Sub-Threshold Design: The Challenges of Minimizing Circuit Energy

B. H. Calhoun[1], A. Wang[2], N. Verma[3], and A. Chandrakasan[3]
[1]University of Virginia, [2]Texas Instruments, [3]Massachussetts Institute of Technology
bcalhoun@virginia.edu, aliwang@ti.com, {nverma,anantha}@mtl.mit.edu

ABSTRACT
In this paper, we identify the key challenges that oppose sub-threshold circuit design and describe fabricated chips that verify techniques for overcoming the challenges.

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

General Terms
Performance, Design, Reliability

Keywords
Sub-threshold digital circuits, low voltage memory, dynamic voltage scaling, process variations, sub-threshold logic

1. INTRODUCTION
Sub-threshold operation for digital circuits first was shown as the means to minimizing CMOS V_{DD} in 1972 [1]. Analog sub-threshold circuits subsequently received a lot of attention for low power applications (e.g. [2][3]). Interest in digital sub-threshold was revived in the late 1990s [4], and a multiplier was demonstrated operating in sub-V_T at 0.475V that used body bias to balance p/n currents [5]. A sub-V_T ring oscillator also employed body biasing and functioned at 80mV [6].

The primary motivation for using sub-V_T circuits is to reduce energy. Analysis of energy contours in [7] demonstrated that minimum energy operation occurs in the sub-threshold region. Once $V_{DD}<V_T$, delay increases exponentially with additional voltage scaling. Leakage current integrates over the longer delay until leakage energy per operation exceeds the active energy and causes the minimum point. Models capture this effect and illustrate the impact of various parameters in [8][9].

The potential for minimizing energy at the cost of speed degradation defines the set of applications for which sub-threshold circuits are well-suited. First, energy-constrained applications such as wireless sensor nodes, RFID tags, or implants are dominated by the need to minimize energy consumption. Speed is a secondary consideration for this class of applications, so sub-V_T circuits offer a good solution. Secondly, many burst-mode applications require high performance for brief time periods between extended sections of low performance operation. Sub-threshold circuits can minimize energy for computations executed during the low performance slots. Finally, the parallelism inherent in many signal processing and communications circuits can be exploited to scale voltages into sub-V_T, providing a low energy solution for throughput-centric applications (e.g. [10]).

This paper describes the key challenges that confront sub-threshold circuit designers and presents chips that overcome the challenges.

ISLPED'06, October 4–6, 2006, Tegernsee, Germany.

2. Sub-Threshold Logic: FFT Processor
Static CMOS gates continue to function in sub-V_T, but some challenges make logic design more difficult. First, CMOS processes are designed with strong-inversion operation in mind, so the ratio of drive current in sub-V_T is frequently imbalanced relative to the case where pMOS and nMOS are symmetrical. The shaded region in Figure 1 shows the operational range for a ring oscillator in 0.18µm CMOS at the worst-case corners. V_{DD} is minimized when the p/n sizing ratio is 12, which indicates that the process is imbalanced such that p/n current is 1/12 relative to the symmetric case. This unfriendly sort of technology imbalance can aggravate process variations and even require different circuit designs for different imbalance scenarios. In addition, the low V_{DD} results in a reduced I_{on}/I_{off} ratio that can reduce robustness, especially for circuits with parallel leakage paths [11].

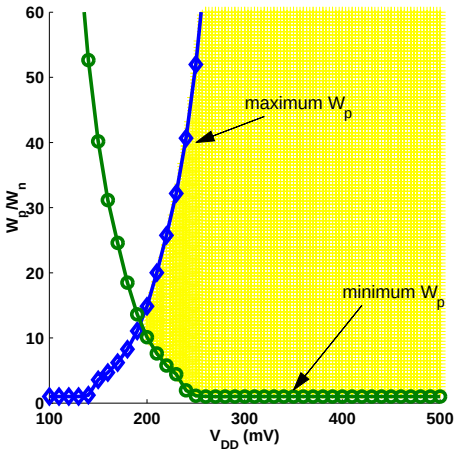

Figure 1: Minimum achievable voltage for 10%-90% output swing for 0.18µm ring oscillator at worst case process corners (simulation).

