Sub-threshold Circuit Design with Shrinking CMOS Devices

Benton H. Calhoun, Sudhanshu Khanna, Randy Mann, and Jiajing Wang
Charles L. Brown Department of Electrical and Computer Engineering
University of Virginia, Charlottesville, VA, USA
<bcalhoun, sk4fs, rwm3p, jjwang>@virginia.edu

***Abstract*— This paper examines the impact of technology scaling to 22nm on sub-threshold circuit design and proposes several solutions for sub-threshold circuits in new processes. To maintain energy-efficient sub-threshold operation, we must reduce variation and suppress leakage current. To combat random variation and minimize energy for nodes below 45nm, we show that special strategies are needed for different categories of sub-threshold circuits.**

I. Introduction

Sub-threshold (sub-V_T) design has proven useful for ultra-low-power (ULP) and low-energy applications since dynamic energy consumption is reduced quadratically with V_{DD} and minimum energy operation usually occurs in the sub-threshold region (e.g. [1][2][3]). Reduced on-off current ratios and heightened sensitivity to variations are the primary challenges for sub-V_T circuits, and these challenges increase as CMOS devices continue shrinking. This paper examines how sub-V_T circuits scale to the 22nm process node.

To motivate this investigation, we first consider three ways to use sub-V_T circuits. The obvious first category of use is energy-constrained applications that permit low performance, which is where sub-V_T circuits most commonly appear (e.g. microsensors, implants, RFIDs, etc.). The second category is energy-constrained portable devices that must occasionally support high performance. Ultra dynamic voltage scaling (UDVS) from strong inversion (for high speed) to sub-V_T supports this type of bursty operation [6]. The final category uses sub-V_T as low overhead support for high performance applications, such as standby management when strong inversion circuits are asleep. For example, a sub-V_T controller and sensors in [16] implement a closed-loop V_{DD}-scaling system to aggressively reduce SRAM leakage. Strictly ULP applications are often asleep, and sleep mode leakage makes older technologies consume less overall energy for this category of sub-V_T use [4][5]. For that reason, burst mode and standby support circuits are the most compelling drivers for sub-V_T operation in processes below 45nm.

To examine how deeply scaled processes impact sub-V_T design, we introduce low power (LP) predictive technology models (PTMs) in Section II. We then show how variations impact functionality (Section III), minimum energy operation (Section IV), and memory (Section V) as technologies scale. We propose general and specific strategies for increasing the robustness of sub-V_T circuits to 22nm.

II. Modeling Advanced Low Power Processes

For advanced CMOS nodes (below 100nm), micro-electronics suppliers offer at least two technology options to cover the broad high volume application space. A high-performance (HP) process targets the microprocessor and gaming market and a low-power (LP) process supports hand held, battery powered, and low standby power applications. Using sub-V_T circuits exhibits a concern for power that is most compatible with LP processes, but, at the time of this writing, published predictive technology models (PTMs) only offer HP technologies [8].

To investigate scaling of LP processes, we generated customized LP PTMs for nodes from 90nm to 22nm for a conventional poly silicon/nitrided-silicon dioxide stack that are consistent with published LP technology data [9][10][11]. Table I provides the essential metrics and scaling assumptions for our LP models. Because variation is an intrinsic property of the technology and a fundamental limiter to scaling, we incorporated variability into the models: local V_T variation due to random dopant fluctuations (RDF) (based on published measurements [12]), global V_T variation (affects P/N ratio), and channel length (L) variation (global and local). We matched T_{ox} and channel doping values to industrial trends, which resulted in a constant relationship between σ_{VT} and $(WL)^{-0.5}$ across processes [13].

TABLE 1 – Key metrics for our LP technology models 90nm-22nm

	Node	90		65		45		32		22	
	Device Type	N	P	N	P	N	P	N	P	N	P
Target values	Vdd (V)	1.2		1.2		1.1		1.1		1	
	Tox (nm)	2.2		2		1.8		1.6		1.4	
	Lpoly (nm)	80		56		39		27		19	
	Ion (uA/um)	420	180	600	300	620	300	700	380	720	380
	Ioff (pA/um)	15		250		400		1000		2000	
	HVT Ion(uA/um)	370	130	400	210	410	210	440	340	450	340
	HVT Ioff(uA/um)	4		10		30		50		150	
Tuned model values	Ion (uA/um)	422	178	606	305	615	309	709	381	725	385
	Ioff (pA/um)	17	16	250	219	477	409	947	965	1858	1915
	HVT Ion(uA/um)	383	149	409	220	425	229	444	330	469	331
	HVT Ioff(pA/um)	4	4	10	9	36	35	45	62	183	172

This work was funded in part by the SRC, FCRP/DARPA C2S2, and the NSF WiCAT center.

