Characterizing and Modeling Minimum Energy Operation for Subthreshold Circuits

Benton H. Calhoun and Anantha Chandrakasan

Massachusetts Institute of Technology, Cambridge, MA, 02139

{bcalhoun, anantha}@mtl.mit.edu

Abstract

Subthreshold operation is emerging as an energy-saving approach to many new applications. This paper examines energy minimization for circuits operating in the subthreshold region. We show the dependence of the optimum V_{DD} for a given technology on design characteristics and operating conditions. Solving equations for total energy provides an analytical solution for the optimum V_{DD} and V_T to minimize energy for a given frequency in subthreshold operation. SPICE simulations of a 200K transistor FIR filter confirm the analytical solution and the dependence of the minimum energy operating point on important parameters.

Categories and Subject Descriptors

B.6.3 [**Design Aids**] - optimization

General Terms

Performance, Design, Reliability

Keywords

Subthreshold circuits, Subthreshold model, Minimum energy point, Energy model

1. INTRODUCTION

The increasing attention on power consumption in circuit design has motivated a significant investigation of the optimum design for minimizing energy or power. Almost all of these efforts have targeted high performance, strong inversion operation. Emerging applications such as distributed sensor networks or medical applications have low energy as the primary concern instead of performance, with the eventual goal of harvesting energy from the environment (e.g. [1]). Minimum energy operation for low performance situations occurs in the subthreshold region [2][3]. This paper analyzes energy minimization for circuits operating in the subthreshold region. We determine the optimum supply voltage, V_{DD}, for a fixed threshold voltage, V_T, when minimizing energy, not performance, is the constraint and show its dependence on design characteristics, operating conditions, and architecture. We also provide an analytical solution for V_{DD} and V_T to minimize energy for a given frequency.

Many previous works address power minimization for a given performance constraint. Early solutions for minimizing the Energy-Delay-Product (EDP) [3] are extended to account for process variation [4] and buffering options [5]. Measurements

of a test chip with adaptive supply and body bias display a minimum power point for a given performance and show how forward-biased diode currents (from body biasing) can make the theoretical optimum unreachable [6]. Analytical expressions for the optimum (V_{DD}, V_T) point to minimize power at a given performance are shown for transregional models based on fitted [7] or physical [8] parameters. Derivations of the sensitivities of energy and delay to different parameters support a methodology for building optimum energy circuits [9]. Taken together, these and other works give thorough attention to power optimization for strong inversion circuits. Optimizing subthreshold circuits has received less attention.

Subthreshold operation is currently used for some low power applications such as watches [10] and hearing aids [11]. Emerging ultra-low power applications such as distributed sensor networks are a natural fit with subthreshold circuits. Special circuit techniques for improving robustness in deep subthreshold have been explored [12][13]. Examining the energy-delay contours over V_{DD} and V_T shows that minimum energy operation occurs in the subthreshold operation regime for low-to-medium performance systems, and the optimum point changes depending on activity factor and threshold variation [2].

2. MODELING FOR SUBTHRESHOLD OPERATION

This section describes the models used for subthreshold energy minimization. The optimization methods mentioned above all use (1) as the basic equation for modeling subthreshold current and total off current.

$$I_{sub} = I_o e^{\frac{V_{GS} - V_T}{n V_{th}}} \tag{1}$$

I_o is the drain current when $V_{GS} = V_T$, n is the subthreshold slope factor (1 is ideal), and V_{th} is the thermal voltage, kT/q. Some of the citations supplement this model to account for DIBL or for low V_{DS} roll-off or both.

