Power Minimization using Simultaneous Gate Sizing, Dual-Vdd and Dual-Vth Assignment*

Ashish Srivastava Dennis Sylvester David Blaauw
University of Michigan, EECS Department, Ann Arbor, MI 48109
{ansrivas, dennis, blaauw}@eecs.umich.edu

Abstract

We develop an approach to minimize total power in a dual-Vdd and dual-Vth design. The algorithm runs in two distinct phases. The first phase relies on upsizing to create slack and maximize low Vdd assignments in a backward topological manner. The second phase proceeds in a forward topological fashion and both sizes and re-assigns gates to high Vdd to enable significant static power savings through high Vth assignment. The proposed algorithm is implemented and tested on a set of combinational benchmark circuits. A comparison with traditional CVS and dual-Vth/sizing algorithms demonstrate the advantage of the algorithm over a range of activity factors, including an average power reduction of 30% (50%) at high (nominal) primary input activities.

Categories and Subject Descriptors: B.6.3 Design Aids

General Terms: Algorithms, performance

Keywords: Power dissipation, optimization, multiple voltages

I. Introduction

The well-known power management gap defined in the International Technology Roadmap for Semiconductors states that an 800X reduction in standby mode power and a 20X reduction in dynamic power are required compared to continued extrapolation of recent power consumption trends [1]. The best known method to attack this gap is the use of multiple supply and threshold voltages on a chip. Previous works [2][3] have shown that using two supply and threshold voltages provides substantial improvement in power dissipation and the use of additional voltages results in small power improvements which is hard to justify the additional costs associated with multiple supply and threshold voltages. Dual-Vdd designs have shown significant improvements in power dissipation in the range of 40-50% [3,4], but these improvements have been expected to decrease with process scaling [5]. Reference [2] shows that an additional threshold voltage can be used to maintain the power reduction with scaling dimensions.

Multiple Vdd designs impose the constraint that gates operating at a lower supply voltage cannot fan-out to gates operating at a higher supply voltage without level concerting the low Vdd signal to a high Vdd signal. Two approaches that obey this constraint have been proposed in the literature. Clustered Voltage Scaling (CVS) [6] allows only one transition from high Vdd to low Vdd gates along a path, and level converts low Vdd signals to high Vdd at the flip-flops and does not require any stand alone level conversion circuit within the combinatorial network. Extended CVS allows for level conversion on paths in between flip-flops and thus can improve the achievable power reduction since the low Vdd assignment problem becomes less constrained. In [7] the authors address the problem of power optimization using simultaneous Vdd and Vth assignment. The approach for dynamic power dominated systems fails to consider that assigning a gate to high Vth negatively impacts the extent to which other gates in the circuit can be assigned to low Vdd and thus fails to consider the optimization of total power. The approach for leakage dominated systems assigns gates to high Vth in the order of their level from the outputs, this unnecessarily limits the achievable power savings. Reference [8] uses a Lagrangian multiplier based optimization followed by heuristic clustering for dual-Vdd and dual-Vth assignment. The approach is used to perform module level power optimization using path enumeration and it cannot be extended to perform gate level power optimization due to its computational complexity. Reference [9] proposed a new method for slack redistribution to solve the leakage power optimization problem with dual-Vth and sizing by iteratively formulating and solving a linear program. However, the extension to dual-Vdd assignment is formulated as an integer linear program, which results in unreasonable runtimes.

Thus we see that all previous approaches fail to solve the assignment problem of all the three design variables (drive strength, supply and threshold voltage) in a computationally efficient to manner to provide minimization of total power. In this work we describe an approach to simultaneously perform gate-level sizing, Vdd, and Vth assignment in a dual-Vdd/Vth environment to minimize total power consumption (defined as the sum of static and dynamic power). Since our algorithm enables simultaneous optimization of total power using Vdd and Vth allocation and sizing we refer to the complete algorithm as VVS.

