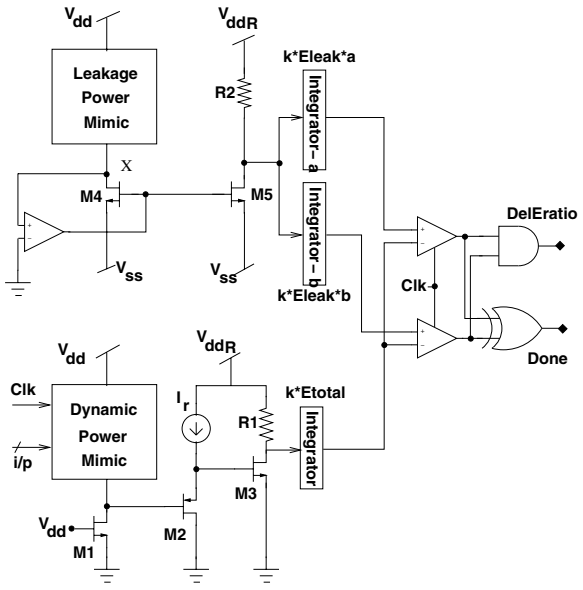


Figure 8. $I_{leak} - I_{sw}$ comparator.

controller each iteration must be followed by a change in control to the power controller(to change V_{dd} and V_{th}). In algorithm proposed in [5] in each iteration either V_{dd} or V_{th} is changed. So the convergence time is

Tc = No.of V_{dd} iterations*time taken per V_{dd} step + No.of V_{th} iterations*time taken per V_{th} step

In the proposed algorithm V_{dd} and V_{th} are changed simultaneously. So the convergence time is

Tc= No.of iterations * Max(time taken per V_{dd} step, time taken per V_{th} step)

Time taken per V_{dd} and V_{th} step depends on the response time of the power regulators, it can be in the order of microseconds. Hence in [5] algorithm even if equal no.of V_{dd} and V_{th} steps are assumed, the proposed algorithm performs better.

3.2 Architecture of the Logic block

The logic block to compute the sum of sixteen 16 bit numbers is similar to the SAD computation architecture used in [6]. It uses 6 stages of carry save adders and the final carry propagation is done using a Hans Carlson adder. A pipelining of 12 FO4 delay per stage found to be optimal [7] is used.

3.3 Delay and Power monitoring Circuits

The Delay monitoring circuit given in [5] is used. The monitoring circuit to measure the total energy consumed by the circuit is given in fig.8. The SAD computing structure is highly regular and is composed of adders, flops and wires only. Due to the regularity of the structure we believe that the power mimic made of adders, flops and wires in a proportionate ratio should mimic the total circuit power across wide range of V_{dd} and V_{th} satisfactorily. The mimic used is shown in fig.9. The voltage drop across the sleep transistor is proportional to the current flowing through it and hence can serve as a measure of the instantaneous current flow. The total energy measurement is done by switching the inputs of the power mimic using a counter and integrating the drop across the sleep transistor over a few cycles. Since the voltage across the sleep transistor is close to ground a PMOS level shifter is used to shift the voltage up by V_{th}. The shifted voltage is amplified and integrated to obtain a voltage proportional to the total energy consumed. The leakage power measurement circuit proposed in [5] is used. The Op-Amp shown maintains the node x at ground potential. The leakage current of the mimic flows through M4 which is mirrored to M5 producing a proportional voltage across R2. This voltage is then integrated and compared with the total energy using a clocked comparator to generate $delEratio$. A small window for comparison is provided with scaling factors aand b. The circuit proposed uses a negative voltage V_{ss} of -0.75V. The linearity of the total energy and leakage energy are shown in fig.10(a) and fig.10(b).

