ILP Based Leakage Optimization During Nano-CMOS RTL Synthesis: A DOXCMOS Versus DTCMOS Perspective

Saraju P. Mohanty
Department of Computer Science and Engineering
University of North Texas, Denton, TX 76203, USA.
Email: saraju.mohanty@unt.edu

Bijaya K. Panigrahi
Department of Electrical Engineering
Indian Institute of Technology, New Delhi - 110016, India.
Email: bkpanigrahi@ee.iitd.ac.in

***Abstract*—In this paper, an integer linear programming (ILP) based algorithm is presented that considers resource constraints and optimize leakage delay product (LDP) using a precharacterized register transfer level (RTL) library. For nanoscale CMOS (nano-CMOS) circuits leakage is a predominate form of power dissipation. Leakage optimization at the early stage of design cycle, such as during high-level synthesis is quite few. Two techniques, dual-T_{ox} (DOXCMOS) and dual-V_{th} (DTCMOS) technology are explored during the high-level synthesis for leakage optimization. The leakage is assumed to be sum of gate-oxide leakage and subthreshold leakage. Register transfer level (RTL) components are characterized for DOXCMOS and DTCMOS technology accounting for process variations, which is an important issue for nanoscale circuits. Experiments were performed on several high-level synthesis benchmark circuits, which show an average reduction of 79% gate leakage and 76% of subthreshold leakage for DOXCMOS and DTCMOS technology, respectively. It is observed that DOXCMOS technology based optimization out performed the results from DTCMOS technology based optimization.**

***Keywords*-Integer Linear Programming; Leakage Optimization; Nanoscale Circuit Optimization; Register Transfer Level Optimization; Low-Power High-Level Synthesis**

I. Introduction

CMOS device scaling is performed aggressively to achieve higher integration density and more functionality. Industry has now reached to a point where $32nm$ CMOS devices are reality and production ready. The transistor count in the integrated circuits (chips) has reached one billion mark. For examples, Intel Core i7 (Quad) has 0.731B transistors (Intel is supposed to reach 1B transistor count chip production mark by this year end with another processor), NVIDIA GT200 has 1.4B transistors, and Xilinx Virtex-5 has 1.1B transistors. Power dissipation is an important axis in the design space exploration for nanoscale CMOS technology based design of these chips. In short channel nanoscale CMOS circuits both leakage power is as important as the dynamic power dissipation [1], [2]. Hence, optimization of leakage has become important. The leakage optimization can be carried out at different levels of design abstraction from system level to physical, however, it is proposed to optimize leakage at the architectural level as the possibility of reduction is higher and circuit complexity lower which will ensure faster solutions.

The leakage current in short channel nano-CMOS device has diverse forms, such as reverse biased diode leakage, subthreshold leakage, gate-oxide leakage, hot carrier gate current, gate induced drain leakage (GIDL), channel punch through, and (reverse-biased) drain-substrate and source-substrate junction band-to-band tunneling [1], [3], [4]. Each one of them has diverse origins and they flow between different terminals and in different operating conditions of a nano-CMOS transistor. In this paper, the optimization of subthreshold leakage and gate-oxide leakage is proposed in order to contain the leakage of short channel devices with ultra thin oxide thickness in both their ON and OFF states.

Due to demand of portable systems, thermal considerations, environmental concerns and reliability issues, the need for low power synthesis is increasing. These factors affect the battery life, cooling and packaging costs, and use of natural resources. For energy efficient (longer battery life) and high performance (smaller delay) circuits, the leakage delay product (LDP) has to be reduced. Although tons of research works addressing analysis of leakage have been presented in literature, the research works on leakage reduction is quite few in number, particularly at higher-level of circuit abstraction [3], [5], [6].

