Dual-V_T SRAM Cells with Full-Swing Single-Ended Bit Line Sensing for High-Performance On-Chip Cache in 0.13 μm 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_{cc} for core and low V_{cc} 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_{cc} dictated by gate-oxide wear-out considerations.

In this paper, we evaluate different dual-V_T cells and cache design choices for high performance microprocessors with a single V_{cc} in a 0.13 μm technology generation. We examine the impact of low-V_T on cell read stability, and investigate different techniques to recover stability with minimal cell area increase. We investigate the effects of excessive bit line leakage on delay for differential low-swing sensing and on noise margin for single-ended full-swing sensing. Different techniques are evaluated to recover noise margin with minimal delay degradation. We also compare cell stability, area impact for stability recovery, noise margin, performance at adequate noise margin, leakage power, total power and energy-delay product of the different dual-V_T design choices for a large cache with single-ended full-swing sensing.

2. DUAL-V_T CELLS WITH DIFFERENTIAL SENSING

Bit line delays for the dual-V_T cells DVTC and DVTC2 (Figs. 1b, 1c) and the low-V_T cell LVTC (Fig. 1d) degrade compared to the high-V_T cell HVTC (Fig. 1a) for differential sensing (Fig. 2a) with 128 rows per bit line pair. Even though one of the bit lines in the complementary pair discharges faster due to larger drive current through the low-V_T pass transistor, the combined leakage current through a large number of low-V_T pass transistors on the other bit line effectively slows down the differential swing development rate. The number of rows per bit line has been reducing by 2x every two generations in order to compensate for slower bit line delay scaling rate. As a result, the number of rows is 64 or less in 0.13 μm technology generation. Reducing the number of rows from 128 to 64 significantly alleviates the adverse impact of leakage on bit line delay in differential sensing (Fig. 2b). On the other hand, for 64 rows or less, the bit line swing development rate may be fast enough such that comparable delay is achieved by a single-ended full-swing sensing scheme. Consequently, single-ended sensing is emerging as an attractive alternative for on-chip cache [8].

ISLPED '00, Rapallo, Italy.

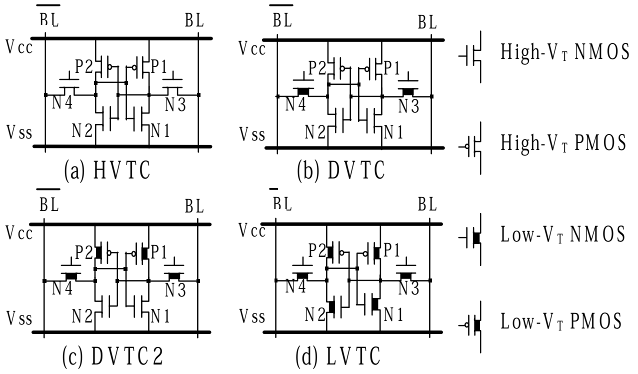
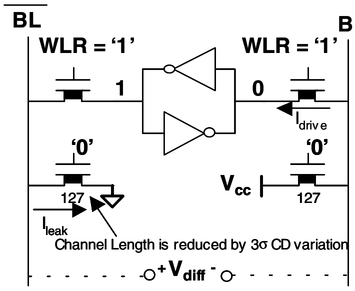


Figure 1. Different 6T SRAM Cell Designs.



(a)

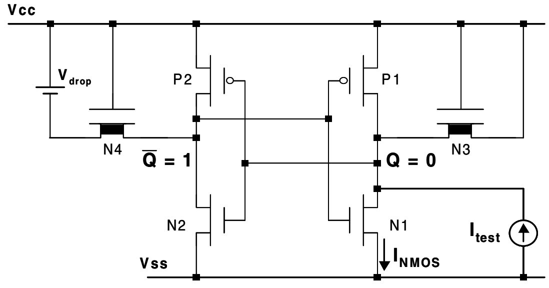
	(+): Delay Improvement			
	HVTC	DVTC	DVTC2	LVTC
128 rows/bitline	Ref	-11.9%	-11.9%	-9.0%
64 rows/bitline	Ref	6.4%	6.4%	8.9%