None of the strong inversion optimization work accounts for gate leakage even though gate leakage contributes significantly to total leakage in deep submicron technologies. Figure 1 shows that gate current comprises 40% of off current for a 90nm inverter at nominal V_{DD}. Since gate current depends strongly on V_{DD}, it becomes insignificant for subthreshold operation, so the following analysis justifiably equates total current to subthreshold current.
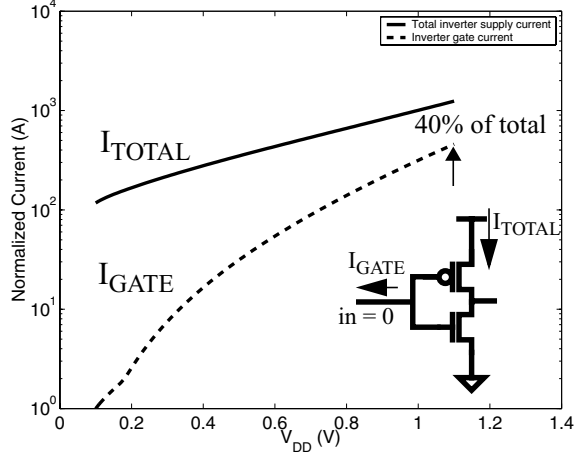
Figure 1. Contribution of gate leakage to total inverter off current (simulation) in 90nm process (in = 0) is 40% of total current at nominal V_{DD}, but diminishes rapidly with V_{DD}.

2.1 Solving for the Minimum Energy Point

In this section, we derive a closed form solution for the optimum V_{DD} and V_T for a given frequency and technology operating in the subthreshold regime ($V_T > V_{DD}$). Equation (2) shows the propagation delay of a characteristic inverter with output capacitance C_g in subthreshold:

$$ t_d = \frac{KC_g V_{DD}}{I_{o,g} e^{\frac{V_{GS} - V_{T,g}}{N_S}}}, \quad (2) $$

where K is a delay fitting parameter, and $N_S = nV_{th}$. The expression for current in the denominator of (2) models the on current of the characteristic inverter, so it accounts for transitions through both NMOS and PMOS devices. Thus, the terms $I_{o,g}$ and $V_{T,g}$ are fitted parameters that do not correspond exactly with the MOSFET parameters of the same name. Operational frequency is simply:

$$ f = \frac{1}{t_d L_{DP}}, \quad (3) $$

where L_{DP} is the depth of the critical path in characteristic inverter delays. Dynamic (E_{DYN}), leakage (E_L), and total energy (E_T) per cycle are expressed in (4)-(6), assuming rail-to-rail swing ($V_{GS} = V_{DD}$ for "on" current).

$$ E_{DYN} = C_{eff} V_{DD}^2 \quad (4) $$

$$ E_L = W_{eff} I_{o,g} e^{\frac{-V_{T,g}}{N_S}} V_{DD} t_d L_{DP} = W_{eff} KC_g L_{DP} V_{DD}^2 e^{-\frac{V_{DD}}{N_S}} \quad (5) $$

$$ E_T = E_{DYN} + E_L = V_{DD}^2 \left(C_{eff} + W_{eff} KC_g L_{DP} e^{-\frac{V_{DD}}{N_S}} \right) \quad (6) $$

Equations (4)-(6) extend the expressions for current and delay of an inverter to arbitrary larger circuits. This extension sacrifices accuracy for simplicity since the fitted parameters cannot account for all of the details of every circuit. Thus, C_{eff} is the average total switched capacitance of the entire circuit, including the average activity factor over all of its nodes. Likewise, W_{eff} estimates the average total width that

contributes to leakage current. Treating this parameter as a constant ignores the state dependence of leakage. Solving this set of equations provides a good estimate of the optimum for the average case and shows how the optimum point depends on the major parameters. Differentiating (6) and equating to 0 allows us to solve for V_{DDopt}:

$$ \frac{\partial E_T}{\partial V_{DD}} = 2C_{eff} V_{DD} + 2KC_g L_D W_{eff} V_{DD} e^{\frac{-V_{DD}}{N_S}} $$
$$ + \frac{-KC_g L_D W_{eff} V_{DD}^2}{N_S} e^{\frac{-V_{DD}}{N_S}} = 0 \quad (7) $$

The following equation gives an analytical solution for V_{DDopt} (see Appendix regarding the Lambert W function and its constraints):

$$ V_{DDopt} = N_S(2 - lambertW(\beta)), \quad (8) $$

$$ \beta = \frac{-2C_{eff}}{W_{eff} KC_g L_{DP}} e^2 > -e^{-1}. \quad (9) $$

Now, substituting (2) into (3) gives V_{Topt} for a given f:

$$ V_{Topt} = V_{DDopt} - N_S \ln\left(\frac{fKC_g L_{DP} V_{DDopt}}{I_{o,g}} \right). \quad (10) $$

Equations (8) and (10) give the optimum supply voltage and threshold voltage for subthreshold circuits consuming the minimum energy for a given frequency. Some ultra-low power applications, such as energy scavenging sensor nodes, might consider minimizing energy to be more important than any performance requirement. Assuming a standard technology where V_T is fixed (i.e. - no triple wells for body biasing), the problem becomes finding the optimum V_{DD} and frequency to minimize energy for a given design. The optimum V_{DD} for minimizing energy per cycle in this scenario still is given by (8), and the optimum frequency is given by (3).