II. Algorithm Description

We employ a two stage sensitivity based approach to minimize total power while assigning the drive strength and the supply and threshold voltage to each of the gates in the circuit. Throughout the flow of the VVS algorithm a front is maintained which is located at the interface between the low and high Vdd gates. Similar to CVS we do not allow level conversion within the logic itself and hence we must strictly observe the topological constraint imposed in dual-Vdd designs.

DAC 2004, June 7–11, 2004, San Diego, California, USA

*This work was supported in part by funding from NSF, SRC, GSRC, IBM and Intel.

In the first stage of the VVS algorithm, called the backward pass, Vdd assignment and sizing are combined to minimize total power while we move the front from the primary outputs towards the primary inputs. The second stage, or the forward pass, uses the optimal location of the front found in the first stage as the starting point for the optimization and then relies on both Vdd and Vth assignment along with sizing to further reduce total power while the front is moved back towards the primary outputs. The timing constraints on the design remain fixed throughout the flow of the algorithm.

i) Backward Pass

We define the *backward front* to consist of all gates operating at high Vdd that do not fanout to any gate operating at high Vdd. Thus, assigning any gate on the backward front to low Vdd will not violate the topological constraint in dual-Vdd designs. A simple CVS procedure is first used to assign gates on the front to low Vdd as long as the circuit meets timing. In the CVS procedure, an initialization procedure is used to create a list of primary outputs of the design that represents the backward front of the design. A predictive metric is then used to order gates in this list. This metric could be based on simple parameters such as the fanout capacitance or the slack of the gate for example. The gate with the maximum value for the predictive metric is selected as the candidate gate, which is then assigned to low Vdd if the timing constraints are not violated. Gates are identified that can be added to the backward front as a result of the assignment and added to the backward front.

At the end of CVS, none of the gates on the backward front can be assigned to low Vdd without violating the timing constraints. Figure 1 shows the scenario at this stage of the algorithm. Gates 1-3 have been set to low Vdd by CVS and gates 4, 5, and 8 now form the backward front. Gate sizing is then employed to compensate for the delay increase arising from the assignment of a gate to low Vdd. After a candidate gate on the backward front is assigned to low Vdd, a sensitivity measure to upsizing for all of the gates in the circuit is calculated which is used to identify gates to be up-sized. Let ΔD represent the change in delay and ΔP the change in power dissipation due to upsizing. The sensitivity of each gate in the circuit to up-sizing is computed as

$$Sensitivity = \frac{1}{\Delta P}\sum_{arcs}\frac{\Delta D}{Slack_{arc} - S_{\min} + K} \quad (1)$$

where Smin is the worst slack seen in the circuit and K is a small positive quantity for numerical stability purposes. The form of the sensitivity measure gives a higher value to gates lying on the critical paths of the circuit. The arcs represent the falling and rising arcs associated with each of the inputs of the gate. The gate with the maximum sensitivity is then selected and sized up. This process is repeated until all slacks in the circuit become positive. It is important to note that the sensitivity calculation does not require a full circuit timing analysis, which would otherwise make the runtime prohibitively large. The sensitivity measure is similar to that employed in [10] to perform sizing in dynamic circuits.

The number of up-sizing moves allowed to meet timing is fixed to a constant large number to avoid pursuing bad solutions that could also possibly result in overly large area increases. However, we do allow moves that result in a net increase of total power in an attempt to allow the flow of the algorithm to escape local minima. Due to the topological constraints imposed on low Vdd assignment, if a gate is not assigned to low Vdd none of the gates in its input cone can be assigned to low Vdd. Hence a steepest decent approach with no means to get out local minima will likely become stuck in a local minima that is far from the global minimum. The ordering of the gates on the backward front using the predictive metric associated with each gate is heuristic and is used to steer the flow of the algorithm in the right direction. At all points during the first stage the best-seen solution is saved and this solution is restored at the end of the first stage. The end of the first stage is signaled when the list containing the gates on the backward front becomes empty or else none of the gates in the list can be assigned to low Vdd without violating timing (even with the maximum allowed amount of upsizing).

ii) Forward Pass

At the end of the first stage the front between high and low Vdd gates is in the best position in terms of the total power dissipation for a dual Vdd, single Vth environment. The second stage, or forward pass, is then used to move the front forward towards the primary outputs in conjunction with high Vth allocation and possible gate upsizing to minimize the total power in a dual-Vth scenario.

