Improving Energy Efficiency in Sensing Subsystems via Near-Threshold Computing and Device Aging

James B. Wendt and Miodrag Potkonjak
Computer Science Department
University of California, Los Angeles
{jwendt, miodrag}@cs.ucla.edu

***Abstract*— Near-threshold computing (NTC) has been shown to achieve maximal energy gains while maintaining performance metrics more favorable than subthreshold operation of circuits. Sensor subsystems are perfect candidates for NTC since low energy is often a crucial sensor system design desideratum and speed is rarely required. However, process variation affects threshold voltages (V_{th}), thus forcing higher-than and lower-than desired supply voltage (V_{dd}) and ground voltage (V_{gnd}), respectively, and results in higher switching energy. We propose to use device aging to slow those gates on the lower tail of the V_{th} distribution in order to increase V_{gnd} and reduce overall energy consumption. We present a methodology for sensor subsystem aging and observe a reduction in energy consumption by 50% over traditional settings when applying NTC and a further 15% reduction by incorporating our sensor subsystem aging algorithms. Furthermore, we apply our developed approach for device aging at the slice-level granularity to minimize energy expenditure in FPGA systems. Through application of our techniques to simulations of FPGA designs we achieve a factor of up to $6\times$ over traditional operation.**

I. Introduction

Near-threshold computing (NTC) is a technique applied to integrated circuits (ICs) in which the supply voltage is set to approximately the threshold voltage of the transistors. This has been shown to achieve energy savings comparable to subthreshold operation while maintaining more favorable performance characteristics [1]. Sensor systems are ideal for NTC because speed is rarely of great benefit while low energy is absolutely paramount [2] [3] [4] [5] [6] [7]. However, the main obstacle to effectively apply NTC is process variation. A single gate with a low V_{th} forces operation of the circuit at a low V_{gnd} and a single gate with a high V_{th} forces operation of the circuit at a relatively high V_{dd}. Since the switching energy of the gates in a circuit is a function of $(V_{dd} - V_{gnd})^2$, the tails on the low ends and high ends of the threshold voltage distribution are ultimately responsible for superfluous energy consumption.

In order to reduce this effect, we propose the use of device aging to effectively shift the V_{th} values of the lower gates to higher values [8]. While this slows individual gates, we benefit from an overall energy reduction over the entire circuit by enabling the increase of V_{gnd}. Furthermore, the majority of gates are often not on the critical path, so aging often has no impact on circuit delay. However, if it is crucial that a gate on the critical path is aged (thus increasing circuit delay), we subsequently increase V_{dd} to return the circuit to its original delay, before returning to aging minimum V_{th} gates.

Furthermore, we recognize that application specific integrated circuit (ASIC) solutions are often not the most cost efficient means in which to realize sensor subsystem designs. Hence, we propose new aging and placement techniques for field programmable gate array (FPGA) energy minimization that capitalizes on the FPGA block architecture.

II. Related Work

A. Near-Threshold Computing

Energy consumption in modern CMOS circuits can be minimized quadratically by lowering supply voltage relative to ground. Consequently, voltage scaling techniques have become one of the most effective methods and widely researched topics in IC power reduction [9].

Operating in the near-threshold region ($V_{dd} \sim V_{th}$) has been demonstrated to offer more balanced trade-offs between energy reduction ($10\times$) and performance degradation ($10\times$) [1]. Iteration 0 in Figure 1 depicts a traditional NTC-operated circuit in which the inherent process variations determines the maximum and minimum values for V_{gnd} and V_{dd}, respectively.

NTC has also introduced numerous design challenges. Due to near-threshold operation and the variability of circuit threshold voltages due to process variation, NTC circuits are highly susceptible to this variability. Post-silicon solutions dealing with these variations include employing dual supply voltages, soft-edge clocking, body-biasing, gate sizing, and threshold voltage assignment [10] [11] [12].

