Analytical Modeling and Reduction of Direct Tunneling Current during Behavioral Synthesis of Nanometer CMOS Circuits

Saraju P. Mohanty, Valmiki Mukherjee, and Ramakrishna Velagapudi
Department of Computer Science and Engineering
University of North Texas, Denton, TX 76203.
Email : {smohanty,vm0058,rv0063}@unt.edu

Abstract

Gate oxide direct tunneling current is the major component of static power dissipation of a CMOS circuit for low-end technology, where the gate dielectric (SiO$_2$) thickness is very low. This paper presents a novel direct tunneling current reduction method during behavioral synthesis of nanometer CMOS circuits. We provide analytical models to calculate the direct tunneling current and the propagation delay of behavioral level components. We then characterize those components for various gate oxide thicknesses. We also provide an algorithm for behavioral scheduling for minimizing the overall tunneling current dissipation of datapath circuits. The algorithm explores dual oxide thickness option for reducing direct tunneling current. We have carried out extensive experiments for various behavioral level benchmarks under various resource constraints and observed significant reductions in tunneling current.

1 Introduction

There has been a significant increase in the demand for low power and high performance digital VLSI circuits. The designers are implementing very high-order scaling of both device dimensions and the supply voltage. As a result, there has been a drastic change in the leakage components of the device both in the inactive as well as active modes of operation. The dynamic power consumption remains almost unchanged, but the leakage power dissipation increases significantly and becomes a major contributor to the total power dissipation as the technology changes [1]. This calls for a greater need for a considerable reduction in leakage power, which continues to dissipate even when a device is not doing any useful job. The leakage current in short channel nanometer transistor has diverse forms, such as reverse biased diode leakage, subthreshold leakage, SiO$_2$ tunnel current, hot carrier gate current, gate induced drain leakage, channel punch through current [2]. While biased diode leakage and SiO$_2$ tunnel current flow during both active and sleep mode of the circuit, the other currents flow during the sleep mode only.

Several methods have been proposed in literature for reducing sleep mode leakage, such as use of multiple threshold CMOS [3, 4], body-biasing [5], and state assignment [6]. However, the leakage during active mode of a device has not got much attention, which is a prominent component of leakage for low-end nanotechnology [7]. As per ITRS high performance CMOS circuits will require gate oxide thickness of $0.7nm - 1.2nm$ in near future [8]. Such ultra-thin oxide devices will be more susceptible to new leakage mechanisms due to tunneling through gate oxide leading to gate oxide current [9]. The probability of electron tunneling is a strong function of the barrier height (i.e., the voltage drop across gate oxide) and the barrier thickness. Thus, reduction of active leakage power dissipation is the need of the technology.

Let us assume that V_{dd} - supply voltage of a transistor, T_{ox} - gate SiO$_2$ thickness, W_{gate} - gate width. Then, the gate oxide tunneling current is expressed as follows [1, 10] :

$$I_{ox} \quad = \quad KW_{gate}\left(\frac{V_{dd}}{T_{ox}}\right)^2 \exp\left(-\alpha\frac{T_{ox}}{V_{dd}}\right) \qquad (1)$$

Where, K and α are experimentally derived factors. From the Eqn. 1 it is observed that the following possible options are available for reduction of gate leakage power consumption, (i) decreasing supply voltage, (ii) increasing gate oxide thickness, and (iii) decreasing gate width. Decreasing power supply voltage is used as a popular option to reduce dynamic power consumption [11], and it will continue playing its role in the reduction of leakage power as well. Increase in the gate SiO$_2$ thickness leads to the increase in propagation delay. Moreover, reduction of gate width may not be an attractive option as gate leakage current is only linearly dependent on it. Thus, we conclude that considering gate oxide of multiple thicknesses may be able to reduce the oxide tunneling current while maintaining the performance.

[0] A short 2-page version of this work will appear in ISVLSI 2005.