The prior research in high-level synthesis mostly considered dynamic power and few of them have addressed leakage. In [7], dual-V_{th} technique is proposed for subthreshold leakage reduction. In [8], [9], MTCMOS approach is used for reduction of subthreshold current. In [10], power island partitioning has been used to reduce subthreshold leakage. In [11], a heuristic based approach using dual-V_{th} library is proposed for subthreshold leakage reduction. First ever high-level synthesis works optimizing gate leakage is presented in [3] that uses dual-T_{ox} technology. It may be noted that none of these research account process variations in their optimization, thus are limited in terms of accuracy.

In this paper, a resource constrained integer linear programming bases (ILP) algorithm is proposed that optimizes

This research is supported in part by NSF award numbers CCF-0702361 and CNS-0854182.

leakage delay product (LDP) of nano-CMOS datapath circuits. The leakage power includes both gate-oxide leakage and subthreshold leakage. The algorithm considers an unscheduled data flow graph (DFG), and schedules each of its node at appropriate control steps and simultaneously binds them to the best available resource so as to achieve the desired optimization. The algorithm uses a precharacterized register transfer level (RTL) library for optimization. The RTL library is constructed for two nano-CMOS technology, dual-T_{ox} (DOXCMOS) and dual-V_{th} (DTCMOS), accounting process variation. The optimization is performed for these two technology and a comparative study of the two is performed for their efficacy in reducing leakage.

II. Leakage Modeling in Nano-CMOS Devices

Gate-oxide leakage is due to the quantum mechanical tunneling of carriers (electrons or holes) across gate-oxide potential barrier. For direct tunnelling, the tunneling probability of an electron is affected by barrier height, structure and thickness. The current density of a MOS is expressed by the following [4], [12], [13], [3], [6]:

$$J_{gate_{MOS}} = \frac{q^3 V_{ox}^2}{16\pi^2 \hbar \phi_B T_{ox}^2} * exp\left[-\frac{4\sqrt{2m_{eff}}{\phi_B}^{1.5} T_{ox}}{3\hbar q V_{ox}} * \left\{ 1 - \left(1 - \frac{V_{ox}}{\phi_B} \right)^{1.5} \right\} \right], \quad (1)$$

where, q is electronic charge, V_{ox} is voltage across the gate oxide, $\hbar$ is Planck's constant, T_{ox} is electrical equivalent oxide thickness, ϕ_B is barrier height for the gate dielectric, and m_{eff} is effective carrier mass.

The gate-oxide leakage current is divided into five components, such as I_{gs} and I_{gd} (components due to the overlap of gate and diffusions, called edge direct tunneling), I_{gcs} and I_{gcd} (components due to tunneling from the gate to the diffusions via the channel) and I_{gb}, the component due to tunneling from the gate to the bulk via the channel [5], [14], [15]. Depending on the biasing conditions either source or the drain component dominate the total gate leakage. The gate-oxide leakage of NMOS is more than that of the PMOS, however, PMOS gate-oxide leakage is not insignificant and need to be accounted for accurate estimation and optimization [2]. The gate-oxide leakage for a device is then calculated using the following expression:

$$I_{gate_{MOS}} = \left| I_{gs} + I_{gd} + I_{gcs} + I_{gcd} + I_{gb} \right|. \quad (2)$$

The direction and values of these components are different for different operating states of the device.

The subthreshold leakage current in a CMOS transistor is given by the following expression [16], [1], [4]:

$$I_{sub_{MOS}} = \left(\mu_0 v_{therm}^2 e^{1.8} \right) \times \left(\frac{\epsilon_{ox} \times W_{eff}}{T_{ox} \times L_{eff}} \right) \times exp\left(\frac{V_{gs} - V_{th}}{v_{therm} S} \right) \times \left(1 - exp\left(\frac{-V_{ds}}{v_{therm}} \right) \right), \quad (3)$$

where μ_0 is the zero bias mobility, ϵ_{ox} dielectric constant of the gate oxide, L_{eff} is the effective channel length, V_{th} is the threshold voltage, v_{therm} is the thermal voltage, S is the subthreshold swing factor, V_{gs} is gate-to-source voltage, and V_{ds} is the drain-to-source voltage.

