Design of Scaled CMOS Circuits in the Nano-meter Regime: Leakage Tolerance and Computing with Leakage

Kaushik Roy
Professor of Electrical & Computer Engineering
Purdue University

VLSI Applications

- Ultralow power applications: medical, space, specific sensor network etc.
- General purpose computing: internet server, database server, real-time jobs etc.
- Portable applications: mobile computing, wireless, multimedia etc.

- Different applications have different power-performance demands
- Scaling affects all applications in different ways
Challenges ahead …

in Si nanometer regime

Challenge No. 1: Device Scaling

- **Bulk MOS**
  - Retrograde Well, Halo
  - Strained Channel

- **UTB-SOI MOS**
  - Fully-depleted ultra-thin body
  - Ground-plane

- **DG-SOI MOS**
  - Planar double-gate structure
  - Independent gate control

- **FinFET**
  - Quasi-planar DG structure
  - Most promising device

- **Tri-gate**
  - Quasi-planar with 3 gates
  - Better area efficiency

- **Vertical MOS**
  - Conduction normal to plane
  - Difficult to fabricate
Scaling & Ion/Ioff

- Increasing leakage
- Increasing process variations
- Short Channel Effects

\[
\frac{I_{ON}}{I_{OFF}} = 10^9
\]

\[
\frac{I_{ON}}{I_{OFF}} = 10^5
\]

\[
\frac{I_{ON}}{I_{OFF}} = 10^4
\]

Silicon micro electronics
Silicon nano electronics
Non-Silicon technology

- Carbon Nanotubes
- Molecular transistors
- Molecular RTDs

2. Power & Power Density

- Increased Average Power
  - Battery Life
  - Cooling Cost

- Increased Power Density
  - Reliability

Source: Intel
Challenge 3: Process Variations

- Intrinsic parameter variations:
  - Channel length and width
  - Variations due to line edge roughness
  - Threshold voltage (Vt) variations due to random dopant fluctuation

Device parameters are no longer deterministic

Challenge 4: Reliability

Temporal degradation of performance -- NBTI
Power Consumption

- **Leakage Power**
  - Subthreshold, Gate, Junction, GIDL, Punchthrough, ....

- **Dynamic Power**
  - Due to charging/discharging of capacitive load
  - Short-circuit power due to direct path currents when there is a temporary connection between power and ground

Leakage Vs. Dynamic Power (Projection)

Leakage power limits $V_{th}$ scaling
Leakage Power

Scaling and Other Leakage Components

- Leakage Components
  - Subthreshold Leakage
  - Gate Leakage
  - Reverse-biased Junction Band-To-Band-Tunneling (BTBT) Leakage.
  - Others

Long Channel (L > 1 \( \mu \)m)
Negligible leakage

Short Channel (L > 180 nm, Tox > 30\( \AA \))
Subthreshold leakage

Very Short Channel (L > 90 nm, Tox > 20\( \AA \))
Subthreshold + Gate Leakage

Nano-scaled (L < 90 nm, Tox < 20\( \AA \))
Subthreshold + Gate + Jn. BTBT leakage
Leakage Power Consumption

\[ V_{out} = V_{DD} \]

Diode leakage

\[ I_O = i_s (e^{V_q/kT} - 1) \]

Sub-threshold leakage

\[ I_D = k \cdot e^{(V_{gs} - V_t)q / nkT} \left( 1 - e^{V_{ds}q / kT} \right) \]

\[ P_{static} = I_{leakage} \cdot V_{DD} \]

Diode Leakage

- Leakage current through the reverse biased diode junctions
- For typical devices it is between 0.1nA - 0.5nA at room temperature
- For a die with 1 million devices operated at 5 V, this results in 0.5mW power consumption → not much
- Junction leakage current is caused by thermally generated carriers -> therefore is a strong function of temperature
- More important is sub-threshold leakage, gate leakage, and Junction BTBT leakage
Leakage Components

Vt Scaling
- Short Channel Effects
  - Scale Tox
  - Scale Wd-doping ↑
  - Channel Engg. – patches of higher doping in the channel

Subthreshold Leakage ↑
Gate Leakage ↑
Junction BTBT Leakage ↑

S/D Extension
Source/Drain
Gate
P
N+
Retrograde Well
Halo

Jn. Band-To-Band-Tunneling Current

\[ I_{\text{BTBT}} \]

Electron tunneling from VB of p-side to the CB of n-side.

