Dual-$V_T$ SRAM Cells with Full-Swing Single-Ended Bit Line Sensing for High-Performance On-Chip Cache in 0.13 $\mu$m Technology Generation

Fatih Hamzaoglu, Yibin Ye, Ali Keshavarzi, Kevin Zhang*, Siva Narendra, Shekhar Borkar, Mircea Stan** and Vivek De
Microprocessor Research Labs, Intel Corporation, Hillsboro, OR
* Low Power Design Lab, Intel Corporation, Hillsboro, OR
** Department of ECE, University of Virginia, Charlottesville, VA
fatih@virginia.edu

ABSTRACT
Comparisons among different dual-$V_T$ design choices for a large on-chip cache with single-ended sensing show that the design using a dual-$V_T$ cell and low-$V_T$ peripheral circuits is the best, and provides 10% performance gain with 1.2x larger active leakage power, and 1.6% larger cell area compared to the best design using high-$V_T$ cells.

Keywords
Dual-$V_T$, SRAM, Single-Ended Sensing.

1. INTRODUCTION
Technology and supply voltage ($V_{cc}$) scaling continues to improve logic circuit delay by 30% per technology generation. However, the combined delay of bit line and sense-amplifier in high-performance on-chip cache with differential low-swing sensing is not improving at the same rate because the offset voltage of the sense amplifier does not scale [1]. The resulting divergence between logic circuit delay and bit line delay is further magnified by the unavoidable usage of low threshold voltage ($V_T$) transistors in speed-critical paths of microprocessor logic designs [2-5].

Low-$V_T$ devices have been used in the peripheral circuits of cache with high-$V_T$ cells [6]. A dual-$V_T$ cell, with high $V_{ce}$ for core and low $V_{ce}$ for both bit line and word line with under-drive, has also been evaluated for caches with differential low-swing sensing in sub-1V $V_{cc}$ [7]. However, neither of these techniques can improve bit line delay in high-performance microprocessor designs which use a single maximum $V_{ce}$ dictated by gate-oxide wear-out considerations.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ISLPED '00, Rapallo, Italy.
Copyright 2000 ACM 1-58113-190-9/00/0007...$5.00.
The stability simulation results are summarized in Fig. 4. Both of the dual-$V_T$ cells, DVTC and DVTC2, have worse stability than the HVTC cell, mainly because using low-$V_T$ pass transistors increases maximum read current. The DVTC cell has worse stability than the LVTC cell because current sinking capacity of the pull-down NMOS is smaller. Stability of the DVTC2 cell is best among all the dual-$V_T$ cell designs. Its stability is better than the DVTC cell because using low-$V_T$ PMOS in the cell inverters reduces voltage degradation at the stored ‘1’ node. Although current sinking capacity of the pull-down NMOS in the DVTC2 cell is smaller than that in the LVTC cell, their stabilities are comparable because weaker pull-down NMOS in the DVTC2 cell reduces voltage degradation at the stored ‘1’ node. Pull-down NMOS widths are increased by requisite amounts in the dual-$V_T$ and low-$V_T$ cells to obtain the same stability as the original high-$V_T$ cell. Although this is the most effective way to recover stability, the cell areas increase by 0.8% to 1.6% (Fig. 4b).
4. IMPACT OF LEAKAGE ON NOISE MARGIN IN DUAL-VT CACHE

Different dual-VT cache designs (Fig. 5a) are compared with a conventional design where all transistors, both in the cells and in the peripheral circuits are high-VT. All of the designs use single-ended, full-swing bit line sensing. Schematic of the top half of a single bit line column, containing 64 rows of cells and a dynamic sense amplifier on the 'read' bit line, is shown in Fig. 5b. For single-ended sensing, excessive bit line leakage through the low-VT pass transistors in the cells and degraded noise margins of low-VT peripheral circuits can induce significant noise at the output of the sense amplifier when reading a '1'. Noise simulation conditions (Fig. 6a) which maximize the noise induced by bit line leakage are used. The robustness of the design is considered unacceptable if the overall noise margin degrades by more than 50% due to this additional noise. When low-VT devices are used only in the peripheral circuits (HVTC_LVTP), the noise margin does degrade, but by an amount less than that required to fail the aforementioned robustness criterion (Fig. 6a). However, when we use low-VT devices in the pass transistors of the cells as well as in the peripheral circuits (DVTC_LVTP, DVTC2_LVTP & LVTC_LVTP), bit line voltage droop from the precharged Vcc level due to excessive leakage through the pass transistors is large enough to turn-on the low-VT column select transistors. As a result, a significant portion of bit line noise propagates to the output of the low-VT, dynamic sense amplifier and the design fails to satisfy the robustness criterion (Fig. 6a).

Seven different techniques are evaluated for recovering noise margins of the dual-VT and low-VT designs with minimal delay degradation (Figs. 6b & 6c). A ‘Quality Factor’ (QF) is used here as a metric to compare effectiveness of these techniques for improving either noise margin or delay. QF is defined as the ratio of “% noise change” to “% delay change”, resulting from the usage of a technique. Simulation results (Figs. 6b & 6c) show that replacing the dynamic sense amplifier with a static one (schematic in Fig. 7) is the technique with highest value of QF, and thus, is best for improving noise margin with minimal delay penalty. Increasing width of the column select transistor, on the other hand, is best for improving delay with minimal noise margin degradation.

Figure 4. (a) Stability Ranking of SRAM Cells with no Resizing, and (b) Cell Area Increase for Stability Recovery.

Figure 5. (a) Different Cache Design Choices, and (b) SRAM Cells and Peripheral Circuits.

