An Adaptive Write Word-Line Pulse Width and Voltage Modulation Architecture for Bit-Interleaved 8T SRAMs

Daeyeon Kim[1, 3], Vikas Chandra[2], Robert Aitken[2], David Blaauw[1], Dennis Sylvester[1]
[1] University of Michigan, Ann Arbor, MI
[2] ARM Inc., San Jose, CA
[3] Currently with Intel Corporation, Hillsboro, OR
daeyeonk@umich.edu, {vikas.chandra, rob.aitken}@arm.com, {blaauw, dmcs} @umich.edu

ABSTRACT
We propose an adaptive WWL pulse width and voltage modulation architecture for low voltage bit-interleaved 8T SRAMs. The 8T bitcell offers improved read/write margins but suffers from write and half select concerns when bit-interleaved [1]. Also, low voltage operation leads to a long-tailed write time distribution, requiring large timing margins and limiting V_{min}. To minimize timing margins and reduce V_{min}, both WWL pulse width and voltage level are adaptively modulated by monitoring written values through the read path. In a 65nm CMOS prototype chip, V_{min} is lowered from 700mV to 500mV using this technique, providing 2.55× leakage power reduction and 2.4× active power reduction.

Categories and Subject Descriptors
B.3.1 **[Memory Structures]**: Semiconductor Memories – *Static Memory (SRAM)*; B.7.1 **[Integrated Circuits]**: Types and Design Styles – *Advanced Technologies, Memory Technologies, VLSI (very large scale integration)*; B.8.0 **[Performance and Reliability]**: General

Keywords
SRAM, Low Voltage Design, Noise Margin

1. INTRODUCTION
Low voltage operation is an effective way to reduce power consumption due to the resulting quadratic power savings. Variability at low voltage is a challenge however, and low voltage SRAM in particular is vulnerable to variation and functional failures due to the use of minimum feature sizes in a bitcell and large SRAM array sizes. The 8T bitcell [2] (Figure 1) improves low voltage operation by allowing separate optimization of read and write paths. Bit-interleaving is essential to avoid soft errors, particularly at low voltages, but induces half select disturb [3] as the 6T portion of the 8T bitcell is optimized for write. Since half select and writability create a double-sided constraint on WWL pulse width and WWL voltage, WWL control is key to reducing V_{min} and maximizing yield for bit-interleaved 8T SRAMs.

ISLPED'12, July 30–August 1, 2012, Redondo Beach, California, USA.

To mitigate the trade-off between write and half select, this paper proposes adaptive WWL voltage modulation with regenerative half selected bit-lines. In addition, SRAM bitcells exhibit a wide write time distribution with a long tail at low voltage (Figure 2). Excessive margins are required due to the few bitcells with long write times. To minimize this impact, we propose a multi-cycle

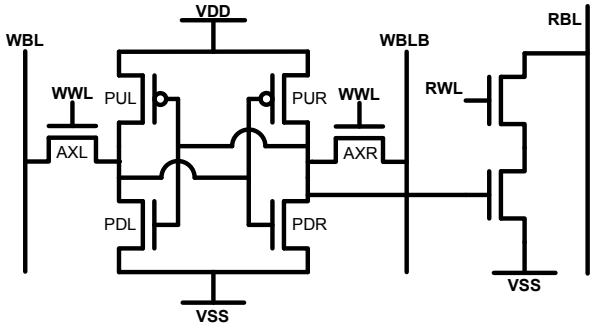

Figure 1. The 8-T bitcell schematic is shown. Write and read are separately optimized.

