Modeling and Reduction of Gate Leakage during Behavioral Synthesis of NanoCMOS Circuits

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

Elias Kougianos
Department of Engineering Technology
University of North Texas, Denton, TX 76203.
Email: eliask@unt.edu

***Abstract*— For a nanoCMOS of sub-$65nm$ technology, where the gate oxide (SiO_2) thickness is very low, the gate leakage is one of the major components of power dissipation. In this paper, we provide analytical models to describe the tunneling current and propagation delay of behavioral level components considering various physical effects in the absence of foundry data. Subsequently, we explore the use of multiple oxide thickness resources as a technique for the reduction of gate leakage. In particular, we introduce a behavioral datapath scheduler that maximizes the utilization of higher gate oxide thickness resources. We characterize behavioral components for both $65nm$ and $45nm$ technologies in order to study the trend of tunneling current as technology scales, and provide them as inputs to the scheduler. We carried out extensive experiments for several benchmarks and observed significant reduction in gate leakage.**

I. Introduction

Several issues such as battery life, reliability, thermal considerations, and environmental concerns have driven the need for low power designs. With such aggressive technology scaling both static and dynamic power have become equally contributing factors for the total power dissipation of a CMOS circuit [1], [2]. In a short channel nanometer transistor, several forms of leakage exist, such as reverse biased diode leakage, subthreshold leakage, gate tunneling current, hot carrier gate current, gate induced drain leakage and channel punch through current [3]. Of all these leakage mechanisms, SiO_2 tunneling current that flows during both active and sleep modes of the circuit is a significant component for low-end nanoCMOS technology (i. e. sub-$65nm$) using ultra-thin gate oxide [4], [2], [5]. Thus, the major sources of power dissipation in a nano-CMOS circuit can be summarized as [5], [6], [7]:

$$P_{total} = P_{dynamic} + P_{short} + P_{static} + P_{oxide} \quad (1)$$

Power reduction in general can be achieved at various levels of design abstraction, such as system, architectural, logic and transistor level. Dynamic power management (DPM) techniques, dynamic voltage (frequency) scaling (DVS), and clock gating are popular system level methods [8]. Similarly, multiple voltage (Multi-V_{dd}) techniques have been explored for behavioral level dynamic power minimization [9], [10]. Moreover, multiple threshold (Multi-V_{Th}) options have been proposed for reduction of subthreshold current [11], [12]. Recently, a Dual-T_{ox} (gate oxide thickness) method is proposed as a transistor level method for tunneling current reduction [4], [13]. Moreover, transistor sizing has been used as attractive option for power reduction [13], [14]. In this paper we propose to use resources of multiple gate oxide thicknesses (Multi-T_{ox}) during behavioral synthesis for reduction of gate leakage.

II. Related and The Proposed Research

Low power behavioral synthesis research works have mostly considered dynamic power reduction and few of them have dealt with leakage. At the same time, few logic and transistor level research works focus on reduction of gate leakage.

In [11], Khouri and Jha have proposed Dual-V_{Th} techniques for subthreshold leakage analysis and reduction during behavioral synthesis. Gopalakrishnan and Katkoori [15] also use Multi-V_{Th} approach for reduction of subthreshold current during high-level synthesis. Mohanty et. al. [5] have introduced models and a datapath scheduling algorithm for reduction of tunneling current.

In [16], Lee et al., developed a method for analyzing gate oxide leakage current in logic gates and suggested pin reordering to reduce it. Sultania et al., in [4], developed an algorithm to optimize the total leakage power by assigning dual T_{ox} values to transistors. In [13], [17], Sirisantana and Roy use multiple channel lengths and multiple gate oxide thickness for reduction of leakage.

Contributions of this Paper: The contributions of this paper are two fold. First, we develop models for direct tunneling current and propagation delay calculations of functional units. Subsequently, we assume that such functional units are made available as standard cells and introduce an algorithm for scheduling of the datapath operations such that overall tunneling current of a datapath circuit is minimal. We assume that all transistors used in a functional unit (such as adder, subtractor, etc.) have oxide of equal thickness, but the thicknesses of different functional units may differ. The functional unit using higher oxide thickness transistors dissipates less tunneling power, but has larger delay. We may use such a functional unit in the off-critical path of a circuit, to achieve the conflicting objective of power reduction and maintaining performance. On the other hand, a functional unit which uses lower oxide thickness transistors exhibits less delay and is suitable to be utilized in the critical path of a circuit. As the oxide thickness we are dealing with is very low it may not remain constant during the fabrication, hence our algorithm takes process variation into consideration.

III. First Principle Analytical Models

In the absence of foundry rules we need models to characterize the behavioral components for design space exploration. Such models bridge the architectural level abstraction with transistor level and help in quicker decision making at behavioral level before laying out the design in silicon. We use a top-down design synthesis with three level hierarchy to form the models. At the top level of hierarchy we have behavioral components such as adders, subtractors, multipliers, etc. They in turn make use of logic level components which are derived from a set of equations available for various transistor characteristics. The models are developed from first principles using standard notations [5] considering various physical effects. Finally, we express the tunneling current and propagation delay of each behavioral unit in terms of gate oxide thickness in order to facilitate the behavioral synthesis.