(b)

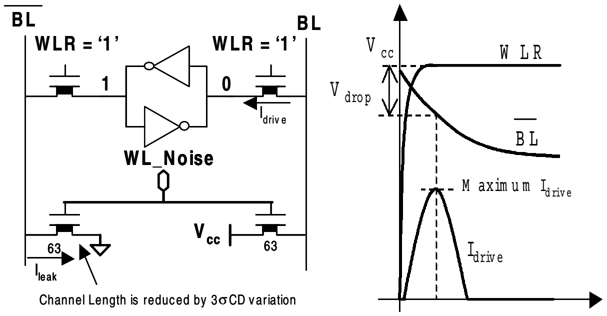
Figure 2. (a) Differential-Sensing for Dual-V_T Cell, and (b) Effect of Bit Line Leakage on Delay.

3. STABILITY OF DUAL-V_T CELLS

Read stability of a cell is measured as the ratio I_{trip} / I_{read}. I_{trip} is the current through the pull-down NMOS on the stored '0' side when the state of the cell is reversed by a current I_{test} injected externally at the stored '0' node (Fig. 3a). I_{read} is the maximum current through the pass transistor during a 'read' operation. Cell stability can degrade due to larger saturation drain current through the pass transistor, or reduction in the current sinking capacity of the pull-down NMOS -- both on the stored '0' side of the cell. Reduction in voltage at the stored '1' node degrades the current sinking capacity of the pull-down NMOS. Deviation of the stored '1' node voltage from V_{cc} in response to excursion of the stored '0' node voltage away from ground is governed by relative strengths of the PMOS and NMOS devices in the cell inverters. In addition, excessive bit line leakage through cells sharing the same bit line pair as the cell being read can cause the bit line voltage on the stored '1' side to droop by an amount larger than the pass transistor V_T. As a result, voltage of the stored '1' node reduces (Fig. 3b). To assess the worst-case impact of bit line leakage on cell stability, the following conditions are used during stability simulation (Fig. 3b) -- (1) channel lengths of all pass transistors are reduced by an amount equal to 3σ CD variation since transistor leakage current increases exponentially at smaller lengths, (2) noise is applied to all the 'off' word lines, and (3) a bit line droop, equal to that when read current through the pass transistor is maximum, is applied statically to the bit lines on the stored '1' side.



(a)



(b)

Figure 3. (a) Stability Measurement Technique, and (b) Voltage Droop in the Complementary Bit Line due to Worst-Case Pass Transistor Leakage.

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).

BEST ⟶ WORST

	HVTC	DVTC2	LVTC	DVTC
Stability Ranking	1	2	2	3

(a)

	DVTC2	LVTC	DVTC
Cell Area Increase	0.79% - 1.58%	0.79% - 1.58%	1.58%

(b)

