Part II

Leakage Reduction Techniques
Leakage Reduction Mechanisms

\[ I_{\text{leakage}} = I_0 \exp \left( nV_{\text{th}} \right) * \left( 1 - \exp \left( \frac{-V_{\text{ds}}}{V_{\text{th}}} \right) \right) \]

- **Increase** \( V_t \)
  - dual threshold Voltage/ MTCMOS/ VTCMOS
- **Increase** \( V_S \)
  - source biasing, self reverse biasing, stack effect
- **Decrease** \( V_G \)
  - Super cut-off CMOS
- **Decrease** \( V_{\text{DS}} \)
  - not practical (CMOS output full rail)
Standby and Active Leakage

- $V_t$ scaling causes exponential increase in leakage currents
- Dynamic power reduced with supply scaling
- Standby periods can be long (Burst Mode operation - cell phone, pager)
- Standby leakage problem more immediate
- Active leakage control can become important too
Source Biasing Principle

- $V_t$ shift due to body effect $\gamma$
- $V_{GS}$ becomes negative
- switched source impedance, self reverse biasing, stack effect

$V_D = V_{DD}$
$V_G = 0$
$V_S > 0$ (source bias)
Switched Source Impedance

- R as source impedance
- Estimate >1000X reduction in standby leakage


Fig. 1. Principle of switched-source-impedance CMOS circuit: (a) schematic circuit diagram, (b) mechanism of subthreshold-current reduction.
Self Reverse Biasing

- Off device as source impedance
- Applied to DRAM decoded drivers
- Only one driver turns on (small $M_C$ needed)
- $M_{D1}$-$M_{D2}$ reverse biased during standby state
- 1000X reduction

Fig. 3. Subthreshold-current-reduced Decoded-Driver by self-reverse biasing. $W_C$: Gate width of $M_C$. $V_{TC}$: Threshold voltage of $M_C$. $\Delta V_{SRB}$: Self-reverse biasing voltage at the steady state. $W_D$: Gate width of $M_C$. $V_{TD}$: Threshold voltage of $M_D$. $I$: Operation current. $i$: Subthreshold current. $n$: Number of Decoded-Driver. $C$: Parasitic capacitance of common source.

Stack Effect By Vector Activation

Activate with predetermined input vector stored in latch

1
0

High leakage vector

1

Low leakage vector

~10X current reduction for 2 stack

Eg. 32-bit static CMOS Kogg-Stone adder ~2X reduction in total leakage current

- limited by number of stacks available
- proper choice of activating vector (NP-hard algorithm -> use of heuristics)
- internal node settling time can be long
- single stacks are still HIGH leakage

Stack Forcing Principle

0.13 μm; 1.5 V

10-30X leakage reduction
~100% higher delay

Force low-\(V_T\) stacks in non-critical paths to reduce leakage
Stack Forcing Effectiveness

32-bit \( \mu \)P instruction decode block 0.13 \( \mu \)m; 1.5 V

Frequency of operation: 1.0 GHz
Active power @ 10% activity: 45.9 mW
All Low-Vt leakage: 39.1 mW
Dual-Vt leakage: 9.0 mW
Forced stack in low-Vt: 13.2 mW
High-Vt usage: 94.2%
Forced stack usage: 70.2%

Leakage power reduction
4.3X with dual-\( V_t \), 3X with stack forcing

Leakage Control Stack Devices

- Single $V_t$ leakage reduction mechanism
- Insertion of extra stack devices (in addition to vector activation)
- Sleep devices can be shared among several gates
- Gives further 35% - 90% reduction compared to state dependence alone
- Boils down to single $V_t$ version of MTCMOS (to be discussed)

Dual $V_t$ CMOS

\[
I_{\text{leakage}} = I_0 \exp \left( nV_{th} \right) \left( 1 - \exp \left( -V_{ds} / V_{th} \right) \right)
\]

- Dual $V_t$ more effective at reducing leakage currents than source biasing
- Multiple threshold technologies more common
- For $S=85$ mV/Decade
  - each 255mV shift = 3 orders of magnitude reduction
- Low $V_t$ device= fast, high leakage
- High $V_t$ device= slow, low leakage
- Achieves both Active and Standby leakage reduction
Dual $V_t$ Gate Partitioning

A simple approach: Use Low $V_t$ cells for time-critical paths to improve performance

  Use of LVT in 4% of standard cells yield 6.5% performance improvement

  LVT + HVT improves performance by 12.5% (all LVT causes standby current to be so large as to cause thermal runaway)
Dual \( V_t \) Optimization

- Initially assume all LVT (for best performance)
- Some non-critical gates can be made HVT
- Choice of high \( V_t \) determines mixture
- Proposal for breadth first search algorithm to assign optimal high \( V_t \) value.