In the hierarchical modeling, we assume that datapath units are constructed using universal NAND logic gates as they exhibit minimal tunneling current compared to other logic gates [18]. Let us assume that there are total n_{total} NAND gates in the network of NAND gates constituting a n-bit functional unit out of which n_{cp} are in the critical path. In this model we do not consider the effect of interconnects and focus on the gate leakage and propagation delay of the functional units only. It may be noted that this assumption does not affect the tunneling current values as oxide tunneling happens only in the transistors not in the interconnects.

A. Tunneling Current (Gate Leakage)

The gate tunneling mechanism in a CMOS can be either Fowler-Nordheim (FN) tunneling or direct tunneling (DT); both differ in the form of potential barrier [3]. We consider the tunneling to be direct with trapezoidal potential barrier, which is the case for sub-$65nm$ technology [19].

We calculate the tunneling current of a n-bit unit as:

$$I_{DT}\text{FU} = \sum_{j=1}^{n_{total}} Pr_j \sum_{\text{MOS}_i \in \text{ NAND}_j} Pr_i \; I_{DTi}. \quad (2)$$

In the above, Pr_j is the probability that input of the NAND gate is at logic "0", which can be obtained by logic level estimations. The contributions of the NMOS and PMOS tunneling depend on the probability of the input signal being at logic "1" and "0", respectively. The average tunneling current for a NAND logic gate is calculated as [4]:

$$I_{DT}\text{NAND} = \sum_{\text{MOS}_i \in \text{ NAND}} Pr_i \; I_{DTi}, \quad (3)$$

where Pr_i is the probability that inputs of the MOS that are connected in parallel i.e. PMOS are at logic "0".

For direct tunnelling the tunneling probability of an electron is affected by barrier height, structure and thickness and for a MOS it is expressed by Eqn. 4 [20], [3], [21]:

$$I_{DT} = \frac{WL\; q^3 V_{ox}^2}{16\pi^2 \hbar \phi_B T_{ox}^2}\; exp\left[-\frac{4\sqrt{2m_{eff}}\;{\phi_B}^{1.5} T_{ox}}{3\hbar q V_{ox}} \left\{1-\left(1-\frac{V_{ox}}{\phi_B}\right)^{1.5}\right\}\right]. \quad (4)$$

The voltage across the MOS gate dielectic V_{ox} is expressed as, $V_{ox} = (V_{gs} - V_{fb} - \psi_S - V_{poly})$ [22], [3]. The voltage across the polysilicon depletion region V_{poly} is expressed as $V_{poly} = \left(\frac{{\epsilon^2}_{ox}\; V_{ox}^2}{2q\;\epsilon_{Si}\;N_{poly}T_{ox}^2}\right)$ [3]. From these two equations we obtain a quadratic equation in terms of the variable V_{ox}, which is solved to the following:

$$V_{ox} = \frac{\sqrt{1-2(V_{fb}+\psi_S-V_{gs})\left(\frac{{\epsilon^2}_{ox}}{q\;\epsilon_{Si}\;N_{poly}T_{ox}^2}\right)}-1}{\left(\frac{{\epsilon^2}_{ox}}{q\;\epsilon_{Si}\;N_{poly}T_{ox}^2}\right)}. \quad (5)$$

The flat-band voltage V_{fb} can be derived from MOS capacitance-voltage (C-V) characteristics or using the expression $\left(\frac{qN_{channel}T_{ox}^2}{2\epsilon_{Si}}\right)$. It may be noted that the effective values of W, L, may be different from original values due to depletion and need to be taken into consideration [7], [23]. The effective gate oxide thickness T_{ox} is a quadratic function of the physical oxide thickness, T_{oxp} [7], [23]. After solving the quadratic equation and taking polysilicon depletion into consideration we obtain the following expression:

$$T_{ox} = 0.5T_{oxp}\left(1+\sqrt{1+4\frac{\epsilon_{ox}}{\epsilon_{Si}}\frac{X_{poly}}{T_{oxp}}}\right). \quad (6)$$

The polysilicon depletion depth is calculated as [7]:

$$X_{poly} = \frac{\epsilon_{ox}}{\epsilon_{Si}}T_{oxp}\left(\sqrt{1+\frac{2{\epsilon^2}_{ox}(V_{gs}-V_{fb}-\psi_S)}{q\epsilon_{Si}N_{poly}{T^2}_{oxp}}}-1\right). \quad (7)$$

The Fermi-level ϕ_F is calculated as $\left[\frac{2kT}{q}\; ln\left(\frac{N_{channel}}{n_i}\right)\right]$; for strong inversion surface potential ψ_S is $2\phi_F$ [20], [24], [25].

