ILP based Gate Leakage Optimization using DKCMOS Library during RTL Synthesis

Saraju P. Mohanty
Dept. of Computer Science and Engineering
University of North Texas, Denton, TX 76207.
Email-ID: smohanty@cse.unt.edu

Abstract

In this paper dual-K (DKCMOS) technology is proposed as a method for gate leakage power reduction. An integer linear programming (ILP) based algorithm is proposed for its optimization during architectural synthesis. The algorithm uses device-level gate leakage models for precharacterizing register-transfer level (RTL) datapath component library and minimizes the leakage delay product (LDP). The proposed algorithm is tested for several circuits for $45nm$ *CMOS technology node. The experiments show that average gate leakage reduction are* 67.7% *and* 80.8% *for* SiO_2*-SiON and* SiO_2*-*Si_3N_4*, respectively.*

1 Introduction and Motivation

In order to meet higher packaging density and performance, VLSI industry has resorted to aggressive technology scaling. Intel has prototyped a processor called *Penryn* using transistors of $45nm$ technology [2]. In nanoscale CMOS transistors leakage is a prominent form of power dissipation. The three major sources of leakage in a nanoscale CMOS circuit are: (i) gate tunneling leakage, (ii) subthreshold leakage, and (iii) (reverse-biased) drain-substrate and source-substrate junction band-to-band tunneling [4, 10, 15]. Of all these leakage mechanisms, SiO_2 tunneling current that flows during both active and sleep modes of a device is significant for sub-$65nm$ technology node with *ultra thin* oxide thickness [4, 15, 10].

Although research work addressing analysis of gate leakage has been presented in literature, research on its reduction is quite few in number [13, 14]. Due to demand of portable systems, thermal considerations, environmental concerns and reliability issues, the need for low power synthesis is increasing. For energy efficient (longer battery life) and high performance (smaller delay) circuits, the leakage delay product (LDP) has to be reduced. It is well known that high-level design space exploration can lead to larger savings in power dissipations.

2 The Proposed and Background Research

The contributions of this paper is the introduction and use of DKCMOS technology for gate leakage optimization of datapath circuits for specified constraints. An ILP based algorithm is proposed that schedules operations of a sequencing data flow graph (DFG) using precharacterized RTL library. The RTL library is constructed for three nano-CMOS technology, classical SiO_2 device based modules, and two nonclassical high-K based modules. The algorithm minimizes the leakage delay product (LDP) of datapath circuits for given resource constraints. The prior research in behavioral synthesis mostly considered dynamic power and few of them have dealt with leakage [9, 7, 16].

3 Dual-K CMOS (DKCMOS) Technology

Materials such as, ZrO_2, TiO_2, BST, HfO_2, Al_2O_3, SiON, and Si_3N_4, have been investigated for use in CMOS technology [8, 17]. Significant progress has taken place in the development high-K dielectric deposition techniques [8]. Thus, the fabrication of high-K dielectric based devices and circuits is a reality. While the materials research is in full swing, there is no research addressing automatic design or synthesis of circuit using high-K devices.

In the proposed DKCMOS technology illustrated in Fig. 1, the circuit has either high-K or SiO_2 devices. Fig. 1(a) shows a nominal logic gate with all SiO_2 devices. In Fig. 1(b), the high oxide leaky devices, which are NMOS can be constructed with high-K dielectric and is more close to dual-V_{Th} technology. In Fig. 1(c) the logic-gate is made of all high-K devices. DKCMOS technology based optimization can be performed during synthesis at various levels of design abstraction, such as RTL, logic, and physical synthesis. Depending on the the design abstraction, SiO_2 devices, logic gates, or RTL components, would be selectively replaced with high-K devices, logic-gates, or RTL components for gate leakage reduction while maintaining performance. The granularity of optimization process will vary, but in essence at the lowest level of design abstraction, two islands, one SiO_2 and one high-K will be created. This may increase the cost of fabrication by couple of masks. In this paper it is proposed that a mix of RTL units of type (a) and type (c) can serve gate leakage and performance trade-offs and can go well with industry trend. During the high-level synthesis, selection of high-K and SiO_2 RTL modules can be performed for gate leakage and performance trade-offs.

