Utilizing Quantum Dot Transistors with Programmable Threshold Voltages for Low-Power Mobile Computing

SHUO WANG and JIANWEI DAI
University of Connecticut
EL-SAYED HASANEEN
El-Minia University
LEI WANG and FAQUIR JAIN
and
University of Connecticut

15

Power consumption poses one of the fundamental barriers for deploying mobile computing devices in energy-constrained situations with varying operation conditions. In particular, leakage power is projected to increase exponentially in future semiconductor process nodes. This challenging problem is pressing for renewed focus on power-performance optimization at all levels of design abstract, from novel device structures to fundamental shifts in design paradigm. In this article, we propose to exploit the programmable threshold voltage quantum dot (QD) transistors to reduce leakage thereby improving the energy efficiency for mobile computing. The unique programmability and reconfigurability enabled by QD transistors extend our capability in design optimization for new power-performance trade-offs. Simulation results demonstrate the significant leakage reduction over conventional techniques.

Categories and Subject Descriptors: B.7.1 [**Integrated Circuits**]: Types and Design Styles

General Terms: Design, Performance

Additional Key Words and Phrases: Low power, threshold voltage and quantum dot transistor

ACM Reference Format:

Wang, S., Dai, J, Hasaneen, E.-S., Wang, L., and Jain, F. 2009. Utilizing quantum dot transistors with programmable threshold voltages for low-power mobile computing. ACM J. Emerg. Technol. Comput. Syst. 5, 3, Article 15 (August 2009), 19 pages.
DOI = 10.1145/1568485.1568489 http://doi.acm.org/10.1145/1568485.1568489

This article is an extended version of "Programmable Threshold Voltage using Quantum Dot Transistors for Low-Power Mobile Computing" by Shuo Wang et al., published in *Proceedings of the International Symposium on Circuits and Systems (ISCAS)* © IEEE 2008.
This research was supported by the NSF grant CCF-0621947 and the University of Connecticut Faculty Research Grant 446751.
Author's address: S. Wang; email: Shuo.wang@engr.uconn.edu.

DOI 10.1145/1568485.1568489 http://doi.acm.org/10.1145/1568485.1568489

1. INTRODUCTION

Fueled by the advance in semiconductor technology, battery-powered portable computing devices are now embedded into our daily life with interface to various mobile and wireless networks. Some future applications envisioned include wearable computers, smart homes, and ambient intelligent systems. It is well known that energy efficiency is a major challenge that mobile computing must surmount in order to realize its full potential. The aggressive scaling of semiconductor process allows dramatic improvement in silicon capacity but also gives rise to an exponential increase of leakage power. It is projected that subthreshold leakage will contribute to 54% of the total power consumption at the 65nm process node [Narendra et al. 2003]. This problem will influence the design of mobile computing from various aspects. In particular, since mobile computing devices need to frequently stay in the standby mode, a considerable portion of the battery power is wasted by leakage. Hence, aggressive leakage reduction is essential to extending the operational life of mobile computing systems.

Many techniques have been developed in modeling, characterizing, and reducing leakage at different levels of design abstract [Mukhopadhyay et al. 2003], [Hamzaoglu et al. 2002]. At the circuit design level of interest for this paper, solutions such as VTCMOS [Kuroda et al. 1996], MVCMOS [Roy et al. 1999], and MVTMOS [Wei et al. 1999] have been proposed. The basic idea of these techniques is to partition a complex circuit based on power and performance requirements, and assign threshold voltages correspondingly by substrate biasing or utilizing dual-V_{th} CMOS process. However, the effectiveness of these techniques is limited by the available choice/range of threshold voltages. For example, dynamic body bias proposed in Tschanz et al. [2003] can only change the threshold voltage by less than 100 mV, and the widely used dual-V_{th} technique can only assign two fixed threshold voltages, which is unlikely to make an optimal design. Given the trend of increase in system complexity, new solutions need to be sought from novel device structures to fundamental shifts in design methodologies to further improve the performance-power trade-offs for mobile computing.

In this article, we propose to exploit the quantum dot (QD) transistors to address the challenging problem of leakage reduction in mobile computing systems. One unique feature of QD transistors is that their threshold voltages can be programmed in a much larger range [Compagnoni et al. 2005; Fernandesa et al. 2001]. Furthermore, very fine-grained V_{th} programming can be obtained in practice and thus V_{th} can be approximated as a continuous variable for engineering purpose. This allows new opportunities for new levels of power-performance trade-offs beyond that allowed by the existing solutions. Furthermore, the reconfigurability of QD transistors introduces new design space extension for energy-efficient mobile computing under various operational modes. In particular, the QD transistors can be utilized to not only compensate for the variations induced by the design and fabrication [Bowman et al. 2002; Samaan 2004; Qi et al. 2005], but also make systems adaptive to the uncertainties in ambient conditions encountered by mobile computing applications. Some preliminary

results were presented in Wang et al. [2008]. In this article, we extend our past work by making the following new contributions. First, we discuss the device modeling (in Section 2.1) and provide the details on the programming circuit (in Section 2.2). Second, we develop a formal design methodology, and in particular perform an analysis on the reconfigurability of our technique for dealing with various operational modes of mobile computing (in Section 3.3). Finally, we conduct a complete set of studies on the performance of the proposed technique with new results.

The rest of this article is organized as follows. In Section 2, we review the basic operation of QD transistors. In Section 3, we present the proposed approach for aggressive leakage reduction using QD transistors. Section 4 evaluates our technique in the design of LSM adaptive filters, which are a key component of baseband DSP in wireless mobile computing devices. The conclusions are given in Section 5.

2. QUANTUM DOT (QD) TRANSISTORS

Recently, quantum dot (QD) transistors have gained a lot of attention. With thin oxide thickness, QD transistors can achieve high-speed V_{th} programming, high density, and good scalability as compared to conventional floating gate devices. Several studies have been reported [Tiwari et al. 1996; Tiwari 1996; Tang et al. 2000; Normand et al. 2001; Liu et al. 2000; Nogami et al. 2003]. Due to the feasibility of fabricating QD transistors using conventional CMOS process, QD transistors have been integrated with CMOS circuits in many applications, such as non-volatile memory, tunable gain operational amplifier, and programmable reference circuits [Hasaneen 2004]. In this article, we propose to exploit QD transistors, for example, as sleep transistors, to address leakage reduction problem in mobile computing. For designs using QD transistors as sleep transistors, quantum dot charges are configured to adjust the threshold voltages for performance-power trade-offs; whereas for flash memory transistors, the quantum dot charges are programmed to represent different states stored in the memory. In this section, we will review the basic operation of QD transistors and discuss the supporting circuits for V_{th} programming.

2.1 Device Structure and Modeling

QD devices have the single-transistor structure with discrete quantum dots embedded within the gate dielectric. There are two gates in a QD transistor. One is control gate, which acts as the external gate of the device, and the other is a floating gate formed by discrete quantum dots distributed on a thin tunneling oxide. Figure 1 shows the cross-section view of a QD transistor. The thickness between the floating dot gate and the top control gate is typically in the range of several nanometers, which is sufficient to keep the electrons and holes from tunneling between the control gate and the dots during normal operation.