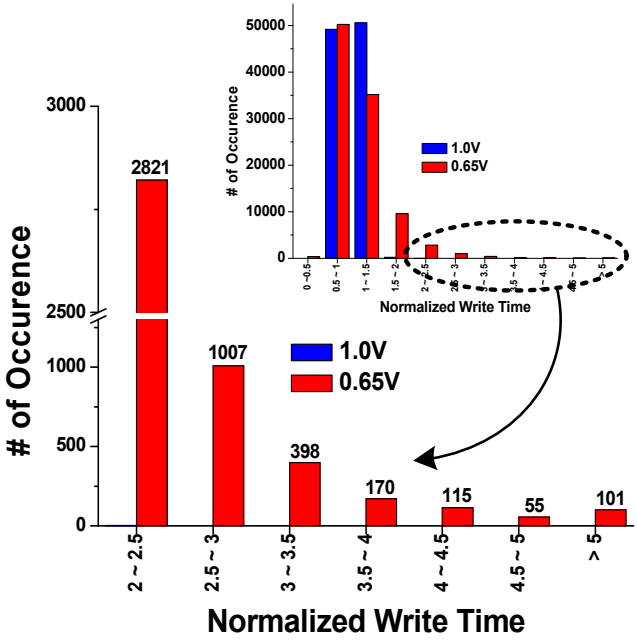

Figure 2. 10K Monte Carlo simulation results of write time distribution at 0.65V and 1.0V. The distribution at 0.65V is wider with a long tail, increasing design margins.

write operation where the number of cycles is adaptively determined via write completion detection, similar to the scheme used in [4]. Overall performance is improved using the multi-cycle write scheme.

The importance of dynamic write stability has been previously discussed [5]. A replica-based adaptive stability enhancement technique [6] was proposed. While [6] is replica-based, this work describes an *in situ* approach (i.e., it uses actual bitcells in the SRAM array, eliminating mismatch issues). Adaptive WWL boosting for 8T SRAMs [7] was also proposed but it does not use a bit-interleaved array and hence is susceptible to soft errors. Half select prevention using a pre-read technique was proposed in [8] but does not address the large write delay margins observed at low voltages in SRAM.

2. 8-T SRAM OPERATIONS AT LOW VOLTAGES

Variation increases as the supply voltage decreases. Figure 3 depicts variation as supply scales. The distribution of FO4 delay is measured using 100K Monte Carlo simulations. At low voltage, performance degrades by larger variation as well as smaller I_{on}. The performance degradation by larger variation limits lowering the supply voltage so variation compensation techniques are required for low voltage operation.

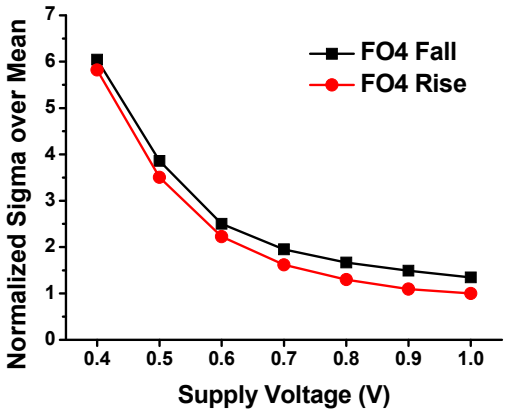

Figure 3. More variation exists at lower voltage.

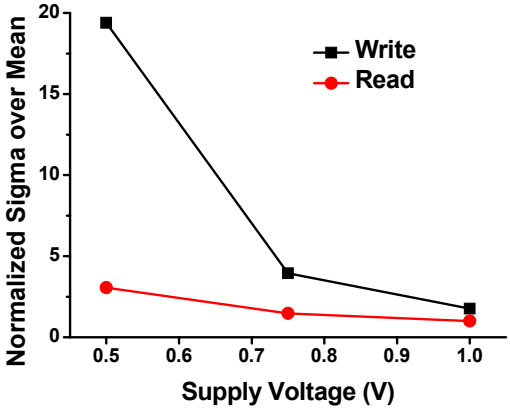

Figure 4. Write operation is critical operation at low voltage because it is more vulnerable to variation.

As already discussed above, the 8T SRAM bitcell is a good candidate as a SRAM bitcell at low voltage since write operation and read operation can be separately optimized. Between read and write, write operation is a critical operation at low voltage because it has more variation. Figure 4 shows 40K Monte Carlo simulation results and it clearly depicts that write operation is more vulnerable to variation. Also, there are five write failures out of 40K at 0.5V while there is no read failure. Because of the five write failure at 0.5V, it is not possible to lower the supply voltage down to 0.5V.

To analyze the write operation of 8T SRAM at low voltage, write time is simulated using Monte Carlo simulation. Figure 5 describes the definition of write time used in this work. WWL is turned on to start write operation. After some time, two internal nodes in an SRAM bitcell are crossed each other. Write time is defined as a time between WWL on and two internal nodes crossing. For successful write operation, WWL pulse width must be larger than this write time. If the write operation fails even though write time is infinitely long, static write failure happens.

