A Variation-tolerant Sub-threshold Design Approach

Nikhil Jayakumar
nikhil_at_ee_dot_tamu_dot_edu

Sunil P. Khatri
sunil_at_ee_dot_tamu_dot_edu

Department of Electrical Engineering, Texas A&M University, College Station, TX 77843

ABSTRACT

Due to their extreme low power consumption, sub-threshold design approaches are appealing for a widening class of applications which demand low power consumption and can tolerate larger circuit delays. However, sub-threshold circuits are extremely sensitive to variations in supply, temperature and processing factors. In this paper, we present a sub-threshold design methodology which dynamically self-adjusts for inter and intra-die process, supply voltage and temperature (PVT) variations. This adjustment is achieved by performing bulk voltage adjustments in a closed-loop fashion, using a charge pump and a phase-detector.

Categories and Subject Descriptors: B.7.1 [Integrated Circuits]: VLSI

General Terms: Design

Keywords: Sub-threshold, Body-biasing, Self-adjusting, Variation-tolerant

1. INTRODUCTION

In traditional digital VLSI design, the sub-threshold region of operation is not utilized beneficially. Circuit operation is based purely on linear or saturation mode currents, and sub-threshold currents are viewed as an attendant evil, since they contribute towards leakage power consumption when the device is in stand-by. In our approach, we turn this problem into an opportunity. We exclusively utilize sub-threshold leakage currents to implement circuits. This is achieved by actually setting the circuit power supply VDD to a value less than or equal to V_T. This choice results in dramatically smaller conduction currents and power, but also larger circuit delays as well.

This paper is organized as follows: Section 2 discusses some previous work in this area. In Section 3 we describe our method to design process, temperature and voltage insensitive sub-threshold circuits. In Section 4 we present experimental results to validate our idea. Conclusions and future work are discussed in Section 5.

DAC 2005, June 13–17, 2005, Anaheim, California, USA.

2. PREVIOUS WORK

In [1, 2, 3], the authors discuss sub-threshold logic for ultra-low power circuits. They state that their approach would be useful for applications where speed is of secondary importance. In one of two proposed approaches, they describe circuitry to stabilize the current of their circuit across process and temperature variations. In these papers, the idea of using sub-threshold circuits was introduced from a device standpoint, and candidate compensation circuits were proposed. However, no systematic design methodology was provided, which addresses the issue of delay compensation across process, temperature and supply variations.

In [4], the authors report a sub-threshold implementation of a multiplier. The methodology utilizes a leakage monitor, and a circuit which compensates the sub-threshold *current* across process and temperature variations. In contrast, our approach compensates circuit *delay* directly, by phase locking it to a beat clock. In [5], a dynamic substrate biasing technique is described, as a means to make a design insensitive to process variations. The approach is described in a bulk CMOS context in contrast to our sub-threshold approach. Further, the technique of [5] matches the circuit delay to that of the critical paths (which needs to be found up-front). The dynamic biasing is not performed on a per-region basis, making it susceptible to intra-die variations.

3. OUR APPROACH

There is a large and growing class of applications where power consumption is a primary design criterion, and speed is less critical. For such applications, we propose to use sub-threshold circuits as a means to reduce power consumption. Section 3.1 discusses this opportunity, by quantifying the power improvements and speed degradation of the sub-threshold design approach. However, this opportunity has some accompanying problems. The main problem is that the sub-threshold circuits are very sensitive to process, temperature and voltage variations. In Section 3.2, our solution to these issues is presented.

3.1 The Opportunity

We performed SPICE [6] experiments to compare the delay of a circuit implemented using sub-threshold CMOS logic versus traditional CMOS logic. Our goal was to compare the delay and power values of both schemes, for a given Deep Sub-micron (DSM) process technology. The device technologies we used were the Berkeley Predictive Technology Model [7] 0.1μm and 0.07μm processes. For these processes, V_{T_N} and V_{T_P} are respectively 0.261V and -0.303V (for the

0.1μm process) and 0.21V and -0.22V (for the 0.07μm process).

