LOW POWER NONVOLATILE SRAM CIRCUIT WITH INTEGRATED LOW VOLTAGE NANOCRYSTAL PMOS FLASH

Shantanu Rajwade*, Wing-kei Yu, Sarah Xu, Tuo-Hung Hou, G. Edward Suh and Edwin Kan

School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853
*contact: srr77@cornell.edu

ABSTRACT

This paper presents a new nonvolatile SRAM design that incorporates low-voltage nanocrystal PMOS Flash transistors. The design enables global store, restore and erase operations with negligible penalty on regular SRAM operation. Store/erase operations also do not consume much power even considering charge pump circuits. Circuit simulations based on experimental I-V characteristics demonstrate that 10 μs store/erase operation at ± 6 V is sufficient for correct restoration of the stored bit even under reasonable process variation.

I. INTRODUCTION

Low-power systems with unreliable power supplies can greatly benefit from nonvolatile (NV) SRAMs. Instance check-pointing can eliminate static power consumption and enable continuous operation across power supply failures. Several designs that incorporate resistive (Re) RAM [1-3], phase change (PC) RAM [4], magnetic (M) RAM [5] and ferroelectric (Fe) memory [6] with SRAMs have been previously proposed. Although ReRAM and PCRAM provide low voltage operation, they rely on accurate high current pulses to switch their nonvolatile states. This puts additional burden over the on-chip power supply as well as increases design complexity of peripheral circuitry. In order to be truly viable, NV-SRAMs must not only be CMOS compatible but also have minimal performance and power overheads.

In the last decade, NAND Flash memory has proved to be the most important driving force in enabling high density nonvolatile storage for mobile applications. However, due to the high voltage program and erase operation for these devices, they are only available in data storage chips instead of embedded with high-performance logic circuits. We propose a novel NV-SRAM design which integrates low voltage nanocrystal (NC) Flash [7] in every SRAM cell. The circuit enables regular SRAM operation in stable power supply. On sensing power failure or scheduled check point, the controller initiates a global store operation that performs back up of the current state with minimal power dissipation. During subsequent power recovery, the saved state in every cell is restored before regular operation. The design thereby enables seamless computation over power interruptions and shows great potential for embedded SoC applications that inherit unstable power supply.

The paper is organized as follows: Section II discusses NC Flash briefly; Section III describes NV-SRAM design and operation, Section IV evaluates performance characteristics, Section V studies the dependence of store/erase time on process variation to ensure correct restoration and finally, Section VI validates the low power advantage of using NC Flash.

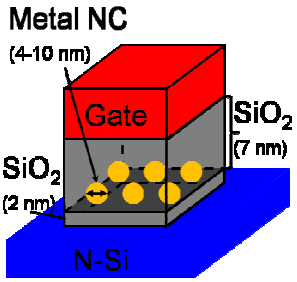

Figure 1: Schematic of NC PMOS Flash device incorporating metal NCs as nano-floating gates

II. NANOCRYSTAL FLASH

To ensure ten year retention of the stored charge, scaling of tunnel oxide in conventional polysilicon floating gate Flash is restricted to 7-9 nm owing to stress induced leakage currents. This constraint can however be eliminated by use of discrete charge storage floating gates like nitride traps [8] or nanocrystals (NCs) [9]. Reducing tunnel oxide thickness to 2-3 nm can bring down program/erase voltages of such devices below 8 V [7]. Further, metal NCs provide significant electric field enhancement in the tunnel oxide during program/erase assisting faster store/erase operation [10]. On chip high voltage generation of such magnitudes is achievable through charge pump circuits. Since the high voltages generated by charge pumps fall only across the high

impedence gates of Flash devices, they can be designed efficiently for minimal power dissipation.