B. Propagation Delay

The critical path delay of a n-bit functional unit using NAND gates as building blocks can be calculated as:

$$T_{pd}\text{FU} = \sum_{i=1}^{n_{cp}} 0.5\left(n_{fan-in}T_{pd}\text{NMOS} + T_{pd}\text{PMOS}\right). \quad (8)$$

The effective fan-in factor is calculated for short channel devices with velocity saturation and strong inversion [6], [26]:

$$n_{fan-in} = 1+\left\{\frac{\left(2-\sqrt{2}\right)(n_{series}-1)V_{dsSat}}{V_{dd}+V_{Th}-0.5V_{dsSat}}\right\}\left(1+\frac{T_{ox}}{\epsilon_{ox}}\sqrt{\frac{qN_{channel}\epsilon_{Si}}{2\psi_S}}\right), \quad (9)$$

where n_{series} is the number of series connected MOS devices.

We use α-power and physical-α-power models to compute the propagation delay (T_{pd}) of a MOS as [7], [27], [28],

$$T_{pd} = \frac{0.5C_LV_{dd}}{I_{DSat0}} + T_T\left\{\frac{0.5-\left(\frac{V_{dd}-V_{Th}}{V_{dd}}\right)}{\alpha+1}\right\}. \quad (10)$$

Here, I_{DSat0} is the saturation drain current of the MOS for $V_{gs} = V_{dd}$. The saturation drain current is given by [7]:

$$I_{DSat} = \frac{W}{L}\left(\frac{V_{gs}-V_{Th}}{V_{dd}-V_{Th}}\right)^{\alpha}\left[\frac{\mu_0 C_{ox}V_{dsSat0}(V_{dd}-V_{Th}-0.5\eta V_{dsSat0})}{\{1+\theta(V_{gs}-V_{Th})\}\left\{1+\frac{\mu_0 V_{dsSat}}{v_{sat}L(1+\theta(V_{gs}-V_{Th}))}\right\}}\right]. \quad (11)$$

The zero bias mobility can be calculated as $\mu_0 = \frac{\mu_{sub}}{\left\{1+\left(\frac{Q_B\mu_{sub}}{\epsilon_{ox}v_{norm}}\right)\right\}}$, where the depletion charge density Q_B is

calculated as $\sqrt{2q\epsilon_{Si}N_{sub}\psi_S}$ [29], [25]. The transition time model is given in Eqn. 12 [7],

$$T_T = \frac{C_L V_{dd}}{I_{D\,Sat0}}\left[\frac{0.9}{0.8}+\frac{V_{ds\,Sat0}}{0.8V_{dd}}\left\{\frac{V_{dd}-V_{Th}-0.5\eta V_{ds\,Sat0}}{V_{dd}-V_{Th}} ln\left(\frac{10V_{ds\,Sat0}(V_{dd}-V_{Th})}{V_{dd}(V_{dd}-V_{Th}-0.5\eta V_{ds\,Sat0})}\right)-1\right\}\right]. \quad (12)$$

The constant modeling carrier saturation velocity is [7], [28]:

$$\alpha = \frac{1}{ln(2)} ln\left\{\frac{2V_{ds\,Sat0}(V_{dd}-V_{Th}-0.5\eta V_{ds\,Sat0})}{V_{ds\,Sata}(V_{dd}-V_{Th}-\eta V_{ds\,Sata})}\right\}. \quad (13)$$

Here, $V_{ds\,Sat0}$ and $V_{ds\,Sata}$ are the saturation drain voltages for $V_{gs}=V_{dd}$ and $V_{gs}=\left(\frac{V_{dd}+V_{Th}}{2}\right)$, respectively. The saturation drain voltage $V_{ds\,Sat}$ is given by the following [7], [28]:

$$V_{ds\,Sat} = \frac{v_{sat}L}{\mu_0}\{1+\theta(V_{gs}-V_{Th})\}\left[\sqrt{1+\frac{2\mu_0(V_{gs}-V_{Th})}{v_{sat}L\eta\{1+\theta(V_{gs}-V_{Th})\}}}-1\right]. \quad (14)$$

The effective V_{Th} in all of the above equations is [3]:

$$V_{Th} = V_{fb}+\frac{2kT}{q}ln\left(\frac{N_{sub}}{n_i}\right)+\frac{1}{C_{ox}}\sqrt{2q\epsilon_{Si}N_{sub}\left\{\frac{2kT}{q}ln\left(\frac{N_{sub}}{n_i}\right)+V_{bs}\right\}}, \quad (15)$$

where the effective oxide thickness for the C_{ox} calculation is performed assuming strong inversion. The mobility degradation factor θ is computed as $\left(\frac{\mu_0}{2T_{ox}v_{norm}}\right)$, where T_{ox} is calculated using Eqn. 6. The subthreshold slope factor η is calculated as $\left[1+\sqrt{\frac{q\epsilon_{Si}N_{channel}T_{ox}^2}{2\epsilon_{ox}^2(\psi_S-V_{bs})}}\right]$ [7], [28].