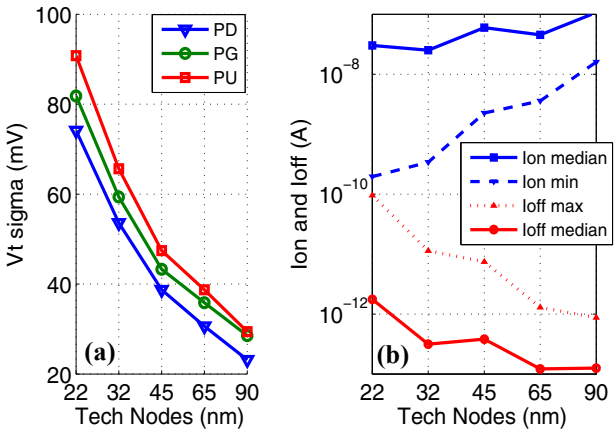

Figure 1 – (a) Impact of scaling trends on SRAM device Vt sigma and (b) PD I_{on} and I_{off} trends for 1000 M-C cases, V_{DS}=0.5V.

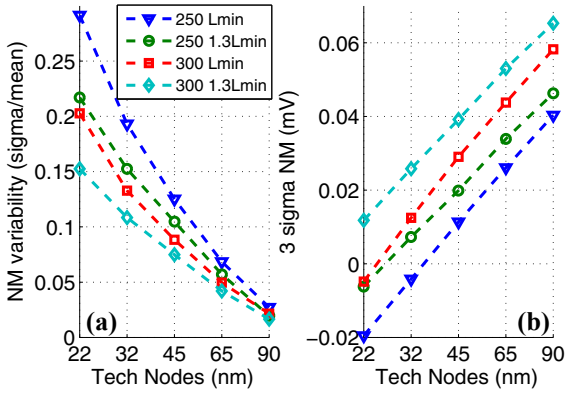

Figure 2 – (a): NM Variability at 250mV and 300mV (b) 3 sigma worst case NM as functions of scaling (using SVT LP models)**.**

The 3σ value for both local and global L variation equals 10% of the target ($L_{physical}$) for each node. The 3σ variation in global V_{TO} for NMOS and PMOS was 30mV. The model to target comparison is shown in Table 1 for standard V_T (SVT) and high V_T (HVT) versions of the LP PTMs.

The LP PTMs capture three key factors that limit scaling in general and sub-V_T operation in particular: (1) variability, (2) leakage and (3) sub-threshold device characteristics. Figure 1(a) shows the σ_{VT} value for each of the HVT transistors in 6T SRAM bit-cells from 90nm to 22nm. Figure 1(b) shows how growing sub-V_T slope decreases the mean I_{on}/I_{off} with scaling and how variability makes this ratio much worse. Despite selecting bitcell dimensions to match optimized values from industry publications (e.g. [10][11]), the predicted variation in the devices becomes prohibitive below 45nm. Sub-V_T operation will reduce the gate-tunneling leakage and gate induced drain leakage (GIDL) contributions that are rising sharply for conventional operation with scaling. However, the on/off current ratio coupled with variability will become the dominant limiter for low V_{DD} operation.

III. Impact of Scaling on Functionality

A. *Noise Margin in Sub-V_T logic circuits*

This section shows the impact of local variation on sub-V_T logic gates. RDF becomes the biggest source of local variation for scaled nodes, and this fact becomes more evident when narrow devices are used. For a large circuit, parameters like delay and leakage current see an averaging effect and are less affected by RDF. However, circuit failure due to inadequate noise margin (NM) in any of the gates has no averaging effect, so RDF threatens basic functionality.

