A Leakage Current Replica Keeper for Dynamic Circuits

Yolin Lih, Nestoras Tzartzanis, and William W. Walker

IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 42, NO. 1, JANUARY 2007

By Eric Liskay
Introduction

- **Dynamic Logic**
  - Clock signal is used to evaluate combinational logic
  - Two Phases
    - Clock is low: Pre-charge / Setup Phase
    - Clock is high: Evaluation Phase
  - Output can decay if clock speed is too slow
  - Good for high-speed OR and AND-OR Gates and multi-port memories
Introduction

- Keepers
  - Needed with dynamic gates to maintain a high state during evaluation
  - Without a keeper, a minimum clock frequency must be maintained at all times, or the clock must be stopped only in the pre-charge state
    - This would make two-phase dynamic design impossible and complicating design-for-test methodologies such as scan and IDDQ

Fig. 1. Generalized conventional dynamic gate topology.
Introduction

- Conventional Keeper
  - A small pFET pull-up transistor that must satisfy two constraints
    1. In the slow-pFET/fast-nFET (sPfN) process corner, the keeper must source enough current to overpower the nFET logic stack leakage current
    2. In the fast-pFET/slow-nFET (fPfN) process corner, the keeper must be weak enough so that a single nFET leg can pull the dynamic node quickly enough through the switching threshold of the succeeding static gate in order to meet the delay specifications.
  - Contradictory, results in maximum gate width
Introduction

• The Problem
  • Threshold voltages \( (V_t) \) are scaled down along with smaller CMOS processes
  • To satisfy constraint 1, there needs to be an exponential increase in keeper transistor width because of the exponential dependence of off-current on \( V_t \)
  • To then satisfy constraint 2, the maximum dynamic gate logic width must decrease
  • Memories are getting larger as the technology scales
  • Partitioning of read paths requires narrower AND-OR and OR
  • Performance scaling of memories at the same rate as the basic (FO4 inverter) delay becomes impossible
  • The result is “the demise of the dynamic gate”
Introduction

- Possible Solution - Conditional keeper circuit
  - Requires an additional five extra FETs in each gate plus a portion of the shared delay circuit
  - Extra pFET controlled by a delayed clock signal and the dynamic node
  - This pFET is conditionally enabled during evaluation after a time set by the delayed clock signal if the dynamic node has not already been discharged
Introduction

- **Possible Solution - Adaptive keeper circuit**
  - Keeper transistors are conditionally enabled based on a circuit that estimates the process corner or temperature
  - Adds four pFETs per gate plus a portion of the shared circuitry that generates the control signals
  - Neither of these options tracks the two critical process corners (fast-P/slow-N and slow-P/fast-N)
The LCR Keeper

- **Presented in this paper: Leakage Current Replica Keeper**
  - Overhead is only one pFET per dynamic gate plus a portion of the shared current mirror circuit
  - Uses a conventional analog current mirror that can track any process corner, voltage, or temperature
    - Can be shared between multiple dynamic gates
    - Cannot track on-die variations

Fig. 4. LCR keeper dynamic gate topology.

\[
\text{Fig. 4. LCR keeper dynamic gate topology.}
\]
Circuit Design Principles

Fig. 1. Generalized conventional dynamic gate topology.

- **AND-OR Dynamic Gate**
- **N+1 pulldown legs**
- **DN: Dynamic Node**
- **Inputs:** $A_0,...,A_n$ and $B_0,...,B_n$
**Circuit Design Principles**

![Generalized conventional dynamic gate topology.](image)

**Example Application**

- Single-ended read bit-line in a memory
- DN is the bit-line
- $A_0,...,A_n$ are the read word lines
- $B_0,...,B_n$ represent the bits stored in the cells
- Worst case leakage current in the pull-down network occurs when $A_0-A_n = 0$ and $B_0-B_n = 1$. 

---

*Fig. 1. Generalized conventional dynamic gate topology.*
The LCR Keeper

Fig. 4. LCR keeper dynamic gate topology.

- The current mirror tracks the leakage current and copies it into the dynamic gate through $p1$
- Mirror is constructed to draw a current of $sf \cdot I_{\text{leak}}$
- $nrpl$ is used in the current mirror as the replica of the worst case leakage current
  - Its size is derived from the sizes of nFETS $n0$ to $nn$
  - Assuming $p1$ and $p3$ have the same dimensions, the width of $nrpl$ is set to be equal to the sum of the widths of $n0, \ldots, nn$ times the safety factor $sf$
The LCR Keeper