C. Function Fitting for Characterization

We consider the functional units of 16-bit size whose structural information is obtained from [30]. We used the parameters from BSIM4 models [31] and also from [25], [24] with appropriate units. It is assumed that the probability of logic "1" and logic "0" is the same. For a given length L, the width of the transistors is chosen as $W_{NMOS}=4L$, $W_{PMOS}=8L$ to ensure smooth current flow between NMOS and PMOS. While changing the oxide thickness the channel length of the transistor is changed proportionately to avoid impact on its functionality [4]. The plots in Fig. 1 confirm that there is decrease in the tunneling current and increase in the propagation delay as the oxide thickness increases. It is also observed that there is increase in the tunneling current as technology scales from $65nm$ to $45nm$, which is consistent with the ITRS prediction trend [19]. We present the tunneling current and propagation delay of various units as functions of gate oxide thickness in Table I.

IV. Synthesis Methodology

There are several steps involved in behavioral synthesis, such as compilation, transformation, datapath scheduling, functional unit allocation, operation binding, connection allocation and architecture generation. Scheduling and binding are the major phases of low-power behavioral synthesis. We assume that the target architecture datapath is specified as a sequencing data flow graph (DFG). Each vertex of the DFG represents an operation and each edge represents a dependency. The DFG does not support the hierarchical entities and the conditional statements are handled using comparison operation. Also, each vertex has attributes that specify the operation type. The delay of a control step is dependent on the delays of the functional unit, the multiplexer, and register.

The proposed behavioral scheduler when used along with a leakage-delay estimator generates a circuit which dissipates minimal gate leakage. The estimator uses analytical models introduced in the previous section and calculates the values for different functional units. It also calculates total gate leakage and critical path delay of a circuit represented as a DFG. The combined reduction of gate leakage and critical delay translates to reduction of the tunneling current-delay-product, which is the objective of the scheduler for minimization. Assuming N_c- number of control steps and $n_{FU\,c}-$ number of resources active in any control step c, the tunneling current-delay-product can be calculated as,

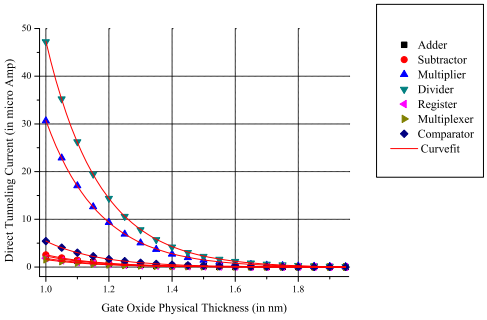
$$CDP = \sum_{c=1}^{N_c}\sum_{r=1}^{n_{FU\,c}} I_{DT\,FU}(c,r) * T_{pd\,FU}(c,r). \quad (16)$$

Here, $I_{DT\,FU}(c,r)$ is tunneling current of the r-th functional unit active in step c with delay $T_{pd\,FU}(c,r)$.

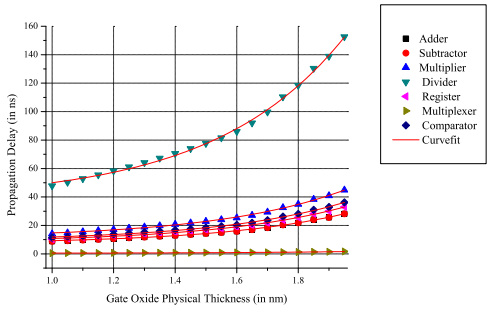
We assume that all the transistors inside a resource have same oxide thickness, which may differ for different resources. However, to take the process variation into account we assume that a given oxide thickness T_{oxp} can take any value in the range $(T_{oxp}-\Delta T_{oxp}, T_{oxp}+\Delta T_{oxp})$. We assume such variation to be Gaussian [32]. It may be noted that as we maintain constant $\left(\frac{L}{T_{ox}}\right)$ ratio and constant $\left(\frac{W}{L}\right)$ ratio, all three process parameters T_{ox}, L, and W have Gaussian variation.

The scheduler algorithm heuristic is presented in Fig. 2, which is developed based on the datapath scheduler in [10]. The inputs to the behavioral scheduler are an unscheduled DFG, the resource constraints that include a number of different resources made of transistors of different oxide thickness. The scheduler time stamps the operations such that more low oxide thickness resources are active in the critical path and more high oxide thickness resources are active in the off-critical path of the datapath circuit. The scheduler attempts to assign higher intrinsic leakage functional units (such as multiplier and divider) with higher oxide thickness. This is in accordance with our conclusions from the analytical model where it is observed that multiplier and divider units dissipate much more tunneling current compared to adder and subtractor units. At the same time it is observed that adder and subtractor units have less delay compared to the multiplier and divider. Thus, the scheduler attempts to operate the higher intrinsic leakage units of the highest thickness to reduce the tunneling and at the same time lower intrinsic leakage units of lowest thickness to compensate the delay increase.