We now define the *forward front*, which consists of all gates that are operating at low Vdd and have all of their fanins operating at high Vdd. In Figure 1, assuming that upsizing in the backward pass allows us to further assign gates 4, 5 and 8 to low Vdd, these same three gates would now form the forward front.

Importantly, assigning a gate on the forward front to operate at high Vdd will not lead to a violation of the topological constraint. We now calculate 1) a sensitivity measure for gates on the forward front with respect to high Vdd operation, and 2) a sensitivity measure for all gates in the circuit with respect to upsizing. Both these sensitivities are calculated as the ratio of the change in delay to the change in power dissipation as a result of the corresponding operation. The gate with the maximum sensitivity is then either assigned to high Vdd or up-sized based on the operation to which the maximum

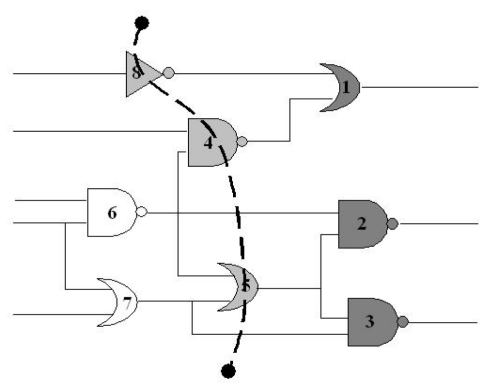

Figure 1 Backward front for an example circuit at the end of CVS.

sensitivity corresponds. Once a gate is up-sized or reset to high Vdd operation, timing slack has been created in the circuit. To exploit this slack and reduce total power, the next step begins by computing the sensitivity of all gates in the circuit with respect to operation at high Vth. This sensitivity is calculated as

$$Sensitivity = \Delta P\sum_{arcs}\frac{Slack_{arc}}{\Delta D} \quad (2)$$

Based on this sensitivity measure gates are assigned to high Vth as long as the timing constraints of the design are met. This set of moves (assignment to high Vdd or upsizing a gate followed by the associated high Vth assignments) is then accepted if the total power is found to decrease, otherwise the moves are reversed. The best-seen solution is always maintained and restored at the end of the forward pass. The pseudo-code for this stage of the algorithm is shown below:

```
Forward Pass
{
Calculate sensitivity of gates on forward front to high Vdd
    operation
Calculate sensitivity for all gates to up-sizing
Set gate to high Vdd or upsize based on maximum sensitivity
Calculate sensitivity of gates to high Vth operation
Set gates to high Vth while timing is not violated
If total power increase
    reverse moves
}
```

This two-stage VVS algorithm allows us to make intelligent choices to trade-off dynamic power for leakage power in order to obtain a reduction in the total power dissipation. The algorithm is effectively directed to automatically provide either more leakage or dynamic power reduction based on the initial design point. The two-stage algorithm can easily quantify the impact of setting a gate to high Vth on the extent to which other gates in the circuit can be assigned to low Vdd. In other words, we can independently judge the impact of Vth and Vdd assignment on total power, something that is difficult to achieve in a flow that simultaneously assigns low Vdd and high Vth as in [7].