The Monte Carlo simulation results with 100K iterations of write time as supply scales are shown in Figure 6. At 1.0V, the worst case write time is ~2.2× larger than typical. If the WWL pulse width is 2.2× larger than typical write time, high yield is expected.

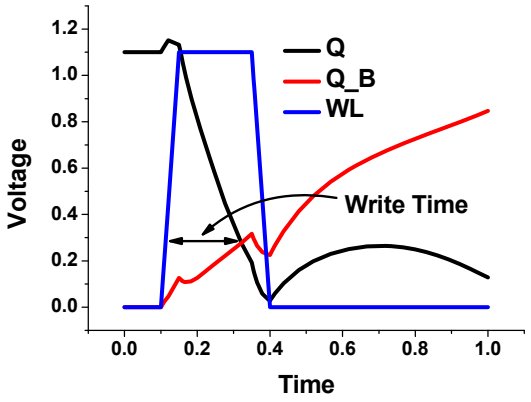

Figure 5. SRAM write time is a time difference from WWL on to two internal nodes crossing.

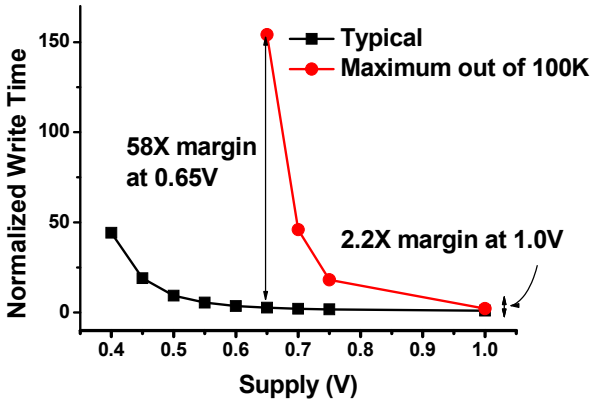

Figure 6. Write time degrades as supply scales. However, the degradation of the worst case write time is much worse than typical cases. Below 0.65V, write failure happens (static write failure).

However, the required margin at low voltage is much larger than nominal voltage. At 0.65V, at least 58× margin is required for the successful write of all 100K iterations. Below 0.65V, the worse case is static write failure: write operation cannot be done even with infinitely long WWL pulse. Based on this simulation, the V_{min} of this SRAM bitcell is 0.65V because of static write failures under 0.65V.

As the long tailed distribution of write time is shown in Figure 2, only a few bitcells need excessive margin. This observance leads the adaptive multi cycle write scheme to increase overall performance while keeping the iso-yield.

3. ADAPTIVE WWL PULSE WIDTH AND VOLTAGE MODULATION

We propose an adaptive WWL pulse width and voltage modulation architecture for bit-interleaved 8-T SRAMs. There are two major WWL parameters: pulse width and voltage level. First, WWL pulse width is modulated by an adaptive multi-cycle write structure. Figure 7 shows the implementation details for adaptive WWL pulse modulation. To write data to worst-case bitcells, frequency is not lowered, but instead additional cycles are used in those cases until write completion is detected. Second, two WWL voltages are applied to fix static failures while avoiding half select disturb. Longer WWL pulses induce more half select disturbs yet still cannot fix static write failures. To address this, a low voltage WWL pulse is initially used to write the majority of bitcells while all half selected bitcells are read. In any subsequent cycles, the WWL voltage is boosted (Figure 8) to write challenging bitcells. Also, half select disturbs are mitigated as half selected bitcells experience only weak read disturb (by a low voltage WWL) during the first cycle while in subsequent cycles half select disturbs are eliminated by driving bit-lines to their appropriate values based on read-out data from the first cycle.