Equation (8) shows that the optimum V_{DD} value is independent of frequency. Instead, it is set by the relative significance of dynamic and leakage energy components as expressed in equation (9). For a single inverter with activity factor of one, C_{eff}, W_{eff}, and L_{DP} all equal one, and β does not meet the constraint in equation (9). Mathematically, this means that the derivative of E_T never equals zero. Physically, the leakage component for the single inverter with high activity remains insignificant compared to dynamic energy over all supply voltages. The true optimum V_{DD} in this case is the lowest voltage for which the circuit functions.
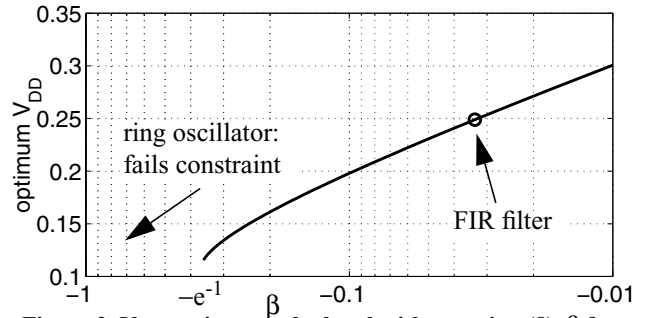

Figure 2. V_{DD} optimum calculated with equation (8). β for ring oscillator (L_{DP}=20) fails constraint. β for 8x8 parallel FIR filter also shown.
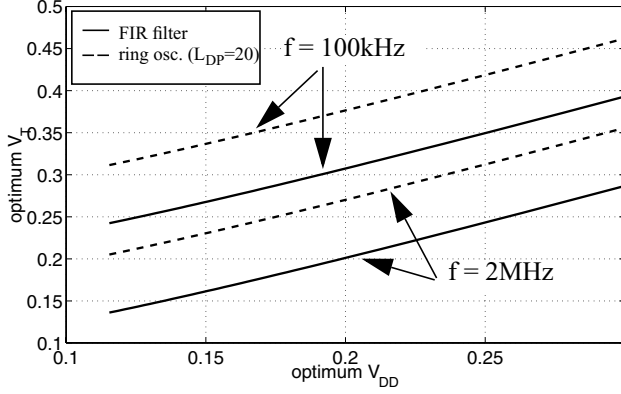
Figure 3. V_T optimum calculated with equation (10) for different design, frequency pairs. The 2MHz design for the FIR filter produces V_T at or below V_{DD}, so the assumption of subthreshold operation becomes invalid.

The leakage energy of a design increases compared with the average inverter in two ways. First, the ratio of C_{eff}/W_{eff} decreases, indicating that a greater fraction of the total width is idle and thereby drawing static current without switching. Secondly, L_{DP} can increase. The larger resulting period gives more time for leakage currents to integrate, raising the leakage energy component. Figure 2 shows the optimum V_{DD} values versus β. For calibration, two example circuits are shown on the plot. An 8-bit, 8-tap parallel FIR filter has a β value that gives an optimum V_{DD} at 250mV. In section 3, we compare this analytical result with simulation. The figure also shows that a ring oscillator fails to meet the constraint for the same reason as the inverter we described. Circuits with higher relative leakage energy have less negative β and thus higher optimum V_{DD}.

The optimum V_T depends only on the desired frequency of operation and the depth of the critical path (L_{DP}). Figure 3 shows the optimum V_T curves corresponding to the ring oscillator and the filter at two frequencies. As described in section 3, the range of optimum supply voltage for a given design can be quite large based on operating parameters. When subthreshold operation can no longer provide the current drive necessary for meeting a speed requirement, then the optimum V_T approaches the optimum V_{DD}, and our assumption of subthreshold operation is no longer correct. This is the case for the FIR filter in Figure 3 at 2MHz.

