13.3 Capacitive-Coupling Wordline Boosting with Self-Induced V_{CC} Collapse for Write V_{MIN} Reduction in 22-nm 8T SRAM

Jaydeep Kulkarni, Bibiche Geuskens, Tanay Karnik, Muhammad Khellah, James Tschanz, Vivek De

Intel, Hillsboro, OR

High-performance microprocessors and SoCs include multiple embedded memory arrays used as register files and low-level caches that typically share the same supply voltage as the core [1]. The desire for wide voltage range operation to optimize power and performance dictates the need for SRAM arrays that can achieve both high performance and low minimum voltage of operation (V_{MIN}). The 8T bitcell (Fig. 13.3.1) is commonly used in these applications because its decoupled read and write ports offer fast read (RD) and write (WR) operations with generally lower V_{MIN} than the 6T bitcell. However, process variations result in mismatches between the pull-up and access devices limiting write V_{MIN}, and/or between read port and keeper transistors limiting read V_{MIN}. Traditional device up-sizing provides diminishing returns at a large area and power cost [2]. In addition to cell upsizing, dynamic assist techniques have been used for V_{MIN} reduction in 6T and 8T arrays—examples include temporary collapse of bitcell voltage for write V_{MIN} reduction and boosting read and write wordlines requiring careful design of the embedded charge pump and the level shifters [2-4]. In contrast, this paper describes a new capacitive-coupling (CC) write wordline boost which employs intrinsic coupling capacitance between write bitlines (WBL) and accessed write wordline (WWL) to boost WWL without the need for a charge pump or complex level shifters. The scheme has a built-in self-induced V_{CC} collapse (SIC) allowing the cell voltage to partially collapse during the write operation, further improving write V_{MIN}. The technique is implemented in a 12KB, 8T cell macro with cell area of 0.238μm^2, fabricated in a 22nm CMOS technology (Fig. 13.3.7).

CC WWL boost relies on the intrinsic coupling capacitance from device and interconnect to the WWL. This coupling capacitance is found in two places (Fig. 13.3.1): the first is at the WWL interface to the PMOS/NMOS devices of the final WWL driver (C_1), while the second is at the WWL interface to the bitcell NMOS write pass devices (C_2 and C_3). To enable use of the first capacitance, the input of the WWL driver is transitioned low to create a rising transition on the WWL. After a short programmable delay (T_1), the input is decoupled from the WWL driver by driving the boost signal low, turning off the top PMOS (P_1) as a result, but without turning on the bottom NMOS (N_1)—effectively floating the WWL and exciting the coupling capacitance between the driver's PMOS transistor and the WWL. This coupling creates ~3 to 5% boost of the floating WWL. To enable use of the second capacitance, both WBL and WBLx are pre-discharged and, depending on data polarity, one is brought high after the WWL has been floated from the first step. This adds ~17 to 20% coupling to the WWL, contributed equally by each bit on the WWL, bringing the total boosting to ~20 to 25% of V_{CC}. The scheme is scalable to any number of bits per WWL with the natural scaling of the WWL driver size and the per-bit coupling capacitance. A beneficial side effect of pre-discharging both the WBL and WBLx prior to the write operation is a self-induced collapse (SIC) of the virtual bitcell voltage when the WWL is asserted. After a short programmable delay (T_2), the appropriate WBL is transitioned high, ending SIC, and completing the write operation. While SIC is inherent with the CC boost technique, it can also be used alone as a low-overhead V_{MIN}-reduction technique, and is an effective alternative when WWL boosting violates gate-oxide reliability limits at high voltage. The SIC-only mode is enabled by keeping the boost signal high (Fig. 13.3.1), ensuring that the WWL is not floated or boosted. A replica delay circuit (Fig. 13.3.2) is implemented to ensure that the WWL has reached V_{CC} before it is floated. This replica delay circuit tracks WR CLK to WWL delay across process, voltage, and temperature variations. Additional programmability is provided to configure the T_1 and T_2 delay elements for maximum boosting benefit, and to adjust the amount of SIC.