One unique feature of QD transistors is that the threshold voltage can be programmed by charging and discharging the embedded quantum dots. To obtain a high threshold voltage, a positive voltage relative to the source and drain is applied at the gate to make the potential energy of the quantum dots below

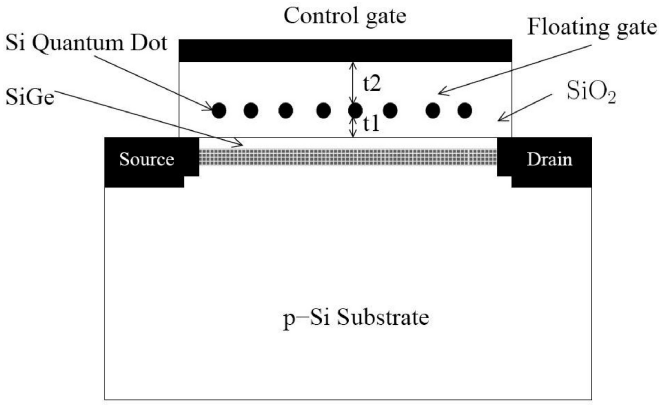

Fig. 1. Cross-section schematic of a quantum dot transistor.

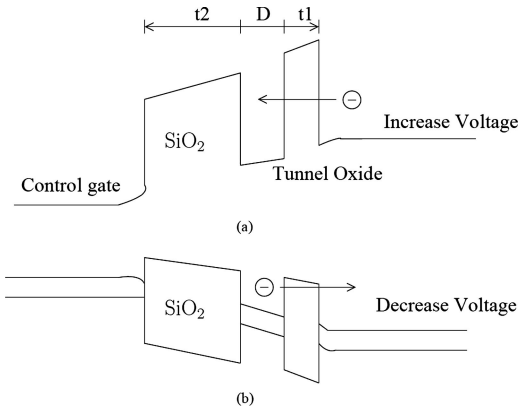

Fig. 2. Programming threshold voltage for a quantum dot transistor: (a) write process and (b) erase process.

that of the channel. The electrons are injected from the inversion channel to the quantum dots thereby increasing the threshold voltage. This operation is called the write/programming process. The reverse process, called the erase process, reduces the threshold voltage by applying a relative negative voltage at the gate. The electrons are tunneling back into the channel as the potential energy of the gate is raised. These two processes in the n-channel QD transistor are illustrated in Figure 2.

The write/erase characteristics are a function of control gate voltage and tunnel oxide thickness [She et al. 2003]. Typically, for a 2.0nm tunnel oxide, the write speed can reach 100ns at $10V$ programming voltage. The threshold voltage is scarcely degraded after more than 10^9 write/erase cycles, which is beyond the lifetime of most mobile computing devices. In our recent work reported in Hasaneen et al. [2004], we have fabricated the QD transistors in an equivalent 90nm process.

An analytical model to quantify the operation of QD transistors was presented in Hasaneen et al. [2004]. Consider the current density j after applying a high voltage at the gate

$$j = q \times n_{dot} \times P, \tag{1}$$

where P is the tunneling transition rate of electrons from the quantum well channel to the quantum dots, n_{dot} is the dot density, and q is the charge of an electron. The transition rate P is given by the transfer Hamiltonian form as

$$P = \frac{4\pi}{\hbar} \sum |\langle \psi_d | H - H_w \psi_w |^2 (f_w - f_d) \delta(E_d - E_w), \tag{2}$$

where H is the total Hamiltonian; H_w, ψ_w, E_w, and f_w are the Hamiltonian, wave function, energy, and Fermi function for the occupation probability of the states in the well; and ψ_d, E_d and f_d are the corresponding parameters of the quantum dots, respectively.

It was shown that by solving the Schrödigner and Poisson equations, we can obtain the wave functions of quantum well and quantum dots. Thus, the current density in (1) can be calculated. This allows us to determine the total quantum dot charge, expressed as

$$Q = \int_0^{t_w} j(t) A dt, \tag{3}$$

where t_w is the duration of charging/discharging time and A is the quantum dot capture area. Here, the current density $j(t)$ is time-dependent. Due to the charge Q stored in the quantum dots, the shift in threshold voltage can be expressed as

$$\triangle V_{th} = \frac{Q}{C} = \frac{\int_0^{t_w} j(t) A dt}{C}, \tag{4}$$

where C is the capacitance between the control gate and quantum dots. It follows that the threshold voltage V_{th} of a QD transistor can be obtained as

$$V_{th} = V_{th0} + \triangle V_{th}, \tag{5}$$

where V_{th0} is the intrinsic threshold voltage (i.e., no charge in the quantum dots). Past work [Compagnoni et al. 2005; Fernandesa et al. 2001] has demonstrated a large range of threshold voltage shift. This feature is exploited in the following sections for aggressive leakage reduction.

Note that this analytical model can be integrated with the BSIM model[1] to assist design automation.

2.2 Programming Circuits

The threshold voltage of a QD transistor can be changed by applying voltages with different amplitudes and durations at the control gate. Many circuits have been proposed to generate and program this kind of voltages on-chip. In Blyth et al. [1991], Tran et al. [1996], and Rolandi et al. [1998], feedback control was

[1]http://www-device.eecs.berkeley.edu/~bsim3/bsim4.html.

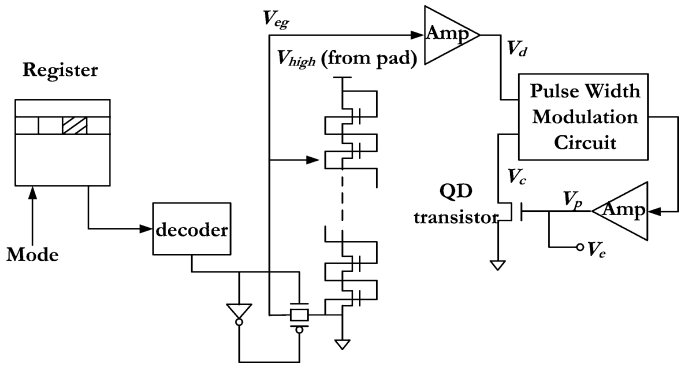

Fig. 3. Logic diagram of a voltage controller.

used in the programming loop consisting of write and read/verify periods to complete the programming. A method based on the pulse width modulation (PWM) Kinoshita et al. [1999, 2001] was also developed with high precision.

To achieve fast programming, a high voltage is needed at the control gate of the QD transistor. This high voltage can be obtained by charge pump amplifiers [Hooper et al. 2005], which are commonly employed to generate high voltages on-chip for programming floating gate devices. By careful place and route and employing mixed-signal design techniques, the PWM and charge pump amplifiers can be integrated reliably with digital circuits.