BTBT Current density depends on Junction field \( \xi \), junction voltage \( V_{\text{app}} \), band-gap \( E_g \).

\[
J_{h\rightarrow b} = A \frac{\xi V_{\text{app}}}{E_g^{1/2}} \exp\left(-B \frac{E_g^{3/2}}{\xi}\right)
\]

\[
A = \frac{\sqrt{2m^* q^3}}{4\pi^2 \hbar^2}, \quad \text{and} \quad B = \frac{4\sqrt{2m^*}}{3q\hbar}
\]

High BTBT in scaled devices
- High junction doping: “Halo” profiles
- Small depletion width
- Large electric field
Gate Leakage ($I_{gate}$)

Direct tunneling of electron through gate oxide.

Gate current density depends on oxide thickness, oxide field and voltage drop across oxide.

$$J_{dr} = A_g \left( \frac{V_{ox}}{T_{ox}} \right)^2 \exp \left( - \frac{B_g \left( 1 - \left( \frac{V_{ox}}{\Phi_{ox}} \right)^{3/2} \right)}{V_{ox}/T_{ox}} \right)$$

High Gate leakage in scaled devices:
Low oxide thickness and high oxide field

Components of Gate Leakage

Gate leakage components
- Gate to source/drain overlap region ($I_{gso}$, $I_{gdo}$)
  - Controlled by $V_{gd}$ and $V_{gs}$
- Gate to channel ($I_{gc} = I_{gcs} + I_{gcd}$)
  - Controlled by $V_{ox} = V_{gs} - V_{FB} - \Phi_s - V_{poly}$
- Gate to body ($I_{gb}$)
  - Controlled by $V_{gb}$

Transistor off ($V_g = '0'$) – $I_{gdo}$ and $I_{gso}$ dominates.
Transistor on ($V_g = '1'$) – $I_{gc}$ ($I_{gcs}$ & $I_{gcd}$) dominates.
$I_{gb}$ small compared to others.
Subthreshold leakage ($I_{sub}$)

Exponentially dependence on $V_{gs}$ and $V_{th}$.

$$I_{sub} = \frac{w_{eff}}{L_{eff}} \mu \sqrt{\frac{q \varepsilon_{si} N_{eibf} \varepsilon_m}{2 \Phi_s}} v_T \exp \left( \frac{V_{gs} - V_{th}}{n v_T} \right) \left( 1 - \exp \left( -\frac{V_{ds}}{v_T} \right) \right)$$

Vth modulation

- **Short channel Effect** – $V_{th}$ reduction due to
  - Increase in $V_{ds}$ (DIBL),
  - Reduction in Channel Length ($V_{th}$ roll off),
- **Body effect** – negative $V_{bs}$ increases $V_{th}$
- **Quantum confinement effect** – increases $V_{th}$

$$V_{th} = V_{FB} + \left( \Phi_{s0} - \Delta \Phi_s \right) + \gamma \sqrt{\frac{\Phi_{s0} - V_{bs}}{n v_T}} \left( 1 - \frac{W_{dm}}{W_{eff}} \right) + V_{nce} + V_{QM}$$

$$\Delta \Phi_s = \left[ 2(V_{bi} - \phi_{30}) + V_{as} \right] \times \left[ e^{-L/2L_s} + 2e^{-L/4L_s} \right] \text{ and } I_c = \sqrt{\left( \frac{\varepsilon_{si} \varepsilon_m}{\varepsilon_{so} \eta} \right) T \frac{W_{dm}}{2}}$$

Subthreshold Leakage

Subthreshold leakage reduces with

- **Negative $V_{bs}$, Reduction of $V_{ds}$**
- **Application of Quantum Correction**
Total Leakage

“Sum of Current Source Model”
Voltage Controlled Current Sources describing each leakage comp.

Total Transistor Leakage: \( I_{\text{overall}} = I_{\text{BTBT}} + I_{\text{sub}} + I_{\text{gate}} \)

Leakage Estimation Method

Leakage Table

Logic Circuits

Current Modeling

Logic Gates

Estimated Total Circuit Leakage

INPUT

OUTPUT
Low-Vdd Low-Vt Design

- Stacked CMOS
- Dual-threshold CMOS
- Dynamic-threshold CMOS

Leakage Reduction (Logic & Memory)

Circuit Techniques

- Design Time
  - Dual $V_{th}$
- Run Time
  - Standby Leakage Reduction
  - Natural Stacking
  - Sleep Transistor
  - FBB/RBB
  - Active Leakage Reduction
  - DVTS
Self-Reverse Bias (Source-Biasing, Supply-Gating, Stacking)

- Primary effect:
  - $V_{GS} < 0$
  - move down subthreshold slope

- Secondary effects:
  - Drain Induced Barrier Lowering
  - Body effect

$V_{DS} \downarrow \Rightarrow V_T \uparrow$

$V_S \uparrow \Rightarrow V_T \uparrow$

Leakage Control: Stacking

For M1:
- $V_{gs} = -V_M < 0, V_{bs} = -V_M < 0$
- $V_{ds} = V_{dd} - V_M < V_{dd}$