B. Device Aging

Negative bias temperature instability and hot-carrier injection phenomena can cause significant alterations to the delay characteristics of individual gates [8] [13] [14] [15]. While device aging is widely considered to be a detrimental effect, we utilize it to our advantage in NTC-operated circuits for overall energy reduction. This is possible because switching energy is a function of the difference between V_{dd} and V_{gnd}, which in turn must be set in accordance with the threshold voltage distribution.

C. Gate Level Characterization

Gate level characterization (GLC) techniques can be classified into four major groups: direct measurement approaches, schemes that employ FPGA reconfiguration, approaches that

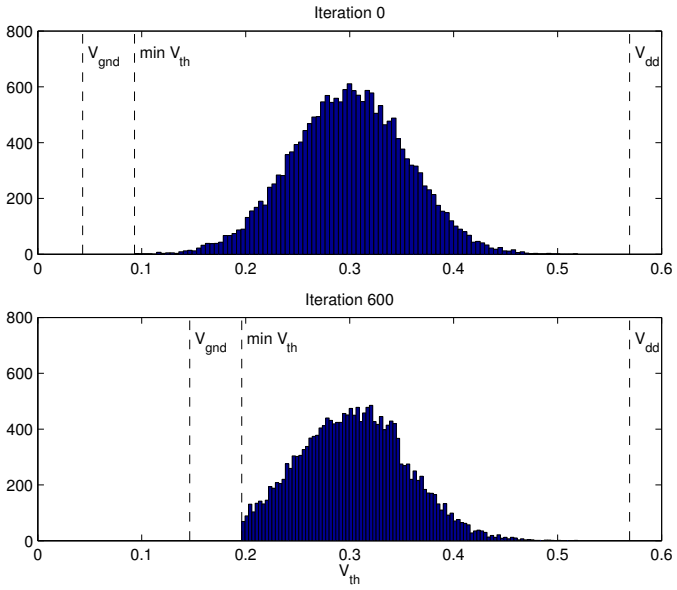

Fig. 1. Gate V_{th} distributions of s38584 [16]. V_{dd} remains at a near-threshold voltage just above the highest V_{th} gate while V_{gnd} is set to a near-threshold voltage just below the lowest V_{th} gate. After aging the lowest 600 gates we can increase V_{gnd} and reduce overall circuit switching energy.

create and observe special IC structures and specialized circuity, and non-destructive techniques that construct global measurements and deduce scaling factors of each gate by solving a system of equations [17] [18] [19].

III. Motivation & Problem Formulation

In this section we present the first approach for energy and power minimization in FGPA-based systems using device aging. We start by explaining the motivation for minimizing energy in FPGA-based sensing systems. We then explain our assumptions and present our problem formulation. Next, we explain the key optimization ideas in the new approach and present a description of our overall approach for energy minimization in FPGA-based systems. Finally, we present a simulation-based study that indicates that energy minimization in FPGAs can be very effective.

There are a number of reasons why studying energy minimization in FPGA-based sensing systems is important. The first is that ASIC- based systems very often have unacceptable, preproduction costs that can reach up to tens of millions of dollars. This means that even before the first ASIC-based sensing system is realized, one has to invest around $30 million dollars just to cover design, testing, and validation.

An alternative to ASIC-based sensing systems is to use more general purpose microprocessors or microcontrollers. However, a serious downfall of these systems is that optimizing general purpose processors for time and energy is often complex and cumbersome.

FPGAs do not impose any non-production costs on the designer and have an unprecedented level of flexibility. Recent developments have put FPGAs at the frontier of integrated systems in terms of both their feature size as well as their potential for integration. For example, FPGA Spartan chips have more than 5.8 million transistors while only Nvidia's graphics processor has a larger number of switching elements. Recently, Xilinx announced that they implemented their newest family of FPGA's at TSMC facilities using 20nm technology. This announcement means that FPGAs now use more advanced and finer technologies than both ASICs and even microprocessors.