Figure 1(b) shows how variation decreases the worst-case I_{on}/I_{off} ratio for a given transistor size, and this can degrade further for logic gates for two reasons. First, the complementary nature of static CMOS gates pits the on-current of PMOS against the off-current of NMOS (or vice versa), so global variations or sub-V_T process imbalance [7] can aggravate I_{on}/I_{off} reduction. Secondly, topological features like stacking or parallel leakage paths degrade gate I_{on}/I_{off}.

We characterize the NM for logic gates using a butterfly curve of a 2-input NAND-NOR gate loop [14] in SVT LP-PTMs (NAND and NOR gates have the worst VOL and VOH respectively from amongst basic logic gates). Figure 2(a) shows the NM variability (σ/μ) from 90nm to 22nm, and Figure 2(b) quantifies the 3σ values of those NMs. At 250mV and minimum L, the NMs verge on failure below 65nm, and there is greater than 10x increase in the variability of NM from 90nm to 22nm.

We now investigate circuit knobs to improve NM with a secondary aim of minimizing energy consumption. NM is a function of V_T variation and hence channel area, W*L. Increasing W decreases σ_{VT} but also increases dynamic and leakage energy. In contrast, increasing L decreases σ_{VT} while lowering leakage current. Thus we consider increasing L with constant W as one powerful knob to improve NM. Raising V_{DD} is a second knob that improves NM by raising I_{on} exponentially. Figure 2 shows that increasing V_{DD} by 20% or upsizing L by 30% increase NM and reduce NM-variability by similar amounts. By using these knobs together (e.g. 300mV, 1.3×L_{min}), we achieve positive 3σ NM at 22nm.

There is a minimum V_{DD} (V_{DDmin}) for any given L that provides a desired NM value. Table 2 shows V_{DDmin}, L combinations for a target NM value of 0.1×V_{DD}. Figure 2 shows that as we scale, the mean V_T decreases, making V_{DD} and L relatively more effective knobs for improving NM. V_{opt} in Table 2 is the voltage that minimizes energy for an adder.

Table 2: V_{DD}, L combinations for achieving NM>=0.1V_{DD} in LP-PTMs

Node (nm)	V_{DDmin} @ L=L_{min}	V_{DDmin} @ L=1.3L_{min}	Optimal V_{DD} (Vopt)
90	150 mV	100 mV	240 mV
65	200 mV	150 mV	250 mV
45	250 mV	200 mV	260 mV
32	300 mV	250 mV	270 mV
22	350 mV	300 mV	300 mV

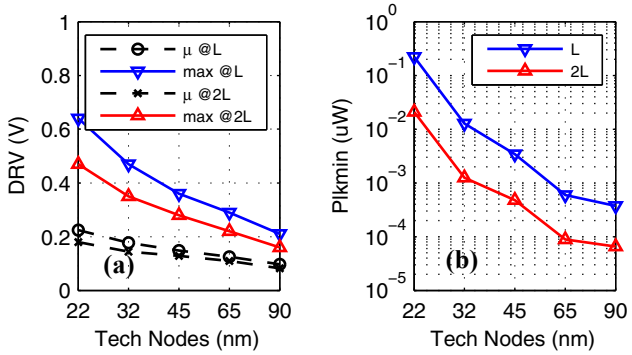

Figure 3 – (a) DRV vs LPPTM nodes (b) Leakage power of a 10-Kb SRAM with minimum V_{DD} set by DRV with L as well as 2L.

In summary, variation and scaling create a V_{DD} floor (V_{DDmin}) due to NM limitations that depends on L. Further, multiple combinations of V_{DDmin} and L can provide a desired NM, and these knobs grow more powerful with scaling.

B. Data retention in Sub-V_T SRAM

SRAM cells use device sizes near the minimum to achieve a competitive density, so V_T variation has an increased impact on SRAM functionality in sub-V_T. Read stability and write ability are the first metrics to fail in 6T bitcells at low voltages, but novel bitcells (>6T) that use various read/write assist circuits have been proposed to overcome these obstacles. As a result, the fundamental limit of V_{DD} scaling in sub-V_T SRAM is hold stability (or data retention). The data retention voltage (DRV) defines the lowest V_{DD} for which a cell retains its data. To maintain data in an entire SRAM array, the minimum V_{DD} (V_{DDmin}) for an error-free SRAM must at least equal the maximum DRV (DRV_{max}) value of all the cells, i.e. the DRV of the worst cell across the chip. DRV_{max} limits the minimum achievable leakage power (Plk_{min}) of an SRAM array. For single-V_{DD} chips with substantial sleep times, it can also limit the minimum achievable energy.