Based on the existing work, a circuit to program QD transistors is shown in Figure 3. This circuit consists of three components: a decoder, a charge pump amplifier, and a PWM circuit. During the design phase, a series of operation points represented by different V_{th} values are derived according to the system requirements (see Section 3). These values are stored in the registers and utilized to program QD transistors. By decoding the information stored in the registers and comparing the reference voltage V_d with the drain voltage V_c, the voltage controller will generate a voltage pulse with required amplitude and duration on the control gate of the QD transistor to attain the desired threshold voltage. This programming circuit can be integrated on-chip because both PWM and charge pump amplifiers can be built along with digital circuits. Note that QD transistors are nonvolatile. Thus, the V_{th} will keep the programmed value as long as the operation point remains unchanged. Many practical design issues related to this programming circuit have been well-studied in the literature. Several optimized designs were studied [Long et al. 2006; Palumbo et al. 2002; Ying et al. 2003] to meet the requirements of power consumption and area overhead. In Lin et al. [2005] and Hasan et al. [2005], the PWM circuit and the charge pump circuit were implemented using standard CMOS process, therefore fully compatible with the QD transistors. The experimental results from an actual chip implementation in Lin et al. [2005] clearly show the feasibility of integrating the programming circuit with digital circuits.

We need to point out that V_{th} programming may take from hundreds of nanoseconds to several microseconds to finish. However, this overhead has

almost no noticeable impact to the performance of mobile computing devices, most of which are operated at the megahertz range.

3. PROGRAMMABLE THRESHOLD VOLTAGE FOR LEAKAGE REDUCTION

In this section, we will explore the new opportunities made available by the QD transistors for leakage power reduction. Specifically, we will study the power-performance trade-offs and reconfigurability of QD-based designs using programmable threshold voltages.

3.1 Power-Performance Trade-offs

Design of mobile computing systems requires various levels of optimization targeting power-performance trade-offs. In this article, we focus primarily on leakage power, which is a critical problem in battery-powered devices such as hand-held and portable devices. A general expression of leakage power is given by

$$P_{leakage} = I_{leakage} V_{dd}, \tag{6}$$

and the leakage current $I_{leakage}$ can be estimated as

$$I_{leakage} = I_0 e^{(V_{gs} - V_{th})/nV_T}, \tag{7}$$

where $I_0 = \mu_0 C_{ox}(W/L)V_T^2 e^{1.8}$, C_{ox} is the gate oxide capacitance, (W/L) is the width to length ratio of MOS devices, V_T is the thermal voltage ($26mV$ at room temperature $T = 300K$), and n is the subthreshold swing coefficient determined by the semiconductor process.

To evaluate the performance, we consider the delay which can be approximated by [Nose et al. 2000]

$$t_p = K \frac{C_L V_{dd}}{\mu C_{ox} \frac{W}{L} (V_{dd} - V_{th})^{\alpha}}, \tag{8}$$

where C_L is the load capacitance, K is a process-dependent constant, and α approaches 1 in nanometer technology.

From (7)–(8), the threshold voltage plays a key role in determining the power-performance trade-offs. Obviously, increasing threshold voltage reduces leakage power at the cost of larger propagation delay. In general, delay slacks should be maximally utilized by either voltage scaling or threshold voltage assignment. Here we consider the threshold voltage assignment, which is commonly employed for leakage reduction. We denote the maximum delay allowed by the system design as T_{max}. The delay slack of a logic path can be expressed as

$$\Delta t_p = T_{max} - t_p - t_m, \tag{9}$$

where t_p is given by (8) and t_m is the design margin for timing robustness. For the sake of simplicity we do not consider various sources of timing noise explicitly but quantify the lumped effect by the parameter t_m. Note that in this article the delay slack is referred to the result of different critical path delays at different

logic blocks, not the delay variations caused by different input combinations at one logic block.

From (7)–(9), the leakage current can be expressed as a function of the delay slack, that is,

$$I_{leakage} = I_0 10^{(V_{gs} - V_{dd} + \frac{\gamma}{T_{max} - \Delta t_p})/S}, \tag{10}$$

where $\gamma = \frac{KC_L V_{dd}}{\mu C_{ox} W/L}$, $S = nV_T \ln(10)$, and α in (8) is selected to be 1.

Clearly, reducing delay slack Δt_p via appropriate V_{th} assignment also leads to significant savings on leakage power. The optimal power-performance trade-offs are achieved when the delay slack is fully utilized, such that

$$V_{th\_opt} = V_{dd} - \frac{\gamma}{T_{max}}, \tag{11}$$

$$\Delta t_{p\_opt} = 0, \tag{12}$$

$$I_{leakage\_opt} = I_0 10^{(V_{gs} - V_{dd} + \frac{\gamma}{T_{max}})/S}. \tag{13}$$

In general, the values of V_{th_opt} in different logic blocks or paths should be different as these blocks/paths are unlikely to have the same delay slack. Ideally, we would like to design the circuits with different V_{th} as well. However, current semiconductor process such as dual-V_{th} only provides a limited set of threshold voltages.

3.2 Design Methodology

We consider a general scenario where the threshold voltage can be chosen from a range between V_{th_min} and V_{th_max}. Note that the existing dual-V_{th} techniques are a special case where the V_{th} can only be selected from a binary set $\{V_{th\_low}, V_{th\_high}\}$. Due to this constraint, a common design practice is to assign V_{th_high} to power-sensitive units and V_{th_low} to logic blocks that are performance-critical, for example, those with sleep transistors or in critical paths.

Unfortunately, the partition of a complex circuit is not always intuitive and the above dual-V_{th} assignment is likely to be suboptimal. Other techniques such as substrate biasing alleviate this problem but are still limited by a narrow range of adjustable threshold voltage. In addition, other related problems such as sleep transistor sizing further complicate the design optimization for power-performance trade-offs.

Using QD transistors, the power-performance trade-offs can be adjusted by programming the threshold voltages of QD sleep transistors. This enables new options of design optimization at multiple levels from individual transistors up to large logic blocks. Furthermore, it is possible to adjust V_{th} to account for the possible performance shifts due to variations in process parameters, supply voltages and temperature. Thus, the inherent constraints in conventional techniques can be relaxed.