III. Implementation and Circuit Issues

The algorithm described in Section II was implemented in C and tested on ISCAS85 benchmark circuits that vary in size from 169 to 2500 gates [11]. The circuits were synthesized using an industrial 0.13µm library with a nominal Vdd of 1.2V and a nominal Vth of ±0.23V (these are fixed throughout) that represent the high Vdd and high Vth respectively. The libraries consist of inverters with twelve different drive strengths and two and three input NAND and NOR gates with seven different drive strengths. We also created duplicate low Vdd libraries in which gate delays are computed with inputs switching at high Vdd rather than low Vdd. This is called the *overdrive* library since the cells in this library are being overdriven at their inputs (as is the case at the boundary of high and low Vdd cells) and hence are faster in one transition direction and slower in the other. This phenomenon can be used to advantage by employing gate libraries with skewed drive strength, although we did not explore such an optimization in this work. All energies (static, short-circuit, and dynamic) and capacitance variations due to varying thresholds [10] are inherently considered using these SPICE-derived library files. It is interesting to note that gates operating at high Vth not only have a smaller input capacitance as compared to low Vth gates but also have a much smaller "internal" power. Internal power accounts for the power dissipation other than those accounted by leakage and the switching output load capacitance [12]. To capture the effect of wire capacitance we approximated this capacitance as

$$C_{wire} = 5 * (1 + 0.4 * (fanouts_{wire} - 1)) fF \qquad (3)$$

where $fanouts_{wire}$ is the number of gates to which the wire connects. Equation 3 is based on the model used in [13] and provides a wire capacitance of 5fF for a gate with one fanout, corresponding to a wire length of approximately 25µm in our technology.

The synthesized design is first sized using a TILOS-like [14] sensitivity-based sizing algorithm to obtain the power-delay curve for the design. The design is then resized from the initial synthesized point to a delay point that is backed off from the minimum achievable delay by a fixed percentage. This is done since the power-delay curve is very steep at the minimum achievable delay and operating at that point is sub-optimal from a power/delay tradeoff perspective. We emphasize that the initial design is operating at the fastest possible combination of Vdd and Vth (high Vdd and low Vth) meaning that the circuit speed at which we perform our optimizations is quite aggressive even at backoff points in the 20% range.[1] Subsequent phases of the algorithm maintain this timing and no further relaxation in timing is used to obtain power improvement.

A. Level Conversion

Since we are employing a CVS-based approach in this work, level conversion is only required at sequential elements. We incorporate the level converter delay penalties by considering that they result in a fixed delay overhead for the circuit. The results in the work are presented using a level conversion penalty of 80ps for a low Vdd of 0.6V. This delay value is chosen based on [15,16], where the authors show that the D-Q delay overhead for a level converting flip-flop can be under two fanout-of-four (FO4) inverter delays for the target technology. In our target technology at nominal high Vdd and high Vth, the FO4 delay is 40ps. The energy penalties of the level converting flip-flops are not considered in our results. When replacing a flip-flop operating at high Vdd with a level converting flip-flop, the energy is reduced since much of the flip-flop's internal capacitance is now toggling at a lower Vdd. While the energy reduction is not quite quadratic with the ratio of (Vdd high/Vdd low), [15] shows that the traditional master-slave level converting FF from [4] demonstrates 40% lower energy than a comparable all-high Vdd master-slave FF when $(\text{Vdd high}/\text{Vdd low})^2$ is 0.5. Therefore, most of the energy savings are preserved and we can consider level conversion energy penalties to be negligible.

B. Switching Activity

The switching activity at each of the circuit primary inputs can be adjusted to obtain a desired initial static vs. dynamic power ratio. We apply a switching activity and state probability at each input which are then propagated through the entire circuit using the approach outlined in [17]. Later we provide results for different circuit activities to demonstrate the efficacy of Vdd and Vth assignment in varying application spaces and how the VVS algorithm efficiently trades off between static and dynamic power.

[1] Indeed, with typical speed differences of 15% between high and low Vth devices, a 20% backoff in our notation would correspond to nearly the fastest possible implementation of the circuit at the high Vdd, high Vth design point.