Figure 6 (a)
5. COMPARISONS OF DELAY, LEAKAGE POWER AND TOTAL POWER

The best technique for noise margin recovery is incorporated into the cache designs with dual-$V_T$ and low-$V_T$ cells by replacing the dynamic sense amplifier with a static one, whose transistor sizes are optimized (Fig. 7) to improve noise margin by the amounts required to meet the robustness criterion. In addition, the best delay improvement technique is applied to all of the cache designs, including the original design with all high-$V_T$ transistors, by increasing the column select transistor widths until further performance gain is marginal.

Simulation results (Fig. 8a) show that using dual-$V_T$ and low-$V_T$ cells with single-ended sensing improves bit line delay by 13% to 16%, compared to a design with high-$V_T$ cell. However, bit line delay improvement is only 6% to 9% when a differential sensing scheme is used. The reason behind this difference is that, while excessive bit line leakage has a direct adverse impact on delay in differential sensing, it degrades only noise margin, not delay, in the single-ended sensing scheme. Performance improvements offered by the dual-$V_T$ and low-$V_T$ cache designs are achieved (Fig. 8b) at the expense of larger leakage power. Leakage power is the dominant component of total active power in a large on-chip cache because, (1) a very small fraction of cells is accessed every cycle, and (2) the junction temperature in a microprocessor is high during active operation. Contributions to active leakage power from the pass transistors in the cell can be reduced by precharging the bit lines only in the one basic sub-block which will be accessed in the next ‘evaluate’ cycle, instead of precharging all the bit lines every cycle. This ‘precharge as needed’ scheme can reduce the active leakage power by 1.6x for designs containing the DVTC cell where leakage through the low-$V_T$ pass transistors is the dominant component of cell leakage power (Fig. 9a).

![Figure 7. Optimized Static Sense Amplifier.](image)

![Figure 8. (a) Bit Line Delay Improvement for Differential and Single-Ended Sensing Schemes, Both with 64 Rows per Bit Line, and (b) Bit Line + Sense Amplifier Delay Comparisons for Different Cache Design Choices.](image)

<table>
<thead>
<tr>
<th>Sensing Scheme</th>
<th>Bit Line Delay Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Differential</td>
<td>Ref 6.4% HVT 6.4% LVTC 8.9%</td>
</tr>
<tr>
<td>Single-Ended</td>
<td>Ref 13.3% HVT 13.3% LVTC 15.7%</td>
</tr>
</tbody>
</table>
Using low-V<sub>T</sub> devices only in the peripheral circuit increases active leakage power of the cache by only 1.5x (Fig. 9b) and standby leakage power by 2.4x (Fig. 9c). At the same time, performance improves by 15% (Fig. 8a). Another 10% performance gain is achieved by using low-V<sub>T</sub> pass transistors in the cell (DVTC) with 1.7x larger active leakage, even when all bit lines are precharged every cycle, and with virtually identical standby leakage. The active leakage power increases by only 1.2x if the ‘precharge as needed’ scheme is used for the bit lines. If low-V<sub>T</sub> is used for devices in the cell inverters as well (DVTC2 and LVTC), active leakage increases by another 2x to 4x, and the standby leakage is another 3x to 8x larger, but virtually no delay improvement is achieved. Using low-V<sub>T</sub> only in the pass transistors causes much smaller increase in leakage power than using low-V<sub>T</sub> in the inverter devices of the cell because of two reasons. First, the column select transistors, which appear in series with the pass transistors, are ‘off’ in more than 99% of the cells. Because column select transistor width on a bit line is significantly smaller than the total width of 64 pass transistors, and because more than one transistor is ‘off’ in a series-connected configuration, the leakage current through the pass transistors is reduced by at least 10x. The second reason is, pass transistors are longer and narrower than the minimum length devices in the cell inverters. The total active power and energy-delay product, including both switching and leakage components, are compared in Fig. 10 for different dual-V<sub>T</sub> design choices. The switching power corresponds to 1 GHz, clock frequency where the cache is accessed every cycle. Clearly, using the DVTC cell with low-V<sub>T</sub> peripheral circuits is the best design choice for single-ended sensing, since it offers 10% performance gain with 8% increase in total power, virtually unchanged energy-delay product, and 1.6% larger cell area compared with the best design using high-V<sub>T</sub> cells (HVTC_LVTP).

6. CONCLUSIONS

We compare cell stability, noise margin, performance and power of different dual-V<sub>T</sub> design choices for large on-chip cache with single-ended, full-swing sensing in a 0.13 µm technology generation. The dual-V<sub>T</sub> design with low-V<sub>T</sub> pass transistors in the cell and low-V<sub>T</sub> peripheral circuits (DVTC_LVTP) provides the best trade-offs in performance, active and standby leakage power, total power, energy-delay product and cell area with adequate noise margin.

7. ACKNOWLEDGEMENTS

The authors would like to thank Dick Hofsheier, Wenliang Chan, Ken Hose and Cheng-Feng Chang of Intel and Stephan Tang of UC, Berkeley for their encouragement and support for this work.

8. REFERENCES


---

**Figure 9.** (a) Reduction in Leakage Component of Active Power (110 °C) achieved by Bit Line Precharge only as Needed, (b) Comparisons of Leakage Component of Active Power (110 °C), and (c) Comparisons of Standby Leakage Power (30 °C).

**Figure 10.** (a) Comparisons of Total Active Power (110 °C), and (b) Comparisons of Energy-Delay Product Including Both Switching and Leakage Components of Energy per Cycle (110 °C).