For M2:
- $V_{gs} = 0, V_{bs} = 0$
- $V_{ds} = V_M < V_{dd}$

✓Negative $V_{gs}$,
✓Negative $V_{bs}$- More Body effect,
✓Reduced $V_{ds}$-Less DIBL
2-T stack has lower subthreshold leakage
Input Vector Control - Subthreshold

Minimum Vgs is For M1:
Vgs_M1 < 0,
Vds_M1 = Vdd - VM

Minimum Vgs is For M2:
Vgs_M2 = 0,
Vds_M2 = Vdd - Vth_M1

‘00’ gives minimum subthreshold leakage.
Turn ‘off’ maximum number of transistors in a stack to reduce subthreshold leakage

Leakage vs. Transistors Off

Leakage [nA]

Number of transistors off in stack
Input Vector Control – Gate Leakage

- $V_g = '0'$ – EDT dominates
  - $I_g = I_{gdo} + I_{gso}$

- $V_g = '1'$ – Gate to Channel tunneling is significant
  - $I_g = I_{gdo} + I_{gso} + I_{gc}$

With ‘00’ –
- $I_{gdo\_M1}(V_{dd}) >> I_{gso\_M1}(V_M) + I_{gdo\_M2}(V_M)$
- $I_{gdo}$ of M1 dominates the total gate current

$$I_{gstack} = W L_{DE} A \left( \frac{V_{dd}}{T_{ox}} \right)^2 \exp \left( \frac{-B \left( 1 - \frac{V_{dd}}{\phi_{ox}} \right)^{3/2}}{V_{dd} / T_{ox}} \right)$$

Input Vector Control – Gate Leakage

With ‘01’ –
- $I_{gdo\_M1}(V_{dd})$ is high.
- $I_{gdo\_M2}(V_{dd})$ is high.
- $I_{gso\_M2}(V_{dd})$ is high.
- Gate to channel leakage of M2 is controlled by:
  - $V_{ox\_M2} = V_{dd} - V_{FB} - \phi_s - V_{poly}$

Total gate current is high.
With ‘10’ the major gate currents are:

- \( I_{gso\_M1}(V_{th}) \)
- \( I_{gdo\_M2}(V_{dd} - V_{th\_M1}) \)
- \( I_{gc\_M1}(V_{gs} = V_{th}) \)

\( I_{gdo\_M2} \) dominates the total current.

\[
I_{\text{gate}} = W_L S_D E A \left( \frac{(V_{dd} - V_{th\_M1})}{T_{ox}} \right)^2 \exp \left( \frac{-B \left(1 - (V_{dd} - V_{th\_M1})/\phi_{ox}\right)^{3/2}}{(V_{dd} - V_{th\_M1})/T_{ox}} \right)
\]

\( I_{g(V_{dd})} > I_{g(V_{dd}-V_{th\_M1})} \)
Rate of change of gate current increases with an increase in \( V_{ox} \) (exponential)

Gate current with ‘10’ is lower than ‘00’
Gate Leakage

Gate leakage increases with
✓ Increase in Vox
✓ Reduction in Tox.

Rate of change of current is higher at higher Vox

Input Vector Control – BTBT

’00’ and ’01’ – drain-substrate BTBT of M1 dominates.
’10’ – additional BTBT components drain-substrate of M2 and source-substrate of M1.

’10’ gives maximum BTBT. However, BTBT is not very sensitive to stacking.
**Input Vector Control – Total Leakage**

Leakage difference between ’10’ and ’00’ =

\[ \Delta I_{\text{leakage}} = I_{\text{10}} - I_{\text{00}}. \]

\[ = (I_{\text{sub-10}} - I_{\text{sub-00}}) + (I_{\text{gdo2-10}} - I_{\text{gdo1-00}}) + (I_{\text{btbt-10}} - I_{\text{btbt-00}}) \]

- Isub-10 > Isub-00
- Igdo2-10 < Igdo1-00
- Ibttb-10 ≥ Ibttb-00

**Supply Gating for Logic**

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-20X Leakage Reduction</td>
<td>Delay/Area Overhead</td>
</tr>
<tr>
<td>Scalable</td>
<td>Floated Output</td>
</tr>
<tr>
<td>Design ease</td>
<td>Can be applied to idle sections only</td>
</tr>
</tbody>
</table>

**How to use supply gating dynamically in active mode?**
Dynamic Supply Gating (DSG): An Example

How to do it for random logic?