IV. Results

All results shown in this section are for a low Vdd and low Vth of 0.6V and 0.12V respectively unless otherwise stated. The high Vdd and Vth are fixed at 1.2V and 0.23V respectively. The circuit delay used to set the timing constraint is set to 20% slower than the fastest possible delay of the circuit (i.e., 20% backoff point). The nominal input activity is adjusted such that leakage power constitutes approximately 8% of the total power dissipation, when the design is synthesized using high threshold voltage gates [18][2]. A 3X higher and 3X lower activity factors are referred to as high and low input activities respectively.

Table 1 shows the results obtained for the ISCAS'85 benchmark circuits for the case of high switching activity and similar results are generated for nominal switching activity. The columns corresponding to the initial power list the actual power numbers. The remaining columns show the percentage reduction in leakage, switching and total power at the end of three distinct phases of the algorithm; 1) CVS only, 2) end of backward pass, and 3) VVS. The results clearly show the advantage offered by each step of the algorithm. At the end of the backward pass CVS coupled with sizing increases the average savings in switching power by approximately 10% from 16.1% to 26.1% for high activities as compared to CVS alone. The leakage power also shows a significant reduction of ~13% for both activity values, which can be attributed to the roughly cubic dependence of leakage power on Vdd [19]. The last phase for the high activity case shows that a small amount of switching power can be traded off to obtain substantial savings in leakage power due to the exponential dependence of leakage current on Vth. If the "internal" power and the change in gate capacitance are neglected a small amount of switching power (~4% on average) is traded-off to obtain the large reduction in leakage power. Interestingly, in reality we find an additional reduction in switching power which can be attributed to the large reduction in "internal" power and gate capacitance due to the assignment of a significant fraction of the gates to high Vth. For the case of nominal activitiy the second pass significantly alters the position of the front of high and low Vdd gates to trade-off dynamic power for leakage power, since larger savings in leakage power for these cases results in a much lower total power dissipation. This important capability leads to a reduction in the total power dissipation of the design and shows that the algorithm is correctly steering towards a proper low-power solution. For the nominal activity case on the average, all of the reduction in switching power is given back to obtain additional 52% savings in leakage power which dominates the total power at such activity levels. The comparison of the two activity cases also shows that a leakage power dominated design shows a much higher reduction in total power. This is expected due to the exponential dependence of leakage current on threshold voltage as compared to a quadratic dependence of switching power on Vdd. The optimization results in an area increase of 14% and 10% for high and nominal activities respectively.

[2] This work shows that leakage power contributes 18% to the total power dissipation at the 0.13µm technology node. Since industrial designs typically employ 10% low Vth devices, leakage power can be estimated to contribute approximately 8% of the total power in designs employing high Vth devices only.

To compare the power reduction achieved using the VVS algorithm and that achieved using a dual-Vth and sizing we implement dual-Vth assignment as a sensitivity-based algorithm similar to the second phase of VVS (Section II) without the availability of a 2^{nd} Vdd. The power reduction using all three design variables simultaneously provides large benefits as compared to a dual Vth and sizing approach for most of the benchmark circuits. On an average the reduction in total power for the proposed approach is nearly double that achieved using a conventional dual Vth and sizing algorithm. For the low activity case, where the leakage power contributes more than 70% of the total power dissipation, most of the gains can be expected from high Vth insertion, and thus we expect both the approaches to perform very similarly. Results obtained using the low activity value show an average difference of 4% demonstrating the fact that the proposed approach efficiently trades-off switching and dynamic power to achieve reduction in the total power dissipation. These results demonstrate that the new single cohesive algorithm effectively seeks out the best power reduction over a range of switching activities and initial switching/leakage power breakdowns as might be found from one functional unit to another in a given design.