Our design methodology for programmable threshold voltage using QD transistors is illustrated in Figure 4 in comparison with the conventional dual-V_{th} MTCMOS [Mutah et al. 1995]. In MTCMOS, the low threshold voltage V_{th_low} is assigned to all the logic blocks, while a high threshold voltage V_{th_high} is used for sleep transistors. The sleep transistors can significantly reduce leakage

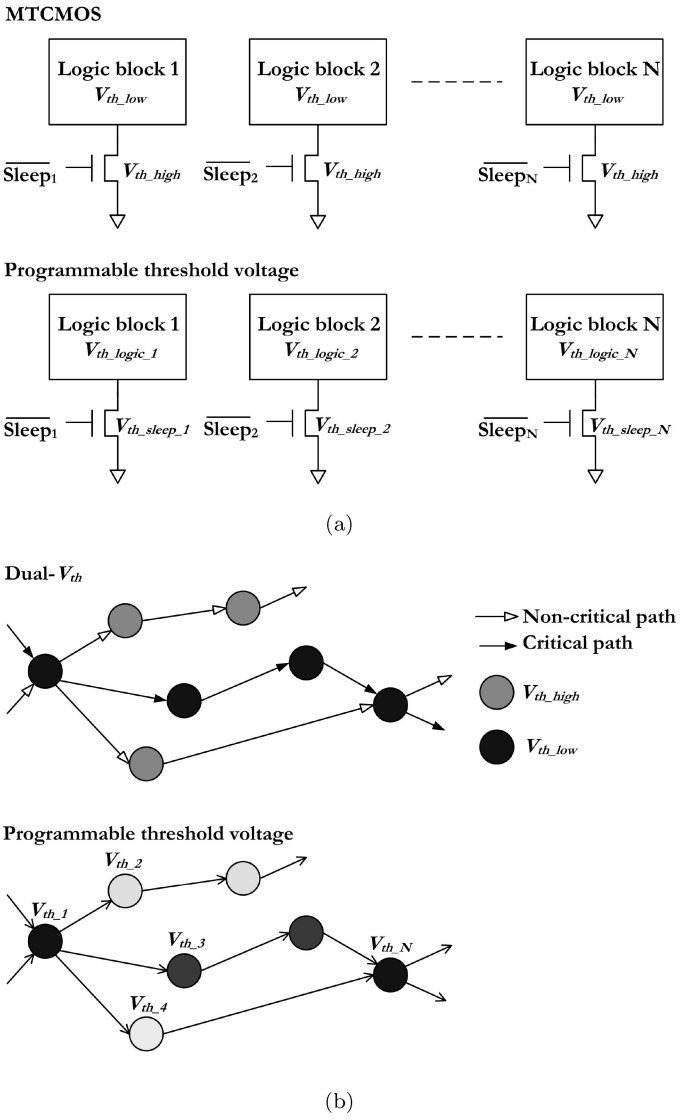

Fig. 4. Illustration of the proposed approach: (a) MTCMOS vs. programmable threshold voltage QD transistors, and (b) dual-V_{th} vs. programmable threshold voltage QD transistors.

power during the standby mode, but they also incur performance penalty during the active mode. Transistor sizing is also difficult as these design metrics are often conflicting with each other. Furthermore, the value of high threshold voltage V_{th_high} may not be suitable for the sleep transistors in different logic blocks at the active mode. Applying the proposed programmable V_{th} technique, we replace the sleep transistors by the QD transistors. Consider the fact that logic blocks usually have different delay slacks in the active mode. We first determine the delay slacks of each logic block from (8)–(9). By assigning individual $V_{th\_sleep\_i}$ according to (11)–(13) to the QD sleep transistors as shown

in Figure 4(a), these delay slacks can be maximally utilized for minimizing leakage power. In addition, V_{th} of the sleep transistors can be programmed to a low value $V_{th\_sleep}^{active}$ at the active mode for high performance and programmed to a high value $V_{th\_sleep}^{standby}$ at the standby mode for low leakage. Thus, the sleep transistors can be made smaller than conventional CMOS sleep transistors due to the extra performance margin rendered by the V_{th} programmability. Besides, the optimal values of $V_{th\_sleep}^{active}$ are determined by (11)–(13), while the values of $V_{th\_sleep}^{standby}$ can be chosen as high as possible because there is no performance requirement during the standby mode.

The proposed technique can be further enhanced by judiciously implementing different threshold voltages to the accessible transistors in the logic blocks according to the delay and leakage power profiles, as illustrated in Figure 4(b). Traditional dual-V_{th} technique assigns low threshold voltage V_{th_low} to the critical paths and high threshold voltage V_{th_high} to other paths. As the threshold voltage of a specific path can only be either V_{th_low} or V_{th_high}, the effectiveness of this technique is limited. Using the proposed technique, however, different threshold voltages can be applied to different paths and even different units in each path according to their delay slacks. This enables further improvement in energy efficiency at the cost of increase in programming complexity.

We need to point out that the programming circuits for QD transistors should not introduce net energy overhead when taking the overall system operation into account. For mobile computing devices where V_{th} programming may be performed once per several hours, the energy overhead related to the programming circuits can be offset by the reduction in leakage power. Also, overheads in chip area and routing can be managed because they are not the limiting factors in the design of mobile computing devices.

3.3 Reconfigurability

Many mobile computing systems need to operate at different modes such as standby, active, and certain intermediate modes in between.[2] To effectively reduce leakage power, reconfiguration of circuit parameters such as threshold voltage in accordance with these operation modes is necessary. The programmable threshold voltage QD transistors provide a viable way for this purpose.

We use the sleep transistor circuit as shown in Figure 4(a) to explain our approach. Consider a logic block with threshold voltage V_{th_logic}. The threshold voltage of the sleep transistor in MTCMOS is $V_{th\_sleep}^{high}$, while the QD sleep transistor has $V_{th\_sleep}^{active}$ and $V_{th\_sleep}^{standby}$ in the active and standby modes, respectively. The selection of $V_{th\_sleep}^{active}$ is based on the delay slack of the logic block during the active mode (see Section 3.1), whereas $V_{th\_sleep}^{standby}$ can be chosen as high as possible to minimize the standby leakage.

[2]http://www.acpi.info/.

The standby leakage current of the logic block is determined by the leakage current flowing through the sleep transistor, which can be expressed as

$$I_{standby\_leakage} = \mu_0 C_{OX} \left(\frac{W}{L}\right)_{sleep} V_T^2 e^{1.8} e^{(V_{gs} - V_{th\_sleep})/nV_T}, \tag{14}$$

where $V_{th\_sleep} = V_{th\_sleep}^{high}$ and $V_{th\_sleep} = V_{th\_sleep}^{standby}$ for the MTCMOS and QD transistors, $\left(\frac{W}{L}\right)_{sleep} = \left(\frac{W}{L}\right)_{sleep}^{MT}$ and $\left(\frac{W}{L}\right)_{sleep} = \left(\frac{W}{L}\right)_{sleep}^{QD}$ are the sizes of the MTCMOS sleep transistor and QD sleep transistor, respectively. Comparing the leakage currents of these two implementations, we obtain

$$\frac{I_{standby\_leakage}^{QD}}{I_{standby\_leakage}^{MT}} = W_{ratio} 10^{(V_{th\_sleep}^{high} - V_{th\_sleep}^{standby})/S}, \tag{15}$$

where $W_{ratio} = \frac{(\frac{W}{L})_{sleep}^{QD}}{(\frac{W}{L})_{sleep}^{MT}}$. The size of sleep transistor is determined by the delay constraint in the active mode. For the MTCMOS sleep transistor, since the threshold voltage is fixed, the size of sleep transistor has to be quite large to compensate for the delay overhead at the active mode. In our approach using QD sleep transistor, threshold voltage can be programmed, that is, $V_{th\_sleep}^{active}$ can be lowered during the active mode. Therefore, the size of QD sleep transistors can be made smaller than conventional sleep transistors ($W_{ratio} < 1$). In addition to the benefit of transistor sizing, $V_{th\_sleep}^{standby}$ can be raised to a value higher than $V_{th\_sleep}^{high}$ in MTCMOS. Consequently, aggressive reduction of standby leakage can be achieved without performance degradation in the active mode as commonly seen in the MTCMOS.