Dynamic Supply Gating for General Circuits

- **Shannon’s expansion:**

  \[ f(x_1, x_2, \ldots, x_n) = x_1 f(x_1 = 1, x_2, \ldots, x_n) + x_1' f(x_1 = 0, x_2, \ldots, x_n) \]

  \[ = x_1 CF_1 + x_1' CF_2 \]

  \[ CF_1 = f(x_1, \ldots, x_i = 1, \ldots, x_n) \]

  \[ CF_2 = f(x_1, \ldots, x_i = 0, \ldots, x_n) \]

  \[ f_1 = x_1 \]

  \[ f_2 = x_1' \]

  \[ x_i \text{ is referred as Control Variable} \]

Control variable selection is important.
Simulation Results

Active Leakage Saving

MCNC Benchmarks, 70nm Process, Vdd=1V, Temp=100°C

Supply-Gating & Test
Iddq Test – Feasible in Scaled Technologies?

- New challenges due to scaling and high integration density
  - Increased number of faults
  - More parametric failures
  - IDDQ test is no longer effective due to increased leakage
  - Yield loss
- Reduction in test power required for mobile devices
- High test coverage needed with reasonable test time because
  - New failure mechanisms have emerged
  - Defect density has increased

An integrated DFT solution is required to reduce test time, test power, while maintaining coverage and alleviating the effects of process variations

Proposed Solution

- Use a Shannon expansion based design and supply gating to
  - Reduce the quiescent current
  - Improve the leakage yield
  - Reduce test power
  - Improve the test coverage/test length
IDDQ Reduction by Cofactor Balancing

- Larger cofactors can consume more standby current
- Change the selection of control variable (CV) for balancing

\[
CV = \max (a + b) \quad CV = \frac{\max (a + b)}{|a - b|}
\]

Improvement in IDDQ Sensitivity

IDDQ Sensitivity (S) = \( \frac{I_f - I_g}{I_g} \)

- \( I_f \) = Faulty IDDQ
- \( I_g \) = Fault free IDDQ

Avg. improvement of 94% in IDDQ sensitivity

MCNC Benchmarks, 70nm Process, Vdd=1V, Temp=100°C

Original

Shannon
IDDQ Distribution Under Process Variation

Improvement of 5% (9%) in parametric yield (for circuit cm150a (pcle), considering leakage bound)

Test Power

- Sources of test power
  - Scan registers
  - Combinational circuits

- Combinational circuit consumes 78% test power

- Advantages of SBS
  - No changes required in scan register and test application procedure
  - Can reduce both switching and leakage power
  - At-speed testing can be performed easily
  - Other techniques can be integrated for power saving in registers
Improvement in Test Power

MCNC Benchmarks, 70nm Process, Vdd=1V, Temp=25°C

Avg. reduction of 50% in test power

Test Coverage/Test Length

- High test coverage is needed because
  - New failure mechanisms have emerged
  - Defect density has increased

- Cost of ATE prohibits exhaustive testing of chip

- Circuits employing BIST for periodic self-test requires high coverages with smaller test time

- Advantages of SBS
  - Reduction in number of faults due to smaller area after multi-level expansion in some cases
  - Increased observability of internal nodes
Improvement in Test Coverage/Test Length

Improvement in Test Coverage/Test Length

MCNC Benchmarks, 70nm Process, Vdd=1V, Temp=25°C

Avg. reduction of 20% (21%) in test time with deterministic (random) patterns

Supply Gating in Scan Design

-- Low-power Scan Operation
Conventional Scan Architecture

- Primary Input
- Primary Output
- Blocking redundant switching in comb. logic
- Latch

High Design Overhead

Any better solution?

First Level Supply Gating (FLS)

- VDD
- INV1
- INV2
- INV3
- MP1
- MN1
- Gating Ctrl
- GND
- Shared First Level Supply Gating Transistor
Results and Comparisons for FLS

- Compared to Nor-based Gating:
  - **Area:** 62% less overhead
  - **Delay:** 94% less

Input Vector Control for Leakage Reduction

- Application of best input during scan shifting can save leakage power
- **About 38% leakage saving** with Mixed VDD/GND FLS over NOR gating
Low-Overhead Delay Fault Testing With Supply Gating
First Level Hold (FLH) for Delay Testing

1. Scan-in V1

2. Apply V1. Hold state for V1
First Level Hold (FLH) for Delay Testing

1. Scan-in V1
2. Apply V1. Hold state for V1
3. Scan-in V2

Embedded latch can be implemented with minimum-sized transistors
• No extra signal; simple control
• Eliminates redundant test power in comb. logic
Results and Comparisons for FLH

- Compared to Enhanced Scan:
  (a) Area: 33% less overhead, (b) Delay: 71% less overhead, (c) Power: 90% less overhead
- Local Fanout Reduction reduces area overhead by ~20%

Gated DeCap: Another Application of Stacking & Leakage Reduction
Decoupling Capacitor (Decap)

Area and power of Decap
- 15-20% of the total chip area (Alpha 21264).
- Total 26W Decap gate leakage power consumption (reported by IBM, 2003).