Fig. 1 shows the schematic of PMOS NC Flash device. The PMOS NC Flash is programmed by applying a high positive voltage (~ 6 V) which puts the device in deep accumulation and enables tunneling of electrons from the n-substrate to the floating NC gates. Trapped electrons in the NCs favor easier depletion in the channel with applied negative bias (- 1 V with respect to bulk), shifting the threshold voltage in the positive direction (V_{TH} of PMOS is negative). The device is erased by applying a high negative voltage (~ -6 V) on the gate which pushes the electrons trapped in the NCs back to the substrate and restores V_{TH}.

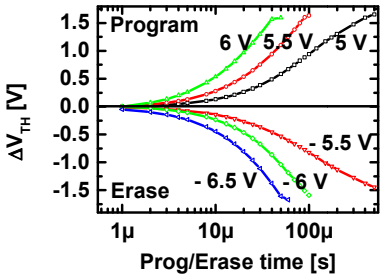

Figure 2: Simulated ΔV_{TH} against program/erase time for NC PMOS Flash incorporated in the proposed NV-SRAM circuit

NC devices are shown to achieve fast program/erase at low operating voltages as well as the ten year retention mark. Memory window (ΔV_{TH}) against the program/erase time is modeled based on experimental measurements and follows the methodology outlined in [11]. Fig. 2 illustrates the simulation of memory window for three different program/erase voltages against time. The device is seen to achieve ΔV_{TH} of 0.5 V with a program (erase) voltage of 6 V (- 6.5 V) in less than 10 µs.

III. NV-SRAM DESIGN AND OPERATION

Fig. 3 presents the NV-SRAM design with NC PMOS Flash transistors. Nodes Q and Qb are loaded with the Flash transistors controlled by the PE (program-erase) signal. They are accessed by the NMOS transistors driven by the EN signal. The circuit operation is classified into four states; REGULAR, STORE, POWER UP and ERASE. The following subsections explain the working of every operation in detail. Circuit simulations are performed in SPICE with 70 nm BSIM3v3 model [12] in low power (high V_{TH}) design at V_{DD} = 1 V operation. PMOS Flash transistors were modeled as described in [13] and references therein.

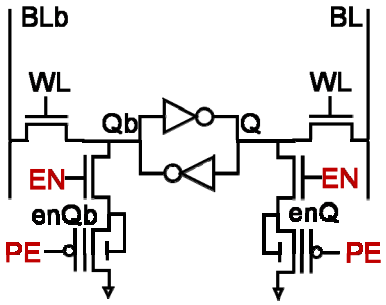

Figure 3: NV-SRAM cell integrating NC PMOS Flash devices controlled by the PE signal. They are accessed through NMOS pass transistors enabled by EN. Nodes enQ and enQb represent bulk terminals of PMOS Flash. Appropriate global switching of PE and EN signals achieves store/erase operation for each cell.

A. REGULAR Operation

The PMOS Flash transistors as well as the NMOS access transistors are turned off in REGULAR operation (EN = 0 V; PE = 1 V). The cell therefore resembles a volatile SRAM barring the extra capacitive loading at nodes Q and Qb.

B. STORE Operation

The memory controller initializes the STORE operation on sensing power supply failure or before entering a deep sleep mode. To handle a power failure, a peripheral circuit needs to detect the fall of V_{DD} below a threshold value and trigger the STORE operation. Modern processors often use such a circuit to generate an exception. We note that SRAM cells hold their state dynamically for more than a millisecond even when V_{DD} drops to ground, which provides sufficient time for the STORE operation. A system can also be designed to maintain its operation for a certain period after a power failure by adding a storage capacitor.

During STORE operation, the WL is disabled and EN is enabled. The body bias of the PMOS Flash devices, at enQ or enQb, are now controlled by the respective Q or Qb node. A suitable program voltage (5 – 6 V) is applied on PE leading to tunneling of accumulated electrons into the floating NC gates. The tunneling flux of electrons is exponentially dependent on the voltage drop between the body and the floating NCs of the PMOS Flash transistors. Therefore, the PMOS Flash with 'LOW' body bias is 'programmed' by exponentially higher electron injection than the other PMOS Flash. V_{TH} of the programmed device shifts significantly in the positive direction. Both the PMOS Flash transistors remain in accumulation (switched off) and the cell does not consume any channel current through the Q and Qb branches, in exception to any subthreshold leakage. At the end

of the STORE operation (~ 10 μs), the Flash transistors acquire a significant V_{TH} difference and the cell is ready to be powered down indefinitely.