- DN must be pulled close to $V_{DD}$ in order to retain the noise immunity of the dynamic gate
- Node voltage $V_{KPR}$ is critical
  - If $V_{KPR}$ is too low, the replica fails to track the dynamic gate leakage current since $V_{DS}$ of $nrpl$ would deviate significantly from the $V_{DS}$ of $n0,...,nn$
  - If $V_{KPR}$ is too high, the sensitivity of the pFET threshold voltage variation increases since $p1$ operates in weak-overdrive
The Safety Factor

\[ sf = \frac{W_{nrpl}}{\sum W_{ni}} \cdot \frac{W_{p1}}{W_{p3}} \]

- Set by ratioing the transistor geometries
- Assumes \( p1 \) and \( p3 \) have the same channel length and that all nFETs have the same channel length
- Forces \( p1 \) into the triode region
The Safety Factor

\[ sf = s f_{\Delta \text{leak}} \times s f_{\text{mirror}} \]

- Margins against on-die random \( V_t \) variation between the transistors in the dynamic gate and current mirror
- \( s f_{\Delta \text{leak}} \) protects against nFET \( V_t \) variation between \( nrpl \) and \( n0, \ldots, nn \)
- \( s f_{\text{mirror}} \) pulls \( pl \) into the triode region and protects against pFET \( V_t \) variation between \( pl \) and \( p3 \)
Deriving The Safety Factor

\[ sf = sf_{\Delta \text{leak}} \times sf_{\text{mirror}} \]

\[ sf_{\Delta \text{leak}} = \frac{\max(I_{dsn_{\text{subVt}}})}{\min(I_{dsn_{\text{subVt}}})}. \]

Maximum over minimum nFET leakage current

\[ I_{dsn_{\text{subVt}}} = \frac{W_n}{L_n} \cdot I_o \cdot \exp\left(\frac{V_{gsn} - V_{tn}}{n \cdot V_T}\right) \cdot \left(1 - \exp\left(-\frac{V_{dsn}}{V_T}\right)\right). \]

Subthreshold drain-to-source current (I_{DS})

\[ sf_{\Delta \text{leak}} = \exp\left(\frac{V_{tn(\max)} - V_{tn(\min)}}{n \cdot V_T}\right). \]

For devices with equal dimensions and V_{DS}
Deriving The Safety Factor

$s_f_{\text{mirror}} = \frac{\max(\text{Idsp}3_{\text{satur}})}{\min(\text{Idsp}1_{\text{triode}})}$. Maximum $p3$ current over minimum $p1$ current

$\text{Idsp}3_{\text{satur}} = \frac{1}{2} \cdot \mu_p \cdot C_{\text{ox}} \cdot \frac{W_p}{L_p} \cdot \left( V_{\text{gsp}3} - |V_{\text{tp}(\text{min})}| \right)^2$

$\text{Idsp}1_{\text{triode}} = \mu_p \cdot C_{\text{ox}} \cdot \frac{W_p}{L_p} \cdot \left( V_{\text{gsp}1} - |V_{\text{tp}(\text{max})} - \frac{V_{\text{dsp}1}}{2}| \right) \cdot V_{\text{dsp}1}$

$s_f_{\text{mirror}} = \frac{\frac{1}{2} \cdot \left( V_{\text{gsp}3} - |V_{\text{tp}(\text{min})}| \right)^2}{\left( V_{\text{gsp}1} - |V_{\text{tp}(\text{max})} - \frac{V_{\text{dsp}1}}{2}| \right) \cdot V_{\text{dsp}1}}$. For equal device dimensions
Deriving The Safety Factor

\[ s_{f_{\text{mirror}}} = \frac{\frac{1}{2} \cdot (V_{\text{gsp3}} - |V_{tp(\text{min})}|)^2}{(V_{\text{gsp1}} - |V_{tp(\text{max})}| - \frac{V_{\text{dsp1}}}{2}) \cdot V_{\text{dsp1}}}. \]

Given \( V_{\text{gsp}} = V_{\text{gsp1}} = V_{\text{gsp3}} = |V_{KPR} - V_{DD}| = V_{DD} - V_{KPR}. \) And \( V_{\text{dsp1}} = |V_{DN} - V_{DD}| = V_{DD} - V_{DN}. \)

\[ s_{f_{\text{mirror}}} = \frac{\frac{1}{2} \cdot (V_{DD} - V_{KPR} - |V_{tp(\text{min})}|)^2}{\left(\frac{V_{DD}}{2} - V_{KPR} - |V_{tp(\text{max})}| + \frac{V_{DN}}{2}\right) \cdot (V_{DD} - V_{DN})}. \]

\[ s_{f_{\text{mirror}}} = \frac{\frac{1}{2} \cdot (V_{DD(\text{max})} - V_{KPR} - |V_{tp(\text{min})}|)^2}{\left(\frac{V_{DD(\text{min})}}{2} - V_{KPR} - |V_{tp(\text{max})}| + \frac{V_{DN}}{2}\right) \cdot (V_{DD(\text{min})} - V_{DN})}. \]