In this research we explore the multiple gate oxide thickness approach for reduction of direct tunneling gate current during behavioral synthesis. The contributions of this paper are of two folds. First, we develop models for direct tunneling current and propagation delay calculation of functional units. We characterize the functional units for various oxide thicknesses and make them available as standard cells. Second, we introduce an algorithm for scheduling of the datapath operations such that overall tunneling current dissipation 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 thickness of different functional units may differ. The functional unit using high oxide thickness transistors dissipates lesser 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 smaller oxide thickness transistors has lesser delay and is suitable to be utilized in the critical path of a circuit.

2 Related Work

Literature in behavioral power reduction techniques have largely focussed on reduction of dynamic power. The few behavioral synthesis static power reduction works deal with reduction of subthreshold current. However, at present, there is hardly any behavioral synthesis addressing methodologies to reduce the tunneling power of a datapath circuit. On the other hand, few logic or transistor level research works focus on addressing reduction of gate tunneling.

In [12, 13], Khouri and Jha have proposed algorithms for subthreshold leakage power analysis and reduction during behavioral synthesis using dual threshold voltage. The algorithms target the least used modules as the candidates for leakage optimization. Gopalakrishnan and Katkoori in [14, 15], also use MTCMOS approach for reduction of subthreshold current during high-level synthesis. They propose binding algorithms for power, delay, and area trade-off. While clique partitioning approach is used in [15], a Knapsack based binding algorithm is proposed in [14].

In [16], Lee et. al. developed a method for analyzing gate oxide leakage current in NOR and NAND gates. They also suggested utilization pin reordering to reduce the gate leakage. Sultania et. al., in [9], introduced algorithm to optimize the total leakage power by assigning dual oxide thickness values to transistors in a given circuit. Their approach produced a tradeoff between thickness and delay values, decreasing leakage current at the cost of some delay penalty. Bowman et. al. implemented an alpha power law MOSFET model to optimize the propagation delay of circuits [17]. They estimated the minimum oxide thickness required for optimal performance of CMOS logic circuits.

3 Analytical Models

In this section, we provide analytical models for direct tunneling current and propagation delay calculation of functional units. We use a top-down approach with three level hierarchy to form the models. At the top level our objective is to prepare a set of characterized cells which are to be used for the behavioral synthesis. These in turn make use of logic level components which are derived from a set of equations available for various transistor characteristics.

The notations used in the modeling are provided in Table 1. In this work, we assumed that datapath functional units (FU) such as adders, subtractors, multipliers, dividers, etc are constructed using universal logic gate 2-input NAND. Let us assume that there are total n_{total} NAND gates in the network of NAND gates constituting a n bit functional unit. Moreover, we also assume that out of total n_{total} NAND gates in the network of NAND gates constituting a n bit functional unit, n_{cp} number of NAND gates are in the critical path. In this model we do not consider the effect of interconnects and focus on the direct tunneling power dissipation and propagation delay of the functional units only.

3.1 Analytical Model for Tunneling Current

We first discuss the high-level analytical models for the tunneling current in FUs and then identify the required terms at logical and transistor level and in turn derive them. We calculate the tunneling current of a n bit functional unit in the following manner.

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

The contributions of the NMOS and PMOS tunneling depend on the probability of the input signal being at logic “1” and “0”, respectively. Here, Pr_j is the probability that input of the NAND gate is at logic “0”, which can be obtained by carrying out logic level estimations; Pr_i is the probability that inputs of the transistors that are connected in the parallel i.e. PMOS are at logic “0”. Now we derive the relation for the I_{DT} for a NAND gate which consequently will be used to calculate tunneling current of FUs. The average tunneling current for a NAND is calculated as [9] :

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

The tunneling mechanism between substrate and gate can be either Fowler-Nordheim (FN) tunneling or direct tunneling, both differ in the form of potential barrier. We consider the tunnelling to be direct with trapezoidal potential barrier. The tunneling probability of an electron is affected by barrier height, structure and thickness and is predominant for thinner gate dielectric. The direct tunneling current is expressed by Eqn. 4 [7, 2, 18].