The NMOS access transistors help decouple the PMOS Flash transistors from Q and Qb. This offers a possibility of making nodes enQ and enQb dynamic by switching EN off. Although this may need further investigation, adjusting the RC time constants of these dynamic nodes may enable simultaneous STORE and REGULAR operation.

In a traditional system with separate SRAM and Flash memory arrays, data must be copied sequentially. Given the limited bandwidth and time, only a small amount of data can be saved on a power failure. On the other hand, the STORE operation is local in the NV-SRAM cells and a large amount of data can be saved simultaneously.

C. POWER UP Operation

The POWER UP operation achieves the restoration of the stored state in every cell immediately after revival of the power supply. Fig. 4(a) shows the detailed sequence of control signals performing this global restoration. The POWER UP sequence begins with the precharge (PreQ) cycle, during which nodes Q and Qb are precharged to $V_{DD}/2$ by enabling all the WLs. This is followed by switching the EN signal as well the power supply V_{DD} in the restore cycle. The EN signal puts the access NMOS transistor in above threshold regime. The resistance of nodes Q and Qb to ground is therefore determined by the leakage through their respective PMOS Flash transistors. The 'programmed' Flash offers lower subthreshold resistance to the ground node. The internal feedback generated between nodes Q and Qb by the cross-coupled inverter pair amplifies this asymmetry to restore 'LOW' state in the 'programmed' PMOS Flash branch. The cell resumes normal (REGULAR) operation from the subsequent cycle.

Notice that the charge pump circuits are not required to be activated during the POWER UP operation.

The time elapsed in amplifying the small resistive difference into rail to rail output depends on the V_{TH} difference between the 'programmed' and the 'non-programmed' PMOS Flash as well as the small signal gain of the inverters. For the low power (high V_{TH}) design in our cells, this time is observed to be 200 - 400 ps, but may be reduced significantly by using high performance transistor design in the inverter pairs.

D. ERASE Operation

The ERASE operation restores the V_{TH} of all PMOS Flash transistors. The V_{TH} difference in 'programmed' and 'non-programmed' PMOS Flash does not participate in the REGULAR operation of the SRAM cell. Therefore, the memory controller has the liberty to perform ERASE during periods of low power utilization.

Figure 4(b) presents the timing diagram of the cell during ERASE operation. A suitable high negative voltage is applied at PE (-6 to -7 V) causing electrons trapped in floating NCs to tunnel back to the substrate. The PMOS Flash device is in deep inversion (low resistance) during this operation. The NMOS access transistor is turned off (EN = 0 V) to ensure no bias current flows in either of the branches. Hence, no additional power is expended on erasing the stored state of the cell except in the high PE voltage charge pump circuits. The ERASE operation (~ 10 - 50 μs) returns the V_{TH} of all PMOS Flash devices to the 'non-programmed' state.

NMOS access transistors are necessary for attaining no channel current during long ERASE operation. In doing so, they also achieve a dual purpose. They decouple the ERASE operation from the normal functioning of the SRAM cell; in other words allow REGULAR operation to run in parallel to ERASE. Fig. 4(b) shows the write (WR) cycles performed in parallel with ERASE. The bulk and source potential of the PMOS Flash devices (nodes enQ and enQb) is held at ground potential due to the highly conductive inversion layer in the device.