The above models are used for each transistor in a circuit level netlist and using SPICE as a nonlinear equation solver the currents are calculated.

III. ILP based Optimization Algorithm

This section discusses the two technology for leakage optimization and the ILP based optimization approach for leakage reduction during RTL synthesis.

A. DOXCMOS and DTCMOS Technology

From Eqn. (1) and (3), it can be observed that leakage of a device is affected by several parameters, such as T_{ox}, ϕ_B, μ_0, ϵ_{ox}, L_{eff}, V_{th}, V_{gs}, and V_{dd}, and operating temperature, etc. Scaling of gate-oxide thickness T_{ox} and threshold voltage V_{th} has been used to reduce gate-leakage and subthreshold leakage, respectively as they affect the leakage more strongly than other parameters. Thus, giving rise to dual-T_{ox} (dual oxide, DOXCMOS) and dual-V_{th} (dual threshold, DTCMOS) technology, respectively.

In DOXCMOS technology, devices of two different oxide thicknesses are used for gate-oxide leakage reduction while preserving the performance [17]. In context of high-level synthesis, resources made of either high-T_{ox} or low-T_{ox} are used selectively for leakage and delay trade-offs.

In DTCMOS technology, additional high-V_{th} sleep transistors are used between supply to ground which are turned-off to discontinue the flow, thus reducing the leakage [18]. In the context of high-level synthesis, resources of high-V_{th} or low-V_{th} are used for leakage and delay trade-offs.

B. Datapath Specifications and Target Architecture

It is assumed that a nano-CMOS circuit is specified by: (i) a sequencing data flow graph (DFG), (ii) a RTL library precharacterized for gate and subthreshold leakage and delay for DOXCMOS and DTCMOS technology, and (iii) a set of resource (R_{con}) and time constraints (T_{con}). The datapath is assumed to be specified as a sequencing DFG, which is a directed acyclic graph. Each vertex of the DFG represents an operation and each edge represents a dependency. Each vertex has attributes that specify the operation type. In the assumed target architecture model, each functional unit feeds one register and also has a multiplexer. A register or a multiplexer is made of same technology as that of the associated functional unit. A controller decides which functional units are active in each control step and the inactive ones are disabled using the multiplexers or sleep signals. The delay of a control step (d_c) is dependent on the delays of the functional unit, the multiplexer, and register.

C. Optimization Problem Statement

RTL synthesis involves various steps: compilation, transformation, datapath scheduling, resource allocation, operation binding, connection allocation and architecture generation. Resource or time constrained scheduling time stamps the variables and operations in the DFG so that the operations in the same group can be executed concurrently. While allocation fixes the number and types of resources to be used in the datapath, the binding process involves attaching operations to functional units and variables to memory units.

The gate or subthreshold leakage optimization problem during high-level synthesis can be formalized as follows: *Given an unscheduled data flow graph (UDFG) $G_U(V, E)$, it is required to find the scheduled data flow graph (SDFG) $G_S(V, E)$ with appropriate resource binding such that the total leakage and delay product (LDP) is minimized under given resource constraints.*

The above can formally be stated as an optimization problem. Let V be the set of all vertices and V_{cp} be the set of vertices in the critical path from the source vertex of the DFG to the sink vertex. $R_{k,t}$ denotes resources (functional units) of type k made up of transistors of a specific T_{ox} or V_{th}. t is called technology index which represents high-T_{ox}/low-T_{ox} for DOXCMOS or high-V_{th}/low-V_{th} for DTCMOS. c is a clock cycle in the total number of clock cycles N in a DFG. The optimization problem can then be stated as follows:

$$\text{Objective Function: Minimize } \left(LDP(\text{DFG})\right), \tag{4}$$

$$\text{Constraints: Allocated } \left(R_{k,t}\right) \leq \text{Available } \left(R_{k,t}\right),\ \forall c \in N. \tag{5}$$