Our objective is to develop an approach that minimizes energy spent in FPGA-based sensing systems. We employ two generic mechanisms. The first is to use device-based aging for the alternation of FPGA component speeds. Specifically, we consider slices and blocks as basic structures in FPGAs and simultaneously age all transistors in a particular slice. The key observation here is that the FPGA will operate in NTC mode and therefore have a very low energy budget.

The second degree of freedom is that in almost all situations, the size of the actual design is significantly smaller than the capacity of the full FPGA. Therefore, we can move each program component by the same horizontal and vertical shift, thus moving the boundaries of the program design to a number of locations. In such a way, we can place the program in a position to avoid using slices that have transistors with threshold voltages that are not amenable for energy efficiency. Specifically, our goal is to select slices in such a way that the highest threshold voltage in any of them is as low as possible while the lowest is as high as possible. In this type of system, we reduce the voltage swing and hence reduce energy quadratically. These two mechanisms, device aging and selection of slices that only have favorable threshold voltages, are combined in such a way that achieves provably optimum minimum energy.

The key observation is that all slices can be normalized by setting them such that the highest threshold voltages among all of them are equal. This is accomplished using non-uniform device aging as dictated by process variation.

An additional and important observation is that transistor-level characterization of slices can be accomplished without using any already proposed and effective but often expensive GLC techniques. We employ a new binary search technique that distinguishes between correctly operating slices and incorrectly operating slices due to either high threshold voltages that prevent low supply voltage or low threshold voltages that prevent high ground voltage. Once we characterize each slice, we search for slices that are adjacent to one another in order not to jeopardize the critical path of the design while excluding slices which have gates with too low of a threshold voltage in their structure.

Our process variation model has two components. In the first we use Asenov's model indicating that threshold voltages are proportional to the number of dopants, and therefore, follow an independent Gaussian distribution [8]. The second component corresponds to the effective channel length of transistors and follows Cline's model that includes information about correlations that are typical for modern integrated circuits [20].

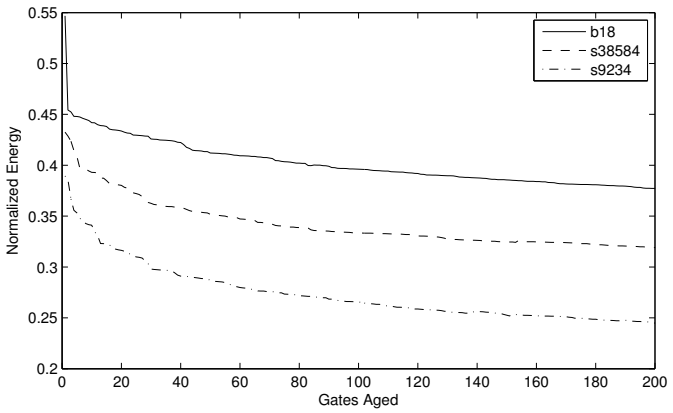

Fig. 2. ASIC energy reduction applied to three benchmark circuits [16] [21] through iterative minimum V_{th} aging.

IV. Energy Reduction

In this section we describe in detail our techniques for energy minimization in sensing subsystems for both ASIC and FPGA systems utilizing device aging and near-threshold computing.

A typical post-fabrication distribution of threshold voltages for the s38584 benchmark circuit is depicted in Iteration 0 of Figure 1. Ultimately, the left and right tails of the distribution are most responsible for forcing the relatively high and low setting of V_{dd} and V_{gnd}, respectively. In order to minimize this voltage swing and thus reduce circuit switching energy we aim to increase the threshold voltages of the gates that comprise the lower tail, thus enabling a higher setting of ground. This is achieved by iteratively aging gates corresponding to the minimum threshold voltage across the circuit to that gate's maximum extent. In the case that a gate on the critical path is selected and subsequently aged, thus increasing the overall circuit delay, we adjust the supply voltage until we return to the original delay before returning to aging.

Figure 2 depicts the energy savings relative to normal superthreshold operation. NTC alone provides 45-60% energy savings while our techniques reduce energy consumption by an additional 15%.