Table 1: Scheme of operation for proposed NV-SRAM cell

Operation	V_{DD}	WL	PE	EN	Notes
REGULAR	1 V	1/0 V	1 V	0 V	6 % performance penalty to read/write
STORE	1 V	0 V	5 to 6 V	1 V	t_{PROG} ~10 – 100 μs; minimal power overhead
POWER UP	1 V (delay)	1 V	0 to -1 V	1 V	V_{DD} enabled 1 clock cycle after PreQ
ERASE	1 V	1/0 V	-6 to -7 V	0 V	t_{ERASE} ~10 – 100 μs; can run parallel to REGULAR operation; minimal power overhead

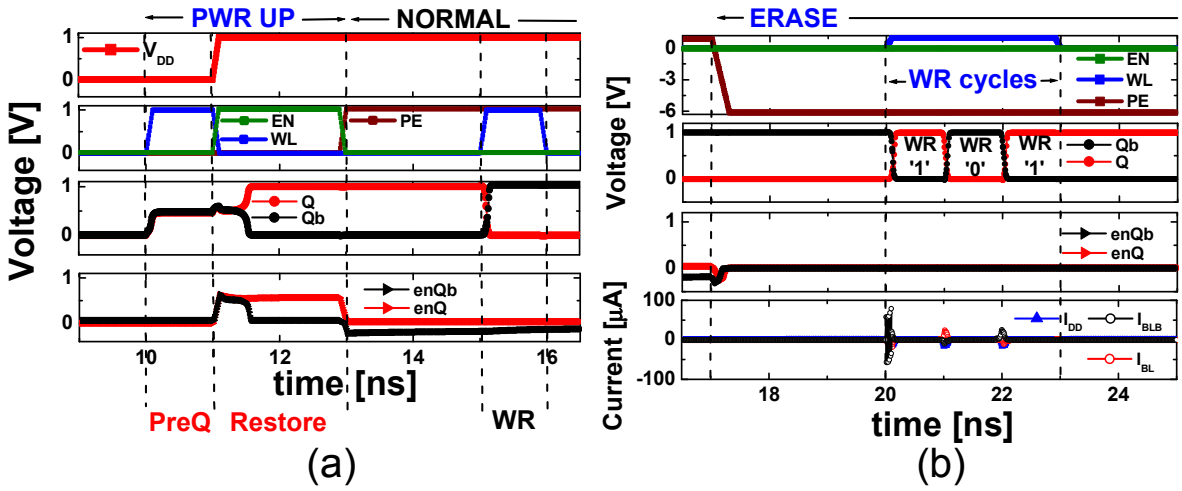

Figure 4: (a) Timing diagram for the POWER UP operation. The cell stores bit '1' (Q = 'HIGH') before powering down. Operation begins with the PreQ cycle that precharges nodes Q and Qb to $V_{DD}/2$. In the next cycle, WL is disabled; EN and V_{DD} are switched on. The 'programmed' PMOS Flash presents lower resistance to ground by higher leakage. This difference in node potentials is amplified by the inverter pair to restore the cell to the correct state. (b) Timing diagram for the ERASE operation. A high negative voltage is enforced on the PE node to initiate tunneling of electrons trapped on floating NCs back to the channel. NMOS access transistors are switched off (EN = 0 V); so the nonvolatile branches do not consume bias current. The controller can perform the REGULAR operation in parallel as shown by write (WR) cycles executed on the cell during ERASE.

Table I summarizes the various operating modes of the proposed NV-SRAM.

It should be noted that PMOS Flash cannot be replaced with conventional NMOS Flash device. PMOS (NMOS) Flash maintains a high resistance state during program (erase) and a low resistance state during erase (program). This inherent difference makes PMOS inevitable to disable any static current during STORE and ERASE operation.

IV. PERFORMANCE EVALUATION

The NV-SRAM design achieves nonvolatile backup and restore by means of globally controlled STORE, POWER UP and ERASE operations at negligible power penalty. The cell footprint is larger by 4 minimum sized transistors. Circuit simulations were performed at 1 GHz, although the operating frequency is not seen to be restricted by the capacitive loading at Q and Qb nodes. Frequency of operation is set by the low power (high V_{TH}) transistor switching time. The capacitive loading however is seen to result in 6 % penalty to inverter switching speed in REGULAR operation.