Figure 2 shows the impact of level conversion on the achievable power reduction for the different cases of input activities studied. The level conversion penalty is assumed to be a fixed delay overhead. Though power reduction is smaller for increasing level conversion penalty all cases show smaller sensitivity to level conversion penalty as it becomes a larger fraction of the circuit delay (i.e., clock cycle). Also the impact on the low activity circuits is significantly smaller as compared to the high activity circuits. This is due to the fact that most of the power reduction in low activity is due to high Vth assignment which is not affected by level conversion penalties. Figure 3 shows the change in power savings as we vary the extent of the backoff from the best delay point on the power-delay curve. The backoff is expressed as a percentage of the value of the minimum achievable delay. The figure clearly shows a marked fall in power reduction for very small backoff values since we move to the steeper region of the energy-delay curve and a large amount of upsizing must be initially performed to meet the delay target. This reduces the available upsizing moves in the circuit and hinders the assignment of gates to low Vdd or high Vth.

V. Conclusions

We have presented the VVS algorithm that combines gate sizing with Vdd and Vth assignment to minimize the total power dissipation and provides the designer with a single approach to minimize total power across a range of circuit parameters. The efficacy of the proposed algorithm was demonstrated on a set of ISCAS benchmark circuits. The new algorithm is compared with traditional CVS and dual-Vth with sizing algorithms to show the advantage of a single complete optimization approach. The impact of level conversion penalties and different timing constraints has also been quantified.

Table 1: Power savings at various phases of the algorithm for high input switching activity

	Initial Power (uW)			% Savings compared to initial design								
				CVS only			Backward Pass			VVS		
Circuit	Leakage	Switching	Total	Leakage	Switching	Total	Leakage	Switching	Total	Leakage	Switching	Total
c432	35.4	81.7	117.1	0.5%	1.9%	1.5%	0.5%	1.9%	1.5%	57.8%	6.0%	21.7%
c880	48.9	140.1	188.9	20.6%	19.8%	20.0%	20.6%	19.8%	20.0%	44.0%	22.9%	28.4%
c1908	75.3	202.7	278.0	5.4%	5.6%	5.5%	5.4%	5.6%	5.5%	44.1%	7.4%	17.4%
c2670	100.0	248.9	349.0	20.3%	21.4%	21.1%	20.2%	37.8%	32.7%	20.2%	37.8%	32.7%
c3540	131.6	302.6	434.2	3.4%	6.5%	5.6%	2.8%	26.4%	19.2%	49.4%	26.1%	33.2%
c5315	210.9	413.8	624.7	21.2%	25.4%	23.9%	18.9%	50.5%	39.9%	19.0%	50.7%	40.0%
c6288	544.3	1716.2	2260.5	1.1%	15.7%	12.2%	1.0%	15.8%	12.2%	20.3%	19.4%	19.6%
c7552	214.9	521.4	736.3	30.2%	32.7%	32.0%	36.4%	50.8%	46.6%	36.6%	51.2%	46.9%
Average				12.8%	16.1%	15.2%	13.2%	26.1%	22.2%	36.4%	27.7%	30.0%

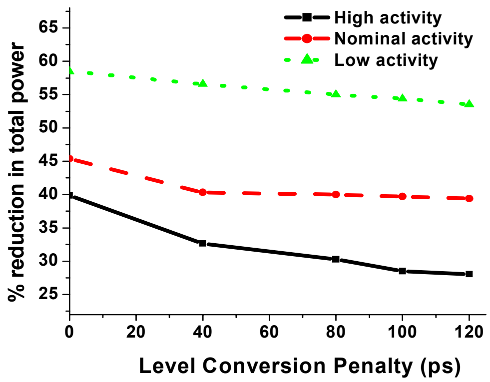

Figure 2: Impact of level conversion on power reduction

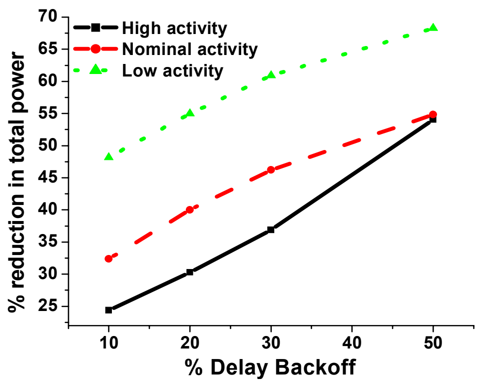

Figure 3: Dependence of power reduction on the amount of backoff

References

[1] International Technology Roadmap for Semiconductors, 2001.