In order to apply our aging-based energy minimization methods to the FPGA we must first characterize the maximum and minimum threshold voltages of each slice. Existing GLC solutions are expensive and overly detailed for our purposes. Thus, we introduce a new and inexpensive binary search technique for FPGA slice characterization. For the purposes of energy minimization of the FPGA using NTC, we need only to know the minimum and maximum threshold voltages within a single slice. We measure these values by varying the supply and ground voltages until the slice is operational.This technique is described in detail in Algorithm 1.

Herein, we denote the maximum and minimum threshold voltages at a particular slice i as $V_{th}^{max}(i)$ and $V_{th}^{min}(i)$. Once all FPGA slices are characterized we perform coarse grained per slice aging by stressing all gates in a particular slice uniformly. This is performed on each slice individually in order to match $V_{th}^{max}(i)$ to the FPGA-wide maximum threshold voltage. This will simultaneously increase the values of V_{th}^{min} across all slices non-uniformly according to each slice's particular V_{th}^{max}. This change is depicted in Figure 3.

Now, in order to minimize energy, we find a subset of physically adjacent tiles (as required by the layout of the program) whose minimum V_{th}^{min} among all covered slices is largest among all possible placements, thus reducing the voltage swing and quadratically reducing energy consumption. This is possible because most programs use a fraction of the total FPGA space. We iteratively disable slices containing the minimum V_{th}^{min} over all slices, until it is no longer possible to position the program on the FPGA board without covering a disabled slice. The last disabled slice that prevented the program from being positioned is then re-enabled. We subsequently iterate over all possible valid program assignments. The final assignment corresponds to the position in which the slices covered by the program at that location has a minimum V_{th}^{min} that is largest among all possible positions, thus yielding the smallest operational voltage swing between V_{dd} and V_{gnd} and providing the lowest energy utilization.

Algorithm 1 FPGA Slice Characterization

Require: $0 < \epsilon < V_{gnd} \leq V_{dd}$
$\delta \leftarrow +1$
$\Delta \leftarrow V_{dd}$
while $\epsilon < \Delta$ **or** $!circuit.IsOperational(V_{dd}, V_{gnd})$ **do**
  $V_{dd} \leftarrow V_{dd} + \delta\Delta$
  if $\delta = +1$ **and** $circuit.IsOperational(V_{dd}, V_{gnd})$ **then**
    $\delta \leftarrow -1$
    $\Delta \leftarrow \Delta/2$
  else if $\delta = -1$ **and** $!circuit.IsOperational(V_{dd}, V_{gnd})$ **then**
    $\delta \leftarrow +1$
    $\Delta \leftarrow \Delta/2$
  end if
end while
$\delta \leftarrow +1$
$\Delta \leftarrow V_{gnd}$
while $\epsilon < \Delta$ **or** $!circuit.IsOperational(V_{dd}, V_{gnd})$ **do**
  $V_{gnd} \leftarrow V_{gnd} + \delta\Delta$
  if $\delta = +1$ **and** $!circuit.IsOperational(V_{dd}, V_{gnd})$ **then**
    $\delta \leftarrow -1$
    $\Delta \leftarrow \Delta/2$
  else if $\delta = -1$ **and** $circuit.IsOperational(V_{dd}, V_{gnd})$ **then**
    $\delta \leftarrow +1$
    $\Delta \leftarrow \Delta/2$
  end if
end while
return V_{dd}, V_{gnd}