V. PROCESS VARIATION AND CIRCUIT PERFORMANCE

In the event of power suspension, the memory controller must utilize the remaining stored energy efficiently to maximize the backup volume of the current system state. In other words, time expended in high voltage STORE operation must be minimized. Minimum time for STORE operation, that ensures correct restoration of the stored bit in the subsequent POWER UP operation, is determined by the degree of matching in transistor pairs within a cell.

Transistors in a single cell are subject to area and V_{TH} mismatch as a result of process variation. Therefore, difference in resistance to ground node seen from Q and Qb nodes during POWER UP operation must be sufficient to overpower any opposing latch-up condition occurring in the cross-coupled inverter pair due to device mismatch. Fig. 5(a) shows the dependence of minimum time for program to compensate for opposing latch-up in the worst-case V_{TH} mismatch (strong-p and weak-n or vice versa) in inverter transistors. Fig. 5(b) illustrates the same constraint under worst case area mismatch in inverter transistors.

Fig. 5(c) and 5(d) demonstrate the minimum time required in STORE operation under V_{TH} and area mismatch in PMOS Flash transistor for correct restoration of the stored bit. Since program time has an inverse exponential relation with program voltage, V_{TH} variation is seen to be critical to minimum program time and especially serious to lower program voltages.

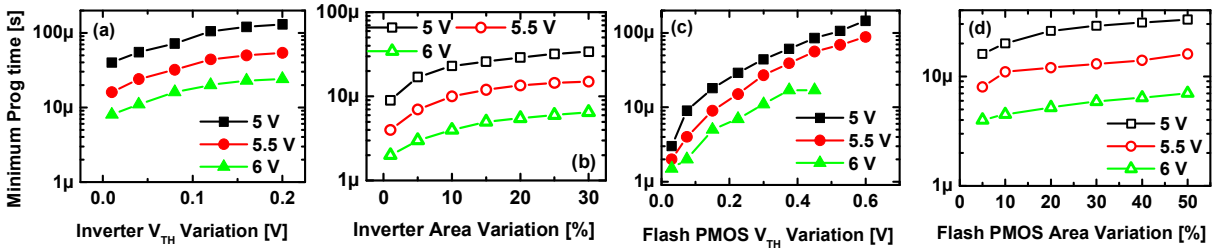

Figure 5: (a) Dependence of minimum program time (STORE operation) on worst-case (strong-p, weak-n) V_{TH} mismatch in the inverter pair. 15 µs of STORE operation at 6 V ensures correct restoration of stored bit even under opposing 0.1 V worst case mismatch. (b) Dependence of minimum program time (STORE operation) on worst case area mismatch in inverter pair. Area variation is seen to be less critical than V_{TH} mismatch. (c) Dependence of minimum time for STORE operation on V_{TH} mismatch in PMOS Flash transistors. (d) Dependence of minimum time for STORE operation on area mismatch in PMOS Flash transistors.

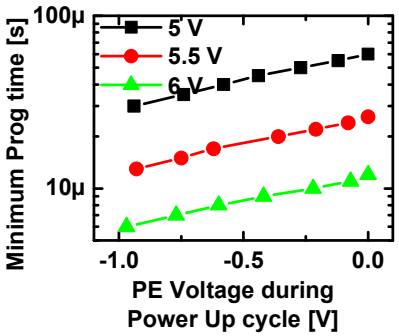

Figure 6: V_{TH} mismatch in inverters can be desensitized by enlarging the asymmetry in conductance to ground node seen from Q and Qb during POWER UP. This asymmetry is increased by applying a small negative voltage on PE while still maintaining the Flash device in subthreshold region. Simulation here shows reduction in minimum time of STORE operation with decreasing PE bias.