Our comparison of traditional versus sub-threshold circuit delays is shown in Table 1. For each process, we constructed a 21-stage ring oscillator circuit using minimum-sized inverters. From this circuit, we computed the delay, power and power-delay product for both design styles. Simulations were performed for a junction temperature of 120°C. Observe that for both the *bsim*70 and *bsim*100 processes, impressive power reductions are obtained, and the power-delay product (with zero body bias) is about 20× improved, over the traditional design style. The delay penalty can be further reduced by applying a slightly positive body bias. When the body is biased to VDD (which is set at V_T in these simulations), the delay can be brought down by a factor of two, while the power-delay product still remains around 10× better. At this operating point, we still achieve upwards of 100× power reductions.

3.2 Our Solution

We propose a technique that uses self-adjusting body bias, to *phase lock* the circuit delay to a *beat clock*. This phase locking is done for a group of spatially localized Programmable Logic Arrays (PLAs). The circuit consists of a network of interconnected, medium-sized PLAs[1]. Spatially localized PLAs are clustered, and each cluster of PLAs shares a common Nbulk node. This Nbulk node is driven by a bulk bias adjustment circuit (one per PLA cluster), whose task it is to synchronize the delay of a representative PLA in the cluster, to a globally distributed *beat clock (BCLK)*.

3.2.1 PLA design

The PLAs in our design operate in their sub-threshold region of conduction. Figure 1 illustrates the schematic of the PLAs used in our design. All the PLAs in our design are of the precharged NOR NOR type, and have a fixed number of inputs (12), outputs (6) and cubes (12). In the precharge phase (when CLK signal is low), all wordlines and outputs of the PLA are pulled high. The inputs are applied to the PLA in the evaluate phase (when CLK signal is high). For each PLA, we have one maximally loaded wordline (annotated as "dummy wordline" in Figure 1) which we design to *always* switch low in the evaluate phase of the clock, and effectively act as a timing reference for the PLA. Its signal is inverted to produce a delayed clock signal (D_CLK) which is used to gate the GND signal for the OR plane, thereby ensuring that all outputs fire simultaneously after a delay slightly larger than the delay to pull D_CLK high. This delayed clock signal is also connected to PMOS pullups at each output line which serve to precharge (pullup) the output lines during the precharge phase. Each PLA has a single output (the "completion" signal in Figure 1) which *always* switches in each cycle, signaling the completion of the PLA computation. This signal is used to phase lock the PLA delay with the $BCLK$ signal. Also, the PLA has keeper devices on the outputs and wordlines, to ensure that these signals are not inadvertently pulled low.

3.2.2 Self-adjusting Bulk-bias Circuit

Our self-adjusting body bias scheme controls the substrate voltage of a cluster of PLAs in a closed-loop fashion, by ensuring that the delay of a representative PLA in the cluster is phase locked to the $BCLK$ signal. The phase detector and charge pump circuits for our design are shown in Figure 3. The NAND gate in this figure detects the case when the completion signal is too slow, and generates low-going pulses in such a condition. These pulses are used to turn on the PMOS device of Figure 3, and increase the *Nbulk* bias voltage, resulting in a speed-up in the PLA. The waveforms of the signals for this case are shown in Figure 4. Note that in general, $BCLK$ is derived from CLK, having coincident falling edges with CLK but a rising edge which is delayed by a quantity D from the rising edge of CLK. This quantity D is the delay which we want for the evaluation of all PLAs. If the completion has not occurred by the time $BCLK$ rises, a downward pulse is generated on the *pullup* signal, which forces charge into the Nbulk node, resulting in faster generation of *completion*. Note that at this time, *pulldown*, the signal which is used to bleed off charge from *Nbulk*, is low. The NOR gate in Figure 3 generates high-going pulses to turn on the NMOS transistor when the PLA delay is less than D. These pulses drive the NMOS device in Figure 3, bleeding charge out of *Nbulk* and thereby slowing the PLA down.

We plotted the variation of sub-threshold circuit delay[2]