Leakage Power of Decap

Gate leakage current of Decap increases exponentially with gate-oxide thickness scaling

<table>
<thead>
<tr>
<th>Year</th>
<th>Gate length (nm)</th>
<th>Oxide thickness (nm)</th>
<th>Gate leakage (μA/μm)</th>
<th>Supply voltage (V)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>65</td>
<td>1.3</td>
<td>0.01</td>
<td>1.2</td>
</tr>
<tr>
<td>2004</td>
<td>37</td>
<td>0.9</td>
<td>0.10</td>
<td>1.0</td>
</tr>
<tr>
<td>2007</td>
<td>25</td>
<td>0.6</td>
<td>1.00</td>
<td>0.7</td>
</tr>
<tr>
<td>2010</td>
<td>18</td>
<td>0.5</td>
<td>3.00</td>
<td>0.6</td>
</tr>
<tr>
<td>2013</td>
<td>13</td>
<td>0.4</td>
<td>7.00</td>
<td>0.5</td>
</tr>
<tr>
<td>2016</td>
<td>9</td>
<td>0.4</td>
<td>10.00</td>
<td>0.4</td>
</tr>
</tbody>
</table>
(a) Conventional NMOS Decap  
(b) NMOS Decap with control gate

- The gate and the channel of M1 constitute a capacitor.
- M2 is turned off when Decap is unnecessary (FU is idle).

Leakage Current Distribution in GDecap

- When M2 is turned on, Decap M1 is enabled.
- When M2 is turned off
  - V_GND is increases
  - Potential drop across the gate-oxide of M1 decreases.
  - Gate leakage of M1 is reduced exponentially.

*Stack Effect (Again)!*
Control Scheme of GDecap

Sizing-up of Control Gate M2

- M1: Width = 11625nm; Length = 700nm for 20x20μm².
- Maintaining the effectiveness of Decap. Noise threshold: 10% of $V_{DD}$ (1.1V) at 70nm Tech.
Layout of GDecap

GDecap
Area Overhead:
6.78%

Conventional Decap

Leakage Power Saving of GDecap in PLB Pipeline

Average Decap leakage power reduction:
Mod. PLB – 41.7% (FU gated ratio: 55.15%)
0.037% worst-case IPC degradation in Mod. PLB.
Leakage & Body Bias

- Sub-threshold leakages decreases with RBB
- Band-to-band tunneling increases with RBB
- Gate Leakage insensitive to body bias

Results for 70nm nmos

BSIM3 device augmented with voltage-controlled current sources for gate leakage and BTBT

OBB and Doping Profile

- Optimal body bias for leakage minimization depends on device structure and doping profile

Doping profiles 17-20 vary in depth of peak halo doping (nm)

I-V curve for each doping profile
OBB Selection Circuit

- Body bias minimizes leakage when BTBT leakage is approximately equal to the sub-threshold leakage.

Adjust body bias until \( V(A) = V(B) \). Leakage current on the left side of the current mirror is twice the leakage current on the right side.

Leakage Reduction with OBB

- Leakage savings ranged from 14-55% compared to zero body bias case for nominal 70nm and 50nm transistors in Taurus device simulations.

<table>
<thead>
<tr>
<th>Tech.</th>
<th>Temp (°C)</th>
<th>( V_B ) (V)</th>
<th>( I_{OFF} ) (normalized)</th>
<th>( I_{ON} ) (normalized)</th>
<th>( I_{ON}/I_{OFF} )</th>
<th>Leakage Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>70nm</td>
<td>25</td>
<td>0</td>
<td>1</td>
<td>97115</td>
<td>97115</td>
<td>43%</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>-0.16</td>
<td>0.57</td>
<td>91005</td>
<td>159657</td>
<td></td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>0</td>
<td>5.14</td>
<td>120673</td>
<td>23477</td>
<td>55%</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>-0.20</td>
<td>2.30</td>
<td>118269</td>
<td>51421</td>
<td></td>
</tr>
<tr>
<td>50nm</td>
<td>25</td>
<td>0</td>
<td>1</td>
<td>3478</td>
<td>3478</td>
<td>45%</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>0.15</td>
<td>0.55</td>
<td>3992</td>
<td>7258</td>
<td></td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>0</td>
<td>2.51</td>
<td>4044</td>
<td>1611</td>
<td>14%</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>0.09</td>
<td>2.15</td>
<td>4286</td>
<td>1993</td>
<td></td>
</tr>
</tbody>
</table>
Variation Effects on OBB

• Optimal Body Bias is affected by variations in:
  – Supply voltage
  – Gate length
  – Doping Profile
  – Temperature

Variation in Supply Voltage

Variation in Halo Doping Location

Variation Reduction with OBB

• OBB selector circuit automatically adjusts to process and operating conditions to reduce variation in leakage
  – Leakage values determined for 50nm transistors with Gaussian distributed parameter variations. Spread of leakage values reduced with OBB compared to ZBB
Variation Reduction Results