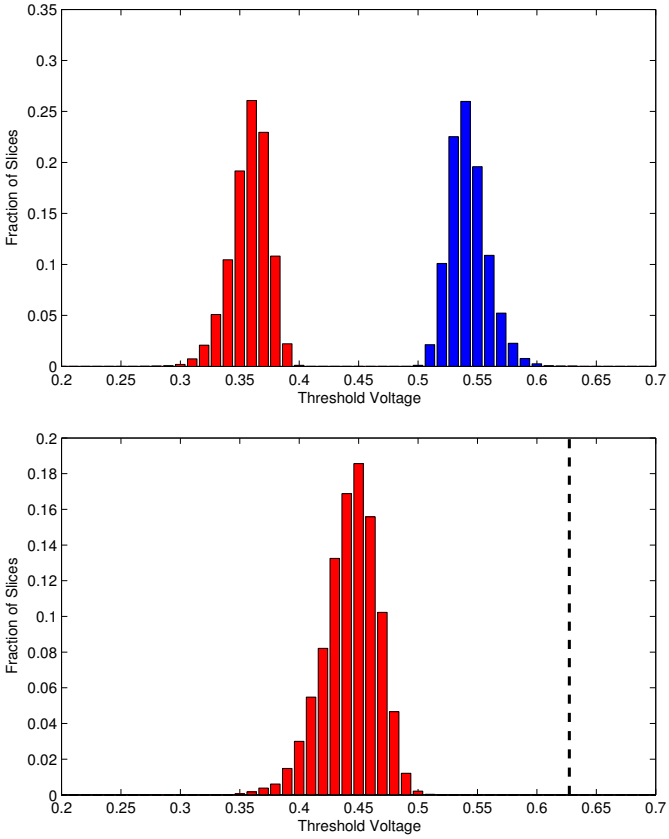

Fig. 3. (a) Distributions of maximum and minimum threshold voltages found in slices of a typical FPGA board. (b) Distribution of minimum threshold voltages across FPGA slices after aging. The dotted line represents the maximum threshold voltage of each slice.

We compare the energy savings of our techniques for aging and placement against an NTC-operated FPGA with and without aging. We present our results in Figure 4. Our technique achieves energy savings of up to 35% against the aged FPGA and 75% to 84% against the traditionally operated non-aged FPGA.

V. Conclusion

We have developed a new approach for the creation of energy efficient sensor systems that employs device aging within the near-threshold computing paradigm. It is applied to both ASIC, where aging is conducted at the gate-level granularity, as well as to FPGA-based subsystems, where aging is applied at the slice-level granularity. We have achieved energy reduction of up to $2.9\times$ over popular ASIC benchmark circuits and up to $6\times$ over traditional operation of FPGAs.

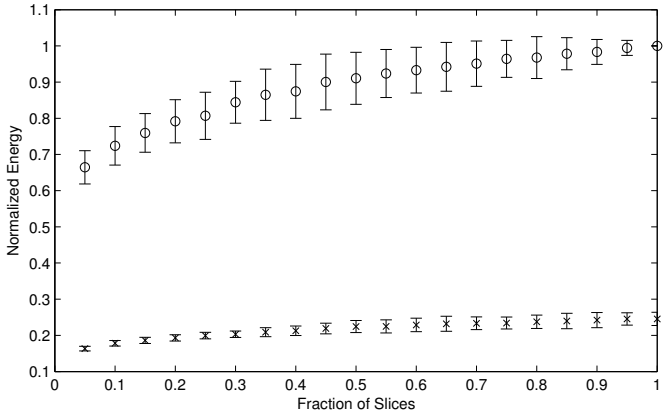

Fig. 4. FPGA energy consumption of varying sized programs as compared to (o) the case in which the FPGA slices are aged and operated at the maximum V_{th}^{max} and the minimum V_{th}^{min} over all slices; and (x) the case in which the FPGA is operated without aging at the maximum V_{th}^{max} and minimum V_{th}^{min} over all slices. Error bars correspond to the standard deviation over all simulated FPGA instances.

References

[1] R. G. Dreslinski, et al., "Near-threshold computing: reclaiming Moore's law through energy efficient integrated circuits," *Proceedings of the IEEE*, vol. 98, no. 2, pp. 253-256, 2010.

[2] J. B. Wendt and M. Potkonjak, "Medical diagnostic-based sensor selection," *IEEE Sensors*, pp. 1507-1510, 2011.