A Dual $V_t$ Partitioning Algorithm


- Initially all low $V_t$
- For each node (gate) in graph calculate
  - Arrival time, Departure time, Propagation delay, graph level
- From output to input (back tracing level-by-level)
  - Determine slack availability for each node in level
  - Gates with enough slack set to high $V_t$
  - level = level -1
- Simulate and reiterate with other $V_t$ choices

\[ \Delta X \text{ (slack)} = T_{\text{max arrv}}(Y) - T_{\text{dep}}(X) + \Delta Y \]

If $\Delta X > 0$, $X$ can be made high $V_t$

Update graph with new $T_p$, $T_{\text{dep}}$, $\Delta X$

Move to next node/ level
Advanced Dual $V_t$ Optimization

- Gate level dual $V_t$ -> transistor level dual $V_t$
  
  L. Wei, "Mixed-$V_{th}$ (MVT) CMOS Circuit Design Meth. for Low Power Appl," DAC 1999
  
  - Improved leakage reduction
  - More involved partitioning algorithm (traverse transistors level by level)

- Combine dual $V_t$ with transistor sizing:
  
  S. Sirichotiyakul, D. Blaauw, "Stand-by Power Minimization through Simultaneous $V_t$ Selection and Circuit Sizing," DAC 1999

  - High $V_t$ to low $V_t$ with same sizing can be too fast
  - Low $V_t$ increases node capacitance seen by crossing paths
  - Use “Dominant Leakage State” + probability to estimate total leakage
  - Too complex to optimize: use heuristic approach
  - 1) Choose some $V_t$ low for performance 2) resize circuit to win back area 3) repeat

![Diagram](image.png)

Figure 4. $V_t$ selection and redistribution of area, two views
CAD for Dual $V_t$ Optimization

- Leakage reduction principle simple
- Difficult to optimally choose parameters $V_{th}$, $V_{tl}$, $V_{DD}$, device selection, transistor sizing
- Need to develop fast, efficient CAD tools
**MTCMOS Principle**

**Active Mode**
Low $V_t$ circuit operation (or combined)

**Standby Mode**
Disconnect power supplies through High $V_t$ devices
- For $S=85$ mV/Decade, $\Delta V_t = 225$ mV
  - ~1000X reduction

Use of LVT sleep with +/- gate (Super Cut-off/ Multi-Voltage CMOS, M. Stan, ISLPED 1998)

For fine grain sleep control
Sequential circuits must retain state

MTCMOS Sleep Sizing

Virtual Ground Bounce
- Gate drive decreases
- Body effect increases $V_t$
- Reverse conduction noise concerns

Main Design Issue in MTCMOS
- Properly size sleep transistor
Input Vector Impact

<table>
<thead>
<tr>
<th>Vector</th>
<th>CMOS Delay</th>
<th>% Degr (W/L=5.4%)</th>
<th>% Degr (W/L=18%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2.58 ns</td>
<td>15.4%</td>
<td>4.6%</td>
</tr>
<tr>
<td>B</td>
<td>2.59 ns</td>
<td>4.7%</td>
<td>1.6%</td>
</tr>
</tbody>
</table>

Hierarchical Sizing Approach

- Compute effective sleep resistor for each gate
  - Sets Maximum Gate Degradation
  - Overall delay is guaranteed
- Mutual exclusive gates can share common sleep transistor
- Applied at multiple hierarchical levels

J. Kao, et al., “MTCMOS Hierarchical Sizing Based on Mutual Exclusive Discharge Patterns,” DAC 1998
Gate Clustering

- Cluster gates that have partially overlapping discharge currents
- Group such that Current Budget (to ensure $V_x$ gives 5% degredation) is maintained
- Use operations research clustering techniques
  - Bin Packing
    - assign gate ($I_j$) to bins so that total current $< I_{\text{max}}$ and #bins minimized
  - Set Partitioning
    - Extension of BP including cost function related to routing complexities

Assumption on predictability of current discharge patterns...

MTCMOS Sleep Sizing TBD

- Need for improved sleep transistor sizing algorithms
- Static, functional timing techniques to better characterize MTCMOS discharge patterns
- Apply ideas from similar CAD research on power /gnd noise
  - + many others …
MTCMOS Sequential Circuits

• MTCMOS Combinational Circuits
  – Simple operation
  – Difficulty in sizing/ distributing sleep transistors

• MTCMOS Sequential Circuits
  – Virtual power/ gnd disconnected during sleep
    • Nodes will float
  – Techniques needed to maintain state
    • Need always powered circuits
    • Must avoid sneak leakage paths
Basic MTCMOS Latch

- Use of always powered CMOS gates

Balloon Flip Flop

- HVT storage “balloon” decoupled from LVT logic
- LVT blocks can share common virtual pwr/gnd
- Elimination of sneak leakage paths
- Complicated signalling


Fig. 9. A balloon circuit applied to a DFF circuit (clock-dependent type).
Sneak Leakage Paths

• Sneak paths from MTCMOS/ CMOS interaction
• Leakage currents from $V_{cc}$ to ground without passing through off high $V_{t}$ device
• Need to utilize:
  – both polarity sleep devices
  – local sleep (non shared sleep devices)
  – novel structures