[2] A. Srivastava and D. Sylvester, "Minimizing total power by simultaneous Vdd/Vth assignment," *Proc. ASP-DAC*, pp. 400-406, 2003.

[3] K. Usami, *et al*., "Automated low-power technique exploiting multiple supply voltage applied to a media processor," *IEEE JSSC,* March 1998.

[4] M. Takahashi, *et al*., "A 60-mW MPEG4 video codec using clustered voltage scaling with variable supply-voltage scheme," *IEEE JSSC,* pp. 1772-1780, Nov. 1998.

[5] M. Hamada, Y. Ootaguro, and T. Kuroda, "Utilizing surplus timing for power reduction," *Proc. CICC,* pp. 89-92, 2001.

[6] K. Usami and M. Horowitz, "Clustered voltage scaling technique for low-power design," *Proc. ISLPED*, pp. 3-8, 1995.

[7] K. Roy, L. Wei, and Z. Chen, "Multiple-Vdd & multiple Vth CMOS (MVCMOS) for low power applications," *Proc. ISCAS,* pp.366 –370, 1999.

[8] Y. S. Dhillon, *et al.,* "Algorithm for achieving minimum energy consumption in CMOS circuits using multiple supply and threshold voltages at the module level," *Proc. ICCAD*, pp.693-700, 2003.

[9] D. Nguyen, *et al.,* "Minimization of Dynamic and Static Power Through Joint Assignment of Threshold Voltages and Sizing Optimization," *Proc. ISLPED*, pp. 158-163, 2003.

[10] S. Sirichotiyakul, *et al.,* "Stand-by Power Minimization through Simultaneous Threshold Voltage Selection and Circuit Sizing," *Proc. DAC*, pp. 436-441, 1999.

[11] F. Brglez and H. Fujiwara. "A neutral netlist of 10 combinational benchmark circuits and a target translator in Fortran," *Proc. ISCAS*, pp. 695-698, May 1985.

[12] User Guide. In *Library Compiler User Manual, Synopsys, Inc.,* 2003.

[13] D. Sylvester and K. Kuetzer, "System level performance modeling with BACPAC-Berkeley advanced chip performance calculator," *Int. Workshop on System-Level Interconnect Prediction,* pp. 109-114, 1999.

[14] J. Fishburn and A. Dunlop, "TILOS: a posynomial programming approach to transistor sizing", *Proc. ICCAD,* pp.326-328, 1985.

[15] M. R. Bai and D. Sylvester, "Analysis and design of level-converting flip-flops for dual-Vdd/Vth integrated circuits," *IEEE Intl. Symp. on System-on-Chip*, pp. 151-154, 2003.

[16] F. Ishihara, F. Sheikh, and B. Nikolic, "Level conversion for dual-supply systems," *Proc. ISLPED*, 2003.

[17] S. Ercolani, *et al*., "Estimate of signal probability in combinational logic networks," *Proc. European Test Conference*, pp.294-299, 1989.

[18] J. Kao, S. Narendra and A. Chandrakasan, "Subthreshold leakage modeling and reduction techniques," *Proc. ICCAD,* pp. 141-148, 2002.

[19] R. K. Krishnamurthy, *et al*., "High-performance and low-power challenges in sub-70nm microprocessor circuits," *Proc. CICC,* pp.125-128, 2002.