It is interesting to note that the QD transistors can be reconfigured to address the leakage problem in the intermediate modes of mobile computing, which may impose quite different power and performance requirements from either standby or active mode. The conventional MTCMOS technique lacks such flexibility in dealing with different power-performance trade-offs.

We now derive the reconfiguration range related to the intermediate modes under different performance requirements for active leakage reduction. Consider the V_{th_opt} expressed in (11), where leakage power is minimized while satisfying the delay constraint. From (7) and (8), the optimal operation point (11)–(13) is achievable as long as

$$T_{max} \in \beta \cdot \left[\frac{1}{1 - V_{th\_min}/V_{dd}}, \frac{1}{1 - V_{th\_max}/V_{dd}}\right], \tag{16}$$

where $\beta = \gamma / V_{dd}$ and γ is defined in (10). The lower bound of T_{max} determines the highest speed that can be achieved in an intermediate mode by trading power for performance. This value can be utilized as a key design specification for the most performance-critical mode. On the other hand, the upper bound of T_{max} corresponds to the maximal leakage reduction. This value is useful in configuring the most power-sensitive mode other than the standby mode. The values of T_{max} between the most performance-critical mode and the most

power-sensitive mode can be configured by programming V_{th} according to the workloads and power-performance requirements of the intermediate modes. For a typical 90-nm process with $V_{dd} = 1.2V$, $V_{th\_min} = 0.2V_{dd}$, and $V_{th\_max} = 0.5V_{dd}$, the range of T_{max} can be calculated as $[1.25\beta, 2\beta]$. This range represents the optimal power-performance points of different operation modes in mobile computing systems.

Note that the proposed technique can be applied in combination with other low-power techniques such as dynamic voltage scaling. This provides new design options beyond the conventional approaches.

4. APPLICATION TO BASEBAND DSP FOR MOBILE COMPUTING

In this section, we evaluate the proposed technique in an application of baseband DSP for wireless mobile computing.

The major function of baseband DSP is to overcome considerable signal distortions caused by wireless communication channels. These signal distortions include propagation loss (due to channel attenuation) and crosstalk noise (generated by the mismatch between adjacent channels). Advanced baseband DSP techniques are needed for signal recovery. For example, feed-forward equalizers (FFE) remove the intersymbol interference (ISI) introduced by the channel attenuation, and echo/NEXT cancellers perform crosstalk suppression. Most of these signal processing tasks utilize the least mean squared (LMS) adaptive filters [Haykin 1996], which involve intensive filtering operations thereby requiring effective energy reduction. Since wireless communication channels are non-deterministic, the configuration of the LMS adaptive filters changes during the operation of mobile computing devices. For example, the number of active taps and data precision needed for a required algorithmic performance are determined by the physical channel conditions such as the distance to the base station. As a result, the critical path delay in the integrated circuit implementation of LMS filters varies as a function of the number of active taps and data precision. This allows us to employ the programmable threshold voltage QD transistors for power-performance optimization. To demonstrate the feasibility of implementing multiple V_{th} QD devices in a VLSI circuit, we apply QD devices as sleep transistors to reduce leakage power in embedded DSP at various operational modes. For future work, we plan to investigate effective CAD algorithms so that the proposed technique can be applied to more complex circuits.

As discussed in Section 3, QD transistors allow the reconfiguration of V_{th} at different operation modes, that is, high V_{th} at the standby mode and low V_{th} at the active mode for the same sleep transistor. In addition, the value of low V_{th} can be further adjusted to exploit the variations in critical path delay caused by changes in physical channel conditions (e.g., distance of wireless channel to the base station). This is important for mobile computing, where a considerable amount of battery power is wasted by leakage. Thus, to effectively minimize leakage power, hardware implementations need flexibility in dealing with the uncertainties in wireless channels. To demonstrate our approach, we simulate an LMS adaptive filter in a 90nm process with the default configuration of

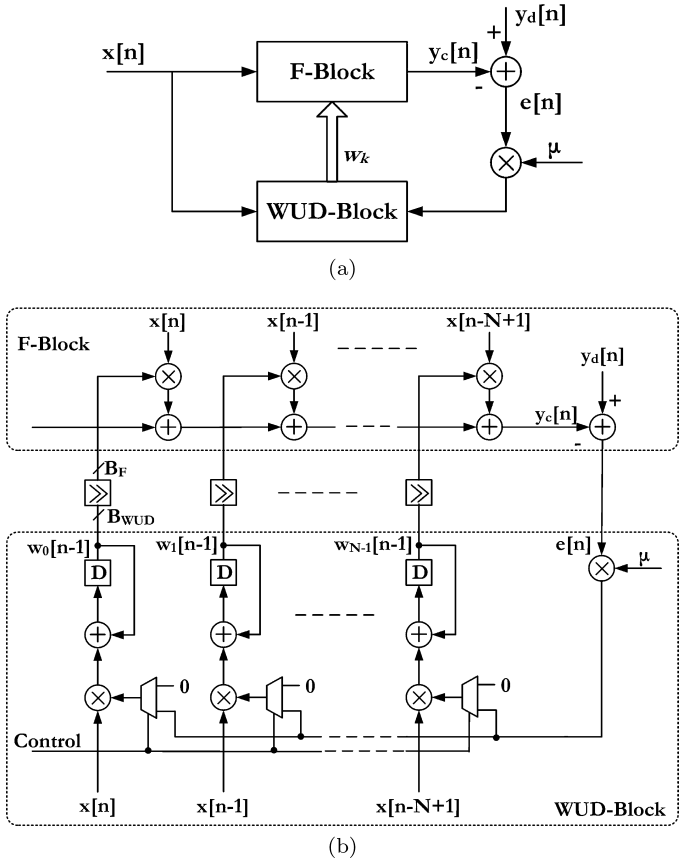

Fig. 5. LMS adaptive filter: (a) functional block diagram and (b) VLSI architecture.

40 taps and 8-bit inputs and coefficients. The block diagram and VLSI architecture of the adaptive filter are shown in Figure 5. The QD transistors with programmable V_{th} are utilized as the common sleep transistors at the power rails of the baseband DSP. We use QD transistors modeled and fabricated by Hasaneen et al. [2004] with device characteristics given by Figures 6 and 7, where the channel length and width of the QD transistors are 100nm and 500nm, respectively, and the corresponding V_d and control gate voltages are determined by device characteristics as discussed in Hasaneen et al. [2004]. All the simulation results are obtained from the BSIM model and Cao et al. [2000] with an extension to QD transistors.