Table 1: Parameters used in the modeling of tunneling current and propagation delay

V_{dd}	Supply voltage in Volt (V)
V_{gs}	Gate-to-source voltage in V
V_{Th}	Threshold voltage in V
V_{fb}	Flat-band Voltage in V
V_{ox}	Voltage across the gate oxide in V
V_{poly}	Voltage across the polysilicon in V
V_{bs}	Body-to-source voltage in V
V_{dsSat}	Saturation drain voltage in V
I_{DT}	Gate oxide direct tunneling current in A
I_{DSat}	Saturation drain current in A
ϕ_B	Barrier height for the gate dielectric in eV
ϕ_F	Fermi-level in V
ψ_S	Surface potential in V
C_L	Output load capacitance in F
C_{ox}	Gate capacitance in $\frac{F}{m^2}$
Q_B	Depletion charge density in $\frac{Coulomb}{m^2}$
μ_{sub}, μ_0	Bulk mobility in $\frac{cm^2}{V-s}$
θ	Mobility degradation factor per V
v_{sat}	Electron saturation velocity in $\frac{cm}{s}$
v_{norm}	Proportionality constant with unit $\frac{cm}{s}$
α	Physical constant modelling carrier saturation velocity
$N_{channel}$	Channel doping concentrations per cc
N_{poly}	Polysilicon gate doping concentrations per cc
N_{sub}	Substrate doping concentrations per cc
n_i	Intrinsic concentration in per cc
T_{ox}	Electrical equivalent oxide thickness in nm
L	Channel length of MOSFET in nm
W	Width of MOSFET in nm
ϵ_{ox}	Permittivity of SiO$_2$ in $\frac{F}{m}$
ϵ_{Si}	Permittivity of Si in $\frac{F}{m}$
q	Electronic charge in $Coulomb$
h, $\hbar$	Planck's constant in Joule-Sec ($J-s$)
T	Temperature in Kelvin (K)
k	Boltzmann's constant in $\frac{J}{K}$
m_o	Rest mass of electron in Kilogram (Kg)
k_m	Constant for mass calculation, 0.19 for electron and 0.55 for hole
m_{eff}	$= k_m m_o$, Effective mass in kilogram (Kg)
η	Subthreshold slope factor
T_T	Transition time in s
T_{pd}	Propagation delay in s

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

The voltage across the MOSFET gate dielectic V_{ox} is expressed as follows [19, 2].

$$V_{ox} = V_{gs} - V_{fb} - \psi_S - V_{poly} \quad (5)$$

The voltage across the polysilicon depletion region V_{poly} is expressed as below [2].

$$V_{poly} = \frac{{\epsilon^2}_{ox}\ V_{ox}^2}{2q\ \epsilon_{Si}\ N_{poly} T_{ox}^2} \quad (6)$$

We plug-in Eqn. 6 in Eqn. 5 and get a quadratic equation in terms of variable V_{ox}. By solving this quadratic equation we obtain the following expression for V_{ox}.

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

The flat-band voltage V_{fb} can be derived from MOSFET capacitance-voltage (C-V) characteristics or using the expression $\left(\frac{qN_{channel}T_{ox}^2}{2\epsilon_{Si}}\right)$. The Fermi-level ϕ_F is calculated as $\left[\frac{kT}{q}\ ln\left(\frac{N_{channel}}{n_i}\right)\right]$ [7, 20, 21]. It may be noted that the effective values of W, L, may be different from original values due to the depletion and needs to be taken into consideration [17, 22].

3.2 Analytical Model for Propagation Delay

We now discuss the model that is going to be used for propagation delay calculation of functional units of a datapath. We calculate the critital path delay of a n bit functional unit using the above NAND gates as building blocks using the following.

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

The n_{fan-in} is the effective fan-in factor and is calculated for short channel devices with velocity saturation and strong inversion as shown below [23, 24].

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

Here, n_{series} is the number of series connected MOSFET and surface potential ψ_S is assumed to be twice of Fermi-level ϕ_F for strong inversion.