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

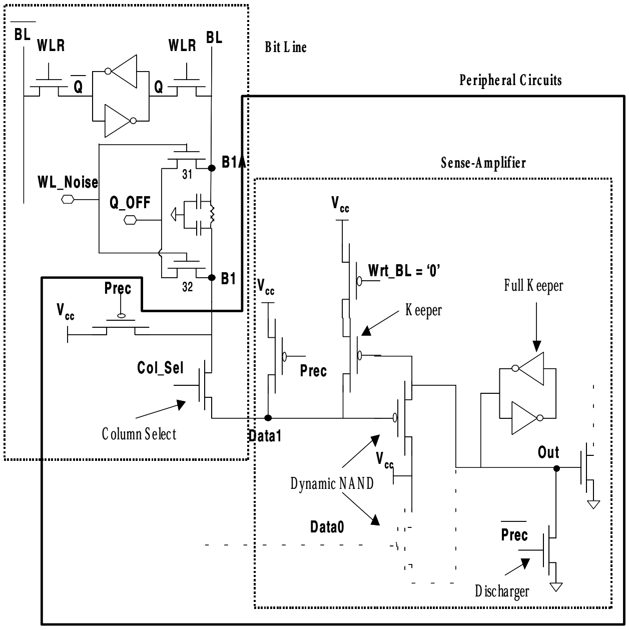
4. IMPACT OF LEAKAGE ON NOISE MARGIN IN DUAL-V_T CACHE

Different dual-V_T cache designs (Fig. 5a) are compared with a conventional design where all transistors, both in the cells and in the peripheral circuits are high-V_T. 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-V_T pass transistors in the cells and degraded noise margins of low-V_T 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-V_T 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-V_T 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 V_{cc} level due to excessive leakage through the pass transistors is large enough to turn-on the low-V_T column select transistors. As a result, a significant portion of bit line noise propagates to the output of the low-V_T, 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-V_T and low-V_T 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.

	HVTC_HVTP	HVTC_LVTP	DVTC_LVTP	DVTC2_LVTP	LVTC_LVTP
SRAM Cell	HVTC	HVTC	DVTC	DVTC2	LVTC
Peripheral	High-V_T	Low-V_T	Low-V_T	Low-V_T	Low-V_T

(a)



(b)

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

BEST ⟶ WORST

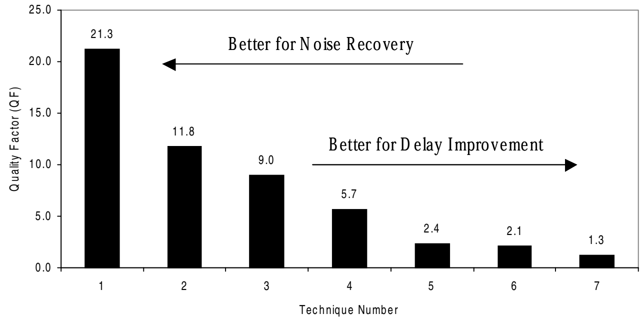
	HVTC_HVT	HVTC_LVT	DVTC_LVT	DVTC2_LV	LVTC_LVT
Noise Margin	1	2	3	3	3
Pass/Fail	Pass	Pass	Fail	Fail	Fail

	WL_Noise	Q_OFF	Q	Vcc	Pass Xtor Channel Length
Noise (Read '1')	Applied	0	1	Vccmax	Reduced by 3σCD Variation

Fig. 6 (a)

		(+): Improvement		
		% Noise Change	% Delay Change	Quality Factor
	Original DVTC_LVTP or DVTC2_LVTP or LVTC_LVTP	Ref	Ref	NA
1	Use LVT Static NAND (w/o Full Keeper and Discharger)	68.0	-3.2	21.3
2	HVT_PMOS for the Keeper	-33.0	2.8	11.8
3	Reduce Full Keeper NMOS Length	67.3	-7.5	9.0
4	HVT_PMOS at the Dynamic NAND	65.7	-11.6	5.7
5	Increase Keeper PMOS Width	67.3	-28.6	2.4
6	Use HVT_NMOS in Column Select	18.6	-9.0	2.1
7	Increase Column Select Width	-13.4	10.6	1.3

(b)



(c)

Figure 6. (a) Noise Margin Ranking and Simulation Conditions, (b) Noise Recovery & Delay Improvement Techniques, and (c) Ranking of Techniques Based on Quality Factor.

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.