Figure 7 shows that the subthreshold current can be significantly reduced by programming the threshold voltages of QD transistors. In techniques using sleep transistors (e.g., MTCMOS), the high threshold voltage of the sleep transistor introduces virtual ground bounces and hence performance degradation during active mode. By using QD sleep transistors, the size of QD sleep transistors can be made smaller than conventional sleep transistors. Thus, the introduced parasitic capacitances on the virtual ground node will

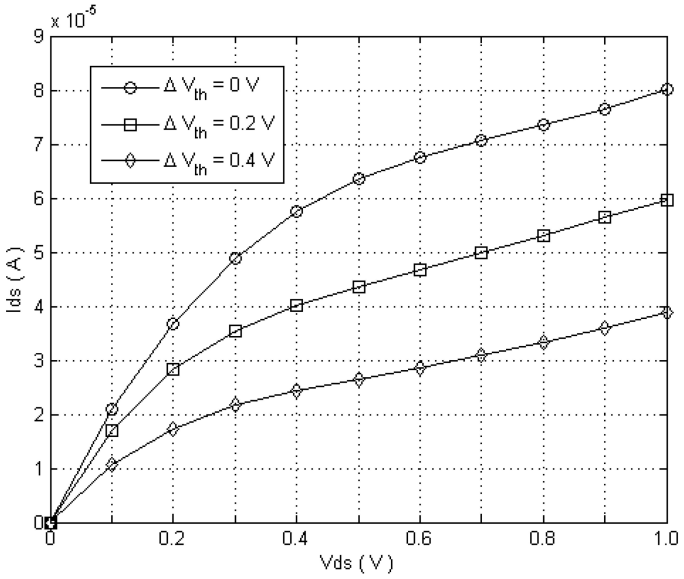

Fig. 6. Device characteristics of QD transistors.

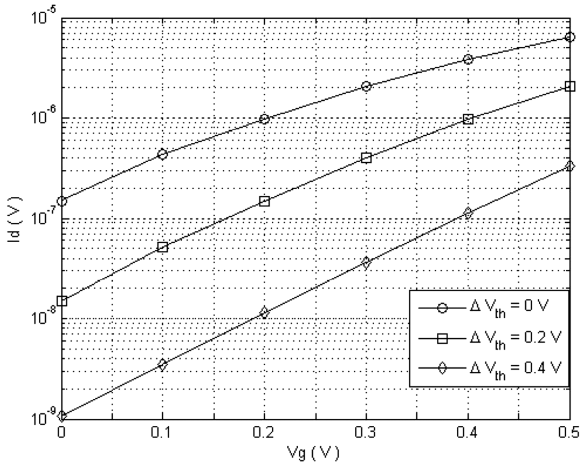

Fig. 7. Subthreshold characteristics of QD transistors.

also be smaller than the existing solutions such as MTCMOS. This effect is evaluated and shown in Figure 8 for a 4-bit adder module. Figure 9 shows the reduction in standby leakage current of the baseband DSP as compared to the conventional dual-V_{th} design. Both DSP modules are designed with the same speed at the active mode under the default configuration. The high V_{th} in the dual-V_{th} implementation is 38% of V_{dd} and the V_{th} of QD sleep transistors is programmed from 28% to 60% of V_{dd}. These results lead to the following observations. First, because V_{th} can be programmed to a low value during the active mode, the constraint on sleep transistor sizing can be relaxed, that is, the size of

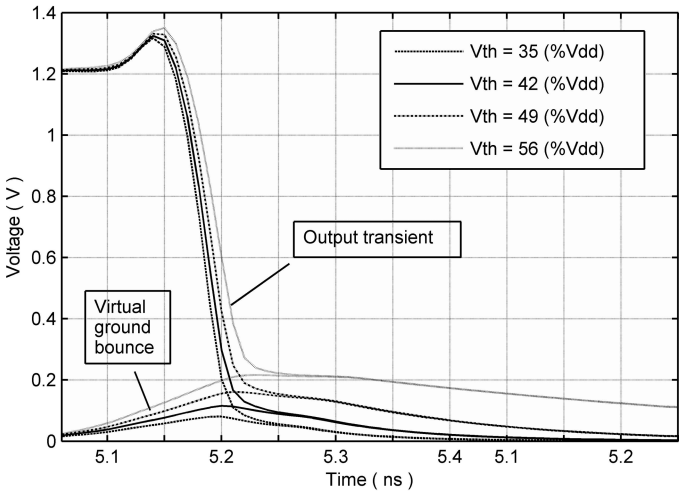

Fig. 8. Virtual ground vs. output delay.

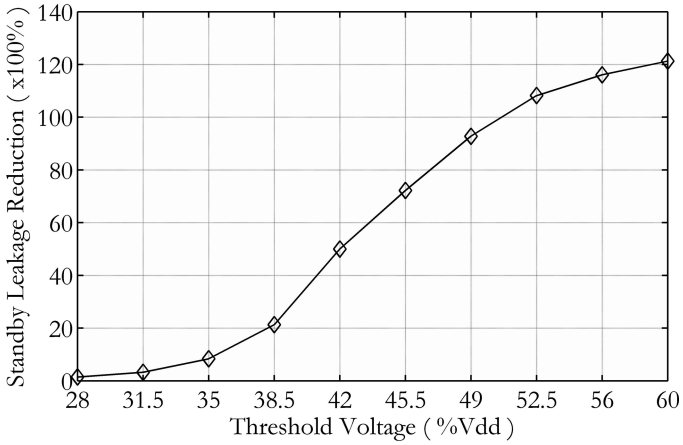

Fig. 9. Comparison of leakage power in the standby mode between the conventional sleep transistor circuit and QD transistor circuit.

QD sleep transistors can be reduced. Also, V_{th} can be increased in the standby mode to achieve more aggressive leakage reduction. Second, the high V_{th} of QD sleep transistors at the standby mode can be lowered while achieving the same level of leakage reduction. This is because the smaller size of QD sleep transistors results in larger equivalent resistance, which helps to further reduce the standby leakage current (even with a smaller V_{th} than the conventional technique). As shown in Figure 9, the QD sleep transistor design achieves as much as $120X$ standby leakage reduction over the conventional dual-V_{th} sleep transistor design.