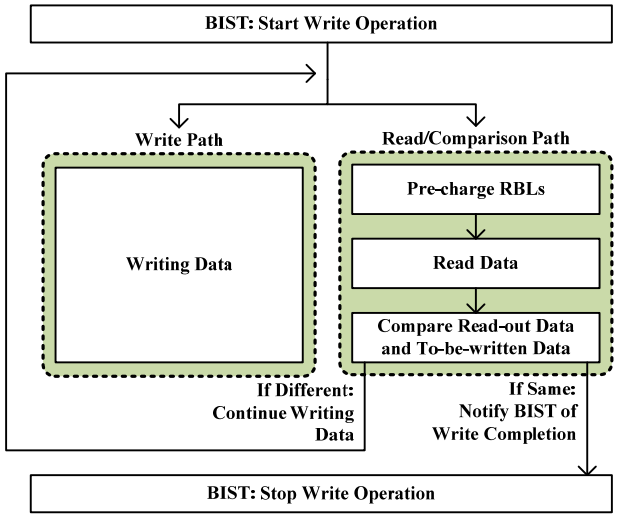

Figure 7. The adaptive WWL pulse width modulation scheme is shown. Write completion is monitored every cycle. When write completion is detected, a BIST stops the write operation.

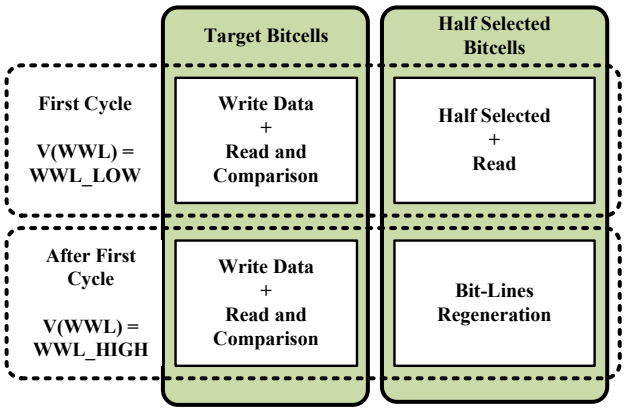

Figure 8. The adaptive WWL voltage level modulation scheme is shown. WWL is boosted after the first cycle but half selected bitcells are not disturbed due to bit-line regeneration.

Figure 9 shows a sample timing diagram when a write operation is performed in two cycles. While WWL is on for multiple cycles, RWL toggles every cycle for reading the data repetitively. When read-out data and to-be-written data are the same, WRITE_DONE signal is asserted and the BIST stops the write operation.

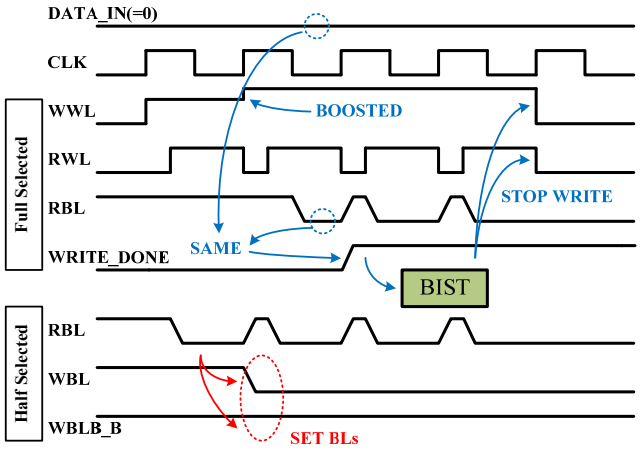

Figure 9. Sample timing diagram when write operation is performed in two cycles.

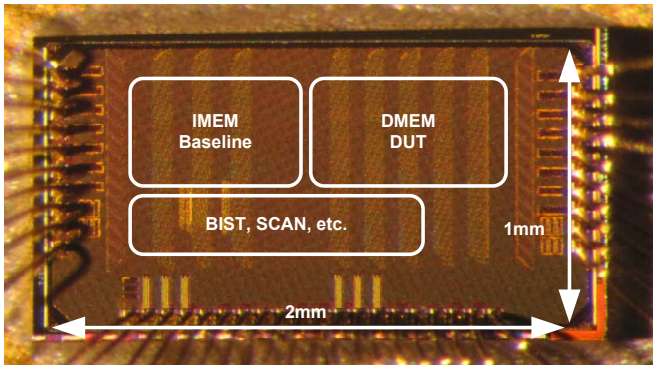

Figure 10. Die photo of the prototype chip fabricated in 65nm CMOS.