A 0.18µm CMOS FFT processor uses circuits that account for these challenges: static CMOS logic is used for robustness, gates with parallel leakage paths are redesigned, large stacks are avoided to improve I_{on}/I_{off}, and a register-file memory uses logic-based structures. The chip is fully functional for 128, 256, 512, and 1024 FFT lengths (8-bit and 16-bit precision) at V_{DD} from 180mV to 900mV [11]. Figure 2 shows the measured energy consumption for 8-bit and 16-bit processing as a function of voltage. 8-bit processing has a lower activity factor and thus has lower switching energy. However, because the leakage energy is the same for both 8-bit and 16-bit processing, the minimum

energy point increases to 400mV from 350mV. At the 16-bit optimum, the chip runs at 10 kHz and consumes 155nJ/FFT, which is 350X more energy efficient than a typical low-power microprocessor and 8X more energy efficient than a standard ASIC implementation [11].

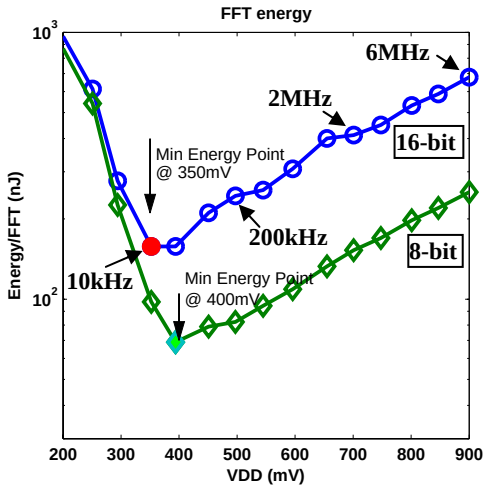

Figure 2: Measured energy per 8- and 16-bit FFT vs. V_{DD}.

3. Scaling Performance: Ultra-DVS

Burst mode applications cannot exclusively utilize sub-threshold operation because they require periodic high speed functionality. Traditional dynamic voltage scaling (DVS) could be extended to include sub-threshold operation, but the overhead of providing the necessary voltages can be large. Adjustable DC-DC converters tend to have limited efficiency over broad voltage ranges, and they take 100s of micro-seconds to switch. An alternative implementation method called local voltage dithering (LVD) offers a reduced overhead means for implementing ultra-DVS (UDVS) down to the sub-threshold region. LVD uses power switches to select from among two or more V_{DD} supplies at the local block level [12]. Figure 3 shows an example system that has 3 V_{DD}s. As the required rate (normalized frequency) for processing incoming data changes, each block spends a different fraction of its operating time at different voltage levels. The averaging effect of this dithering produces an energy consumption profile that nears the optimal (e.g. infinite voltage levels) profile.

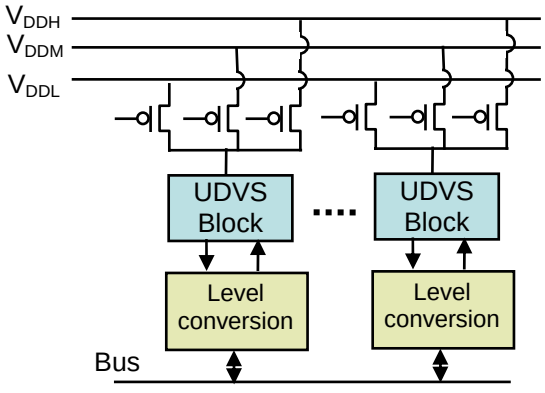

Figure 3: Example UDVS system using LVD and three V_{DD}s.

A 90nm CMOS test chip uses LVD to implement UDVS for 32-bit Kogge-Stone adders [12]. Measurements from the chip show that high rate (e.g. >0.1) dithering can occur in 1 cycle due to the local granularity of the headers. Figure 4 shows an example energy profile for a UDVS system using energy measurements from the test chip. For high rates, the blocks dither between the top two supplies (1.1V and 0.8V in the figure) to achieve near-optimal energy consumption. When performance requirements relax for low rate operation, the blocks can hop to the V_{DD} that gives minimum energy operation (330mV for the 90nm adder block) to achieve 9X savings in energy consumption.