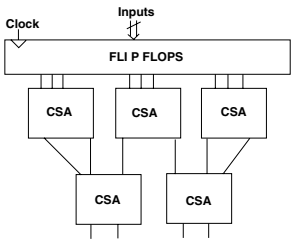
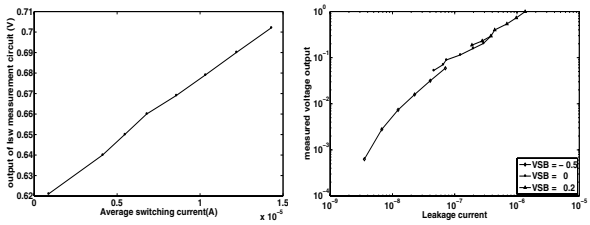
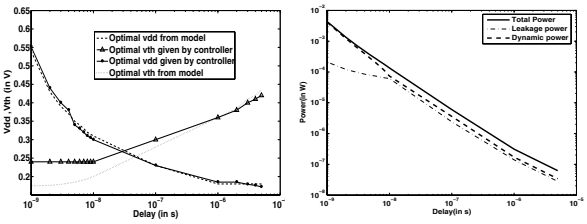


Figure 9. Power Mimic.



(a) Total Energy. (b) Leakage Energy.

Figure 10. Linearity of Measurements



(a) V_{dd}-V_{th} points. (b) Power points from SPICE.

Figure 11. Optimal Results.

3.4 DVFM Controller Simulation results

The DVFM controller with the measurement circuits was simulated with behavioral models for all the digital components of the controller. 65nm PTM model is used to build the logic block and the measurement circuits. An ideal DC-DC converter with voltage resolution of around 10mV was assumed in the behavioral model. The target frequencies Vs the optimal V_{dd}-V_{th} points are given in fig.11(a) and the power consumed by the block at different frequencies is given in fig.11(b). In the super threshold region the optimal V_{th} turns out to be lesser than the V_{th} obtainable by forward body bias. So the controller settles down the minimum obtainable V_{th} and V_{dd} to maintain the delay constraint. That is the controller settles to the reachable optimal point in the V_{dd}-V_{th} plane. The power consumed by the circuit for various frequencies at the optimal V_{dd}-V_{th} given by the controller is given in fig.11(b). The leakage power and the dynamic power are comparable indicating that the optimal points given by the controller are indeed close to the minimal power points.

4 Conclusions

Analytical expressions for optimal V_{dd}-V_{th} valid in both strong inversion and sub threshold regions has been derived. Gate leakage is found to affect the optimal V_{dd}-V_{th} curve very little except for very low frequencies. Minimum energy operation for a given frequency occurs when the leakage energy and the dynamic energy are comparable to each other and this criteria holds quite well for a wide range of frequencies. V_{dd}-V_{th} controller based on a new control algorithm has been designed and simulated using behavioral models. The proposed algorithm performs better than the previously reported algorithms in terms of convergence speed.

References

[1] K. Nose and T. Sakurai. Optimization of V_{dd} and V_{th} for Low-power and high-speed applications.*Proc. Asia South Pacific Design Automation Conference*, pp. 469-474, Jan.2000.

[2] A. Wang, A.P. Chandrakasan and S.V. Kosonocky. Optimal supply and threshold scaling for subthreshold CMOS circuits. *Proc. IEEE Computer Society Annual Symposium on VLSI* 2002, pp. 5-9.

[3] Benton H. Calhoun and Anantha Chandrakasan. Characterizing and Modeling Minimum Energy Operation for Subthreshold Circuits. *Proceedings of the 2004 International Symposium on Low Power Electronics and Design* 2004, pp. 90-95.

[4] Christian.C.Enz, Francois Krummeenacher and Eric A.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* 1995, pp. 81-114.

[5] Masahiro Nomura et al. Delay and power monitoring schemes for minimizing power consumption by means of supply and threshold voltage control in active and standby modes. *IEEE Journal of solid-state circuits*,41(4):805-814, April 2006.

[6] J. Vanne, E. Aho, T.D. Hamalanen and K. Kuusilinna. A High-Performance Sum of Absolute Difference Implementation for Motion Estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(7):876-883, July 2006.

[7] Kim et al. Total power-optimal pipelining and parallel processing under process variations in nanometer technology *IEEE/ACM International Conference on Computer-Aided Design*, November 2005, pp. 535-540.

[8] Dejan Markovic, Mark Horowitz et al. Methods for true energy performance optimization. *IEEE Journal of solid-state circuits*, 39(8):1282-1293, August 2004.

[9] PTM. *http://www.eas.asu.edu/ptm/.*