2.2 Numerical Model

The analytical model presented above estimates the minimum energy operating point and gives intuition, but it is only valid for subthreshold operation. This section describes the transregional model in [7] that we modified to be more detailed in the subthreshold region, as shown in (11).

$$I_D = \begin{cases} I_o e^{\left(\alpha + \eta \cdot \frac{V_{DS}}{N_S}\right)} \left(\frac{V_{GS} - V_T}{\alpha N_S}\right)^\alpha & (V_{GS} \geq V_T + \alpha N_S) \\ I_o e^{\frac{V_{GS} - V_T + \eta V_{DS}}{N_S}} \cdot \left(1 - e^{\frac{-V_{DS}}{V_{th}}}\right) & (V_{GS} \leq V_T + \alpha N_S) \end{cases} \quad (11)$$
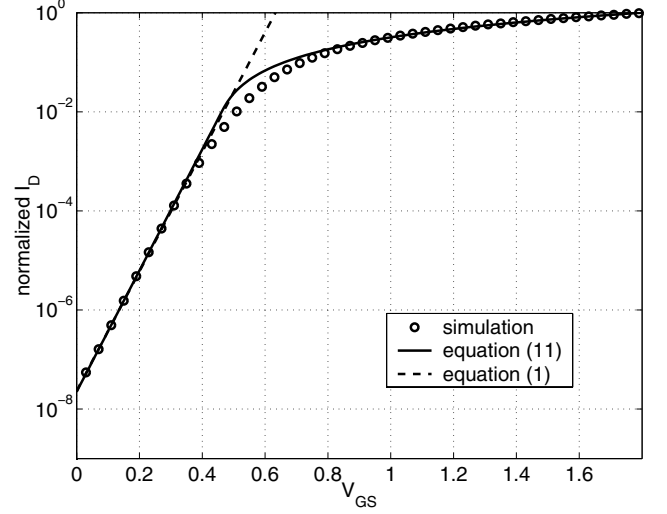

Figure 4. Modeling transistor current. Normalized I_D versus V_{GS} for $V_{DS}=V_{DD}$ for 0.18μm NFET.

The α parameter is the velocity saturation index, and η is the DIBL coefficient. Figure 4 compares the transregional model to an HSPICE simulation of NMOS drain current a in 0.18μm technology. The model is accurate to within a few percent in the subthreshold region and the strong inversion region, but it deviates from the simulation near V_T. The figure also shows that the simple current model in equation (1) matches well over the subthreshold region.

Like the subthreshold model, the transregional model can be extended to estimate the behavior of large circuits. The current used to discharge the output of an inverter is the average of I_{NMOS} with $V_{GS}=V_{DD}$ as V_{DS} drops from V_{DD} to $V_{DD}/2$. Using the transregional model numerically with $|V_{GS}|=V_{DD}$ for each current calculation and with $|V_{DS}|$ shown gives:

$$t_d = \frac{4KC_g V_{DD}}{I_{Dn, Vds = Vdd} + I_{Dn, \frac{Vdd}{2}} + I_{Dp, Vdd} + I_{Dp, \frac{Vdd}{2}}} \quad (12)$$

Determining the parameters C_{eff}, W_{eff}, and L_{DP} and using (11) for current in the equation for E_L provides a numerical model for total energy for arbitrary gates. This model permits analysis across the full range of V_{DD}.