Flash devices are prone to V_{TH} variation due to effective oxide thickness fluctuation in thicker gate stack. During POWER UP, PMOS Flash operates in subthreshold and is the most resistive device in the leakage path. This resistance decreases exponentially with the applied negative bias at PE and may compensate for device mismatch effectively. In other words, the asymmetry in the conductance to ground node can be boosted by applying a small negative bias on PE (but still keeping the device in subthreshold) which helps overpower opposing force resulting from inverter mismatch. This facilitates in bringing down the lower bound on STORE operation time. Fig. 6 illustrates the reduction in program time with applied negative bias at PE during POWER UP for a 0.1 V worst case mismatch in the inverter pair.

Leakage characteristics of NMOS access transistors become critical during ERASE operation. As seen from Fig. 7, low V_{TH} of the access transistor results in exponentially higher power dissipation stemming from the leakage through the device. NMOS access transistors should be designed to provide low leakage at zero bias (V_{TH} > 0.4 V).

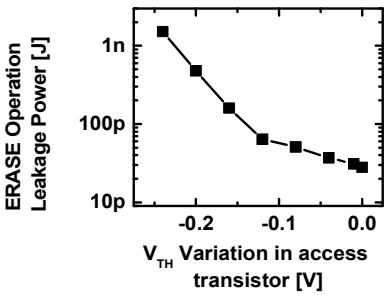

Figure 7: Leakage power during ERASE operation due to V_{TH} variation (V_{TH0} = 0.5 V) in the NMOS access transistor

VI. POWER EVALUATION

The STORE and ERASE operations in the proposed NV-SRAM design do not consume any bias current and therefore offer negligible power overhead at the cell level. Other proposals of NV-SRAM that incorporate ReRAM, PCRAM or MRAM consume high set/reset currents during these operations [14, 15]. Delivering precise high current pulses introduces complexity of current mirrors, temperature compensation techniques and variable I-R drops in the word lines, especially during power instability. Besides, it prohibits parallel program/erase of large memory blocks due to current compliance issues. On the contrary, generation of on-chip high voltages is well established by efficient charge pump designs [16]. Further so, as these high voltages appear only across the high impedence gates of Flash transistors, power consumption can be reduced significantly by means of reduced clock speeds in

charge pumps. This unique advantage in our design presents no additional complexity to the peripheral circuitry.

Table 2 presents a comparative summary of estimated energy dissipation per cell at 70 nm node for STORE operation against proposed NV-SRAM design which includes the high PE generating charge pumps. As seen from the projected data, the PMOS Flash design offers a competitive edge over several other proposed low power nonvolatile architectures. Also, the high voltage charge pump in this design, which accounts for a majority of power overhead can be maintained dynamically, so that there is no instantaneous increase in power consumption during the back-up operation. These unique advantages make NC Flash based NV-SRAM a compelling approach for embedded designs even with the relatively high voltages and program times.

Table 2: Comparative study of estimated power dissipation per cell during STORE operation for various proposed NV-SRAMs

NV-SRAM Design	Estimated STORE operation Power (μW)
ReRAM [1]	24.6
PCRAM [4]	378
MRAM [5]	32.2
FeRAM [6]	0.124
This work	0.075

VII. CONCLUSION

We have proposed a new low power low voltage NV-SRAM design performing global STORE, RESTORE and ERASE operations with minimal power overhead. The minimum time required for critical STORE operation was evaluated under process variations. It offers 10 μs program/erase at ± 6 V operation with negligible cell level power dissipation. The design ensures instant recovery from power instability and shows promise in low power embedded SoC architectures with minimal performance penalty.

ACKNOWLEDGMENT

This work was partially supported by the National Science Foundation under grants CNS-0708788 and CNS-0932069, and an equipment donation from Intel Corporation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or Intel Corporation.