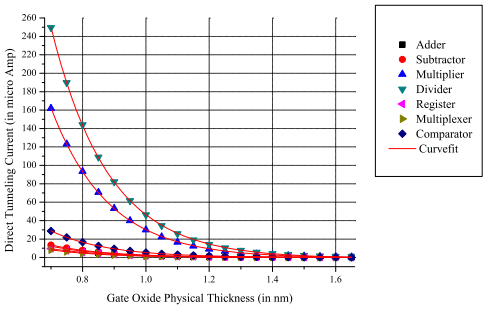
The scheduler performs assignment for all potential off-critical paths and calculates CDP for each assignment for the DFG using Eqn. 16. Once the minimum CDP is obtained a particular vertex is time stamped and the T_{oxp} assignment is accepted. The predecessor and successor time stamps are adjusted accordingly to maintain the data dependency. Gaussian distributed random numbers are generated to take into account



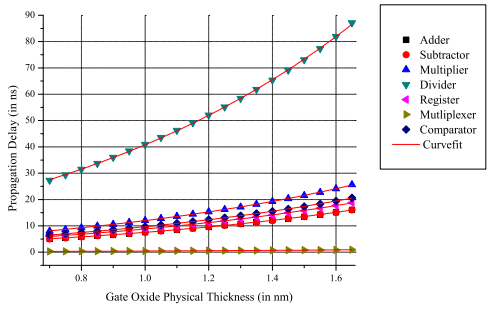
(a) Tunneling Current Versus T_{oxp} for $65nm$ Technology



(b) Propagation Delay Versus T_{oxp} for $65nm$ Technology



(c) Tunneling Current Versus T_{oxp} for $45nm$ Technology



(d) Propagation Delay Versus T_{oxp} for $45nm$ Technology

Fig. 1. Tunneling Current and Propagation Delay Versus Oxide Physical Thickness for $65nm$ and $45nm$ Technology

TABLE I

ANALYTICAL FUNCTIONS TERMS OF T_{oxp} TO BE USED FOR MULTI-T_{ox} BASED LOW POWER BEHAVIORAL SYNTHESIS

	Tunneling Current $I_{DT\text{FU}}$ in μA: $f(T_{oxp}) = Ae\left(-\frac{T_{oxp}}{\alpha}\right) + B$						Propagation Delay $T_{pd\text{FU}}$ in ns: $g(T_{oxp}) = Ae\left(\frac{T_{oxp}}{\alpha}\right) + B$					
	65nm Technology			45nm Technology			65nm Technology			45nm Technology		
	α	A	B	α	A	B	α	A	B	α	A	B
Adder	0.16877	8.64×10^{2}	-7.54×10^{-3}	1.8029	5.93×10^{2}	-5.39×10^{-2}	0.42445	0.21	6.98	1.05039	3.81	-2.33
Subtractor	0.16877	9.66×10^{2}	-8.43×10^{-3}	1.8029	6.63×10^{2}	-6.02×10^{-2}	0.42445	0.21	6.98	1.05039	3.81	-2.33
Multiplier	0.16877	1.15×10^{4}	-1.00×10^{-1}	1.8029	7.92×10^{3}	-7.19×10^{-1}	0.42445	0.34	11.10	1.05039	6.07	-3.71
Divider	0.16877	1.78×10^{4}	-1.55×10^{-1}	1.8029	1.22×10^{4}	$-1.11\times10^{+0}$	0.42445	1.16	37.8	1.05039	20.60	-12.61
Comparator	0.16877	2.05×10^{3}	-1.79×10^{-2}	1.8029	1.41×10^{3}	-1.28×10^{-1}	0.42445	0.28	8.96	1.05039	4.89	-2.99
Register	0.16877	6.86×10^{2}	-5.99×10^{-3}	1.8029	4.71×10^{2}	-4.28×10^{-2}	0.42445	0.25	8.17	1.05039	4.46	-2.73
Multiplexer	0.16877	5.84×10^{2}	-5.09×10^{-3}	1.8029	4.01×10^{2}	-3.64×10^{-2}	0.42445	0.01	0.40	1.05039	0.22	-0.13

the effect of process variation on T_{oxp}; the values are generated in the range $(T_{oxp} - \Delta T_{oxp}, T_{oxp} + \Delta T_{oxp})$. The algorithm picks any one value in that range to replace T_{oxp} under consideration. The algorithm in the final step scans through every clock cycle and finds all the scheduled vertices in each. For a particular type of operation if the critical vertex has higher T_{oxp} than an off-critical vertex then the values of T_{oxp} are swapped between them. This step further compensates the performance degradation due to the use of high leakage resources with higher T_{oxp}. The above described algorithm can be easily used to handle various types of datapath operations, such as multicycling, chaining, and pipelining.

V. EXPERIMENTAL RESULTS

The algorithm was implemented for experiments in the behavioral synthesis framework in [10] and tested with several benchmark circuits for several constraints. We present the results in this section for selected benchmarks and constraints.