The objective function ensures minimization of gate or subthreshold leakage and delay simultaneously. The constraints ensure that the total allocation of the i^{th} resources of type k and made up of transistors of technology index t is less than the total number of same resources available. The LDP of the DFG is estimated as the sum for all control steps using the following expression:

$$LDP(\text{DFG}) = \sum_{c=1}^{N} LDP_c \tag{6}$$

$$= \sum_{c=1}^{N} \sum_{\forall v_{i,c}} P_{leakage}(v_{i,c}) \times d_c, \tag{7}$$

where, $v_{i,c}$ is a vertex v_i scheduled in cycle c of delay d_c. $P_{leakage}(v_{i,c})$ is the leakage of the corresponding resource active due to the execution of the same vertex, which may be either gate leakage P_{gate} or subthreshold leakage P_{sub}. The leakage and delay can either be nominal values or mean values in the case of statistical data accounting nanoscale process variations.

D. Integer Linear Programming (ILP) Formulation

The following notations are assumed in order to formulate an ILP based optimization scheme for the DFG: $M_{k,t}$ - maximum number of $R_{k,t}$ resources, S_i - as soon as possible time stamp for the vertex v_i, E_i - as late as possible time stamp for the vertex v_i, $LDP(i,t)$ - leakage delay product of $R_{k,t}$ used by vertex v_i, $X_{i,t,c}$ - decision variable which takes the value of 1 if v_i is using $R_{k,t}$ and scheduled in control steps c, and $L_{i,t}$ - latency in number of cycles for v_i using $R_{k,t}$. The ILP formulation needs to minimize the LDP while satisfying the resource constraints and data dependency ensuring that every vertex is scheduled in uniquely allowable control steps. The formulations are presented for single cycle scenario which can be easily modified for multicycling through $L_{i,t}$. When a vertex uses a nominal-T_{ox}/V_{th} resource, then it is scheduled in one unique control step. On other hand, when a vertex is using a high-T_{ox}/V_{th} resource it may need more than one clock cycle $L_{i,t}$ for completion, thus restricting the mobility.

(a) *Objective Function*: The objective is to minimize the LDP of the whole DFG over all control steps. This can be expressed using decision variable as follows:

$$Minimize : LDP(DFG), \tag{8}$$

$$Minimize : \sum_{c} \sum_{i} \sum_{t} X_{i,t,c} * LDP(i,t). \tag{9}$$

(b) *Uniqueness Constraints*: These constraints ensure that each vertex v_i is scheduled in the appropriate control step within the mobility range (S_i, E_i) being assigned the resource $R_{k,t}$. A vertex may be operated with more than one clock cycle sometimes depending on the delay of a resource. These constraints are represented as, $\forall i,\ 1 \leq i \leq V$,

$$\sum_{c} \sum_{t} X_{i,t,c} = 1. \tag{10}$$

(c) *Precedence Constraints*: These constraints guarantee that for a vertex v_i, all its predecessors are scheduled in earlier control steps and its successors are scheduled in later control steps. These constraints are modeled as, $\forall i, j, v_i \in Pred_{v_j}$,

$$\sum_{t} \sum_{d=S_i}^{E_i} d * X_{i,t,d} - \sum_{t} \sum_{e=S_j}^{E_j} e * X_{j,t,e} \leq -1. \tag{11}$$

(d) *Resource Constraints*: These constraints ensure that each control step needs resources not exceeding available number of resources (which are specified by allocation). These can be enforced as, $\forall t$ and $\forall c$, $1 \leq c \leq N$,

$$\sum_{i \in R_{k,t}} X_{i,t,c} \leq M_{k,t}. \tag{12}$$

E. The Leakage Optimization Algorithm Flow

The flow of the proposed optimization approach is presented in Algorithm 1. The inputs to the algorithm are an unscheduled data flow graph (UDFG), the resource constraints, the delay of each resource, the multiplexer, the register for different technology. For given resource constraints,

the algorithm determines a RTL implementation that has minimum LDP. The resource constraints are expressed as number of different types of resources made of transistors of each technology index.