Figure 3(a) shows the DRV across LP processes from Monte-Carlo (M-C) simulation of a 10-Kb SRAM, assuming independent V_T variation on each device in the 6T cell (Fig. 1(a)). DRV mean and max values increase with scaling even beyond the logic NMs due to smaller devices and larger numbers of SRAM cells. For our LP-PTM nodes, HVT V_{TSAT} remains around 0.5V, so the worst DRV value exceeds V_T for even a modestly sized SRAM at 22nm. V_T variation thus defines the minimum voltage at which the technology can reliably store data. This voltage floor sets a practical lower bound on V_{DD}, although this floor can be adjusted by transistor upsizing, redundancy, and error correction techniques. As with NM, increasing L reduces DRV and cell leakage current, but upsizing L is limited by area and lower on-current. Figure 3(a) shows that doubling L in each bitcell FET effectively reduces the DRV_{max} value and leakage (Figure 3(b)). Low-V_{DD} SRAM is severely DRV-limited below 45nm. Increasing L toward 2L lowers DRV_{max} and gives comparable or lower leakage power than minimum L bitcells at the previous node.

IV. Minimum Energy Analysis for Sub-Vt Logic

Energy reduction usually motivates sub-V_T circuit design. This section shows the impact of scaling and V_{DD}, L optimization on energy. The V_{DD} point (V_{opt}) that minimizes energy per operation (E_{min}) for a given circuit and technology node typically is below V_T, but higher leakage power relative to active power pushes V_{opt} higher [1]. To see scaling effects on E_{min}, we first examine a 32-bit Kogge-Stone adder and assume that any sleep mode is optimized separately.

Figure 4(a) summarizes energy at V_{opt} across technology. The reduction of 2.75X in E_{min} from 90nm to 22nm results from lower dynamic energy. Leakage energy remains nearly constant as the leakage current goes up and delay decreases with scaling. Figure 4(b) shows the dynamic, leakage and total energy for the adder ($L=L_{min}$) versus V_{DD} at 90nm and 22nm. The high logic depth of the adder (~30 FO1 inverters) makes delay variation due to RDF small. Going from 90nm to 22nm, V_{opt} increases because decreasing dynamic energy coupled with increasing I_{off} with scaling leads to a larger contribution of leakage power to the total.

Previously, we saw that the choice of L allows different V_{min} values based on NM (or DRV). For a given process and V_{DD}, changing L produces less than 10% change in energy numbers. In some cases, $V_{DDmin} < V_{opt}$, meaning that operation at V_{opt} is possible with adequate NM. However, for 32nm and 22nm, $V_{DDmin} > V_{opt}$ at $L=L_{min}$, which means the adder must operate at a higher than optimal energy point due to NM limits. Using $L=1.3L_{min}$ brings $V_{DDmin} < V_{opt}$, enabling minimum energy operation. This shows how L helps reduce energy in scaled nodes. If SRAM is added to the circuit, it will impose a potentially higher V_{DDmin} due to the DRV, but it will also increase V_{opt} due to the additional leakage. This tradeoff between V_{opt} and V_{DDmin} is crucial at scaled nodes, and L and V_{DD} remain the most powerful knobs for tuning it.

V. Targeted Variation Reduction Methods

We have proposed that V_{DD} and L are strong knobs for reducing variation in general sub-V_T circuits, but the heightened impact of variation in scaled processes often demands more drastic measures. We argue that, to maintain the viability of sub-V_T operation at scaled nodes, context specific variation reduction methods are additionally necessary. These usually leverage the exponential impact of voltage (V_{GS}) on the I_{on}/I_{off} ratio in specific circuit contexts.