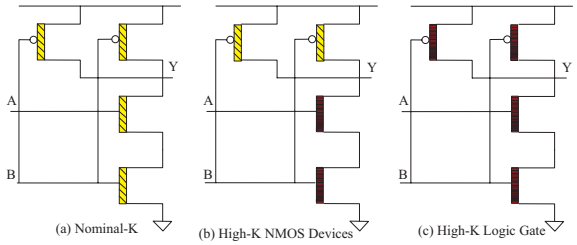


Figure 1. The DKCMOS Technology.

4 Gate Leakage Modeling

Gate leakage is due to the quantum mechanical tunneling of carriers 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 [15, 10] the following in which all the associated parameters are implicitly or explicitly affected by the use of high-K:

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

The gate leakage current has 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 [13, 5, 1]. The gate leakage for a device can then be calculated as follows:

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

For SPICE based analysis of high-K non-classical devices, two possible options are: (i) varying the parameter in the model card that denotes relative permittivity (EPSROX) or (ii) finding the equivalent oxide thickness (EOT). An equivalent oxide thickness (T_{ox}^*) is calculated according to the formula [13]: $T_{ox}^* = \left(\frac{\epsilon_{gate}}{\epsilon_{ox}} \right) \times T_{ox}$, where, ϵ_{gate} is the dielectric constant or relative permittivity of a high-K dielectric and ϵ_{ox} is the permittivity of SiO_2.

5 Problem Formulation

The datapath is assumed to be specified as a sequencing DFG. Each vertex of the DFG represents an operation and each edge represents a dependency. The target architecture model assumed is shown in Fig. 2. Each functional unit feeds one register and also has a multiplexer. The register and the multiplexer belong to same island (high-K or low-K) as that of the functional units. The delay of a control step (d_c) is dependent on the delays of the functional unit, the multiplexer, and register. In one scenario, (a) and (b), gate leakage reduction can be achieved by selection of a high-K ALU in 2nd cycle in high-K ALU. In another scenario, (c) and (d), gate leakage can be achieved by by selection of a high-K ALU in 2nd cycle. The number of SiO_2 or high-K resources of the architecture depends on the resource constraints and allocation.

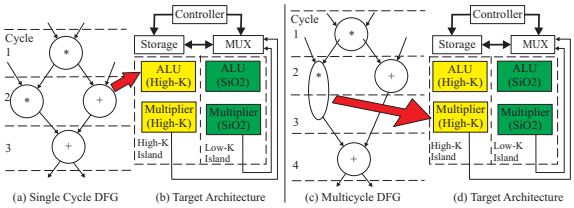


Figure 2. The Target Architecture.

The optimization problem during architectural synthesis can be formalized as: *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 gate leakage and delay product (LDP) is minimized under given resource.*

The above can formally be stated as follows. 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. $R_{t,\mathrm{K}}$ denotes resources of type t and made up of transistors of dielectric K. c is cycle in total N clock cycles in a DFG. The optimization problem can then be stated as:

$$\text{Minimize} : LDP(\text{DFG}), \quad (3)$$

$$\text{Allocated}\left(R_{t,\mathrm{K}}\right) \leq \text{Available}\left(R_{t,\mathrm{K}}\right), \ \forall c \in N. \quad (4)$$

The objective function ensures minimization of gate leakage and delay simultaneously. The constraints ensure that the total allocation of the i^{th} resources (functional units) of type t and made up of transistors of dielectric K is less than the total number of same resources available. Another alternative for the same optimization is minimizing gate leakage (instead of LDP) for specified time and resource constraints. However, considering time as a factor in the objective function under resource constraint can lead to faster convergence compared to both time and resource constraints. The LDP of the DFG is the sum for all control steps expressed as:

$$LDP(\text{DFG}) = \sum_{c=1}^{N} LDP_c = \sum_{c=1}^{N} \sum_{\forall v_{i,c}} P_{gate}(v_{i,c}) \times d_c, \quad (5)$$

where, $v_{i,c}$ is a vertex v_i scheduled in c with delay d_c and $P_{gate}(v_{i,c})$ is the gate leakage of a resource that it is using.