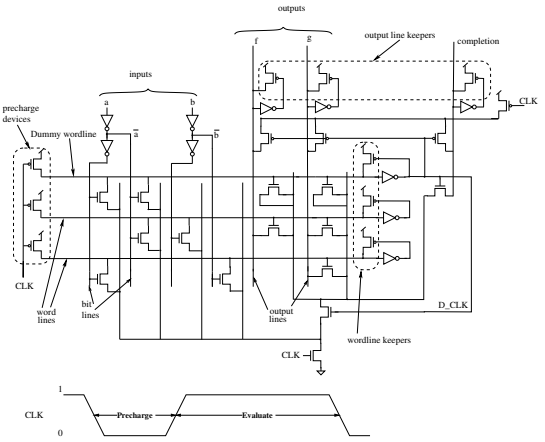

Figure 1: Schematic View of PLA

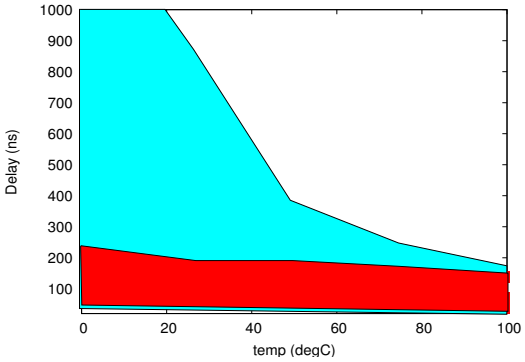

Figure 2: Delay Range with and without Our Dynamic Body Bias Technique

[1]By medium sized PLAs, we mean PLAs that have about 5-15 inputs, 3-8 outputs, and 10-20 rows.

[2]defined as the delay from the start of the evaluation phase of the computation, to the time that the *completion* signal

	Traditional Ckt			Sub-threshold Ckt ($V_b = 0V$)			Sub-threshold Ckt($V_b = VDD$)		
Process	Delay (ps)	Power (W)	P-D-P (J)	Delay ↑	Power ↓	P-D-P ↓	Delay ↑	Power ↓	P-D-P ↓
bsim70	14.157	4.08e-05	5.82e-07	17.01×	308.82×	18.50×	9.93 ×	141.10×	14.43×
bsim100	17.118	6.39e-05	1.08e-06	24.60×	497.54×	20.08 ×	12.00 ×	100.96×	8.20×

Table 1: Comparison of Traditional and Sub-threshold Circuits

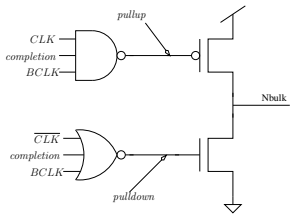

Figure 3: Phase Detector and Charge Pump Circuit

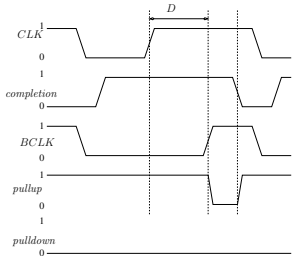

Figure 4: Phase Detector Waveforms when PLA Delay lags $BCLK$

(for a precharged NOR-NOR PLA) against temperature, while varying various process, voltage and temperature parameters. The results are shown in Figure 2. The light area represents the envelope of delays with respect to PVT variations *when no compensation was applied*. Note that the PLA delay varied by more than 2 orders of magnitude. Further, in the light area of the plot, for very low temperatures (to the top and left of the Figure 2) the PLA outputs did not switch at all. The parameters that were varied to compute the envelope were l_{eff} (±5% variation), V_T (±5% variation) and VDD (±10% variation). These variation values represent 3σ variation around the mean, and are obtained from [8]. The dark region of Figure 2 represents the PLA delay variation *after* our self-adjusting body bias technique was applied. The same variations were applied as for the light region. Note the significant reduction in the effect of PVT variations on PLA delay. Also, and importantly, *these adjustments are done in a closed-loop manner during circuit operation*.

There are several observations we can make about this approach:

- The PLAs in our approach operate *just fast enough* to stay synchronized with $BCLK$, thereby minimizing circuit power for a given speed of operation.
- We do not perform bulk voltage control for PMOS devices, since there are very few PMOS devices per PLA, and they are mostly utilized for precharging purposes. It is crucial to perform bulk voltage control for NMOS devices since they are used to perform the computation during the evaluate phase of the clock.
- Sequential designs are implemented using $BCLK$ as the system clock (as well as the clock used to synchronize the delays of the combinational part of the design). Additional margin is included in T_{BCLK}, to account for setup delays of the memory elements and and lock margin. Margin for hold times of the memory elements need not be considered since these elements are latched at the falling edge of $BCLK$.
- The distribution of a sub-threshold VDD signal could be challenging, but this challenge can be addressed by using a high quality power distribution methodology such as a layout fabric [9, 10].
- We use PLAs as the circuit implementation structure because we can design them such that the delay of all outputs is constant, regardless of the input vector applied. Also, design methodologies using a network of medium sized PLAs was shown [10] to be a viable way to perform digital design, resulting in improved area and delay for a design. Finally, a design implemented using such a network of PLAs can be easily mapped into a structured ASIC setting [11].