Algorithm 1 ILP Based LDP Optimization during Low-Power Nano-CMOS RTL Synthesis

1: Preprocess given behavioral description to construct a sequencing data flow graph (DFG).
2: Perform simulations to estimate gate or subthreshold leakage and delay of register-transfer level (RTL) units.
3: Construct resource allocation table and available resource table based on input resource constraints R_{con}.
4: Determine the number of different resources for each technology index t using the resource allocation table.
5: Obtain resource constrained as-soon-as-possible (ASAP) and as-late-as-possible (ALAP) schedules.
6: Construct the mobility graph based on above schedules.
7: Model the ILP formulations of the DFG using AMPL.
8: Obtain the final optimal solution by solving the ILP formulations.
9: Estimate the gate and subthreshold leakage and delay.
10: Postprocess scheduled sequencing DFG to generate gate or subthreshold leakage optimal RTL description.

IV. Experimental Results

This section discusses the method adopted for creating the RTL library. This is followed by the experiments performed on selected high-level synthesis benchmark circuits.

A. RTL Library Characterization

To characterize the leakage and delay of datapath components, their RTL description is synthesized to logic-level netlist. As a trade-off accuracy and time, simulation is performed on the netlist for a series of test vectors. The input test vectors are generated as correlated signals. Then, the results for all of the simulations is averaged over to calculate the leakage and delay of a architectural unit.

For the logic level leakage analysis, it is assumed that in a n-bit functional unit there are total n_{total} NAND gates out of which n_{cp} are in the critical path. The assumption of NAND realization is based on two facts: first it is a universal gate and it is lower leaky logic gate compared to other gates [5], [7]. Once the leakage current per gate is known, the *average* leakage of an n-bit RTL unit is calculated as [19]:

$$P_{gate_R} = \sum_{j=1}^{n_{total}} P_{gate\text{NAND}_i}, \quad (13)$$

$$P_{sub_R} = \sum_{j=1}^{n_{total}} P_{sub\text{NAND}_i}. \quad (14)$$

The critical path delay of an n-bit functional unit (i.e. resource or datapath component) using the above NAND gates as building blocks is calculated as follows:

$$T_{pd_R} = \sum_{i=1}^{n_{cp}} T_{pd\text{NAND}_i}. \quad (15)$$

By probabilistic reasoning it is assumed that there are as many low-to high transitions as there are high-to-low.

Gate and subthreshold leakage for the logic gate for a specific state is calculated using device level models through the use of SPICE. The average delay of a logic gate is calculated as the average of t_{HL} and t_{LH}, high-to-low transition time and low-to-high transition time, respectively.

The process variations are accounted through Monte Carlo simulations while characterizing gate and subthreshold leakage and delay. It is observed that logic level distributions of leakage are lognormal in nature and delay is normal distribution. Using Central Limit Theorem (CLT) the logic level distributions are translated to architectural level. Since a typical resource comprise of hundreds of logic gates, the leakage and delay for a RTL component will be normally distributed. Assuming that the distributions for each gate are statistically independent of each other, the mean and variance of the leakages can be derived as follows:

$$\mu_R = \sum_{i=1}^{n_{total}} \mu_{\text{NAND}_i} \text{ and} \quad (16)$$

$$\sigma_R = \sqrt{\sum_{i=1}^{n_{total}} \sigma^2_{\text{NAND}_i}}. \quad (17)$$

Similar approach is used to account the process variations in delay where n_{total} is replaced with n_{cp} in the expressions.