6 ILP based Gate Leakage Optimization

In order to formulate the ILP based optimization scheme [3, 11], the notations given in Table 1 are used.

6.1 ILP Formulations

(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:

$$Minimize: LDP(DFG), \tag{6}$$

$$Minimize: \sum_{l}\sum_{i}\sum_{\mathrm{K}} X_{i,\mathrm{K},l,(l+L_{i,\mathrm{K}}-1)} * LDP(i,\mathrm{K}). \tag{7}$$

(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 $R_{t,\mathrm{K}}$. A vertex may be operated with more than one clock cycle sometimes depending on the delay of a resource. These constraints are represented as, $\forall i, 1 \leq i \leq V$,

$$\sum_{\mathrm{K}} \sum_{l=S_i}^{S_i+E_i+1-L_{i,\mathrm{K}}} X_{i,\mathrm{K},l,(l+L_{i,\mathrm{K}}-1)} = 1. \tag{8}$$

When a vertex uses a nominal-K resource, then it is scheduled in one unique control step. On other hand, when a vertex is using a high-K resource it may need more than one clock cycle for completion, thus restricting the mobility.

(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 should also ensure the multicycling and are modeled as, $\forall i, j, v_i \in Pred_{v_j}$,

$$\sum_{\mathrm{K}} \sum_{l=S_i}^{E_i} (l + L_{i,\mathrm{K}} - 1) * X_{i,\mathrm{K},l,(l+L_{i,v}-1)} - \sum_{\mathrm{K}} \sum_{l=S_j}^{E_j} l * X_{j,\mathrm{K},l,(l+L_{j,\mathrm{K}}-1)} \leq -1. \tag{9}$$

(d) *Resource Constraints*: These constraints ensure that each cycle uses resources not exceeding available number of resources and are enforced as, $\forall$K and $\forall l, 1 \leq l \leq N$,

$$\sum_{i \in R_{t,\mathrm{K}}} X_{i,\mathrm{K},l,(l+L_{i,\mathrm{K}}-1)} \leq M_{t,\mathrm{K}}. \tag{10}$$

Table 1. Notations used in ILP Formulations

$R_{t,\mathrm{K}}$	: Resource of type t made of transistors of dielectric K.
$M_{t,\mathrm{K}}$	: Maximum number of $R_{t,\mathrm{K}}$.
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.
$LDP(i,\mathrm{K})$	: Gate leakage delay product of $R_{t,\mathrm{K}}$ used by vertex v_i.
$X_{i,\mathrm{K},l,m}$	: Decision variable which takes the value of 1 if v_i is using $R_{t,\mathrm{K}}$ and scheduled in control steps $l \rightarrow m$.
$L_{i,\mathrm{K}}$	: Latency in number of cycles for v_i using $R_{t,\mathrm{K}}$.

6.2 Optimization Algorithm

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 (d_{FU}), the multiplexer (d_{Mux}), the register (d_{Reg}) for different dielectric technology. The resource constraints R_{con} are expressed as number of different resources made of transistors of each dielectric.

Algorithm 1 ILP Based Synthesis for LDP Optimization

1: Preprocess given behavioral description to construct a sequencing DFG.
2: Perform simulations to estimate gate leakage and delay of RTL units.
3: Construct resource allocation table and available resource table based on input resource constraints R_{con}.
4: Obtain ASAP and ALAP schedules of the input DFG.
5: Determine the number of different resources for each K using the resource allocation table.
6: Modify both ASAP and ALAP schedules obtained above using the number of resources found in previous step.
7: Construct the mobility graph based on above schedules.
8: Modify the mobility graph if needed to for multicycling.
9: Fix the total number of clock cycles as the maximum of modified ASAP and ALAP schedules' control step.
10: Model the ILP formulations of the DFG using AMPL.
11: Obtain the final solution by solving the ILP formulations using LP-Solve.
12: Postprocess scheduled sequencing DFG to generate gate leakage optimal RTL.