has switched

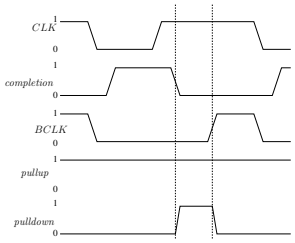

Figure 5: Phase Detector Waveforms when PLA Delay leads $BCLK$

4. EXPERIMENTAL RESULTS

We implemented our technique using PLAs as described in Section 3.2.1. Each cluster consisted of 1000 spatially localized PLAs. PLAs were designed with 12 inputs, 12 rows and 6 outputs. The layout of each PLA occupied slightly over $25\mu \times 15\mu$, so each cluster was of size 0.8mm × 0.5mm. We simulated these PLAs using the the 65nm BSIM4 model cards from [7].

Table 2 reports the PLA delay as a function of several varying parameters. The delay is expressed as a function of l_{eff} and V_T, with varying VDD and V_{Nbulk}. The notation 'S' indicates a slow corner, 'F' indicates a fast corner, and 'T' represents a typical corner. This table represents the PLA delay range that we would encounter if we did not use our technique to perform active compensation. Note that a 'n/a' entry in Table 2 indicates that for the particular set of

Corner	VDD	V_{Nbulk}	0°C	27°C	50°C	75°C	100°C
SS	0.18	0	n/a	685.24	376.84	251.59	169.46
		max	219.34	167.79	126.52	105.11	86.47
	0.20	0	n/a	866.15	376.12	217.01	156.98
		max	138.25	108.54	91.39	77.71	67.94
	0.22	0	n/a	n/a	360.33	204.91	148.71
		max	92.92	78.64	66.41	59.06	51.45
TT	0.18	0	254.45	168.68	139.63	105.60	82.73
		max	113.69	91.07	76.38	63.76	54.50
	0.20	0	189.59	126.91	100.19	82.22	69.11
		max	78.67	64.48	55.88	47.69	42.12
	0.22	0	135.12	102.17	82.68	63.66	59.77
		max	54.55	45.55	40.52	36.45	37.99
FF	0.18	0	88.45	67.41	61.34	46.91	40.20
		max	60.16	46.56	40.51	34.06	30.68
	0.20	0	65.41	52.19	43.11	37.60	33.48
		max	41.33	33.54	29.76	24.91	23.50
	0.22	0	47.53	40.03	34.03	30.45	25.70
		max	28.68	23.58	22.71	22.33	20.56

Table 2: Selecting the Value of D

parameters, the PLA did not switch at all. The magnitude of variations for l_{eff} and V_T are as described earlier in this paper, and are obtained from [8]. Note that for any process and VDD entry at any temperature, the highest speed possible is when V_{Nbulk} is maximum (i.e. set to the value of VDD for that simulation). Also, note that the ratio of the fastest to the slowest delay in this table is as high as 42:1.

Using Table 2, we can find the value of D (the amount by which we delay the rising edge of CLK to obtain $BCLK$, as illustrated in Figure 4). We find the largest delay in the table for all rows with maximum V_{Nbulk}, and add a guardband value to this (to account for lock margin and setup margin for the memory elements). This quantity is the value of D used. When we utilize our approach using self-adaptive body bias, the process variations described above are reduced to the dark region in Figure 2. In other words, our approach is able to **work for all the conditions in Table 2, with a delay contained in the darkened region in Figure 2**. The PLA delays for our approach are very tightly bounded across all these operating conditions.