Now we present the transistor and logic level relations for propagation delay. We consider the alpha-power law and physical-alpha-power model and compute the propagation delay of a MOSFET T_{pd} as follows [17, 25, 26].

$$T_{pd} \quad = \quad \frac{0.5 C_L V_{dd}}{I_{D\,Sat0}} + T_T \left\{ \frac{0.5 - \left(\frac{V_{dd} - V_{Th}}{V_{dd}}\right)}{\alpha + 1} \right\} \tag{10}$$

Here, $I_{D\,Sat0}$ is the saturation drain current of the MOSFET for $V_{gs} = V_{dd}$. The saturation drain current is given by the following equation [17].

$$\begin{aligned} I_{D\,Sat} \quad = \quad & \frac{W}{L} \left(\frac{V_{gs} - V_{Th}}{V_{dd} - V_{Th}} \right)^{\alpha} \\ & \left[\frac{\mu_0 C_{ox} V_{ds\,Sat0} (V_{dd} - V_{Th} - 0.5 \eta V_{ds\,Sat0})}{\{1 + \theta(V_{gs} - V_{Th})\} \left\{ 1 + \frac{\mu_0 V_{ds\,Sat}}{v_{sat} L (1 + \theta(V_{gs} - V_{Th}))} \right\}} \right] \end{aligned} \tag{11}$$

The transition time model is given in Eqn. 12 [17].

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

The constant modeling carrier saturation velocity α is calculated as follows [17, 26].

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

Here, $V_{ds\,Sat0}$ and $V_{ds\,Sata}$ are the saturation drain voltage 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 below [17, 26].

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

The mobility degradation factor θ is computed as $\left(\frac{\mu_0}{2 T_{ox} v_{norm}}\right)$ and η is calculated as $\left[1 + \sqrt{\frac{q \epsilon_{Si} N_{channel} T_{ox}^2}{2 \epsilon_{ox}^2 (\psi_S - V_{bs})}}\right]$, assuming strong inversion [17, 26]. The mobility is calculated using the following expression $\left[\mu_{sub} / \left\{1 + \left(\frac{Q_B \mu_{sub}}{\epsilon_{ox} v_{norm}}\right)\right\}\right]$ [27, 21].

4 Behavioral Scheduler

In this section we present an integrated behavioral synthesis flow in Fig. 1 that can generate RTL for circuits with optimized tunneling current. When the proposed behavioral scheduler is used alongwith the direct tunneling and propagation delay estimator, the system generates a circuit which dissipates minimal tunneling power. The delay-tunneling current estimator uses analytical models introduced in previous section and calculates the values for different functional units. It also calculates the total tunneling current and critical path delay of the circuits when a scheduled data flow graph (DFG) is given to it. We will now introduce an algorithm for behavioral datapath scheduler based on heuristic in [11] such that datapath operations can be assigned functional units of multiple oxide thicknesses.

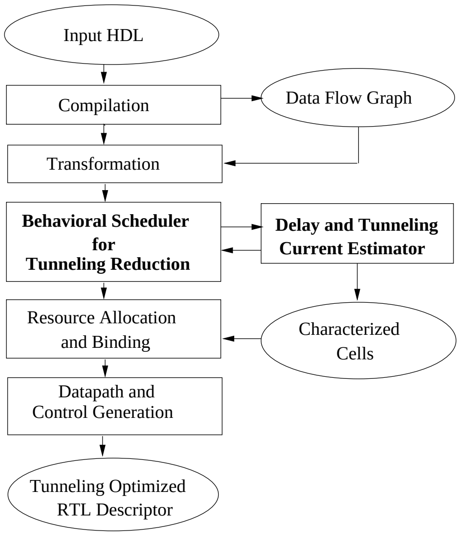
Input HDL
Compilation
Data Flow Graph
Transformation
Behavioral Scheduler for Tunneling Reduction
Delay and Tunneling Current Estimator
Resource Allocation and Binding
Characterized Cells
Datapath and Control Generation
Tunneling Optimized RTL Descriptor