The optimization approach first obtains the sequencing DFG from the behavioral description. The as soon as possible (ASAP) time stamp and the as late as possible (ALAP) time stamp of each vertex for the DFG are obtained to limit the search space for the ILP solutions. The above schedules are then modified to accommodate the resource constraints. This will restrict the mobility of the vertices and further reduce the search space of ILP solutions. The scheduler uses the modeling language AMPL to model the ILP formulations [6]. At this step, the gate leakage, delay, and LDP are calculated using a look up table approach. The operational delay of a resource is assumed as ($d_{FU} + d_{Mux} + d_{Reg} + d_{Conv}$).

7 Register Transfer Level Module Library

A RTL module library is created following a three level hierarchy approach. The top level of hierarchy are the RTL components such as adders, subtractors, multipliers, etc. These in turn use logic gates which are derived from characteristics of CMOS devices.

A logic gate presents different dominant gate leakage paths, depending on the combination of inputs. The gate leakage current for a specific state of a logic gate is then calculated by summing the absolute gate currents over all the MOS devices in the logic gate, as both positive and negative gate current contributes to leakage:

$$I_{gate}\mathrm{Logic}_{\mathrm{state}} = \sum_{\forall \mathrm{MOS}_i} \left| I_{gate}\mathrm{MOS}[i] \right|, \tag{11}$$

where the index i identifies the device within a logic gate.

Let us assume that in a $n-$bit RTL 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 reasons: first it is an universal gate and is a low leaky logic gate compared to other gates [13, 9]. In this model the effect of

interconnect wires is not considered and focus is on the direct tunneling current and delay of the active units only. The *average* gate leakage of a n-bit RTL unit is calculated as:

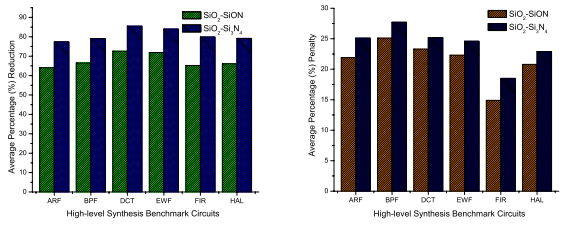
$$I_{gate\mathrm{R}} = \sum_{j=1}^{n_{total}} \text{Prob (state)} \times I_{gate\mathrm{NAND_j\,state}}. \quad (12)$$

The index j runs for all the NAND gates in a RTL unit. The Prob (state) is the probability of occurrence of an input state. The critical path delay of an n-bit RTL unit using the above NAND gates as building blocks is calculated as follows: $T_{pd_{\mathrm{R}}} = \sum_{i=1}^{n_{cp}} T_{pd\mathrm{NAND}_i}$. The average delay of a logic gate is calculated as: $T_{pd_{\mathrm{Logic}}} = \left(\frac{t_{HL}+t_{LH}}{2}\right)$, where t_{HL} and t_{LH} are the propagation delay times for high-to-low and low-to-high transitions, respectively. For a $45nm$ nano-CMOS technology, a RTL library is presented in [12].

8 Experimental Results

The overall design flow is implemented using C and integrated into the synthesis framework in [11]. The algorithm was experimented with various circuits. Two dual dielectric pair SiO_2- SiON and SiO_2- Si_3N_4 were considered. The base case for the experiments was the SiO_2 with a thickness of $T_{ox} = 1.4nm$ corresponding to the nominal case of BSIM4 [1]. For each circuit and each pair of dual-K, several sets of experiments were performed for various resource constraints similar to [10].

