Reduction of Tunneling Current during Behavioral Synthesis of Nanometer Circuits

Saraju P. Mohanty, Valmiki Mukherjee, Ramakrishna Velagapudi, and Hao Li
Dept of Computer Science and Engineering
University of North Texas
Email: smohanty@cs.unt.edu
Power Dissipation Trend

Source: Hansen 2004

- Sub-Threshold
- Tunneling
- Dynamic
- Trajectory, if High-K

Chronological (Year)

Normalized Power Dissipation

Gate Length

Physical Gate Length (nm)

Normalized Power Dissipation


Source: Hansen 2004
Tunneling Paths in an Inverter

- **Low Input**: Input supply feeds tunneling current.
- **High Input**: Gate supply feeds tunneling current.

\[ V_{in} = V_{Low}, \quad V_{out} = V_{High} \quad \text{ON} \]

\[ V_{in} = V_{High}, \quad V_{out} = V_{Low} \quad \text{OFF} \]
Dual-$T_{\text{o}}$?

- Gate oxide tunneling current $I_{\text{ox}}$ [Kim2003, Chandrakasan2001] ($\alpha$ is an experimentally derived factor):

  $$I_{\text{ox}} \propto \left(\frac{V_{\text{dd}}}{T_{\text{o}}}ight)^2 \exp\left(-\alpha \frac{T_{\text{o}}}{V_{\text{dd}}}\right)$$

- Options for reduction of tunneling current:
  - Decreasing of supply voltage $V_{\text{dd}}$ (will play its role)
  - Increasing gate SiO$_2$ thickness $T_{\text{o}}$ (delay increases)

- We believe that combined use of high-$T_{\text{o}}$ resources and low-$T_{\text{o}}$ resources can reduce the gate oxide tunneling current of a datapath with little compromise in circuit performance.
Dual-\(T_{\text{ox}}\) ?

**Assumption:** All transistors of a resource are of equal \(T_{\text{ox}}\).

High \(T_{\text{ox}}\) Resource: Smaller \(I_{\text{ox}}\), but Larger delay

Low \(T_{\text{ox}}\) Resource: Larger \(I_{\text{ox}}\), but Smaller delay
Related Works

Behavioral Level Subthreshold:

- Gopalakrishnan, ICCD2003: MTCMOS approach for reduction of subthreshold current

Logic / Transistor Level Tunneling:

- Lee, TVLSI2004: Pin reordering to minimize gate leakage during standby positions of NOR and NAND gates.
- Sultania, DAC2004: Heuristic for dual $T_{ox}$ assignment for tunneling current and delay tradeoff.
We assumed that resources such as adders, subtractors, multipliers, dividers, are constructed using 2-input NAND.

There are 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.
Analytical Model for Tunneling Current

● The tunneling current of a $n$-bit functional unit:

$$I_{DTFU} = \sum_{j = (1 \to n_{total})} Pr_j \sum_{MOS_i \in NAND_j} Pr_i I_{DTi}$$

$Pr_j$ is the probability that input of the NAND gate is at logic “0”, and $Pr_i$ is the probability that inputs of the transistors that are connected in the parallel i.e. PMOS are at logic “0”.

● The average tunneling current for a NAND is calculated as

$$I_{DTNAND} = \sum_{MOS_i \in NAND} Pr_i I_{DTi}$$
Analytical Model for Tunneling Current

- The direct tunneling current of a MOS:

\[
I_{DT} = \frac{WLq^3V_{ox}^2}{16\pi^2\phi_BT_{ox}^2} \exp\left[-\frac{4\sqrt{2m_{\text{eff}}\phi_B^{1.5}T_{ox}}}{3\hbar qV_{ox}} \left\{1 - \left(1 - \frac{V_{ox}}{\phi_B}\right)^{1.5}\right\}\right]
\]

- The voltage across the MOSFET gate dielectric \(V_{ox}\) is expressed as follows:

\[
V_{ox} = V_{gs} - V_{fb} - \psi_S - V_{poly}
\]
Analytical Model for Tunneling Current

- By solving a quadratic equation we obtain an expression for $V_{ox}$:
  
  $$V_{ox} = \sqrt{1 - 2(V_{fb} + \psi_S - V_{gs})\left(\frac{\varepsilon_{ox}^2}{q\varepsilon Si N_{poly} T_{ox}^2}\right)} - 1$$

- The flat-band voltage $V_{fb}$ can be obtained using the expression $(qN_{channel} T_{ox}^2 / 2\varepsilon_{Si})$.

- $\psi_S = 2 *$ Fermi-Level, for strong inversion.
Analytical Model for Propagation Delay

- The critical path delay of a $n$ bit functional unit using the NAND gates as building blocks:
  \[ T_{pd_{FU}} = \sum_{I = (1 \rightarrow ncp)} I \cdot 0.5(n_{fan-in} T_{pd_{NMOS}} + T_{pd_{PMOS}}) \]
- $n_{fan-in}$ is the effective fan-in factor.
- Using physical-alpha-power model the delay of a MOS, where $I_{DSat0}$ is the saturation drain current of the MOS for $V_{gs} = V_{dd}$.

\[
T_{pd} = \frac{0.5 \cdot C_L V_{dd}}{I_{DSat0}} + T_T \left\{ 0.5 - \left( \frac{V_{dd} - V_{th}}{V_{dd}} \right) \right\} \alpha + 1
\]
Cell Characterization: 45nm Tech

Direct Tunneling Current Versus Oxide Thickness

- Dielectric Thickness (nm) vs. Direct Tunneling Current
- Add/Sub and Mult curves

Propagation Delay Versus Oxide Thickness

- Dielectric Thickness (nm) vs. Propagation Delay
- Add/Sub and Mult curves
Dual-T<sub>ox</sub> Based Behavioral Synthesis

- Input HDL
- Compilation
- Transformation
  - Behavioral Scheduler for Tunneling Reduction
  - Delay and Tunneling Current Estimator
  - Characterized Cells
- Resource Allocation and Binding
- Datapath and Control Generation
- RTL Descriptor
Dual-$T_{OX}$ Assignment: Basis

- **Observation**: Tunneling current of Functional Units increases and propagation delay decreases as oxide thickness decreases.

- **Strategy**: Maximize utilization of high-$T_{OX}$ high leaky resources (e.g. multipliers) and low-$T_{OX}$ low leaky resources (e.g. adder-subtractor) to improve chances of tunneling current reduction with minimal performance degradation.
The library consists of Multipliers and Adder-Subtractor units, characterized for the 45 nm technology.

We used $T_1 = 1.4$ nm, and $T_2 = 1.7$ nm to perform our experiments.

The value of $T_1$ is chosen as the default value from the BSIM4.4.0 model card and value of $T_2$ is intuitively chosen.
Conclusions and Future Works

- Tunneling current is a major component of total power consumption of a low-end CMOS nanometer circuit.
- Dual-$T_{OX}$ approach results significant reductions in tunneling current with minimal performance penalty.
- Development of optimal assignment algorithm is under progress.
- Tradeoff of tunneling, area and performance needs to be explored.
- Dual-$T_{OX}$ based design may need more masks for the lithographic process during fabrication.