If we consider supply voltage variation...
The Safety Factor

- **Example**
  - 1.2V 90nm CMOS process
  - Assume $V_{tn} = 120\text{mv} \pm 40\text{mV}$, $V_{tp} = 150\text{mv} \pm 25\text{mV}$
  - Ratio $p3$ and $nprl$ for nominal $V_{KPR} = 0.9\text{V}$
  - Minimum voltage on node DN is set to 1.1V
  - $sf = 7.78 \cdot 2.04 = 5.87$
  - With a 5% supply voltage variation ($V_{DD}=1.17\text{-}1.23\text{V}$)
    - $sf_{\text{mirror}} = 5$, thus $sf = 38.9$

$$sf_{\Delta \text{leak}} = \exp \left( \frac{V_{tn(\text{max})} - V_{tn(\text{min})}}{n \cdot V_T} \right)$$

$$sf_{\Delta \text{leak}} = \exp \left( \frac{160\text{ mV} - 80\text{ mV}}{1.5 \cdot 26\text{ mV}} \right) = 7.78$$

$$sf_{\text{mirror}} = \frac{1}{2} \cdot \left( V_{DD} - V_{KPR} - |V_{tp(\text{min})}| \right)^2$$

$$sf_{\text{mirror}} = \frac{1}{2} \cdot \left( 1.2\text{ V} - 0.9\text{ V} - 0.125\text{ V} \right)^2$$

$$sf_{\text{mirror}} = \frac{1}{2} \cdot (1.2 \text{ V} - 0.9 \text{ V} - 0.175 \text{ V} + \frac{1.1}{2}) \cdot (1.2\text{ V} - 1.1\text{ V})$$

$$sf_{\text{mirror}} = 2.04.$$
Circuit Simulation

Fig. 6. Experimental setup for worst case noise in pull-down nFET network.

- LCR keeper structure evaluated against the conventional keeper
- Same 1.2V 90nm CMOS process at 110°C
- Target application is a multiport memory
- Wide AND-OR structure with 1μm pull-down nFETs
- Both the conventional and the proposed keeper structures were sized to sustain a maximum voltage drop of 10% of \( V_{DD} \) on the dynamic node \( DN \) under the pull-down current conditions for the worst-case with slow-P and fast-N transistors
Simulation results show that the conventional keeper has more delay than the LCR keeper and fails to switch with more than 24 legs.

For 16 and fewer legs, the fast-P/slow-N is slower.

For more than 16 legs, the slow-P/fast-N becomes slower.

- Larger keeper current required to compensate for leakage.
- Current ratio remains almost constant.
SRAM Macro Using LCR

- 1024-word 72-bit 3W/4R SRAM
- Designed and fabricated on same process using LCR keepers in its single-ended dynamic read path
- 32-word x 144-bit blocks
- Only one current mirror needed for the entire SRAM

Fig. 9. Micrograph of the 1024 x 72 3W/4R SRAM macro with block diagram overlay.
SRAM Macro Using LCR

- SRAM storage cell with single-ended reads and differential static writes
- Read and Write decoders implemented using static CMOS
- Previous SRAMs in this technology designed using high-Vt nFETs in the dynamic gates
- LCR allows us of low-Vt nFETs to increase speed
Fig. 11. SRAM single-ended four-stage domino read path.
SRAM Macro Using LCR

- First stage - local bit lines connect 16 cells through an AND-OR dynamic gate
- Second stage - an eight-leg AND-OR dynamic gate used for column select with the local bit line inputs
- Third stage - an OR dynamic gate to select from one of four blocks
- Fourth stage - a five-input dynamic OR gate with one of the inputs fed from the top global bit line and the other from the bottom blocks
SRAM Macro Using LCR

- Due to the differing number of legs and different nFET sizes used in the pull-down stages, ratioing was used to divide the replica current resulting in safety factors between 6 and 10 (very aggressive).
- Used 2x minimum channel length for $p1$ and $p3$ to eliminate channel-length modulation and reduce $V_t$ variation.
SRAM Macro Results

- LCR keeper design compared compared against simulated conventional design without KPR-gated transistors and with high-\(V_t\) dynamic FETs
- Delay from CLK to RWL is identical since both circuits use static logic
- Delay from RWL to RDO improved by 19%
- Overall access time improved by 7.6%

<table>
<thead>
<tr>
<th>Delay Path</th>
<th>Conventional</th>
<th>LCR</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>(CLK \rightarrow RWL)</td>
<td>631ps</td>
<td></td>
<td>same</td>
</tr>
<tr>
<td>(RWL \rightarrow RDO)</td>
<td>415ps</td>
<td>336ps</td>
<td>-19%</td>
</tr>
<tr>
<td>(CLK \rightarrow RDO)</td>
<td>1046ps</td>
<td>967ps</td>
<td>-7.6%</td>
</tr>
</tbody>
</table>

TABLE I
SIMULATED SRAM ACCESS TIME, CONVENTIONAL VERSUS LCR (TYPICAL PROCESS, 1.2 V, 125°C)
SRAM Macro Results

- The measured maximum clock frequency difference between fast-P/slow-N and slow-N/fast-P process corners is 50 MHz with a 32-MHz standard deviation.