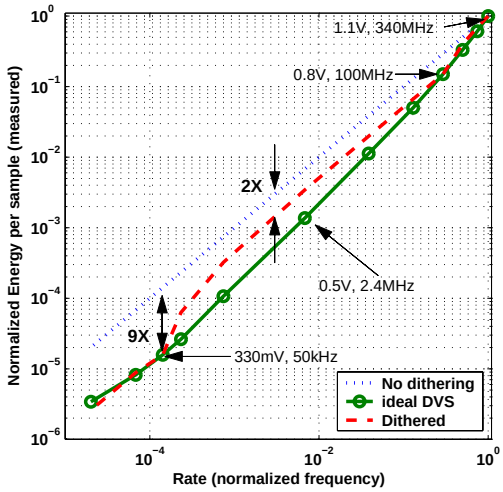

Figure 4: Energy profile based on 90nm chip measurements for example 3-V_{DD} system.

4. Sub-Threshold SRAM

SRAM is an important component of many ICs, and it can contribute a large fraction of the active and leakage power consumption. It is important to have sub-V_T compatible SRAMs for sub-V_T systems. However, the nature of SRAM circuits makes them a melting pot of all of the major sub-V_T challenges.

Random variation fundamentally affects the geometry and threshold voltage of CMOS devices and is increasingly prominent in scaled technologies. The large array nature of SRAM implies that extreme tails of the distributions limit yield. The problem is exacerbated in sub-V_T, where device strength depends exponentially on threshold voltage, and, in the presence of variation, relative strengths cannot be guaranteed by sizing. As a result, the widely used 6T SRAM cell, which relies on ratioed operation and is used to maintain density, fails to operate in sub-V_T. Figure 5a,b show the read/hold and write static nose margins [13] respectively for a typical 6T cell and for the 3σ case. At reduced voltages, read margin is negative and write margin is positive, indicating failure for both operations.

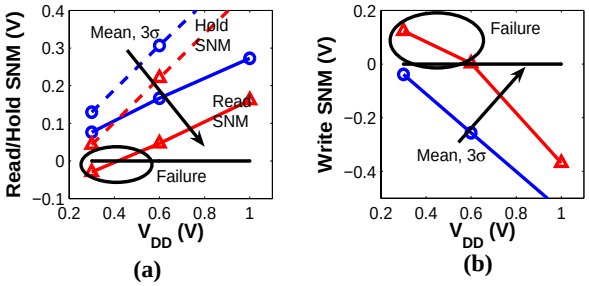

Figure 5: Simulated SNM for (a) read/hold and (b) write.

The increased impact of variation on device strength in sub-V_T also has a limiting effect on SRAM performance and integration. SRAM cell read current, I_{RD}, decreases exponentially in sub-V_T, but the speed is ultimately set by the weakest cell in the array. Figure 6a plots I_{RD} for cells on the weak side of the distribution normalized to the mean (i.e. $I_{RD}/\mu(I_{RD})$). The limiting effect of cell strength variation is amplified in sub-V_T where cells can be over an order of magnitude weaker than the mean.

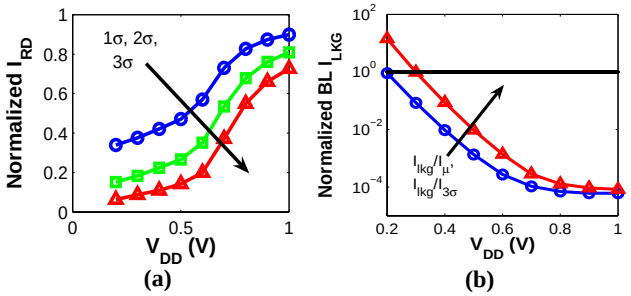

Figure 6: Effect of cell variation on (a) worst case read current and (b) bit-line leakage.

Parallel leakage also limits voltage scaling for SRAM. In conventional 6T SRAM, a stored "1" is read dynamically from a precharged bit-line. However, the reduced I_{on}/I_{off} ratio in sub-V_T is lowered even more due to the unaccessed cells sharing the bit-line, which results in a degraded logic level. Sub-V_T bit-line leakage is less problematic at high voltages where the discharge time of an accessed cell is much faster than that of the aggregate unaccessed cells. However, where variation extends the required discharge time, bit-line leakage severely limits the number of cells that can be integrated onto a column. Figure 6b shows the leakage current of 127 unaccessed cells normalized to the drive current of a single accessed cell weakened by variation. Values greater than unity, which occur in sub-V_T, imply that drive current is indistinguishable from leakage, making reliable read accesses impossible.