First we carried out our experiments using resources of two different gate oxide thicknesses. For both $65nm$ and $45nm$ technology we have chosen two different oxide thickness in which higher thickness is 35% more than the lower thickness. A selected set of resource constraints is given in Table II. These are representatives of various forms of the corresponding RTL representation. We have not shown the number of dividers or comparators in the table as there was only one benchmark (HAL) that needed them.

The experimental results are presented in Table III and Fig. 3. The quantities with ST subscript represent results for single thickness and MT subscript represent results for the multiple oxide thickness case. We assume the minimal oxide thickness case with T_{oxp} of $1.0nm$ as the base ST case. The value of

Input : UDFG, Resource Constraints, Analytic Functions for $I_{DT\,FU}$ and $T_{pd\,FU}$, Number of T_{ox}
Output : Scheduled DFG, Tunneling Current and Delay Estimates, Number of Clock Cycles

---

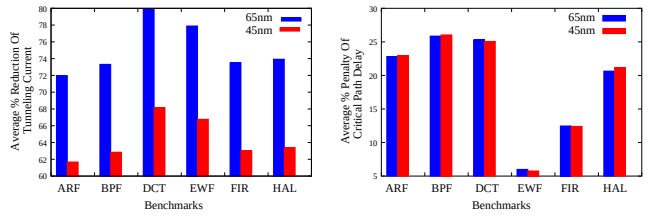
Find total number of FUs of all available oxide thicknesses from the DFG : $G(V, E)$
Get resource constrained as soon as possible schedule S_{ASAP} and as late as possible schedule S_{ALAP}.
Fix the total number of clock cycles as the maximum of S_{ASAP} and S_{ALAP} steps.
Find the vertices in critical path V_c and off-critical path V_{oc} (where, both V_c and $V_{oc} \in V$).
Assume the above S_{ASAP} schedule as the current schedule S_i.
For each $v \in V_c$ assign highest thickness T_{oxp_H} to operations needing high-leaky resources and lowest thickness T_{oxp_L} to operations needing low-leaky resources.
While all $v \in V_{oc}$ of the current schedule S_i are not considerd for time stamping {
    If vertex v is needs a high-leaky resource then assign the highest available thickness T_{oxp_H}.
    Else assign the highest available thickness T_{oxp_L}.
    Generate Gaussian random numbers in the range of $(T_{oxp} - \Delta T_{oxp},\ T_{oxp} + \Delta T_{oxp})$ to take process variation into account.
    Calculate the current delay product of the current schedule CDP_{S_i} for one value from
        the range $(T_{oxp} - \Delta T_{oxp},\ T_{oxp} + \Delta T_{oxp})$ using the analytical functions f and g from Table I.
    For each off-critical vertex V_{oc} (i. e. $v \in V_{oc}$) of the current schedule S_i {
        For every allowable control step c in the mobility range of v {
            Assign next higher thickness if vertex needs high leaky resource and next lower thickness if vertex needs low leaky resource.
            Generate Gaussian random numbers in the range of $(T_{oxp} - \Delta T_{oxp},\ T_{oxp} + \Delta T_{oxp})$.
            Find CDP of the DFG for each case for a values from $(T_{oxp} - \Delta T_{oxp},\ T_{oxp} + \Delta T_{oxp})$. } End For } End For
    Fix time stamp of the vertex with the current T_{oxp} assignment for which CDP is minimum.
    Remove the above time stamped vertex from V_{oc}. } End While
Find all vertices scheduled in each clock cycle. For a particular type of operation in a clock cycle, if critical vertex has higher T_{oxp} than off-critical then swap T_{oxp}.
Calculate gate leakage and delay for the scheduled DFG.

Fig. 2. Heuristic based Mult-T_{ox} Behavioral Scheduling Algorithm

TABLE II
A SELECTED LIST OF RESOURCE CONSTRAINTS USED TO PERFORM OUR EXPERIMENTS

Number of Resources for Various T_{oxp}												Resource Constraint No.
$65nm$ Technology						$45nm$ Technology						
Multiplier		Adder		Subtractor		Multiplier		Adder		Subtractor		
$1.35nm$	$1.0nm$	$1.35nm$	$1.0nm$	$1.35nm$	$1.0nm$	$0.95nm$	$0.7nm$	$0.95nm$	$0.7nm$	$0.95nm$	$0.7nm$	
1	1	2	0	2	0	1	1	2	0	2	0	1
2	1	1	1	1	1	2	1	1	1	1	1	2
2	0	0	2	0	2	2	0	0	2	0	2	3
3	0	1	1	1	1	3	0	1	1	1	1	4

ΔT_{oxp} is assumed to be 10% of the original T_{oxp}. We estimate the critical path delay of the circuit as the sum of the delays of the vertices in the longest path of the data flow graph.



(a) Tunneling Current (b) Critical Path Delay

Fig. 3. Average % Results for Various High-Level Synthesis Benchmarks.