3. MOVEMENT OF MINIMUM ENERGY POINT

This section examines the minimum energy point for a subthreshold system whose energy is the primary constraint. We assume that the system is implemented in a standard technology whose V_T is fixed (i.e. - no triple wells for body biasing). In this scenario, we examine the problem of setting V_{DD} to minimize energy per operation. For comparison, we apply the numerical model and the analytical solution to the parallel 8-bit, 8-tap FIR filter introduced in Figure 2. The filter has 200K transistors and $L_{DP}=140$. The simulations of the filter use netlists extracted from synthesized layout in a 0.18μm process. The synthesis flow incorporates a standard cell library that we modified to enable operation down to 100mV at the typical corner.
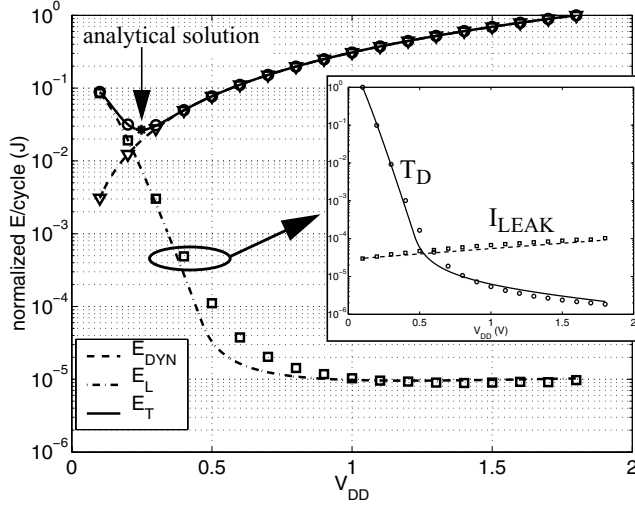
Figure 5. Model versus simulation of FIR filter (200k FETs) showing minimum energy point and contribution of active and leakage energy. Inset shows I_{LEAK} and T_D effect on E_L. Markers are simulation values, lines are model. Analytical solution from eqs. (8) and (6) is shown.

Calibrating the model requires three parameters. First, C_{eff} was determined by measuring average supply current for an extended NanoSim simulation of filtering operations and solving $C_{eff} = I_{avg}/(fV_{DD})$. Simulating to exercise the circuit's critical path provides the logic depth, L_{DP}. Lastly, W_{eff} is determined by simulating the circuit's leakage current in steady state at a low value of V_{DD} to match well in subthreshold and normalizing to the characteristic inverter. Since W_{eff} is a function of circuit state, averaging the circuit leakage current for simulations over many states improves the total leakage estimate.

Figure 5 shows the energy profile of the FIR filter versus V_{DD}. The contributions of active and leakage energy are both shown. The lines on the plot show the numerical model, and the markers show the simulation values. The numerical model matches very closely with simulation. The largest deviations occur in the E_L curve when V_{DD} is near the threshold voltage. This error results from the inaccuracy of equation (11) in the same region. The analytical solution is shown to match the numerical model with less than 0.1% error. Equation (8) provided the optimum V_{DD} for the analytical solution, and substituting this value into equations (4)-(6) gave the total energy. The inset in Figure 5 shows how the delay, T_D, and current, I_{LEAK}, components of leakage energy per cycle (E_L) vary with supply voltage. As V_{DD} reduces, the current decreases due to the DIBL effect, but the delay increases exponentially in the subthreshold region.

Figure 6 shows simulation results for the maximum frequency, set by the critical path delay, of the FIR filter as a function of supply voltage. In the subthreshold region, the performance is anywhere from several hundred Hz to around one MHz. The optimum supply voltage for a given frequency constraint in this range is simply read off of the figure. When energy is the primary constraint, then Figure 5 shows that V_{DDopt}=250mV.
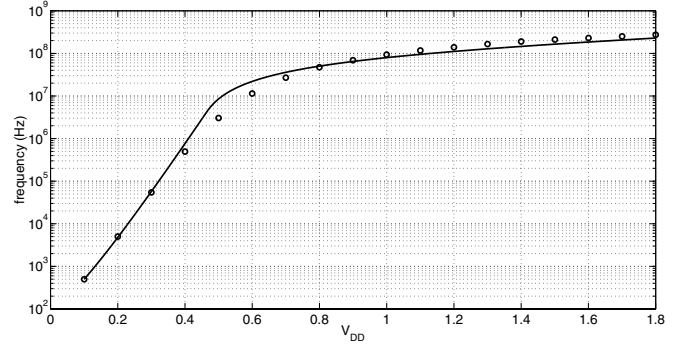

Figure 6. Frequency versus V_{DD} for FIR filter. V_T is ~450mV. Markers are simulation values, lines are numerical model.

The operating frequency at this optimum is 30kHz. The following section demonstrates how this solution can change based on operating parameters.

3.1 Operating Conditions

Figure 5 graphically confirms the trend apparent in equations (8) and (9). Any relative increase in the leakage component of energy per cycle will push the optimum V_{DD} higher, and the frequency at the optimum point also increases. In the figure, this corresponds to any decrease in E_{DYN} or increase in E_L. Likewise, any decrease in E_L or increase in E_{DYN} will lower the optimum V_{DD}. These types of changes can occur for a given circuit without changing its intrinsic attributes.