- OBB reduces mean leakage by 30-37%
- OBB reduces the spread of leakage values by 40-71%

Taurus Device simulation results for 50nm nmos with Gaussian distributed parameter variations

<table>
<thead>
<tr>
<th>Device Variation</th>
<th>Leakage Variation</th>
<th>Leakage Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>µ @ ZBB [A]</td>
<td>µ @ OBB [A]</td>
</tr>
<tr>
<td>Length</td>
<td>1.14e-7</td>
<td>7.97e-8</td>
</tr>
<tr>
<td>VDD</td>
<td>1.20e-7</td>
<td>7.87e-8</td>
</tr>
<tr>
<td>Peak Halo</td>
<td>1.27e-7</td>
<td>7.96e-8</td>
</tr>
</tbody>
</table>

Dual Threshold CMOS

- Low-\(V_{th}\) transistors in critical path for high performance
- Some high-\(V_{th}\) transistors in non-critical paths to reduce leakage
Total Power of 32-bit Adder

- Total power can be reduced by 9% for high activity
- Total power can be reduced by 22% at low activity

Process Variation & Dual-Vt
MTCMOS

- Multi-Threshold CMOS (From S. Mutoh, etc. JSSC 1995)

- In active mode:
  - SL=0, MP and MN are “on” VDDV and VSSV almost function as VDD and VSS.

- In standby mode:
  - SL=1, MP and MN are “off” leakage is suppressed.

MTCMOS (cont’d)

- Only one type of high-Vth sleep control transistor is enough

- NMOS size smaller
  - NMOS insertion is preferable
MTCMOS (cont’d)

- Advantage:
  - Effective for standby leakage reduction
  - Easily implemented based on existing circuits
  - 1-V MTCMOS DSP chip for mobile phone application (1996)

- Disadvantage:
  - Increase area and delay
  - If data retention is required in standby mode, an extra high-\(V_{th}\) memory circuit is needed

SCCMOS

- Super Cut-off CMOS (From H. Kawaguchi, ISSCC, 1998)
- Single-low-\(V_{th}\) circuit
  - Low-\(V_{th}\) sleep control transistor with smaller size
  - Minimal \(V_{dd}\) is lower than that of MTCMOS

- A gate bias generator is required
VTCMOS

• Variable Threshold CMOS (from T. Kuroda, ISSCC, 1996)

• In active mode:
  – Zero or slightly forward body bias for high speed

• In standby mode:
  – Deep reverse body bias for low leakage

• Triple well technology required

DTMOS

• Dynamic Threshold CMOS
  – from F. Assaderaghi, IEDM, 1994

• Vth altered dynamically to suit the operation state of the circuit

• Vdd<0.6V

• Triple well required for BULK silicon technology

• DTMOS in partially-depleted SOI

SiO2

Si
DGDT SOI CMOS

- Double Gate Dynamic Threshold SOI CMOS
- Asymmetrical double gate fully-depleted SOI MOSFET
- Front gate: conducting gate
  Back gate: controlling gate

Design of Nanometer Caches: Low-Leakage
Scaling and Other Leakage Components

- Leakage Components
  - Subthreshold Leakage
  - Gate Leakage
  - Reverse-biased Junction Band-To-Band-Tunneling (BTBT) Leakage.
  - Others

- Intrinsic parameter variations:
  - Channel length and width
  - Variations due to line edge roughness
  - Threshold voltage (Vt) variations due to random dopant fluctuation

Device parameters are no longer deterministic
Leakage Power in Cache

30% of L1 Cache Power
80% of L2 Cache Power

@ 0.13µm process

Cache is large leakage power consuming block in a high performance processor
Solution: Put idle part of the cache in low leakage mode

SRAM Leakage Reduction Schemes

<table>
<thead>
<tr>
<th>Schemes</th>
<th>Source Biasing</th>
<th>Fwd/Reverse Body-Biasing</th>
<th>Dynamic $V_{DD}$ ($V_{BL}$)</th>
<th>Floating Bitlines ($V_{BL}$, $V_{BLB}$)</th>
<th>Negative Word Line ($V_{wl}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$V_{ml}$</td>
<td>$V_{pwell}$, $V_{nwell}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leakage reduction</td>
<td>Sub: ↓↓ Gate: ↓↓</td>
<td>Sub: ↓↓ BTBT: (RBB)</td>
<td>Sub: ↓ gate: ↑ Bitline leak: -</td>
<td>Sub: ↓ Gate: ↓</td>
<td>Sub: ↓ Gate: ↑</td>
</tr>
<tr>
<td>Delay</td>
<td>*Delay increase</td>
<td>No delay increase</td>
<td>No delay increase</td>
<td>No delay increase</td>
<td>No delay increase</td>
</tr>
<tr>
<td>Overhead</td>
<td>Low transition overhead</td>
<td>Large transition overhead</td>
<td>Large transition overhead</td>
<td>*Precharge latency overhead</td>
<td>*Low charge pump efficiency</td>
</tr>
<tr>
<td>Stability</td>
<td>Impact on SER</td>
<td>No impact on SER</td>
<td>*Worst SER</td>
<td>No impact on SER</td>
<td>No impact on SER, voltage stress</td>
</tr>
</tbody>
</table>