Figure 1: Behavioral Synthesis Flow for Tunneling Current Reduction

The scheduling algorithm aims at minimizing the total direct tunneling current of the datapath circuit while maintaining performance. The combined reduction of tunneling power dissipation and maintenance of execution time translates to reduction of the tunneling current-delay-product (CDP). Thus, the objective of the algorithm is to minimize the CDP while assigning a schedule for the DFG. Let us assume, N_c- number of control steps, $n_{FU\,c}-$ number of resources active in any control step c. Then, the tunneling current-delay-product can be calculated as follows.

$$\text{CDP} \quad = \quad \frac{N_c}{f} \sum_{c=1}^{N_c} \sum_{r=1}^{n_{FU\,c}} I_{DT\,FU}(c, r) \tag{15}$$

Here, f is the operating frequency of the datapath circuit, which is determined by the slowest functional unit. And, $I_{DT\,FU}(c, r)$ is the functional unit active in the control step c. The inputs to the algorithm are an unscheduled DFG, the resource constraints that include number of different resources made of transistors of different oxide thickness. The algorithm generates various outputs, such as scheduled DFG with appropriate functional unit assignment to a datapath operation, estimates of current and delay.

The behavioral scheduler takes in the datapath, specified as a sequencing data flow graph (DFG), which is a directed acyclic DFG, as an input. While each vertex of the DFG represents an operation, each edge represents a dependency. The DFG does not support the hierarchical entities and the conditional statements are handled using comparison operation. Each vertex has attributes to specify the operation type.

From this input DFG alongwith the resource constraints determines the resource constrained ASAP (as soon as possible) and ALAP (as late as possible) schedules. In the next step it identifies the critical vertices V_c and the off-critical vertices V_{oc}. To begin with we consider the ASAP schedule as the default schedule. At this point, for each critical vertex V_c we assign the largest gate oxide thickness multiplier unit and smallest gate oxide thickness adder-subtractor unit.

```
Find total number of FUs of all available thickness
        from the DFG : $G(V, E)$
Get resource constrained as soon as possible schedule
         $S_{ASAP}$ and as late as possible schedule $S_{ALAP}$.
Find the vertices in critical path $V_c$ and
         off-critical path $V_{oc}$ (where, both $V_c$ and $V_{oc} \in V$).
Assume above $S_{ASAP}$ schedule as current schedule.
For each $v \epsilon V_c$ assign largest thickness $T_L$ to
       multiplication and smallest thickness $T_S$ for add-sub.
For each $v \epsilon V_{oc}$ of the current schedule $S_i$
     If vertex $v$ is a multiplication then assign the
                multiplier of highest available thickness $T_H$.
     Else assign the adder-subtractor of
                lowest available thickness $T_L$.
     Calculate CDP of the current schedule $CDP_{S_i}$.
     For each off-critical vertex $V_{oc}$
       For each allowable control steps $C_i$
         Assign multipliers of next higher thickness
                or adder-subtractor of next lower thickness.
         Find CDP of the DFG at each case.
       End For
       Fix time stamp of the vertex with the FU
                  assignment for which CDP is minimum.
     End For
End For
```

Figure 2: Behavioral Scheduler Heuristic

The scheduler algorithm heuristic is presented in Fig. 2. The algorithm attempts to assign higher leaky FUs with higher oxide thickness. This is in accordance with our conclusions from the analytical model where it is observed that multiplier units dissipate much more tunneling current compared to adder-subtractor unit. At the same time it is observed that adder-subtractor units have lesser delay compared to the multipliers. Thus, the heuristic attempts to operate the multiplier units of the highest thickness to reduce the tunneling and at the same time adder-subtractor units of lowest thickness to compensate the delay increase as much as possible. The same assignment is carried out in case of all potential off-critical paths and the CDP is calculated at each step. The CDP for the DFG is finally calculated and the FUs with the minimum CDP at each iteration are time stamped. Depending on the availability of the resources the time stamping for the vertices may change. The algorithm attempts all possible time steps for the off-critical path vertices. The algorithm identifies critical and off-critical path vertices using a simple approach. The vertices with same ASAP and ALAP time stamps are the critical vertices which are given more priority over off-critical.