The result of the above approach is a RTL component library which will be used in the optimization algorithm for high-level synthesis. We have presented data for nominal values corresponding to the $45nm$ BSIM4 model in the Fig. 1. Exhaustive data is not provided in this paper due to page limitations and will be presented in longer journal version of this research. For high-T_{ox} devices, the channel length is changed proportionately to avoid impact on its functionality. This constant aspect ratio of ensures constant per width gate capacitance of the transistor, as per fabrication requirements and to reduce short channel impacts [17], [3].

B. Experiments with Benchmark Circuits

The overall design flow is implemented using C and integrated into the high-level synthesis framework in [20]. The algorithm is experimented with various high-level synthesis benchmark circuits [21]. The baseline case for the experiments is the $T_{ox} = 1.4nm$ and $V_{th} = 0.22V$, corresponding to the nominal case of $45nm$ BSIM4 model. The algorithm is experimented on data intensive digital signal processing benchmarks whose applications are immense in

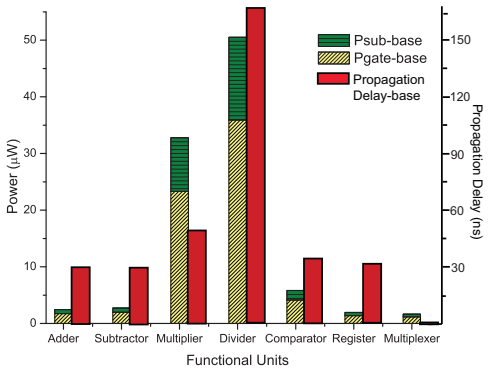


Figure 1. Characterization data for nominal baseline parameters.

day to day life and needed to be low power consuming for environmental friendly. For each benchmark several sets of experiments are performed for various resource constraints.

The percentage reduction in leakage for a particular experiment is calculated as follows: ΔP_{gate} or $\Delta P_{sub} = \left(\frac{P_{base}-P_{final}}{P_{base}}\right) * 100\%$. The percentage penalty in critical path delay for a particular experiment is calculated as follows: $\Delta T_{cp} = \left(\frac{T_{cp_{final}}-T_{cp_{base}}}{T_{cp_{base}}}\right) * 100\%$.

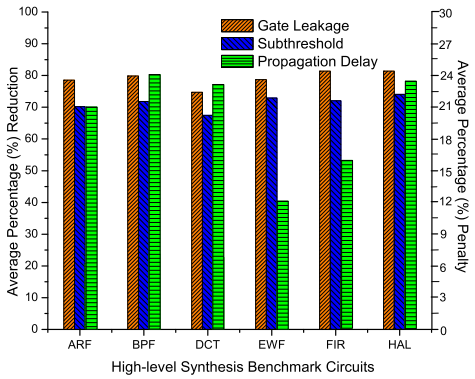
The percentage reduction in gate and subthreshold leakage and critical path delay penalty averaged over all resources constraints is presented in Fig. 2 for high-$T_{ox} = 1.7nm$ and high-$V_{th} = 0.25V$, reasonable for fabrication purpose. For DOXCMOS technology, the average reduction in gate leakage for the benchmarks ranges from 74% to 81%. This is accompanied by subthreshold leakage reduction of 67% to 74% and delay penalty of 12% to 23%. For the DTCMOS technology, the average reduction in subthreshold leakage for all the benchmarks ranges from 70% to 77% accompanied with a gate leakage reduction of 62% to 70% and time penalty of 17% to 27%. Experiments show that as the number of available high-T_{ox}/V_{th} resources increases, the reduction in gate / subthreshold leakage also increases. The benchmarks that had more number of operations needing high leaky resources created opportunity to be replaced with high-T_{ox}/V_{th} resources resulting in more reduction.