Simulation results (Fig. 13.3.3) show the V_{CC} collapse achieved in the SIC mode, and both WWL boost and simultaneous V_{CC} collapse in the CC boost scheme. The amount of WWL boosting varies with process corner and voltage (Fig. 13.3.3), with highest boosting levels observed at low voltage and slow process corner as desired since V_{MIN} impact is largest. Because a component of the coupling capacitance (C_2 in Fig. 13.3.1) depends on whether the same or opposite data is written to a given cell, boost ratio is weakly-dependent on the number of opposite data transitions (Fig. 13.3.3). To minimize this effect, the WBL transition is delayed enough to allow additional V_{CC} collapse, equalizing coupling for all internal bitcell nodes on the selected WWL. Area overhead of the scheme is smaller than required by either cell upsizing or inclusion of a charge pump and level shifters to generate boost voltages [2-3]. For a 0.5KB subarray size, the CC boost technique incurs 11% total area overhead, which is primarily due to the TG-based WWL driver and modified WBL driver supporting data pre-discharge control. This area overhead is amortized for large arrays down to 5%.

To assess V_{MIN} advantage of the scheme for a 1MB array target based on measuring a small sample of 12KB dies with limited bitcell variation, V_{CC} is aggressively lowered, as limited by peripheral logic, to induce random bitcell failures and results are extrapolated to the 1MB failure rate target. Measured P_{FAIL} results vs. V_{CC} (Fig. 13.3.4) show a steep failure rate at low V_{CC}, indicating logic path failures rather than random bit failures. To create random write bit failures as typically observable from a large sample size, 200mV write WL under-drive (WLUD) is used to target array bitcell failures within the V_{MIN} and V_{MAX} limits of peripheral circuits and oxide reliability; respectively. Under this scenario, failure rate is sensitive to both V_{CC} and operating frequency. Write failure rate versus supply voltage for CC boost and SIC-only modes (Fig. 13.3.5) shows V_{MIN} benefit and highlights sensitivity to WWL and WBL timing as parameters T_1 and T_2 are tuned. The slope of write P_{FAIL} curve is governed by the SIC magnitude, WWL boost ratio, and how late WBL transitions with respect to WWL activation. Extrapolating P_{FAIL} data to 1MB array size demonstrates 140mV reduction in write V_{MIN} for CC boost and 80mV reduction for SIC-only mode at optimal timing settings. At lower frequencies, V_{MIN} savings increase to 180mV for CC boost and 130mV for SIC-only mode using optimal delay settings (Fig. 13.3.6). Both SIC and CC boost show an increase in array power when run at the same voltage as baseline (Fig. 13.3.6) due to additional switching of the WBLs, bitcell V_{CC}, and overhead circuitry. However, both techniques enable V_{MIN} scaling beyond the baseline. Total array power savings when operating at lower V_{MIN} (and iso-frequency) are 12% for SIC and 27% for CC boost. Additionally, when the array supply is shared with core logic, improving array V_{MIN} through CC boost enables 29 to 31% total power savings depending on the array activity.

Acknowledgements:

The authors sincerely thank K. Ikeda, T. Hwa Foo, D. Jenkins, D. Finan, T. Nguyen, P. Aseron, and C. Tokunaga for chip implementation and testing, R. Forand and G. Taylor for encouragement and support. This research was, in part, funded by the U.S. Government under contract number HR0011-10-3-0007. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References:

[1] R. Kumar, et al., "A family of 45nm IA processors," *ISSCC Dig. Tech. Papers*, pp. 58-59, Feb 2009.

[2] A. Raychowdhury et al., "PVT & Aging Adaptive Word-Line Boosting for 8T SRAM Power reduction", *ISSCC Dig. Tech. Papers*, pp. 352-353, Feb. 2010.