Sneak Leakage Path From Parallel Combinations

- Need for both polarity high $V_t$ sleep devices
Sneak Leakage Path Through Low Threshold Passgate

- Need for both polarity high \( V_t \) sleep devices
Sneak Leakage Path From Reverse Conduction Paths

- Need for localized, non shared, high $V_t$ sleep devices
Improved MTCMOS Flip Flop

- Careful consideration of sneak leakage paths yields improved implementation
Leakage Feedback Gate

- Sufficient if either VCC or VSS path is cutoff
- Proper cutoff path yields actively driven output
- Low $V_t$ operation + actively driven low leakage state
  - directly imported into CMOS structures
Leakage Feedback Flip Flop

- Virtually no extra loading -> performance is better than standard MTCMOS FF
- Same operation as a CMOS FF
Leakage Feedback Effect

- Output data holds even if input floats
  - held by leakage mismatch
- Potential charge sharing if inputs change
Leakage Induced DC Operating Point

I-V Curves High Vt PMOS and Low Vt NMOS

- Low Vt NMOS
- High Vt PMOS

$V_t = 0.4$

$V_t = 0.2$
MTCMOS / CMOS Interface

- Leakage feedback gate natural interface block between MTCMOS logic and CMOS logic
Dynamic Leakage Feedback FF

- Operates like standard dynamic FF during active mode
- Retains state during the standby mode (held by leakage)
Imbedded Dual $V_t$

- Logic gates have internal HVT and LVT devices
- No extra series HVT transistor required
- Suppose special case:
  - known sleep configuration
  - one transition direction more critical
Dual $V_t$ Domino Gate

- Evaluate through LVT devices
- Precharge through HVT devices

Clock Delayed Domino Logic

(\(\phi_1\) Pipeline Stage)

- Clock path matches evaluate path
- NMOS series transistors can be eliminated
Leakage In Dual $V_t$ Domino Gate

Sleep condition during evaluate mode
Sleep Mode Condition

- Clear pipeline before sleep
- Put gates in EVALUATION mode
- Inputs to all gates must be high

(Dual $V_t$ domino DATAPATH)
Variable Threshold CMOS (VTCMOS)

- Body effect to change device $V_t$
- Standby leakage reduction with maximum reverse bias
- Triple well structure

$$V_t = V_{t0} + \gamma\left(\sqrt{2\phi_B} - V_{BB} - \sqrt{2\phi_B}\right)$$
VT.CMOS Example

T. Kuroda, et al, “A 0.9V, 150Mhz, 10mW, 4mm², 2-DCT Core Processor with Variable Vt Scheme, “ JSSC Nov. 1996

- VT.CMOS principle applied to 4-mm² DCT core processor
- SSB increases Vt (more reverse bias)
- SCI decreases Vt (Standby -> Sleep)
- Leakage reduction
  0.1mA active -> 10nA sleep (2.8v ∆VBB)
  4 orders of magnitude
- Dynamically tunes Vt (by matching leakage current monitor) to minimize Vt variation

Fig. 3. VT block diagram.

Fig. 4. Substrate-bias control in VT.
VTCMOS Pros/Cons

**PROS:**
- Significant standby leakage reduction
- Memory elements retain state
- No transistor sizing/ partitioning required
- **Dynamically tunable** $V_t$ **during runtime**

**CONS:**
- Requires expensive triple well process
- Body factor decreases with scaling

- Dynamically tune $V_t$ so that critical path speed matched clock period
- Reduces chip-to-chip parameter variations
- Reverse bias:
  - Operate only as fast as necessary (reduces excess active leakage)
- Forward bias:
  - Speeds up slow chips
- Standby leakage with maximum reverse bias
- Also known as Adaptive Body Biasing (ABB)
Adaptive Supply & Body Bias (ASB)

- Dynamically tune both $V_{DD}$ & $V_t$ as operating conditions change
- Trade-off between dynamic power ($V_{DD}$ knob), leakage power ($V_t$)
- Minimize total ACTIVE power consumption
  (higher active leakage current at expense of lowering dynamic power)

Power vs. $V_{DD}$ (implicit $V_t$) for fixed frequency

Optimal $V_{DD}/V_T$ Selection

- Optimal $V_{DD}$ & $V_T$ target changes with operating conditions
  - e.g. Varying Workload
- Low frequencies high $V_T$ more optimal
  - reduce leakage at expense of increased dynamic
- High Frequencies low $V_{DD}$ more optimal
  - reduce dynamic at expense of increased leakage
$V_{DD}/V_T$ Optimization vs. DVS

- Dynamic voltage scaling ignores $V_T$ influence
- DVS is sub-optimal over the frequency range
Summary

- Subthreshold leakage currents will grow exponentially
- Need to manage during STANDBY and ACTIVE
- Three main principles
  - Source Biasing
  - Multiple threshold voltage
  - Body biasing
- Need for CAD tools to model leakage currents
- Need for CAD tools to implement leakage reduction principles