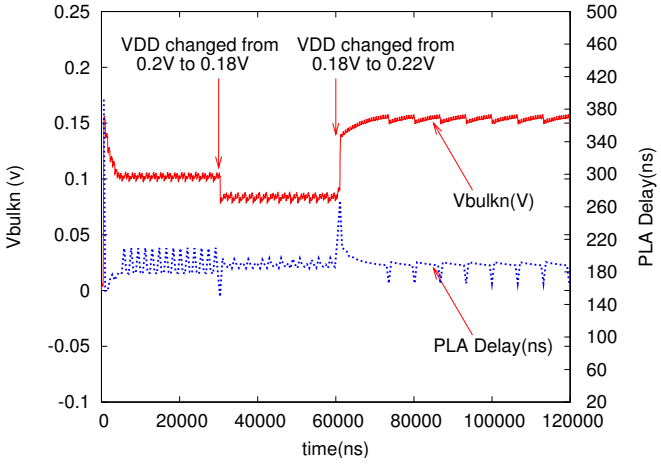

Figure 6: Dynamic Adjustment of PLA Delay and V_{Nbulk} with VDD Variation

Figure 6 describes a SPICE [6] plot of the variation of bulk voltage and PLA delay in our self-adjusting bulk bias scheme. The (higher) solid line represents the value of V_{Nbulk}, while the (lower) dotted line represents the PLA delay. Note that in this figure, the VDD value was initially 0.2V. At time 30,000ns, VDD was changed to 0.22V. Note that in response to this change, our body bias adjustment circuitry modified V_{Nbulk} to a lower value in order to slow the PLAs down. At time 60,000ns, the VDD value was changed to 0.18V, and consequently, our bias adjustment circuit modified V_{Nbulk} to a higher value to speed up the PLAs and keep them phase locked with $BCLK$. Note that in spite of all the changes in VDD, the delay of the PLA stays tightly bounded. This simulation was done for a slow corner, at 27°C.

5. CONCLUSION

We present a practical sub-threshold design methodology, which actively compensates for variations in supply, temperature and process. The power of our approach is its ability to *adapt to inter and intra-die PVT variations*, enabling a significant yield improvement. The design has a global *beat clock* to which the delay of a spatially localized cluster of PLAs is "phase locked". The synchronization is performed in a closed-loop fashion, using a phase detector and a charge pump which drives the Nbulk nodes of the PLAs in the cluster. Our results demonstrate that our technique is able to dynamically phase lock the PLA delays to the beat clock across a wide range of PVT variations, enabling significant yield improvements.

6. REFERENCES

[1] H. Soeleman, K. Roy, and B. Paul, "Robust subthreshold logic for ultra-low power operation," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 9, pp. 90–99, Feb 2001.

[2] H. Soeleman and K. Roy, "Digital CMOS logic operation in the sub-threshold region," in *Tenth Great Lakes Symposium on VLSI*, pp. 107–112, Mar 2000.

[3] H. Soeleman and K. Roy, "Ultra-low power digital subthreshold logic circuits," in *International Symposium on Low Power Electronic Design*, pp. 94–96, 1999.

[4] B. Paul, H. Soeleman, and K. Roy, "An 8x8 sub-threshold digital CMOS carry save array multiplier," in *European Solid State Circuits Conference*, Sept 2001.

[5] J. Tschanz, J. Kao, S. Narendra, R. Nair, D. Antoniadis, A. Chandrakasan, and V. De, "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage," in *IEEE Journal of Solid-State Circuits*, vol. 37, pp. 1396–1402, Nov 2002.

[6] L. Nagel, "Spice: A computer program to simulate computer circuits," in *University of California, Berkeley UCB/ERL Memo M520*, May 1995.

[7] Y. Cao, T. Sato, D. Sylvester, M. Orshansky, and C. Hu, "New paradigm of predictive MOSFET and interconnect modeling for early circuit design," in *Proc. of IEEE Custom Integrated Circuit Conference*, pp. 201–204, Jun 2000. http://www-device.eecs.berkeley.edu/ ptm.

[8] P. Zarkesh-Ha, T. Mule, and J. D. Meindl, "Characterization and modelling of clock skew with process variation," in *IEEE 1999 Custom Integrated Circuits Conference*, 1999.

[9] S. Khatri, A. Mehrotra, R. Brayton, A. Sangiovanni-Vincentelli, and R. Otten, "A novel VLSI layout fabric for deep sub-micron applications," in *Proceedings of the Design Automation Conference*, (New Orleans), June 1999.

[10] S. Khatri, R. Brayton, and A. Sangiovanni-Vincentelli, "Cross-talk immune VLSI design using a network of PLAs embedded in a regular layout fabric," in *IEEE/ACM International Conference on Computer Aided Design*, pp. 412–418, Nov 2000.

[11] N. Jayakumar and S. Khatri, "A METAL and VIA maskset programmable VLSI design methodology using PLAs," in *IEEE/ACM International Conference on Computer Aided Design*, pp. 590–594, Nov 2004.