[3] M. Khellah, et al., "PVT-Variations and Supply-Noise Tolerant 45nm Dense Cache Arrays with Diffusion-Notch-Free (DNF) 6T SRAM Cells and Dynamic Multi-V_{CC} Circuits", *VLSI Circuits Symposium*, pp. 48-49, June 2008.

[4] M. Yuffe et al., "A fully integrated multi-CPU, GPU and memory controller 32nm processor," *ISSCC Dig. Tech. Papers*, pp. 264-265, Feb. 2011.

 978-1-4673-0377-4/12/$31.00 ©2012 IEEE

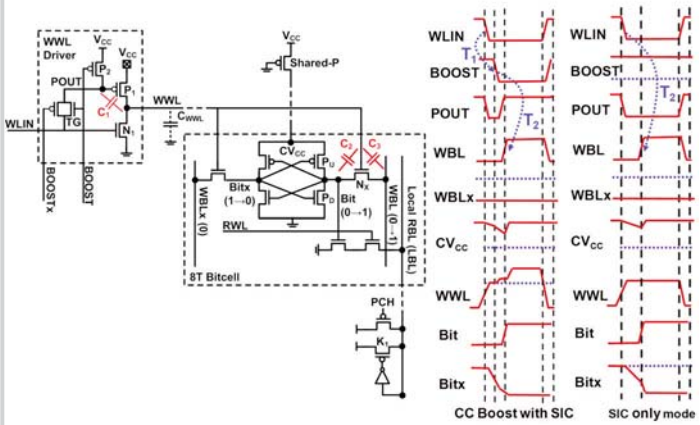


Figure 13.3.1: Proposed Capacitive-Coupling (CC) WWL boosting along with Self-Induced V_{CC} Collapse (SIC) for write V_{MIN} reduction.

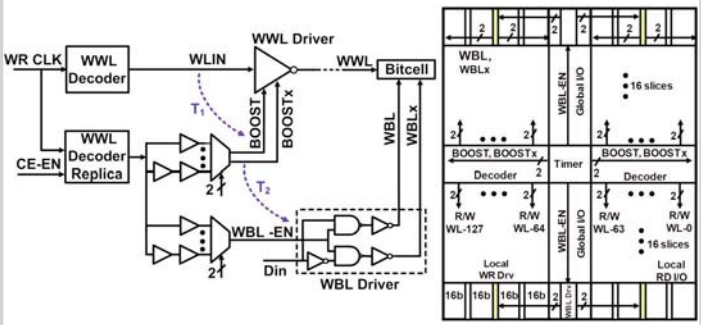


Figure 13.3.2: CC boost circuit with control signal generation and sub-array floor-plan.

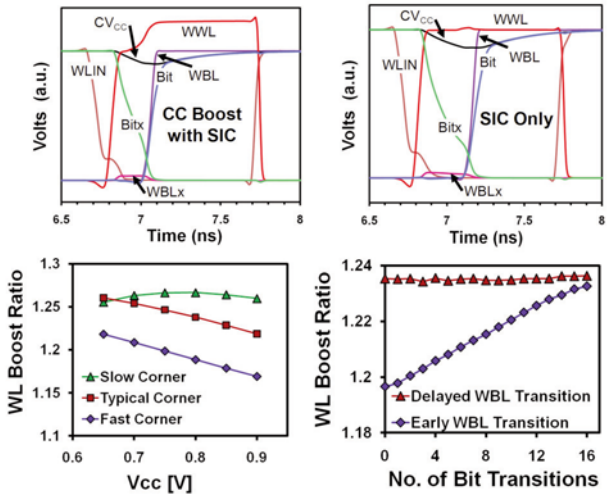


Figure 13.3.3: Simulation results showing CC boost mode and SIC-only mode operation; Boosting ratio variation with V_{CC}, process corner, and data pattern dependence.