For example, consider using the FIR filter in a system whose workload, ω, changes widely. This might be in a video context where the processing per frame depends on the difference between consecutive frames. If the current frame is nearly identical to the previous, then very little work is required. A scene change, on the other hand, could demand the maximum number of computations. Assuming the clock is gated when no computation is required and normalizing to one cycle, C_{eff} per cycle becomes ωC_{eff} in equation (4).

Duty cycle, d, also can vary widely. A lower duty cycle means that the circuit spends more idle cycles (e.g. waiting for data) per active cycle. Consequently, the leakage contribution per operation increases, which corresponds to replacing W_{eff} with W_{eff}/d in equation (5). Normalizing to one cycle, we include duty cycle and workload in the analytical model and solve the equation set again to find the optimum V_{DD}, resulting in a new equation for β.

$$\beta = \frac{-2\omega C_{eff}}{\frac{W_{eff}}{d}KC_gL_{DP}}e^2 > -e^{-1} \qquad (13)$$

Figure 7 and Figure 8 show the effects of workload and duty cycle on the minimum energy and optimum V_{DD} of the FIR filter. The figures compare the numerical result with the analytical model and with simulation. The supply voltage for the simulations was quantized to 100mV increments. The quantization causes most of the error for values of ω and d close to one. The error in modeled energy at low values of ω and d occurs because the optimum V_{DD} has exceeded V_T. Thus the assumption of subthreshold operation implicit to the analytical model becomes invalid. The numerical model is also
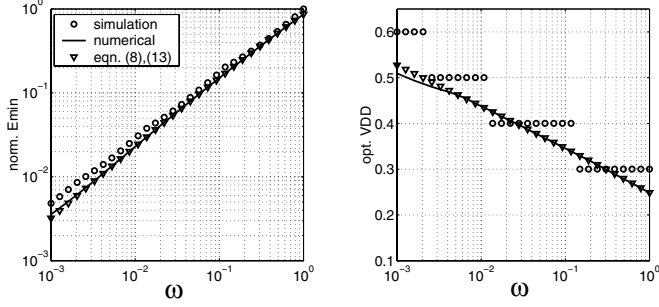
Figure 7. Normalized energy (left) and optimum V_{DD} (right) for FIR versus workload, ω. Simulation V_{DD} quantized to 100mV.
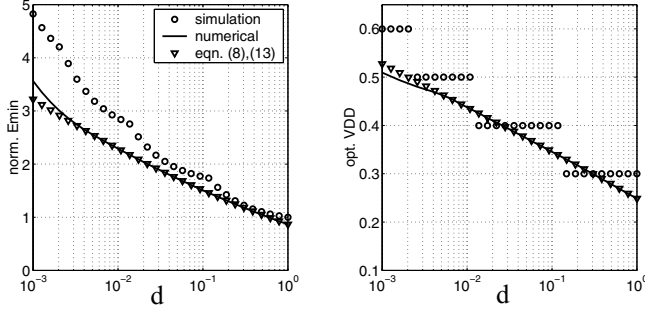

Figure 8. Normalized energy (left) and optimum V_{DD} (right) for FIR versus duty cycle, d. Simulation V_{DD} quantized to 100mV.

less accurate in that region. The analytical result matches the numerical values quite well until V_{DDopt} nears V_T.

Large reductions in either ω or d result in increased optimum V_{DD}, but the total energy per operation (normalized to one cycle) decreases as workload decreases and increases when duty cycle decreases. Clearly, knowing the average workload and duty cycle of a circuit can impact the choice of optimum supply voltage. The operational frequency, and thus the data rate, implicitly changes with V_{DD} in these figures. A system in which these parameters vary widely would benefit from closed-loop tracking of the optimum point since Figure 7 and Figure 8 show a large variation in the minimum energy.