For example, SRAM read and write stability degrade

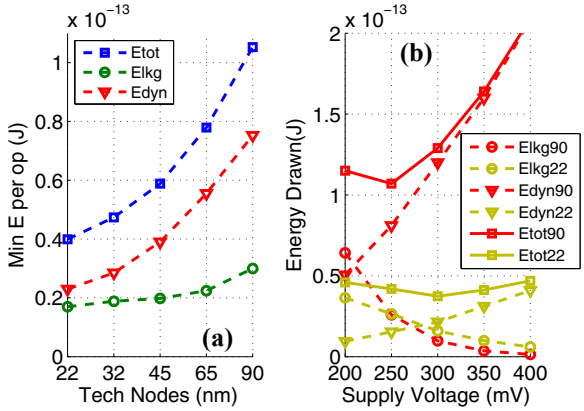

Figure 4 – (a) Min E/op as a function of scaling and (b) Energy components at 22nm and 90nm as a function of V_{DD}.

rapidly due to variation at lower voltages, and effective read/write assist techniques become essential for sub-V_T SRAM [3, 17-20]. Adding a separate read path to the bitcell eliminates the read stability issue, making write failures and read failures due to bitline leakage the major obstacles to robust sub-V_T SRAM operation below 65nm. Wordline boosting or bitcell V_{DD} collapse [17] rapidly improve writability by increasing the passgate to pullup current ratio exponentially. We specifically evaluate the effectiveness of V_{DD} collapse at 22nm. We use the wordline-based write margin metric, which defines the margin between V_{DD} and wordline voltage when the cell nodes flip during wordline voltage sweeping, to identify the writability. Figure 5 shows that the cell write failure rate effectively decreases when using a collapsed cell supply voltage (V_{DDa}) for a 6T bitcell at the 22nm node. Similar voltage knobs (e.g. negative wordline or boosted cell V_{SS} for unselected cells, boosted wordline for selected cells) can reduce BL leakage and/or increase I_{on} to reduce read access failures.

As a second example, consider the offset voltage of voltage mode latching sense amps (SAs), which are important for SRAM reads and many other applications. Straightforward upsizing of the input FETs to a SA actually increases offset. Careful analysis of the equations governing the SA offset show that the common mode voltage (V_{INDC}) impacts the offset more than does sizing, so a joint voltage-sizing co-optimization is the best solution for minimizing offset [15]. Again, circuit-specific use of powerful knobs like voltage produce the best reduction of variation effects.

Finally, sub-V_T circuits can also be utilized within high-performance systems for specific purposes, such as standby power reduction. A canary-replica based feedback loop uses a sub-V_T controller to aggressively scale standby V_{DD} for SRAM without losing data in [16]. The canary replicas use the power of voltage knobs in sub-V_T to induce failure so they can predict the proximity of failure in the core cells. In addition, the controller uses longer transistors, redundancy and majority voting techniques to combat variation effects [16].

VI. CONCLUSIONS

In summary, we refine PTMs to capture the scaling trends of low-power CMOS technologies and their effect on sub-V_T circuits. Variation and leakage current remain the major obstacles to energy savings down to 22nm, but they are amplified. We find that there is a net improvement in minimum energy with scaling in the LP technologies. Although Ioff is increasing with scaling, the improvement in delay results in a net improvement in energy. Variation fundamentally limits logic NM and SRAM DRV. V_{DD} and L knobs for minimum energy become more useful in scaled technologies generally, but targeted application of voltage knobs to specific circuit contexts are also necessary for maintaining sub-V_T functionality.

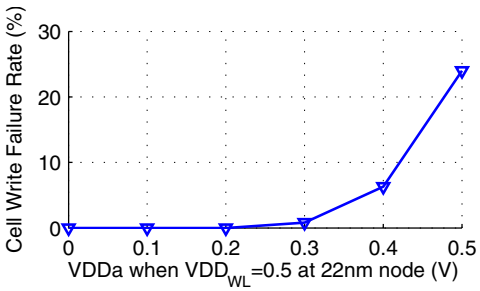

Figure 5 – Cell write error rate with the decease of cell V_{DD} (V_{DDa}) in a 10-Kb SRAM when wordline voltage (VDD_{WL}) is 0.5V for 22nm.