The percentage reduction in gate leakage and critical path delay penalty averaged over all resources is presented in Fig. 3. It is observed that reduction in gate leakage for all the benchmarks ranges from 46% to 83% for SiO_2-SiON and 56% to 92% for SiO_2-Si_3N_4 for different resource constraints considered in the experiments. The corresponding delay penalty are in the range of 8% to 37% for SiO_2-SiON and 11% to 38% for SiO_2-Si_3N_4. Moreover, as the number of available high-K resources increases, the reduction in gate leakage also increases. The benchmarks that had more number of operations needing high leaky resources created opportunity to be replaced with high-K resources resulting in more reduction. The reductions are more for larger benchmark circuits.



(a) Gate Leakage (b) Critical Path Delay

Figure 3. Average experimental results.

9 Summary, Conclusions and Future Works

This paper presents a new process driven technique called DKCMOS for reduction of gate leakage during RTL synthesis. The ILP based algorithm does scheduling and assignment for gate leakage reduction for different resource constraints. Experimental results reveal significant reductions in gate leakage with the use of this technology, thus proving its effectiveness. The proposed work has no architectural synthesis counterpart for fair comparison. Further exploration of this technique is the incorporation of process variation. The ultimate objective is to extend the work on gate leakage current to provide a broader solution to the problem of power dissipation in all its forms at the behavioral level. DKCMOS based designs may need more masks for the lithographic process of circuit fabrication. However, such costs would be compensated by the reduction of energy costs. The area overhead due to the use two separate islands (high-K and low-K) will also investigated in future.

References

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

[2] Meet the World's First 45nm Processor. `http://www.intel.com/technology/silicon/45nm_technology.htm`.

[3] H. Achatz. Extended 0/1 LP formulation for the scheduling problem in high-level synthesis. In *Proceedings of the European Design Automation Conference (EURO-DAC)*, pages 226–231, 1993.

[4] A. Agarwal, S. Mukhopadyaya, A. Roychowdhury, K. Roy, and C. H. Kim. Leakage Power Analysis and Reduction for Nanoscale Circuits. *IEEE Micro*, 26(2):68–80, March-April 2006.

[5] 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 the IEEE Custom Integrated Circuits Conf.*, pages 201–204, 2000.

[6] R. Fourer, D. Gay, and B. Kernighan. *AMPL: A Modeling Language for Mathematical Programming*. Thomson Brooks Cole, 2003.

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

[8] A. Karamcheti, et al. Silicon Oxynitride Films as Segue to the High-K Era. *Semiconductor Fabtech*, 12, 2000.

[9] K. S. Khouri & N. K. Jha. Leakage power analysis and reduction during behavioral synthesis. *IEEE Trans on VLSI Systems*, 10(6):876-885, Dec 2002.

[10] 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.

[11] S. P. Mohanty, N. Ranganathan, & S. K. Chappidi. ILP models for simultaneous energy and transient power minimization during behavioral synthesis. *ACM Trans on Design Auto. of Electronic Systems*, 11(1):186-212, Jan 2006.

[12] S. P. Mohanty, R. Velagapudi, and E. Kougianos. Dual-K Versus Dual-T Technique for Gate Leakage Reduction : A Comparative Perspective. In *Proc. of International Sympo. on Quality Electronic Design*, pages 564–569, 2006.

[13] V. Mukherjee, S. P. Mohanty, & E. Kougianos. A Dual Dielectric Approach for Performance Aware Gate Tunneling Reduction in Combinational Circuits. In *Proc of IEEE International Conf on Computer Design*, pp. 431-436, 2005.

[14] 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 Trans. on Very Large Scale Integration Systems*, 11(4):716–730, Aug 2003.

[15] K. Roy, S. Mukhopadhyay, and H. M. Meimand. Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits. *Proceedings of the IEEE*, 91(2):305–327, February 2003.

[16] 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*, pages 202–207, 2005.

[17] M. Yang and et al. Performance Dependence of CMOS on Silicon Substrate Orientation for Ultrathin and HfO_2 Gate Dielectrics. *IEEE Electron Device Letters*, 24(5):339–341, May 2003.