*Precharge latency overhead
*Low charge pump efficiency
Conventional Cell Leakage Paths

- $V_{dd}$ to ground path
- Bitline to ground path

Gated-Ground (Source-Biased) SRAM

- Gating options: NMOS, Dual-$V_t$, PMOS
Leakage Reduction in Diode Footed Cache

Dashed arrows represent improved leakage components.

Voltages across terminals get reduced by Vd (diode intrinsic voltage).
Reduces gate and subthreshold leakage.

Gated-Ground Transistor Sharing

Virtual $V_{dd}$

Gate Control
16K-Byte SRAM Organization

- Active leakage reduction SRAM
- Distributed sleep transistors
- SRAM block turned on ahead of time
- Self-decay circuit for low dynamic power overhead

2x16K-Byte SRAM Testchip

<table>
<thead>
<tr>
<th>Technology</th>
<th>180nm 6-metal CMOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chip Size</td>
<td>3.3X2.9 mm²</td>
</tr>
<tr>
<td>Supply Voltage</td>
<td>1.8V</td>
</tr>
<tr>
<td>Threshold Voltage</td>
<td>NMOS: 0.53V PMOS: -0.53V</td>
</tr>
<tr>
<td>Read Access Cycle</td>
<td>984MHz @ 1.8V, RT</td>
</tr>
<tr>
<td>Active Current</td>
<td>0.14mW/MHz @ 1.8V</td>
</tr>
<tr>
<td>Standby Current</td>
<td>7.27μA (16KB array)</td>
</tr>
</tbody>
</table>

Kim, Roy, ISSCC’05
Mesured Leakage Reduction

- 94.2% total leakage reduction at VGND=0.9V
- Raising VGND also reduces gate tunneling leakage

Forward Body-Biased Cache (50nm)

- Previous techniques: use circuit/arch. to lower leakage
- This technique: use dev/ckt/arch opt. to lower leakage
- Main idea: high Vt device + forward body-biasing
32x32 Forward Body-Biased Sub-array

Comparison

Conventional  SBSRAM  FBSRAM

\[ V_{T} = 270 \text{mV} \]

- SBSRAM (DRG) has been proven with Si measurements
- Dynamic VDD, RBB SRAM have fundamental design issues
- MEDICI: gate/BTBT leakage is also modeled
SBSRAM and FBSRAM are designed to give iso-leakage savings
64% total leakage reduction including overhead

Another Application: Data Retention Flip-Flop

Cross-coupled inverters are cores of any flip-flops
Cross-coupled inverters retain data under gated ground
Data and clock gating is required to preserve data
Successful fabrication and test:
- 16-bit shift-register based on our data-retention FF

40% power reduction by enabling power-down mode
Computing with Leakage for Ultralow Power: Digital Subthreshold Logic

Subthreshold Operation

\[ I_{DS} \alpha \exp(V_{GS}-V_{TH}) \]

and not \((V_{GS}-V_{TH})\)

\[ C_{GATE} < C_{OX} \]
Is scaling necessary?

Device for sub-threshold operation?

Scaling & Subthreshold Operation

- Reduced $L \Rightarrow$ Reduced capacitance

 Scaling is essential even for subthreshold operation
Proposed device vs. Std. Device

@ *iso-performance* (3.4ns)

<table>
<thead>
<tr>
<th>Technology Node (nm)</th>
<th>Average Power (X 10^-7 J)</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>500mV</td>
</tr>
<tr>
<td>180</td>
<td>420mV</td>
</tr>
<tr>
<td>130</td>
<td>280mV</td>
</tr>
<tr>
<td>90</td>
<td>200mV, 180mV</td>
</tr>
</tbody>
</table>

48%

Raychowdhury, Paul, Roy; IEEE TED, Feb’05, ISLPED’04

Circuit Considerations

- **CMOS-NAND**
  - **PUP**
  - **PDN**

- **Pseudo-NMOS (NAND)**
  - **PUP**
  - **PDN**

Pseudo-NMOS over CMOS
- Less power
- Faster operation
Pseudo-NMOS logic

VTC of an Inverter (350nm Tech)

Std. operation ($V_{dd} = 3.3V$)  

Sub-threshold (0.5V)

Pseudo NMOS logic is good for sub-threshold operation

Improvement Through Circuit Innovation

Pseudo-NMOS over CMOS (sub-threshold)
- Faster operation
- Reasonable power

Pseudo-NMOS logic is suitable for Sub-threshold operation
**Architecture Optimization**

**Pipelining**

- IN → Latch → Logic → Latch → Logic → ... → Logic → Latch → OUT

**Parallelism**

- IN → Logic → Logic → Logic → Control → OUT

---

**Architecture Optimization**

- 5-Tap FIR filter
- 90nm Predictive Tech.