C. Discussions of Results: DOXCMOS Versus DTCMOS

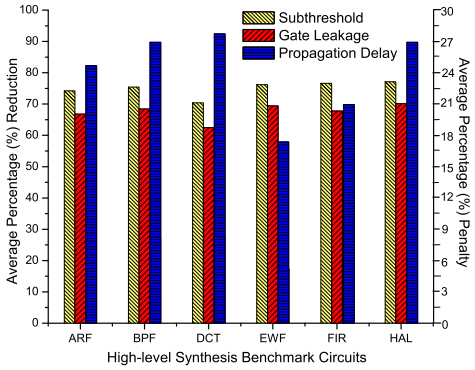
For DOXCMOS technology, the reduction of subthreshold leakage is due to change of V_{th} with T_{ox}, related by:

$$V_{Th} = V_{fb} + 2\phi_F + \left(\frac{T_{ox}}{\epsilon_{ox}}\right)\sqrt{2q\epsilon_{Si}N_{sub}\left(2\phi_F + V_{bs}\right)}. \quad (18)$$

Where V_{fb} is the flat-band voltage, V_{bs} is the body bias, γ_{body} is the body effect coefficient, and ϕ_F is the Fermi-level. Thus, V_{th} and T_{ox} have linear relationship. The increase in T_{ox} increases V_{th} and consequently decrease in subthreshold leakage. The reduction in gate leakage due to increase in V_{th} is not straight forward. It may be attributed to the fact that the gate tunneling current density is reduced due to increase of V_{th}, as voltage across oxide drops by V_{th}. A comparative perspective of the DOXCMOS and DTCMOS technologies is presented in Table I.



(a) DOXCMOS Technology



(b) DTCMOS Technology

Figure 2. Average experimental results for various benchmarks.

V. Conclusions

An ILP based algorithm is presented for leakage optimization during high-level, RTL, or architectural synthesis. The algorithm uses DOXCMOS and DTCMOS technology for leakage and delay reduction under resource constraints. The experiments proved that both the techniques are quite effective. Also, it is observed that the percentage reductions in leakage is higher compared to existing literature [7], [8], [9], [11]. It is observed that DOXCMOS technology outperforms the DTCMOS technology and may be cheaper from fabrication point of view as well as only parameter need to be controlled. Evaluation of the impact of these techniques on the area, capacitance and dynamic power is ongoing research. While the optimization is performed based on the mean value of power and delay distributions, our future optimization will account the variances as well to more accurately account process variations.

Table I
COMPARATIVE PERSPECTIVE OF THE TWO TECHNOLOGY

DOXCMOS Technology	DTCMOS Technology
One parameter varied.	Several parameters varied.
Less area overhead: T_{ox} and L.	More area overhead: additional transistors.
Gate/subthreshold directly affected.	Subthreshold directly and gate indirectly.
Higher reduction.	Lower reduction.

REFERENCES

[1] A. Agarwal, S. Mukhopadyaya, A. Roychowdhury, K. Roy, and C. H. Kim, "Leakage Power Analysis and Reduction for Nanoscale Circuits," *IEEE Micro*, vol. 26, no. 2, pp. 68–80, March-April 2006.

[2] E. Kougianos and S. P. Mohanty, "Metrics to Quantify Steady and Transient Gate Leakage in Nanoscale Transistors: NMOS vs. PMOS Perspective," in *Proceedings of the International Conference on VLSI Design*, 2007, pp. 195–200.

[3] S. P. Mohanty and E. Kougianos, "Modeling and Reduction of Gate Leakage during Behavioral Synthesis of NanoCMOS Circuits," in *Proceedings of the 19th International Conference on VLSI Design*, 2006, pp. 83–88.

[4] K. Roy, S. Mukhopadhyay, and H. M. Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," *Proceedings of the IEEE*, vol. 91, no. 2, pp. 305–327, February 2003.

[5] V. Mukherjee, S. P. Mohanty, and E. Kougianos, "A Dual Dielectric Approach for Performance Aware Gate Tunneling Reduction in Combinational Circuits," in *Proceedings of the 23rd IEEE International Conference of Computer Design (ICCD)*, 2005, pp. 431–436.