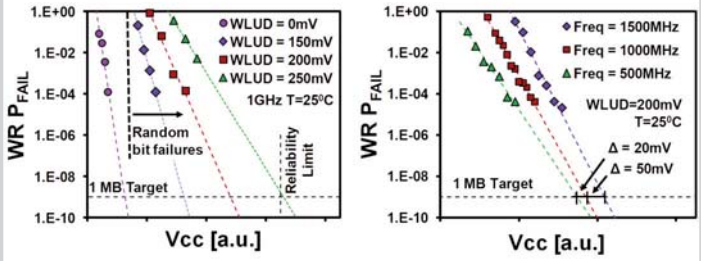
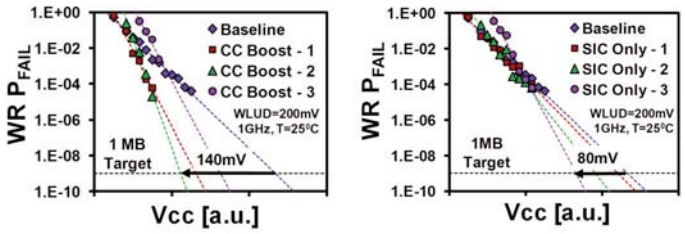


Figure 13.3.4: Measured write failure rate vs. V_{CC} showing the effect of wordline under-drive (WLUD) and frequency on the baseline design.



Delay Setting	CC Boost		SIC Only	
	T_1	T_2	T_1	T_2
1	Min.	2 Inv.	Max.	2 Inv.
2	Min.	4 Inv.	Max.	4 Inv.
3	Min.	6 Inv.	Max.	6 Inv.

Figure 13.3.5: Measured write failure rate vs. V_{CC} for CC boost and SIC-only for different WWL-WBL delay (T_2) settings. Extrapolated to a 1MB array target, CC boost provides 140mV write V_{MIN} improvement, while SIC-only provides 80mV V_{MIN} improvement.

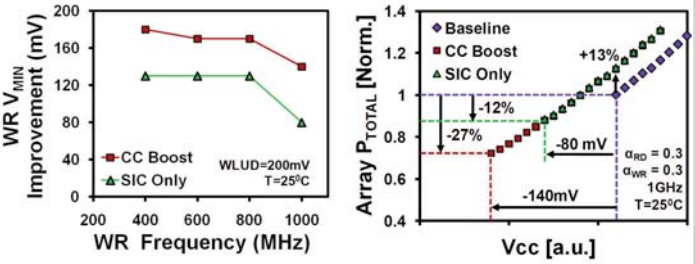
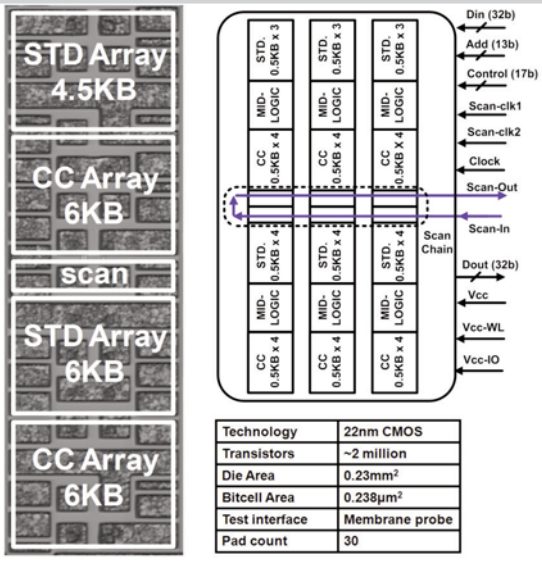


Figure 13.3.6: Measured V_{MIN} savings vs. frequency and total power comparison vs. V_{CC}.



Technology	22nm CMOS
Transistors	~2 million
Die Area	0.23mm^2
Bitcell Area	0.238μm^2
Test interface	Membrane probe
Pad count	30

Figure 13.3.7: Die photograph, pin interface, and chip characteristics.

978-1-4673-0377-4/12/$31.00 ©2012 IEEE