Numerous techniques have been reported to mitigate the low-voltage SRAM problems described above. For instance, reduced bit-line precharge voltages and negative word-line bias for unaccessed cells have been used to increase the read SNM. Similarly, increased word-line bias and negative bit-line voltages have been used to improve the write SNM. While these approaches can improve the situation for sub-V_T SRAM, approaches that address the problems more fundamentally provide a better solution for robust operation in sub-threshold.

A 65nm test chip implements a 256kb memory that overcomes the problems and provides functionality in the sub-threshold region to below 400mV [14]. The SRAM uses a 10T bit-cell, shown in Figure 7. M7-M10 form a read buffer that isolates the internal storage nodes, Q and QB, so that a read upset is not possible. This eliminates the read SNM problem of Figure 5a, and stability is instead limited by the hold SNM. Measurements from the test chip show that the cell can hold data correctly below 250mV. Write operations in Figure 5b fail since the access devices in a 6T bit-cell are too weak to over-power the internal cell feedback, which is made worse by process imbalance that makes pMOS sub-threshold current higher than nMOS by an order of magnitude. Robust write in the new 10T cell is performed by weakening the feedback structure by floating VV_{DD}. Finally, bit-line leakage on RBL is minimized by unconditionally raising the voltage of QBB for unaccessed cells. This relies on either the active pull-up current through M9, or the ratio of its leakage current to that of M10's. In either case, M8's V_{GS} becomes negative, resulting in vanishingly small sub-threshold leakage current to the bit-line. This structure allows 256 bit-cells to be integrated per column.

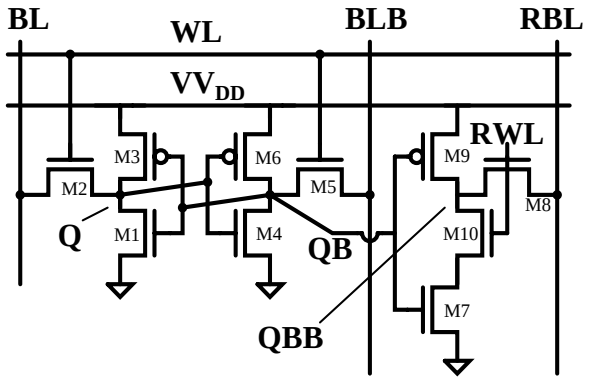

Figure 7: Schematic of 10T sub-threshold bit-cell [14].

5. Conclusions

Numerous problems increase the challenge of designing robust sub-threshold circuits. Some time-testing design practices, such as ratioed write in SRAM, become unreliable due to the exponential dependence of sub-threshold drive current on parameters with large process variations. We have presented an overview of the types of circuits and architectures that overcome these problems and produce working designs. Functional implementations of a sub-threshold FFT processor [11], an energy-scalable UDVS test chip [12], and a sub-threshold SRAM [14] attest that robust sub-threshold systems can practically offer minimum energy operation.

6. ACKNOWLEDGEMENTS

We acknowledge DARPA and Texas Instruments for funding.

7. REFERENCES

[1] Swanson and Meindl, *JSSC*, 1972.
[2] Vittoz and Fellrath, *JSSC*, 1977.
[3] Mead, Addison-Wesley, 1989.
[4] Soeleman and Roy, *ISLPED*, 1999.
[5] Paul, Soeleman, and Roy, *ESSCIRC*, 2001.
[6] Deen, Kazemeini, and Naseh, *ICCDCS*, 2002.
[7] Wang, Chandrakasan, and Kosonocky, *SVLSI*, 2002.
[8] Zhai, Blaauw, Sylvester, and Flautner, *DAC*, 2004.
[9] Calhoun and Chandrakasan, *ISLPED*, 2004.
[10] Sze, Blazquez, Bhardwaj, and Chandrakasan, *ICASSP*, 2006.
[11] Wang and Chandrakasan, *ISSCC*, 2004.
[12] Calhoun and Chandrakasan, *ISSCC*, 2005.
[13] Seevinck, List, and Lohstroh, *JSSC*, 1987.
[14] Calhoun and Chandrakasan, *ISSCC*, 2006.