[6] S. Mukhopadhyay, C. Neau, R. T. Cakici, A. Agarwal, C. H. Kim, and K. Roy, "Gate leakage reduction for scaled devices using transistor stacking," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 11, no. 4, pp. 716–730, Aug 2003.

[7] K. S. Khouri and N. K. Jha, "Leakage power analysis and reduction during behavioral synthesis," *IEEE Transactions on VLSI Systems*, vol. 10, no. 6, pp. 876–885, December 2002.

[8] C. Gopalakrishnan and S. Katkoori, "Knapbind: an area-efficient binding algorithm for low-leakage datapaths," in *Proceedings of 21st International Conference on Computer Design*, 2003, pp. 430–435.

[9] ——, "Resource allocation and binding approach for low leakage power," in *Proceedings of 16th International Conference on VLSI Design*, 2003, pp. 297–302.

[10] D. Dal, A. Nunez, and N. Mansouri, "Power islands: a high-level technique for counteracting leakage in deep submicron," in *Proceedings of the 7th International Symposium on Quality Electronic Design*, 2006.

[11] X. Tang, H. Zhou, and P. Banerjee, "Leakage power optimization with dual-v_{th} library in high-level synthesis," in *Proceedings of the 42nd Design Automation Conference*, 2005, pp. 202–207.

[12] M. Depas, B. Vermeire, P. W. Mertens, R. L. V. Meirhaeghe, and M. M. Heyns, "Determination of Tunneling Parameters in Ultra-Thin Oxide Layer Poly-Si/SiO_2/Si Structures," *Elsevier Solid-State Electronics Journal*, vol. 38, no. 8, pp. 1465–1471, August 1995.

[13] C. H. Choi, K. H. Oh, J. S. Goo, Z. Yu, and W. W. Dutton, "Direct Tunneling Current Model for Circuit Simulation," in *Proceedings of International Electron Devices Meeting*, 1999.

[14] Y. Cao, T. Sato, D. Sylvester, M. Orshansky, and C. Hu, "New Paradigm of Predictive MOSFET and Interconnect Modeling for Early Circuit Design," in *Proceedings of the IEEE Custom Integrated Circuits Conference*, 2000, pp. 201–204.

[15] "Berkley Short-Channel Insulated-Gate Model (BSIM4)," http://www-device.eecs.berkeley.edu/~bsim3/bsim4_get.html.

[16] F. Sill, J. You, and D. Timmerman, "Design of Mixed Gates for Leakage Reduction," in *Proceedings of the 17th Great Lakes Symposium on VLSI*, 2007, pp. 263–268.

[17] N. Sirisantana and K. Roy;, "Low-power Design using Multiple Channel Lengths and Oxide Thicknesses," *IEEE Design & Test of Computers*, vol. 21, no. 1, pp. 56–63, Jan-Feb 2004.

[18] S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, and J. Yamada, "1-V power supply high-speed digital circuit technology with multithreshold-voltage CMOS," *IEEE Journal of Solid-State Circuits*, vol. 30, no. 8, pp. 847–854, Aug 1995.

[19] A. Agarwal, S. Mukhopadhyay, C. H. Kim, A. Raychowdhury, and K. Roy, "Leakage power analysis and reduction: models, estimation and tools," *IEE Proceedings- Computers and Digital Techniques*, vol. 152, no. 3, pp. 353–368, May 2005.

[20] S. P. Mohanty and N. Ranganathan, "A Framework for Energy and Transient Power Reduction during Behavioral Synthesis," *IEEE Transactions on VLSI Systems*, vol. 12, no. 6, pp. 562–572, June 2004.

[21] S. P. Mohanty, N. Ranganathan, E. Kougianos, and P. Patra, *Low-Power High-Level Synthesis for Nanoscale CMOS Circuits*. Springer, 2008, 0387764739 and 978-0387764733.