5 Experimental Results

We characterized functional units, such as adder-subtractor unit, and multiplier unit of 16-bit size. While the adder-subtractor unit is a ripple carry unit and the multiplier is an array multiplier [28]. The units were presented in the form of NAND gates and characterized using the models presented in the previous sections. We used the following parameters for calculation with appropriate units as shown in Table 1, $K_{ox} = 3.9$, $V_{dd} = 0.7$, $V_{gs} = 0.7$, $V_{Th} = 0.22$, $V_{bs} = 0$, $V_{fb} = -1.0$, $\phi_B = 3.15$, $k_{mNMOS} = 0.19$, $k_{mPMOS} = 0.55$, $L = 45$, $W_{NMOS} = 180$, $W_{PMOS} = 360$, $N_{channel} = 1.7 \times 10^{17}$, $N_{poly} = 5.0 \times 10^{19}$, $N_{sub} = 6.0 \times 10^{16}$, $T = 300$, $n_i = 9.5 \times 10^9$, $v_{norm} = 2.2 \times 10^9$, $v_{sat} = 6.4 \times 10^6$, $\mu_{subNMOS} = 750$, and $\mu_{subPMOS} = 250$. Fig. 3 and 4 show the variation of direct tunneling current and propagation delay of the functional units as the oxide thickness changes. It is assumed that the probability of logic "1" and logic "0" is same. While changing the oxide thickness the channel length of the transistor is changed proportionately to avoid impact on its functionality [9].

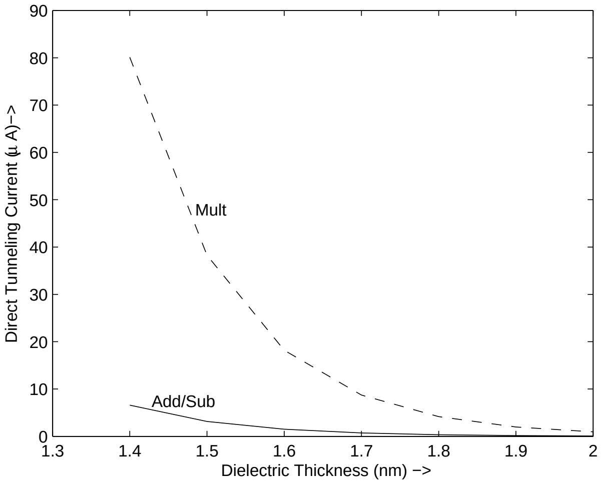

Figure 3: Direct Tunneling Current Versus Oxide Thickness

The algorithm was implemented for experiments in the behavioral synthesis framework proposed in [11] and tested with several behavioral level benchmark circuits for several