The leakage power of the baseband DSP can be further reduced by reconfiguring the critical path threshold voltages under different DSP configurations in response to the changes in wireless channel conditions. Employing the method

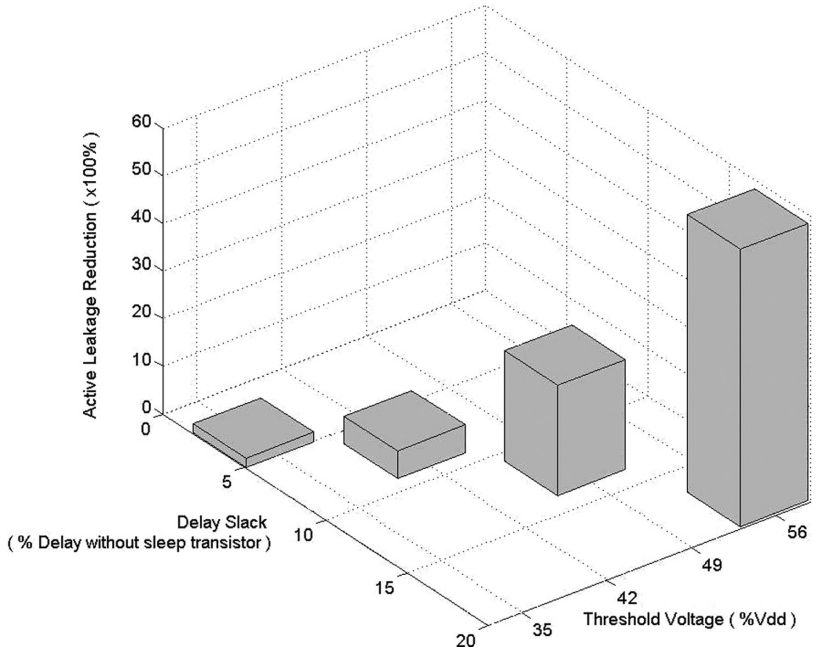

Fig. 10. Optimal operation points in terms of threshold voltages, delay slacks and active mode leakage reduction.

proposed in Section 3.3, we examine the relationship between threshold voltage, propagation delay and leakage current. The optimal operation points are obtained as shown in Figure 10. Consider the situation where the wireless channel is relatively reliable, for example, the mobile computing device is closer to the base station. The critical path of the LMS filter has, for example, a 16% of delay slack due to the smaller number of active taps and/or data precision. By adjusting the V_{th} of QD sleep transistors from 35% to 56% of V_{dd}, we can achieve about $60X$ reduction in active leakage power over the conventional dual-V_{th} sleep transistor design, thereby further improving the energy efficiency for mobile computing. In addition, the baseband DSP can also operate at intermediate modes under different situations by reconfiguring threshold voltages as shown in Figure 10 to achieve optimal power-performance trade-offs.

We need to point out that all the simulations are based on the QD device model [Hasaneen et al. 2004] that is sufficiently accurate for early-phase design exploration. Thus, we expect these results to well predict the leakage reduction in actual systems. Still, some issues related to the physical implementation need to be considered. As mentioned before, the voltage controller for programming V_{th}, if properly implemented, should not introduce net energy overhead. This is because the target applications of the proposed technique is battery-powered mobile computing platforms (e.g., cell phone, PDA), which usually stay idle in 90% of the time and thus V_{th} programming is performed very infrequently (e.g., only during mode switches and maybe once every several hours). Also, V_{th} programming can be finished quickly in hundreds of nanoseconds to

several microseconds. In these cases, the programming overhead can be offset by the gain in leakage reduction. Note that the proposed technique may not be beneficial to other applications (e.g., high-end computing with quickly changing workloads). Further study on this issue needs to consider the usage profiles and operational conditions of various computing platforms, which are currently unavailable. Nevertheless, the results here for embedded DSP circuits complement our studies.

In addition, the programming time has almost no noticeable impact on the performance of mobile computing systems, most of which are operated at the megahertz range. The high programming voltages are only connected to the gates of QD sleep transistors, which can be isolated from the rest of the digital circuits. The possible reliability problems can be controlled by employing mixed-signal IC design techniques. Please note voltage programming circuits are widely used in many circuits utilizing dynamic voltage scaling or similar techniques for energy efficiency. Most of the works show that the power overhead can be managed. Finally, the V_{th} programming requires extra layout area and routing. This is the overhead that we have to pay in order to achieve better power-performance trade-offs. These trade-offs are more critical for mobile computing systems.

5. CONCLUSIONS

Leakage reduction is very important for mobile computing devices due to the limited battery lifetime. In this article, we propose to exploit the programmable threshold voltage QD transistors for new avenues of power-performance optimization. Our work has demonstrated new capabilities for sustainable operation of mobile computing devices in energy-constrained situations with varying conditions. Future work is being directed towards addressing runtime variations in process parameters, voltages, and temperature. The associated architecture and microarchitecture level optimization solutions will also be investigated to exploit the new design options opened up by the QD transistors.

REFERENCES

Anis, M., Areibi, S., and Elmastry, M. 2003. Design and optimization of multithreshold CMOS (MTCMOS) circuits. *IEEE Trans. Comput.-Aid. Des. Integr. Circ. Syst. 22*, 10, 1324–1342.

Blyth, T., Khan, S., and Simko, R. 1991. A nonvolatile analog storage device using EEPROM technology. In *Proceedings of the IEEE International Solid-State Circuits Conference*, 192–193.

Bowman, K., Duvall, S., and Meindl, J. 2002. Impact of die-to-die and withindie parameter fluctuations on the maximum clock frequency distribution for gigascale integration. *IEEE J. Solid-State Circ. 37*, 2, 183–190.

Cao, Y., Sato, T., Sylvester, D., Orshansky, M., and Hu, C. 2000. Dual-threshold voltage techniques for low-power digital circuits. In *Proceedings of the IEEE Custom Integrated Circuits Conference*, 201–204.

Compagnoni, C. M., Ielmini, D., Spinelli, A. S., and Lacaita, A. L. 2005. Optimization of threshold voltage window under tunneling program/erase in nanocrystal memories. In *IEEE Trans. Electron Dev. 52*, 11, 2473–2481.

Fernandesa, A., DeSalvoa, B., Baronb, T., Damlencourta, J. F., Papona, A. M., Lafonda, D., Mariollea, D., Guillaumotc, B., Bessonc, P., Massonb, P., Ghibaudod, G., Pananakakisd, G., Martina, F., and Haukkae, S. 2001. Memory characteristics of Si quantum dot devices with $SiO_2/ALD\ Al_2O_3$ tunneling dielectrics. *International Electron Devices Meeting*, 741–744.

HAMZAOGLU, F. AND STAN, M. R. 2002. Circuit-level techniques to control gate leakage for sub-100nm CMOS. In *Proceedings of the International Symposium on Low Power Electronics and Design*, 60–63.

HASANEEN, E. 2004. Modeling and simulation of floating gate nanocrystal FET devices and circuits. Doctoral dissertation. University of Connecticut.

HASANEEN, E., HELLER, E., BANSAL, R., HUANG, W., AND JIAN, F. 2004. Modeling of nonvolatile floating gate quantum dot memory, *Solid-State Electron. 48*, 10-11, 2055–2059.

HASAN, T., LEHMANN, T., AND KWOK, C. Y. 2005. On-chip high voltage charge pump in standard low voltage CMOS process. *Electron. Lett.*, *41*, 15, 840–842.

HOOPER, M., KUCIC, M., AND HASLER, P. 2005. On-chip analog floating-gate array programming in a submicro standard CMOS process using high voltage charge pumps. In *Proceedings of the International IEEE-NEWCAS Conference*, 147–150.

HAYKIN, S. 1996. *Adaptive Filter Theory*, Prentice Hall.

ISCAS HIGH-LEVEL MODELS GROUP. http://www.eecs.umich.edu/~jhayes/iscas.restore/74283.html.