We observe that for $65nm$ technology, the reduction in tunneling current (ΔI_{DT}) is in the range of $51.66\% - 87.87\%$ with an overall average of 75.08%. The average reduction for each benchmark is very consistent in the range of $71.96\% - 79.85\%$. It can be seen that the reduction in tunneling current is maximum for the DCT and EWF benchmarks, and minimum for ARF benchmark. The delay penalty (ΔT_{pd}) is found to be in the range of $6.0 - 25.86\%$ with an average overall average of 18.83%. The results for the $45nm$ technology are similar to that of $65nm$. The reduction in the tunneling current is decreased by approximately $10 - 12\%$ and the average delay penalty remains approximately same.

We also carried out experiments using functional units

TABLE III
EXPERIMENTAL RESULTS FOR $65nm$ TECHNOLOGY

Bench-marks	Res Con	I_{DT} in μA			T_{pd} in ns		
		ST	MT	ΔI_{DT}	ST	MT	ΔT_{pd}
ARF	1	521.19	251.93	51.66	142.14	189.96	33.63
	2	521.19	161.81	68.95	142.14	167.07	17.53
	3	521.19	89.78	82.77	142.14	174.18	22.53
	4	521.19	80.92	84.47	142.14	167.07	17.53
		Average ΔI_{DT}		71.96	Average ΔT_{pd}		22.81
BPF	1	411.07	157.69	61.63	127.93	169.96	32.85
	2	411.07	123.71	69.90	127.93	159.96	25.04
	3	411.07	87.51	78.71	127.93	154.18	20.52
	4	411.07	69.78	83.02	127.93	159.96	25.04
		Average ΔI_{DT}		73.32	Average ΔT_{pd}		25.86
DCT	1	472.02	84.19	39.43	213.21	269.94	26.60
	2	472.02	84.19	82.16	213.21	269.94	26.60
	3	472.02	121.49	74.25	213.21	261.27	22.53
	4	472.02	90.47	80.83	213.21	267.05	25.24
		Average ΔI_{DT}		79.85	Average ΔT_{pd}		25.29
EWF	1	311.03	37.71	87.87	227.43	250.85	10.30
	2	311.03	70.95	77.18	227.43	239.94	5.50
	3	311.03	95.33	69.35	227.43	233.57	2.70
	4	311.03	70.95	77.18	227.43	239.93	5.50
		Average ΔI_{DT}		77.89	Average ΔT_{pd}		6.00
FIR	1	283.29	115.23	59.32	156.35	183.24	17.20
	2	283.29	58.72	79.27	156.35	180.07	15.17
	3	283.29	67.59	76.14	156.35	159.96	2.30
	4	283.29	58.72	79.27	156.35	180.07	15.17
		Average ΔI_{DT}		73.50	Average ΔT_{pd}		12.46
HAL	1	196.71	77.77	60.46	56.85	79.98	40.67
	2	196.71	59.67	69.66	56.85	67.09	18.00
	3	196.71	34.93	82.24	56.85	67.09	18.00
	4	196.71	32.71	83.36	56.85	60.19	5.88
		Average ΔI_{DT}		73.93	Average ΔT_{pd}		20.64
Overall		Average ΔI_{DT}		75.08	Average ΔT_{pd}		18.83

of three different gate oxide thicknesses. In this scenario for different benchmark circuits the maximum reduction was improved in the range of $3-7\%$ and the average reduction was improved by $2-5\%$. But, there is increase in the average delay penalty for different benchmark circuits in the range of $5-11\%$. This is observed for both $65nm$ and $45nm$ technology.

VI. Conclusions

In this paper we presented a novel technique of Multi-T_{ox} functional units as an attractive option for overall gate leakage reduction of a datapath circuit. However, Multi-T_{ox} based designs may need more masks for the lithographic process of circuit fabrication. We believe such costs would be compensated by the reduction of energy costs. We also presented a comparative view of $65nm$ and $45nm$ technology. The resource selection is being made during scheduling using a heuristic based approach. We are anticipating that use of better optimization techniques may further improve the results. We can also incorporate methods to accurately estimate the logic values for more accurate modeling. We have considered the synthesis of datapath circuits, however the work in principle can be extended to control synthesis. Finally, it is our belief that the proposed Multi-T_{ox} approach can be used along with Multi-V_{dd} and Multi-V_{Th} approaches to provide a solution for total power dissipation of CMOS circuits.

References

[1] D. Sylvester and H. Kaul, "Power-Driven Challanges in Nanometer Design," *IEEE Design and Test of Computers*, vol. 13, no. 6, pp. 12–21, Nov-Dec 2001.

[2] N. S. Kim, T. Austin, D. Blaauw, T. Mudge, K. Flautner, J. S. Hu, M. J. Irwin, M. Kandemir, and N. Vijaykrishnan, "Leakage Current - Moore's Law Meets Static Power," *IEEE Computer*, pp. 68–75, December 2003.

[3] 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.

[4] A. K. Sultania, D. Sylvester, and S. S. Sapatnekar, "Tradeoffs Between Gate Oxide Leakage and Delay for Dual T_{ox} Circuits," in *Proceedings of Design Automation Conference*, 2004, pp. 761–766.