The optimum point also depends on temperature, but the sensitivity over the range of possible operating temperatures is not as large. Figure 9 shows the effect of temperature on the components of energy. The numerical model shown in Figure 9 accounts for temperature dependence by decreasing the effective threshold voltage and increasing mobility (I_{off}) at higher temperatures as in [15]. These changes to the numerical model match well with simulation across most of the temperature range, but they underestimate the leakage energy at high temperatures. The threshold voltage drop accompanying a temperature increase raises the leakage current exponentially. This effect appears in the figure at higher V_{DD} where I_{LEAK} dominates E_L. The lower V_T also causes the delay to decrease, countering the rise in I_{LEAK}, so the total effect on E_L is not so pronounced at lower V_{DD} where the delay component dominates. Consequently, the total leakage energy does not change enormously near the minimum
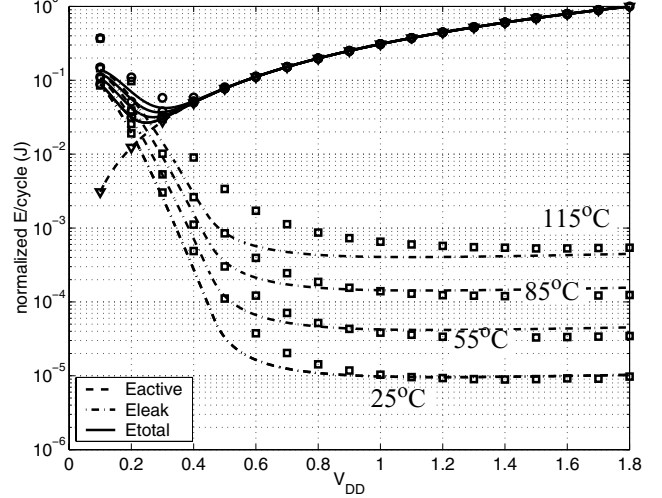

Figure 9. Dependence of minimum energy point on temperature shown in simulation (markers) and by the numerical model (lines). Temperature varies from 25°C to 115°C.

energy point. It does increase by several times, however, and the optimum V_{DD} increases by about 100mV over the full temperature range. The variation would clearly be much greater if the active energy was several orders of magnitude lower, as in the case of reduced workload.

3.2 Architecture

The models we have presented allow a quick assessment of the effect of architecture on minimum energy operation. Traditionally, architectural approaches such as parallelism or pipelining can reduce power for a given performance constraint by operating at a lower voltage. For performance constraints that are met at voltages above the optimum, the same conclusion holds for subthreshold operation. The model can also show how architecture will affect the minimum energy point when performance is not a constraint.

Figure 10 shows the numerical model (a) and simulation (b) of a pipelined implementation of the FIR filter. The model does not account for overhead capacitance, leakage, and delay in the pipeline registers, but it shows the general effect of ideal pipelining. As the number of stages increases, L_{DP} decreases and thus reduces leakage energy per operation (E_L). The total energy per cycle thus is reduced, and the optimum V_{DD} moves to the left. The simulation results in Figure 10 (b) show the same trend, however the overhead active energy makes deep pipelining more costly. The simulation also shows a decrease in active energy for shallow pipelines because of reduced glitching in the multipliers. Thus, shallow pipelining (2-4 stages) can reduce the total energy per operation for a system in a minimum energy scenario.

In contrast, parallelism cannot reduce the energy per operation, but it can increase the operating frequency at the minimum energy point. Once the optimum V_{DD} is known for a functional unit like the FIR filter, parallel copies of the filter will consume the least energy if they operate at the original optimum point. Discounting the overhead of muxing and demuxing indicates that ideal parallelism can increase the operating frequency at the minimum energy point by increasing the number of stages. It cannot, however, decrease
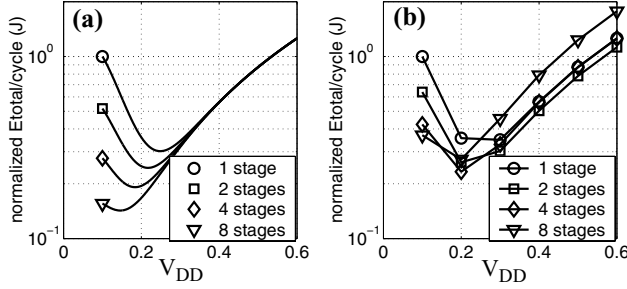
Figure 10. Effect of pipeline on minimum energy and optimum V_{DD} for the FIR filter. Ideal pipelines with numerical model (left) and simulated (right).

the energy per operation. Clearly, the overhead of parallelism in a real system will increase active energy and change the minimum energy point.