**Pipelining**

- Optimum no. of pipeline stages and parallel blocks need to be chosen
Dev/Cir/Arc Co-design: Summary

90nm Predictive Tech.
5-Tap FIR Filter

0.8V
0.7V
0.6V
0.5V
0.4V
0.3V
0.2V
0.15V

CMOS to Pseudo-NMOS

Standard CMOS

Optimal parallelization and pipelining

Device optimization

Under review, TVLSI

Other Device Options

- Improve performance ??
- Reduce Power ??
Underlap DG-SOI

Device Dimension

- $L_{\text{gate}} = 50\text{nm}$
- $L_{\text{un}} = 50\text{nm}$
- $T_{\text{ox}} = 3\text{nm}$
- $T_{\text{si}} = 10\text{nm}$
- $V_{dd} = 200\text{mV}$

$C_G$ reduces by $\sim 10\times$

RO: Delay improved by $40\%$

PDP reduced by $7.3\times$

Power-Throughput Trade-off in SOI and Bulk Technologies

DG SOI is better suited for subthreshold operation
Example Application: Adaptive Filters in Digital Hearing Aid Devices

- Adaptive filters are used to cancel out the annoying high intensity oscillation
  - Acoustic feedback through the human body
  - Hearing aid output leaking into the input again
Prototype Adaptive Filter For Hearing Aid Devices

- Subtracts the unwanted acoustic feedback noise
- Reference signal: delayed error output

Filter Architecture With Single Functional Unit

<table>
<thead>
<tr>
<th># of FU</th>
<th>Algorithm</th>
</tr>
</thead>
</table>
| Single  | \[ W(n+1) = W(n) + \mu e(n)U(n) \]  
|         | \[ e(n) = d(n) - W^T(n)U(n) \]  

\[ W(n) = [a_0(n) a_1(n) \cdots a_{K-1}(n)]^T : \text{Filter coefficients} \]
\[ U(n) = [u(n) u(n-1) \cdots u(n-N+1)]^T : \text{Data input} \]
\[ t_m : \text{Multiplier delay, } t_a : \text{Adder delay, } N : \text{Filter length} \]

LMS filter with a single FU
CLK = 22kHz*34cycle/sample  
= 748 kHz

- Not suitable for ultra-low voltage operation
- LMS algorithm cannot be implemented in a parallel architecture
Filter Architecture With Multiple Functional Units

<table>
<thead>
<tr>
<th># of FU</th>
<th>Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple</td>
<td>[ W(n+1) = W(n) + \mu e(n-N)U(n-N) ]</td>
</tr>
<tr>
<td></td>
<td>[ e(n-N) = d(n-N) - W^T(n-N)U(n-N) ]</td>
</tr>
</tbody>
</table>

\[ W(n) = [a_0(n) \; a_1(n) \; \ldots \; a_{N-1}(n)]^T : \text{Filter coefficients} \]
\[ U(n) = [u(n) \; u(n-1) \; \ldots \; u(n-N+1)]^T : \text{Data input} \]
\[ t_m : \text{Multiplier delay,} \; t_a : \text{Adder delay,} \; N : \text{Filter length} \]


- DLMS algorithm enables parallel architecture
- Trading off area for power

Power Consumption
- Architecture & Logic Styles -

<table>
<thead>
<tr>
<th>Implementation</th>
<th>Clock frequency</th>
<th>Vdd</th>
<th>Energy/Operation</th>
<th># of Transistors</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ Sub-CMOS</td>
<td>748 kHz</td>
<td>650 mV</td>
<td>19.1 nJ</td>
<td>31k</td>
</tr>
<tr>
<td>+ Sub-CMOS</td>
<td>22 kHz</td>
<td>450 mV</td>
<td>2.47 nJ</td>
<td>111k</td>
</tr>
<tr>
<td>+ Sub-Pseudo NMOS</td>
<td>22 kHz</td>
<td>400 mV</td>
<td>1.77 nJ</td>
<td>86k</td>
</tr>
</tbody>
</table>

- Parallel architecture lowers the clock rate, reduces power dissipation by 87%
- Pseudo NMOS logic styles provides another 28% reduction