REFERENCES

[1] B.H. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and Sizing for Minimum Energy Operation in Sub-threshold Circuits," IEEE JSSC, Vol. 40, No. 9, pp. 1778-1786, September 2005.

[2] S. Hanson, B. Zhai, K. Bernstein, D. Blaauw, A. Bryant, L. Chang, K. Das, W. Haensch, E. J. Nowak, D. M. Sylvester, "Ultralow-voltage, minimum-energy CMOS," IBM Journal of Research & Development, Vol. 50, Issue 4/5, pp. 469-490, Jul/Sep2006

[3] B.H. Calhoun and A. Chandrakasan, "A 256kb 65nm Sub-threshold SRAM Design for Ultra-low Voltage Operation," IEEE JSSC, Vol. 42, No. 3, pp. 680-688, March 2007.

[4] M. Hemstead, N. Tripathi, P. Mauro, G. Y. Wei, and D. Brooks, "An ultra-low power system architecture for sensor network applications," ISCA, pp. 208-219, 2005.

[5] B. H. Calhoun, J. Bolus, S. Khanna, A. D. Jurik, A. C. Weaver, T. N. Blalock, "Sub-threshold operation and cross-hierarchy design for ultra low power wearable sensors," ISCAS, 2009.

[6] B. H. Calhoun and A. Chandrakasan, "Ultra-dynamic voltage scaling (UDVS) using sub-threshold operation and local voltage dithering," JSSC, vol. 40, no. 9, pp. 1178-1786, Sept. 2005.

[7] J.F. Ryan, J. Wang, B.H. Calhoun, "Analyzing and modeling process balance for sub-threshold circuit design," GLSVLSI, March 2007.

[8] W. Zhao and Y. Cao, "New generation of predictive technology modeling for sub-45nm early design exploration," IEEE Trans.on Electron Device, vol. 53, no. 11, pp. 2816-2823, Nov. 2006.

[9] Wu, C.C. et al., "A 90-nm CMOS device technology with high-speed, general-purpose, and low-leakage transistors for system on chip applications," IEDM, pp. 65-68, 2002.

[10] Luo, Z. et al., "High performance and low power transistors integrated in 65nm bulk CMOS technology," IEDM, pp.661-664, 2004.

[11] Shien-Yang Wu et al., "A 32nm CMOS Low Power SoC Platform Technology for Foundry Applications with Functional High Density SRAM," IEDM, pp.263-266, 2007.

[12] K. Rochereau, R. Difrenzs, J. Mc Ginley, 0. Noblanc, C. Julien, S. Parihar, P. Llinares, "Impact of pocket implant on MOSFET mismatch for advanced CMOS technology", ICMTS, pp. 123-126, 2004.

[13] Asenov, A.; Brown, A.R.; Davies, J.H.; Kaya, S.; Slavcheva, G. "Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs", IEEE Transactions on Electron Devices, Vol. 50, No. 9, pp. 1837–1852, Sept. 2003.

[14] J. Kwong and A. Chandrakasan, "Variation-driven device sizing for minimum energy sub-threshold circuits," ISLPED, October 2006.

[15] J.F. Ryan and B.H. Calhoun, "Minimizing Offset for Latching Voltage-Mode Sense Amplifiers for Sub-threshold Operation," ISQED, pp. 127-132, 2008.

[16] J. Wang and B.H. Calhoun, "Canary Replica Feedback for Near-DRV Standby V_{DD} Scaling in a 90nm SRAM," CICC, pp. 29-33, 2007.

[17] N. Verma and A. Chandrakasan, "A 65nm 8T Sub-Vt SRAM Employing Sense-Amplifier Redundancy," ISSCC, pp. 328–329, 2007.

[18] T.-H. Kim, J. Liu, J. Keane, and C. Kim, "A high-density subthreshold sram with data-independent bitline leakage and virtual ground replica scheme," ISSCC, pp. 330–606, 2007.

[19] S. P. I. Chang, J. Kim and K. Roy, "A 32kb 10t subthreshold sram array with bit-interleaving and differential read scheme in 90nm cmos," ISSCC, 2008.

[20] B. Zhai, D. Blaauw, D. Sylvester, and S. Hanson, "A sub-200mv 6t sram in 0.13um cmos," ISSCC, pp. 332–606, 2007.