4. CONCLUSIONS

This paper has examined minimum energy operation for subthreshold circuits. We have shown that the minimum energy point depends on the technology, the characteristics of the design, and on operating conditions such as temperature, duty cycle, and workload. The optimum V_{DD} for minimizing energy per operation changes over several hundred millivolts when these parameters vary, pointing to the importance of tracking the optimum point or carefully characterizing a design before choosing V_{DDopt}. We introduced an analytical solution for the optimum V_{DD} and V_T to minimize energy for a given frequency in the subthreshold region. We modified a transregional current equation to provide a numerical model for large circuits. Simulations matched the analytical value and the numerical model within a few percent as long as the subthreshold assumption was valid.

5. APPENDIX

The Lambert W Function, $W = lambertW(x)$, gives the solution to the equation $We^W = x$, just as $W = \ln x$ is the solution to $e^W = x$. For real $x \geq 0$, the equation has exactly one real solution. For real $-e^{-1} < x < 0$, there are exactly two real solutions, called branches. The upper branch increases monotonically in $[-1, \infty)$ for $x \in [-e^{-1}, \infty)$, and the lower branch decreases monotonically in $[-1, -\infty)$ for $x \in [-e^{-1}, 0)$ [14]. For the solution in (8), the argument to lambertW is always negative, so two real solutions exist subject to the constraint $\beta > -e^{-1}$. The lower branch gives the minimum energy solution, and the upper branch solution is the local maximum.

7. REFERENCES

[1] R. Amirtharajah, et al, "A Micropower Programmable DSP Powered using a MEMS-based Vibration-to-Electric Energy Converter," *ISSCC*, 2000.

[2] A. Wang, et al, "Optimal Supply and Threshold Scaling for Subthreshold CMOS Circuits," *Intl. Symp. on VLSI,* pp. 7-11, 2002.

[3] J. Burr and A. Peterson, "Ultra Low Power CMOS Technology," *3rd NASA Symposium on VLSI Design*, pp. 4.2.1-4.2.13, 1991.

[4] R. Gonzalez, et al, "Supply and Threshold Voltage Scaling for Low Power CMOS," *JSSC*, vol 32, no 8, pp. 1210-1216, Aug 1997.

[5] M. Stan, "Optimal Voltages and Sizing for Low Power," *Intl. Conf. on VLSI Design*, pp. 428-433, 1999.

[6] J. Kao, et al, "A 175-mV Multiply-Accumulate Unit Using an Adaptive Supply Voltage and Body Bias Architecture," *JSSC*, vol 37, no 11, pp. 1545-1554, Nov 2002.

[7] K. Nose and T. Sakurai, "Optimization of V_{DD} and V_{TH} for Low-Power and High-Speed Applications", DAC, pp. 469 - 474, 2000.

[8] A. Bhavnagarwala, et al, "A Minimum Total Power Methodology for Projecting Limits on CMOS GSI," *Trans. on VLSI Systems*, vol 8, no 3, pp. 235-251, June 2000.

[9] R. Brodersen, et al, "Methods for True Power Minimization," *ICCAD*, pp. 35 - 42, 2002.

[10] E. Vittoz, "Micropower Techniques," *Design of VLSI Circuits for Telecommunication and Signal Processing*, ed. J.E. Franca and Y.P. Tsividis, ch. 5, Prentice Hall, 1994.

[11] H. Kim and K. Roy, "Ultra-low Power DLMS Adaptive Filter for hearing Aid Applications," *ISLPED*, 2001.

[12] A. Bryant, et al, "Low-Power CMOS at Vdd=4kT/q," *Device Research Conference*, pp. 22 - 23, 2001.

[13] A. Wang and A. Chandrakasan, "A 180mV FFT Processor Using Sub-threshold Circuit Techniques," *ISSCC*, 2004.

[14] R. Corless, et al, "On the Lambert W Function", Advances in Computational Mathematics, vol. 5, pp. 329-359, 1996.

[15] A. Bellaouar, et al, "Supply Voltage Scaling for Temperature Insensitive CMOS Circuit Operation," *IEEE Trans. on Circuits and Systems*, vol. 45, no. 3, pp. 415-417, March, 1998.