- The plot of the effect of temperature on SRAM shows data for the part at room temperature (25°C) and 125°C.
  - $V_{DD} > 1.05$ V, the SRAM operating frequency is reduced at high temperature compared to room temperature as would be expected.
  - $V_{DD} < 1.05$ V, the SRAM operates better at high temperature compared to room temperature.

Fig. 12. Shmoo plot for high-frequency and different process corners (25°C).

Fig. 13. Shmoo plot for varying temperature (typical process).
SRAM Macro Results

- Tested parts from all corners at low frequencies (1-5MHz) at temperature points: 25°C, 62°C and 110°C

- All parts passed the test at 62°C and 110°C

- Subthreshold current increases exponentially with temperature, VKPR is lowered, which reduced the effect of threshold voltage variation

- Parts from all processes except slow showed failures at room temperature

- Failures traced to DC leakage in local or global bit-lines for different parts

- Attributed to variations in $V_t$ between the replica and dynamic gate FETs and the “excessively small” safety factor that was used

<table>
<thead>
<tr>
<th>Voltage (V)</th>
<th>25 °C</th>
<th>62 °C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.30</td>
<td>fffff</td>
<td>1.30</td>
</tr>
<tr>
<td>1.28</td>
<td>fffff</td>
<td>1.28</td>
</tr>
<tr>
<td>1.26</td>
<td>fffff</td>
<td>1.26</td>
</tr>
<tr>
<td>1.24</td>
<td>ffff</td>
<td>1.24</td>
</tr>
<tr>
<td>1.22</td>
<td>ffff</td>
<td>1.22</td>
</tr>
<tr>
<td>1.20</td>
<td>ffff</td>
<td>1.20</td>
</tr>
<tr>
<td>1.18</td>
<td>ffff</td>
<td>1.18</td>
</tr>
<tr>
<td>1.16</td>
<td>ffff</td>
<td>1.16</td>
</tr>
<tr>
<td>1.14</td>
<td>ffff</td>
<td>1.14</td>
</tr>
<tr>
<td>1.12</td>
<td>ffff</td>
<td>1.12</td>
</tr>
<tr>
<td>1.10</td>
<td>ffff</td>
<td>1.10</td>
</tr>
<tr>
<td>1.08</td>
<td>ffff</td>
<td>1.08</td>
</tr>
<tr>
<td>1.06</td>
<td>ffff</td>
<td>1.06</td>
</tr>
<tr>
<td>1.04</td>
<td>ffff</td>
<td>1.04</td>
</tr>
<tr>
<td>1.02</td>
<td>f***</td>
<td>1.02</td>
</tr>
<tr>
<td>1.00</td>
<td>f***</td>
<td>1.00</td>
</tr>
<tr>
<td>0.98</td>
<td>f***</td>
<td>0.98</td>
</tr>
<tr>
<td>0.96</td>
<td>f***</td>
<td>0.96</td>
</tr>
<tr>
<td>0.94</td>
<td>f***</td>
<td>0.94</td>
</tr>
<tr>
<td>0.92</td>
<td>f***</td>
<td>0.92</td>
</tr>
<tr>
<td>0.90</td>
<td>*****</td>
<td>0.90</td>
</tr>
</tbody>
</table>

* = pass, |---| = fail

Frequency (MHz)

Fig. 14. Shmoo plot for low-frequency and varying temperature (typical process)
Conclusions

• LCR keeper requires an overhead of only one FET per dynamic gate + shared replica circuit

• For equal noise margins, LCR keepers result in either more legs or faster gates than conventional keepers

• A fairly large safety factor is needed to account for random on-die variation, especially threshold-voltage variation

  • Aggressively small safety factor used in SRAM macro (6-10) was not sufficient

  • Benefits are still substantial with safety factor as high as 48 according to simulations
Conclusions

• Process variation is expected to increase as the CMOS process scales

• With the LCR keeper it could be possible to use dynamic gates with more legs compared to a conventional keeper

• With the LCR keeper, the use of dynamic logic can be extended until process variation increases limit the maximum number of legs per gate
Further Study

• Possible to statically or dynamically vary the safety factor by adjusting the width of the replica pull-down network in the current-mirror circuit.

• Safety factor can be increased when the clock stops or the frequency is reduced for scan testing.

• Random on-die variation effect can be mitigated by using several parallel-connected current mirror circuits dispersed throughout the covered dynamic gate area instead of using a centralized topology.
Questions?
References


doi: 10.1109/JSSC.2006.885051

URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4039592&isnumber=4039574