[3] J. B. Wendt and M. Potkonjak, "Nanotechnology-based trusted remote sensing," *IEEE Sensors*, pp. 1213-1216, 2011.

[4] J. B. Wendt, S. Meguerdichian, H. Noshadi, and M. Potkonjak, "Energy and cost reduction in localized multisensory systems through application-driven compression," *DCC*, p. 411, 2012.

[5] J. B. Wendt, S. Meguerdichian, H. Noshadi, and M. Potkonjak, "Semantics-driven sensor configuration for energy reduction in medical sensor networks," *ISLPED*, pp. 303-308, 2012.

[6] J. B. Wendt, V. Goudar, H. Noshadi, M. Potkonjak, "Spatiotemporal assignment of energy harvesters on a self-sustaining medical shoe," *IEEE Sensors*, pp. 1-4, 2012.

[7] T. Xu, J. B. Wendt, and M. Potkonjak, "Digital bimodal function: an ultra-low energy security primitive," to appear in *ISLPED*, 2013.

[8] B. Cheng, A. R. Brown, S. Roy, and A. Asenov, "PBTI/NBTI-related variability in TB-SOI and DG MOSFETs," *IEEE Electron Devices Letters*, vol. 31, no. 5, pp. 408-410, 2010.

[9] P. Pillai and K. G. Shin, "Real-time dynamic voltage scaling for low-power embedded operating systems," *ACM SIGOPS Operating Systems Review*, vol. 35, no. 5, pp. 89-102, 2001.

[10] N. A. Conos, S. Meguerdichian, S. Wei, and M. Potkonjak, "Maximizing yield in near-threshold computing under the presence of process variation," to appear in *PATMOS*, 2013.

[11] M. Seok, et al., "CAS-FEST 2010: Mitigating variability in near-threshold computing," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 1, no. 1, pp. 42-49, 2011.

[12] M. R. Kakoee, et al., "Automatic synthesis of near-threshold circuits with fine-grained performance tunability," *ISLPED*, pp. 401-406, 2010.

[13] S. Chakravarthi, V. Reddy, C. F. Machala, and S. Krishnan, "A comprehensive framework for predictive modeling of negative bias temperature instability," *IRPS*, pp. 273-282, 2004.

[14] S. Wei, J. X. Zheng, and M. Potkonjak, "Low power FPGA design using post-silicon device aging," *FPGA*, p. 277, 2013.

[15] S. Wei, J. X. Zheng, and M. Potkonjak, "Aging-based leakage energy reduction in FPGAs," *Field Programmable Logic*, 2013.

[16] F. Brglez, D. Bryan, K. Kozminski, "Combinational Profiles of Sequential Benchmark Circuits," *IEEE International Symposium on Circuits and Systems*, pp. 1929-1934, 1989.

[17] S. Smith, et al., "Comparison of measurement techniques for linewidth metrology on advanced photomasks," *IEEE Transactions on Semiconductor Manufacturing*, vol. 22, no. 1, pp. 72-79, 2009.

[18] J. S. J. Wong , P. Sedcole , and P. Y. K. Cheung, "Self-measurement of combinatorial circuit delays in FPGAs," *ACM Transactions on Reconfigurable Technology and Systems*, vol. 2 no. 2, pp. 1-22, 2009.

[19] Y. Alkabani, T. Massey, F. Koushanfar, and M. Potkonjak, "Input vector control for post-silicon leakage current minimization in the presence of manufacturing variability," *DAC*, pp. 606-609, 2008.

[20] B. Cline, K. Chopra, D. Blaauw, and Y. Cao, "Analysis and modeling of CD variation for statistical static timing," *ICCAD*, pp. 60-66, 2006.

[21] F. Corno, M. S. Reorda, and G. Squillero, "RT-level ITC'99 benchmarks and first ATPG results," *IEEE Design & Test of Computers*, vol. 17, no. 3, pp. 44-53, 2000.