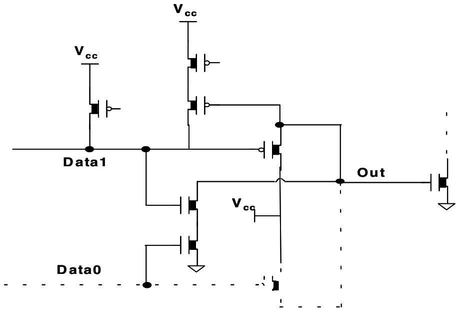
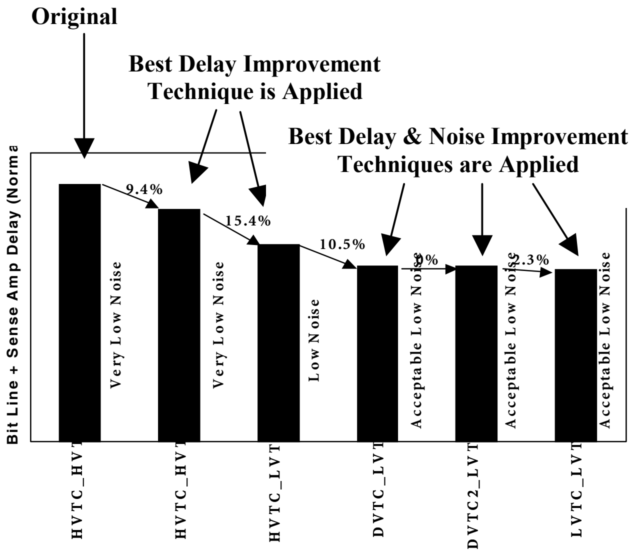


Figure 7. Optimized Static Sense Amplifier.

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).

Sensing Scheme	Bit Line Delay Improvement			
	HVTC	DVTC	DVTC2	LVTC
Differential	Ref	6.4%	6.4%	8.9%
Single-Ended	Ref	13.3%	13.3%	15.7%

(a)



	WL_Noise	Q_OFF	Q	Vcc	Pass Xtor Channel Length
Delay (Read '0')	None	1	0	Vccmin	Drawn

(b)

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.

Using low-V_T 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_T 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_T 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_T only in the pass transistors causes much smaller increase in leakage power than using low-V_T 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_T 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_T 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_T cells (HVTC_LVTP).

6. CONCLUSIONS

We compare cell stability, noise margin, performance and power of different dual-V_T design choices for large on-chip cache with single-ended, full-swing sensing in a 0.13 μm technology generation. The dual-V_T design with low-V_T pass transistors in the cell and low-V_T 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

[1] V. De, S. Borkar, Proc. ISLPED'99, pp. 163-168

[2] S. Thompson et al., 1997 Symp. on VLSI Tech., pp. 69-70

[3] N. Rohrer et al., IEEE ISSCC'98, pp. 240-241

[4] L. Su et al., 1998 Symp. on VLSI Tech., pp. 18-19

[5] L. Wei et al., IEEE Trans. VLSI Systems, 7(1), 1999, pp. 16-23

[6] I. Fukushi et al., 1998 Symp. on VLSI Ckt., pp. 142-145

[7] K. Itoh et al., 1996 Symp. on VLSI Ckt., pp. 132-133

[8] H. Tran, 1996 Symp. on VLSI Ckt., pp. 68-69

HVTC_HVT	HVTC_LVT	DVTC_LVT	DVTC2_LV	LVTC_LVT
1.2x	1.1x	1.6x	1.3x	1.1x

(a)

PRECHARGE SCHEME	HVTC_HVTP	HVTC_LVTP	DVTC_LVTP	DVTC2_LVTP	LVTC_LVTP
Every Cycle	Ref	1.4x	2.4x	4.5x	7.1x
As Needed	Ref	1.5x	1.8x	4.2x	7.2x

(b)

HVTC_HVTP	HVTC_LVTP	DVTC_LVTP	DVTC2_LVT	LVTC_LVTP
Ref	2.4x	2.7x	9.9x	21.8x

(c)

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).

PRECHARGE SCHEME	HVTC_HVTP	HVTC_LVTP	DVTC_LVTP	DVTC2_LVTP	LVTC_LVTP
Every Cycle	Ref	1.2x	1.5x	2.3x	3.2x
As Needed	Ref	1.2x	1.3x	2x	3x

(a)

PRECHARGE SCHEME	HVTC_HVTP	HVTC_LVTP	DVTC_LVTP	DVTC2_LVT	LVTC_LVTP
Every Cycle	Ref	1x	1.2x	1.7x	2.3x
As Needed	Ref	1x	1x	1.5x	2.2x

(b)

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).