[5] S. P. Mohanty, V. Mukherjee, and R. Velagapudi, "Analytical Modeling and Reduction of Direct Tunneling Current during Behavioral Synthesis of Nanometer CMOS Circuits," in *Proceedings of the 14th ACM/IEEE International Workshop on Logic and Synthesis (IWLS)*, 2005, pp. 249–256.

[6] A. J. Bhavnagarwala, B. L. Austin, K. A. Bowman, and J. D. Meindl, "A Minimum Total Power Methodology for Projecting Limits of CMOS GSI," *IEEE Transactions on VLSI Systems*, vol. 8, no. 3, pp. 235–251, June 2000.

[7] K. A. Bowman, L. Wang, X. Tang, and J. D. Meindl, "A Circuit-Level Perspective of the Optimum Gate Oxide Thickness," *IEEE Transactions on Electron Devices*, vol. 48, no. 8, pp. 1800–1810, August 2001.

[8] L. Benini, A. Bogliolo, and G. De Micheli, "A Survey of Design Techniques for System-level Dynamic Power Management," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 8, no. 3, pp. 299–316, June 2000.

[9] W. T. Shiue and C. Chakrabarti, "Low-Power Scheduling with Resources Operating at Multiple Voltages," *IEEE Transactions on Circuits and Systems-II : Analog and Digital Signal Processing*, vol. 47, no. 6, pp. 536–543, June 2000.

[10] 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.

[11] 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.

[12] R. M. Rao, J. L. Burns, and R. B. Brown, "Circuit Techniques for Gate and Sub-Threshold Leakage Minimization in Future CMOS Technologies," in *European Solid-State Circuits Conference*, 2003, pp. 313–316.

[13] 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.

[14] P. Pant, V. K. De, and A. Chatterjee, "Simultaneous power supply, threshold voltage, and transistor size optimization for low-power operation of CMOS circuits," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 6, no. 4, pp. 538–545, Dec 1998.

[15] 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.

[16] D. Lee, D. Blaauw, and D. Sylvester, "Gate Oxide Leakage Current Analysis and Reduction for VLSI Circuits," *IEEE Transactions on VLSI Systems*, vol. 12, no. 2, pp. 155–166, February 2004.

[17] N. Sirisantana, L. Wei, and K. Roy, "High-Performance Low-Power CMOS Circuits using Multiple Channel Length and Multiple Oxide Thickness," in *Proceedings of the IEEE International Conference on Computer Design*, 2000, pp. 227–232.

[18] 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.

[19] "Semiconductor Industry Association, International Technology Roadmap for Semiconductors," `http://public.itrs.net`.

[20] 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.

[21] 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.

[22] E. M. Vogel, K. Z. Ahmed, B. Hornung, P. K. McLarty, G. Lucovsky, J. R. Hauser, and J. J. Wortman, "Modeled Tunnel Currents for High Dielectric Constant Dielectrics," *IEEE Transactions on Electron Devices*, vol. 45, no. 6, pp. 1350–1355, June 1998.

[23] B. Yu, D. H. Ju, W. C. Lee, N. Kepler, T. J. King, and C. Hu, "Gate Engineering for Deep-Submicron CMOS Transistors," *IEEE Transactions on Electron Devices*, vol. 45, no. 6, pp. 1253–1262, June 1998.

[24] S. M. Sze, *Semiconductor Devices : Physics and Technology*, John Willey, 2002.

[25] S. M. Sze, *Pyhsics of Semiconductor Devices*, John Wiley, 1981.

[26] A. J. Bhavnagarwala B. Ausin and J. D. Meindle, "Minimum Supply Voltage for Bulk Si CMOS GSI," in *Proceedings of International Symposium on Low Power Electronic Design*, 1998, pp. 100–102.

[27] T. Sakurai and A. R. Newton, "Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and Other Formulas," *IEEE Journal of Solid-State Circuits*, vol. 25, no. 2, pp. 584–594, April 1990.

[28] K. A. Bowman, B. L. Austin, J. C. Eble, X. Tang, and J. D. Meindl, "A Physical Alpha-Power Law MOSFET Model," *IEEE Journal of Solid-State Circuits*, vol. 34, no. 10, pp. 1410–1414, October 1999.

[29] S. L. Garverick and C. G. Sodini, "A Simple Model for Scaled MOS Transistor that Includes Field-Dependent Mobility," *IEEE Journal of Solid-State Circuits*, vol. 22, no. 1, pp. 111–114, February 1987.

[30] N. H. E. Weste and D. Harris, *CMOS VLSI Design : A Circuit and Systems Perspective*, Addison Wesley, 2005.

[31] 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.

[32] R. Rao, A. Srivastava, D. Blaauw, and D. Sylvester, "Statistical Analysis of Subthreshold Leakage Current for VLSI Circuits," *EEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 12, no. 2, pp. 131–139, Feb 2004.