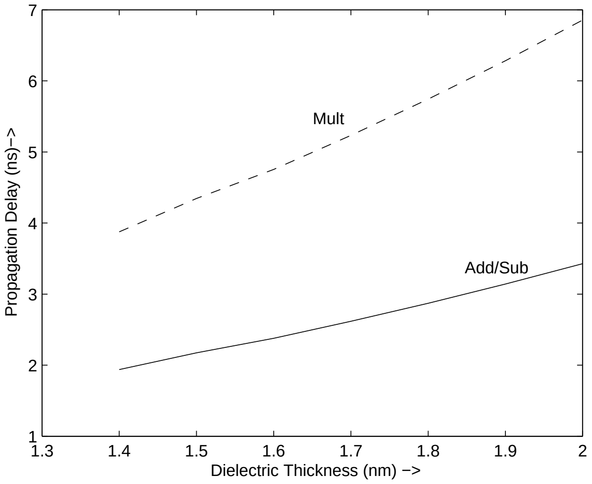

Figure 4: Propagation Delay Versus Oxide Thickness

constraints. However, we have presented the results in this section for selected benchmarks and constraints. A selected set of resource constraint is given in Table 2. These represent the functional units of different thickness available to the behavioral scheduler. The sets of resource constraints were chosen so as to cover functional units consisting of different oxide thickness. These are the representatives of various forms of the corresponding RTL representation. A selected set of benchmarks used are as follows [29]: (i) Auto-Regressive filter (ARF) (total 28 nodes, 16*, 12+, 40 edges), (ii) Band-Pass filter (BPF) (total 29 nodes, 10*, 10+, 9-, 40 edges), (iii) DCT filter (total 42 nodes, 13*, 29+, 68 edges), (iv) Elliptic-Wave filter (EWF) (total 34 nodes, 8*, 26+, 53 edges), (v) FIR filter (total 23 nodes, 8*, 15+, 32 edges), and (vi) HAL differential equation solver (total 11 nodes, 6*, 2+, 2-, 1<, 16 edges).

Table 2: A Selected Resource Constraints used in our Experiments

Number of FUs of Different Oxide Thickness T_{ox}				No.
Multiplier		Adder-Subtractor		
$1.7nm$	$1.4nm$	$1.7nm$	$1.4nm$	
1	1	2	0	1
2	1	1	1	2
2	0	0	2	3
3	0	1	1	4

The experimental results are presented for different constraints for two different oxide thicknesses in Table 3. The results take into account the tunneling current and propagation delay of functional units present in the datapath circuit. The quantities with ST subscript represent the values for single oxide thickness and the multiple oxide thickness are shown with MT subscript. We assume the minimal oxide thickness case with T_{ox} of $1.4nm$ as the base ST case. The reduction in tunneling current is calculated as, $\frac{I_{ST}-I_{MT}}{I_{ST}} * 100\%$. It is observed that while the minimum reduction in the tunneling current is 52.38%, the maximum reduction is 89.10%. The overall reduction for all benchmarks over all constraints is 76.02% in average.

We anticipate that the critical path delay is going to increase due to the use of multiple dielectric as delay increases with the increase in oxide thickness. The time penalty is calculated as, $\frac{T_{pd_{MT}}-T_{pd_{ST}}}{T_{pd_{ST}}} * 100\%$. We used two ways to calculate the critical path delay of the circuit of the benchmarks. In one method, we estimate the critical path delay of the circuit and the sum of the delays of the vertices in the longest path of the data flow graph, which are reported in the results table. The time penalty is found to be in the range of $18.71 - 54.46\%$ with an overall average of 34.58%. In the second method we defined critical path delay as the product of the number of control steps and the inverse of the operating frequency. In this method the maximum time penalty is 35.1%.

We also carried out experiments using functional units of three different oxide thickness. In this scenario the maximum reduction is as high as 90.23% with an average of 2.06%. But, there is increase in the time penalty, which on an average is 11.05%.

6 Conclusions

The direct tunneling current is a significant leakage component and contributes to an appreciable portion of total power consumption of a CMOS nanometer circuits. In this paper we presented a novel technique which utilizes functional units of multiple gate oxide thickness as an attractive option for overall direct tunneling current reduction of a datapath circuit. The functional unit selection is being implicitly made during scheduling and we are in the process of evaluating impact on the area, capacitance and dynamic power. A heuristic based approach is presented here for functional unit assignment. We are anticipating that use of more advanced optimization techniques may be further helpful. We also need to incorporate methods to accurately estimate the logic values for more accurate modeling of the tunneling current and propagation delay. Finally, it is our goal in future to expand on the work done for the tunneling current and to develop a holistic step by step solution to the entire spectrum of power dissipation in the behavioral level. While multiple oxide thickness is highly effective, use of multiple dielectrics using high-K dielectric materials alongwith multiple thickness will be explored in future.