REFERENCES

1. W. Wang, A. Gibby, Z. Wang, T. W. Chen, S. Fujita, P. Griffin, Y. Nishi and S. Wong, “Nonvolatile SRAM cell,” in *IEDM Tech. Dig.*, Dec. 2006, pp. 1-4.
2. X. Xue, G. Jin, J. Zhang, L. Xu, Y. Ding, Y. Xie, C. Zhao, B. A. Chen and Y. Lin, “Nonvolatile SRAM cell based on Cu_xO,” in *ICSICT Dig. Tech. Papers*, Oct. 2008, pp. 869-871.
3. S. Yamamoto, Y. Shuto and S. Sugahara, “Nonvolatile SRAM (NV-SRAM) using Functional MOSFET merged with resistive switching devices,” in *Proc. CICC*, Sept. 2009, pp. 531-534.
4. M. Takata, K. Nakayama, T. Izumi, T. Shinmura, J. Akita and A. Kitagawa, “Nonvolatile SRAM based on phase change,” in *Proc. Nonvolatile Semicond. Mem. Workshop*, Feb. 2006, pp. 95-96.
5. N. Sakimura, T. Sugibayashi, R. Nebashi and N. Kasai, “Nonvolatile magnetic flip-flop for standby-power-free SoCs,” *IEEE J. Solid-State Circuits*, vol. 44, no. 8, pp. 2244-2250, Aug. 2009.
6. T. Miwa, J. Yamada, H. Koike, H. Toyoshima, K. Amanuma, S. Kobayashi, T. Tatsumi, Y. Maejima, H. Hada and T. Kunio “NV-SRAM: A Nonvolatile SRAM with backup ferroelectric capacitors,” *IEEE J. Solid-State Circuits*, vol. 36, no. 3, pp. 522-527, Mar. 2001.
7. J. Lee and D. Kwong “Metal nanocrystal memory with high-κ tunneling barrier for improved data retention,” *IEEE Trans. Electron Devices*, vol. 52, no. 4, pp. 507-511, Apr. 2005.
8. M. H. White, D. A. Adams and J. Bu, “On the go with SONOS,” *IEEE Circuits Devices Mag.*, pp. 22-31, Jul. 2000.
9. S. Tiwari, F. Rana, K. Chan, H. Hanafi, W. Chan and D. Buchanan, “Volatile and nonvolatile memories in silicon with nano-crystal storage,” in *IEDM Tech. Dig.*, Dec. 1995, pp. 521-524.
10. C. Lee, U. Ganguly, V. Narayanan, T. Hou, J. Kim and E. C. Kan, “Asymmetric electric field enhancement in nanocrystal memories,” *IEEE Electron Device Lett.*, vol. 26, no. 12, pp. 879-881, Dec. 2005.
11. T. Hou, C. Lee, V. Narayanan, U. Ganguly and E. C. Kan, “Design optimization of metal nanocrystal Memory–Part I: nanocrystal array engineering” *IEEE Trans. Electron Devices*, vol. 53, no. 12, pp. 3095-3102, Dec. 2006.
12. Device Group, UC Berkeley, “Berkeley Predictive Technology Model.”
13. L. Larcher, P. Pavan, S. Pietri, L. Albani and A. Marmiroli, “A new compact DC model of floating gate memory cells without capacitive coupling coefficients,” *IEEE Trans. Electron Devices*, vol. 49, no. 2, pp. 301-307, Feb. 2002.
14. I. G. Baek, M. S. Lee, S. Seo, M. J. Lee, D. H. Seo, D. Suh, J. C. Park, S. O. Park, H. S. Kim, I. K. Yoo, U. Chung, and J. T. Moon, "Highly scalable non-volatile resistive memory using simple binary oxide driven by asymmetric unipolar voltage pulses," *IEDM Tech. Dig.*, Dec. 2004, pp. 587 - 590.
15. S. Lai, “Current status of phase change memory and its future,” in *IEDM Tech. Dig.*, Dec. 2003, pp. 255-258.
16. K. Choi, J. Park, J. Kim, T. Jung and K. Suh, “Floating-well charge pump circuits for sub-2.0 V single power supply Flash memories,” in *Symp. VLSI Circuits Dig. Tech. Papers*, June 1997, pp. 61-62.