KINOSHITA, S., MORIE, T., NAGATA, M., AND IWATA, A. 1999. New nonvolatile analog memory circuits using PWM methods. *IEICE Trans. Electron.*, *E82-C*, 1655–1661.

KINOSHITA, S., MORIE, T., NAGATA, M., AND IWATA, A. 2001. A PWM analog memory programming circuit for floating-gate MOSFETs with 75-μs programming time and 11-bit updating resolution. *IEEE J. Solid-State Circ.*, *36*, 8, 1286–1290.

KURODA, T., FUJITA, T., MITA, S., NAGAMATU, T., YOSHIOKA, S., SANO, F., NORISHIMA, M., MUROTA, M., KALO, M., KINUGAWA, M., KAKUMU, M., AND SAKURAI, T. 1996. A $0.9V$ $150MHz$ $10mW$ $4mm^2$ $2-D$ discrete cosine transform core processor with variable-threshold-voltage scheme. In *Proceedings of the International Solid-State Circuits Conference*, 166–167.

LIN, Y.-T., CHUNG, W.-Y., WU, D.-S., CHANG, K.-S., AND CHEN, J.-J. 2005. Integrated low-voltage pulse width modulation circuit using CMOS processes. In *Proceedings of the IEEE-NEWCAS Conference*, 163–165.

LIU, A., KIM, M., NARAYANAN, V., AND KAN, E. C. 2000. Process and device characteristics of self-assembled metal nano-crystal EEPROM. *Superlatt. Microstruct. 28*, 5, 393–399.

LONG, C., REDDY, S., PAMARTI, S., HE, L., AND KARNIK, T. 2006. Power-efficient pulse width modulation DC/DC converters with zero voltage switching control. In *Proceedings of the International Symposium on Low Power Electronics and Design*, 326–329.

MUKHOPADHYAY, S. H. AND ROY, K. 2003. Modeling and estimation of total leakage current in nano-scaled CMOS devices considering the effect of process variation. In *Proceedings of the International Symposium on Low Power Electronics and Design*, 172–175.

MUTAH, S., DOUSEKI, T., MATSUYA, Y., AOKI, T., SHIGEMATSU, S., AND YAMADA, J. 1995. 1-V power supply high-speed digital circuit technology with multi-threshold voltage CMOS. *IEEE J. Solid-State Circ. 30*, 8, 847–853.

NARENDRA, S., BLAAUW, D., DEVGAN, A., AND NAJM, F. 2003. Leakage issues in IC design: Trends, estimation and avoidance. In *Proceedings of the International Conference on Computer Aided Design (Tutorial)*, 11.

NOGAMI, M. AND OHNO, A. 2003. Imaging of Si quantum dots as charge storage nodes. *Mater. Sci. Engin. C*. *23*, 6, 1047–1051.

NORMAND, V., KAPETANAKIS, E., TSOUKALAS, D., KAMOULAKOS, G., BELTSIOS, K., VAN, D. B. J. AND ZHANG, S. 2001. MOS memory devices based on silicon nanocrystal arrays fabricated by very low energy ion implantation. *Mater. Sci. Engin. C*, *15*, 1, 303–305.

NOSE, K. AND SAKURAI, T. 2000. Optimization of V_{DD} and V_{TH} for low-power and high-speed applications. In *Proceedings of the Asia and South Pacific Design Automation Conference*, 469–474.

PALUMBO, G., PAPPALARDO, D., AND GAIBOTTI, M. 2002. Charge pump circuits: Power consumption optimization. *IEEE Trans. Circ. Syst. I*, *49*, 11, 1535–1542.

QI, X., LO, S. C., LUO, Y., GYURE, A., SHAHRAM, M., AND SINGHAL, K. 2005. Simulation and analysis of inductive impact on VLSI interconnects in the presence of process variations. In *Proceedings of the IEEE Custom Integrated Circuits Conference*, 309–312.

ROLANDI, P. L., CANEGALLO, R., CHIOFFI, E., GERNA, D., GUAITINI, G., ISSARTEL, C., LHERMET, F., PASOTTI, M., AND KRAMER, A. 1998. M-cell 6 b/cell analog flash memory for digital storage. In *Proceedings of the International Solid-State Circuits Conference*, 334–335.

ROY, K., WEI, L., AND CHEN, Z. 1999. Multiple-V_{DD} multiple-V_{TH} CMOS (MVCMOS) for low power applications. In *Proceedings of the International Symposium on Circuits and Systems*, 430–435.

SAMAAN, S. 2004. The impact of device parameter variations on the frequency and performance of VLSI chips. In *Proceedings of the International Conference on Computer-Aided Design*, 343–346.

SHE, M. AND KING, T. J. 2003. Impact of crystal size and tunnel dielectric on semiconductor nanocrystal memory performance. *IEEE Trans. Electron Dev. 50*, 9, 1934–1940.

TANG, X., BAIE, X., COLINGE, J. P., CRAHAY, A., KATSCHMARSYJ, B., SCHEUREN, V., SPOTE, D., RECKINGER, N., VAN DE WIELE, F., AND BAYOT, V. 2000. Self-aligned silicon-on-insulator nano flash memory device. *Solid-State Electron. 44*, 12, 2259–2264.

TIWARI, S., RANA, F., HANAFI, H., HARTSTEIN, A., CRABBE, E. F., AND CHAN, K. 1996. A silicon nanocrystals based memory. *Appl. Phys. Lett. 68*, 10, 1377–1379.

TIWARI, S. 1996. Silicon nanocrystal memories: devices in the limit of conventional minaturization. In *Proceedings of the International Symposium on the Physics and Chemistry*, 250–259.

TRAN, H. V., BLYTH, T., SOWARDS, D., ENGH, L., NATARAJ, B. S., DUNNE, T., WANG, H., SARIN, V., LAM, T., NAZARIAN, H., AND HU, G. 1996. A 2.5 V 256-level nonvolatile analog storage device using EEPROM technology. In *Proceedings of the International Solid-State Circuits Conference*, 270–271.

TSCHANZ, J. W., NARENDRA, S. G., YE, Y., BLOECHEL, B. A., BORKAR, S., AND DE, V. 2003. Dynamic sleep transistor and body bias for active leakage power control of microprocessors. *IEEE J. Solid-State Circ. 38*, 11, 1838–1845.

WANG, S., DAI, J., HASANEEN, E., WANG, L., AND JAIN, F. 2008. Programmable threshold voltage using quantum dot transistors for low-power mobile computing. In *Proceedings of the IEEE International Symposium on Circuits and Systems*, 3350–3353.

WEI, L., CHEN, Z., ROY, K., YE, Y., AND DE, V. 1999. Mixed-Vth (MVT) CMOS circuit design methodology for low power applications. In *Proceedings of the Design Automation Conference*, 430–435.

YING, T., KI, W., AND CHAN, M. 2003. Area efficient CMOS charge pumps for LCD drivers. *IEEE J. Solid-State Circ. 38*, 10, 1721–1725.

Received June 2008; revised October 2008, March 2009; accepted March 2009 by Krishnendu Chakrabarty