Table 3: Direct Tunneling Current and Propagation Delay of different Benchmark Circuits

Benchmark Circuits	Resource Constraints	Tunneling Current in μA			Critical Path Delay in ns		
		IDT_{ST}	IDT_{MT}	$\%Reduction$	$T_{pd_{ST}}$	$T_{pd_{MT}}$	$\%Penalty$
ARF	1	1360.53	647.79	52.38	34.86	49.72	42.62
	2	1360.53	409.23	69.92	34.92	43.80	33.04
	3	1360.53	218.58	83.93	34.86	45.74	31.21
	4	1360.53	195.12	85.65	32.93	43.80	33.00
	Average Reduction			72.97	Average Penalty		34.96
BPF	1	1073.06	402.36	62.50	30.99	44.48	43.53
	2	1073.06	312.41	70.88	29.05	41.87	44.13
	3	1073.06	216.60	79.81	30.94	40.51	30.71
	4	1073.06	169.67	84.18	29.05	41.87	44.13
	Average Reduction			74.34	Average Penalty		40.62
DCT	1	1232.15	205.58	83.31	52.29	70.65	35.11
	2	1232.15	222.19	81.96	52.29	69.97	33.81
	3	1232.15	304.32	75.30	52.29	68.61	31.21
	4	1232.15	222.19	81.96	52.29	69.97	33.81
	Average Reduction			80.63	Average Penalty		33.48
EWF	1	811.92	88.43	89.10	44.55	62.80	40.96
	2	811.92	176.42	78.27	44.55	58.15	30.52
	3	811.92	240.95	70.32	44.55	54.07	21.36
	4	811.92	176.42	78.27	44.55	58.15	30.52
	Average Reduction			79.00	Average Penalty		30.84
FIR	1	739.51	294.66	60.15	29.05	41.87	44.13
	2	739.51	145.07	80.38	29.05	35.17	21.06
	3	739.51	168.53	77.20	29.05	34.49	18.72
	4	739.51	145.07	80.38	29.05	35.17	21.06
	Average Reduction			74.52	Average Penalty		22.24
HAL	1	513.49	198.67	61.30	13.55	20.93	54.46
	2	513.49	150.76	70.63	11.62	17.63	51.72
	3	513.49	85.26	83.39	13.55	17.63	30.11
	4	513.49	85.26	83.39	11.62	15.02	29.25
	Average Reduction			74.67	Average Penalty		41.38
For all Benchmarks		Average Reduction		76.02	Average Penalty		34.58

References

[1] 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.

[2] 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.

[3] P. Pant, R. K. Roy, and A. Chattejee, “Dual-Threshold Voltage Assignment with Transistor Sizing for Low Power CMOS Circuits,” *IEEE Transactions on VLSI Systems*, vol. 9, no. 2, pp. 390–394, April 2001.

[4] 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.

[5] S. Narendra, A. Keshavarzi, B. A. Bloechel, S. Borkar, and V. De, “Forward Body Bias for Microprocessors in 130-nm Technology Generation and Beyond,” *IEEE Journal of Solid-State Circuits*, vol. 38, no. 5, pp. 696–701, May 2003.

[6] D. Lee and D. Blaauw, “Static Leakage Reduction Through Simultaneous Threshold Voltage and State Assignment,” in *Proceedings of the Design Automation Conference*, 2003, pp. 191–194.

[7] 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.

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

[9] 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.

[10] A. Chandrakasan, W. Bowhill, and F. Fox, *Design of High-Performance Microprocessor Circuits*, IEEE Press, 2001.

[11] 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.

[12] K. S. Khouri and N. K. Jha, "Leakage power analysis and reduction during behavioral synthesis," in *Proceedings of International Conference on Computer Design*, 2000, pp. 561–564.

[13] 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.

[14] 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.

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

[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] 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.

[18] 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.

[19] 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.

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

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] S. P. Mohanty, N. Ranganathan, and V. Krishna, "Datapath Scheduling using Dynamic Frequency Clocking," in *Proceedings of the IEEE Computer Society Annual Symposium on VLSI*, Apr 2002, pp. 65–70.