4. PROTOTYPE IMPLEMENTATION

Figure 10 shows the die photo of the prototype chip fabricated in 65nm CMOS, which contains a 16kb (128×128) bitcell array (Figure 11). For improved performance and functionality, bit-lines have a cascaded structure with several local blocks. Eight bitcells, a pre-charger, keeper, and tri-state buffer comprise a local block. Sixteen local blocks are connected to form a 128-bit tall global bit-line. The 128×128 bitcell array is bit-interleaved with four 32-bit words. A comparator and a WWL width and voltage controller are also implemented; the controller selects the WWL driver output voltage level from two voltage levels.

5. MEASUREMENT RESULTS

Figure 12 shows a measured shmoo plot of the SRAM array V_{min} when V_{SUPPLY}, V_{WWL_HIGH}, and V_{WWL_LOW} are the same. The results clearly show the double-sided constraint on frequency due to write/read and half select disturb. V_{min} considering half select disturbs is measured to be 0.775V. Below this voltage, half select disturb becomes critical while write and read remain functional. Neglecting half select, the array V_{min} would be 0.5V.

By lowering V_{WWL_HIGH} and V_{WWL_LOW} voltage levels together (i.e., conventional WWL underdrive), half select disturb is partially mitigated (Figure 13). Half select disturb can be eliminated at voltages down to V_{SUPPLY} = 0.7V without multi-cycle write, or 0.6V with multi-cycle write, while write operation remains functional. In the 8T bitcell used in this work, the access transistor and pull-down transistor are sized identically, making write operation strong compared to a traditional 6T bitcell. However, at V_{SUPPLY} < 0.6V, write failures occur when WWL is underdriven sufficiently to fix half select disturbs. In other words, there is no feasible WWL voltage level that avoids both write failure and half select disturbs.

Figure 14 analyzes failures at V_{SUPPLY} = 0.5V as WWL voltage (both V_{WWL_HIGH} and V_{WWL_LOW}) decreases. Initially, half select

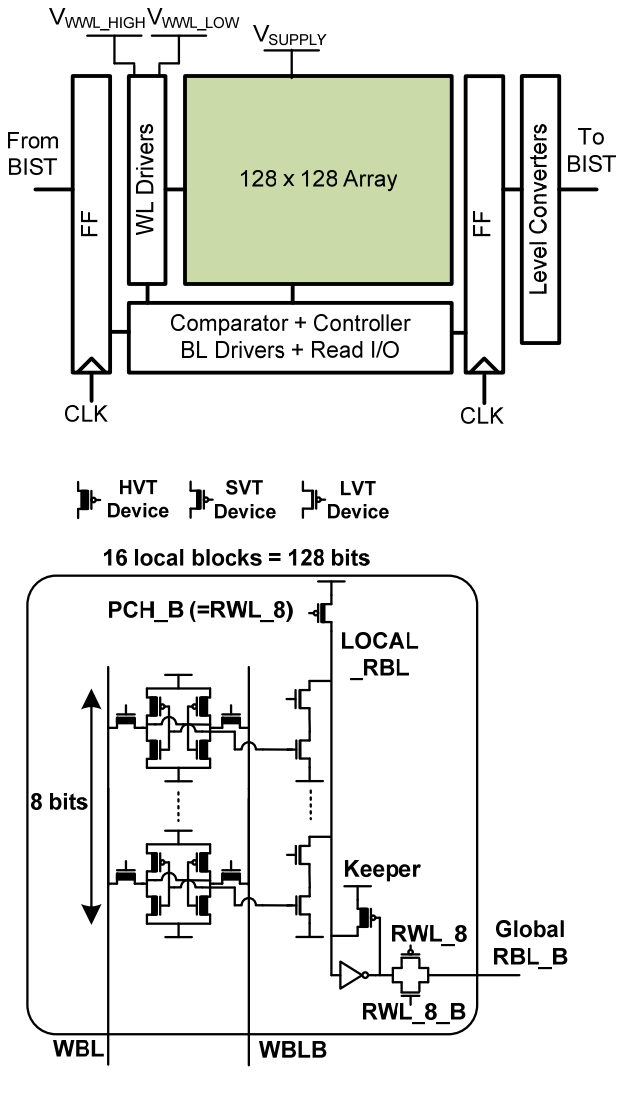

Figure 11. A block diagram of the prototype memory bank is shown at top. A comparator and controller for WWL width and voltage modulation are implemented. A local 8-bit bit-line structure is shown at right with 16 8-bit local blocks. A 128-bit wide row consists of four bit-interleaved 32-bit words.

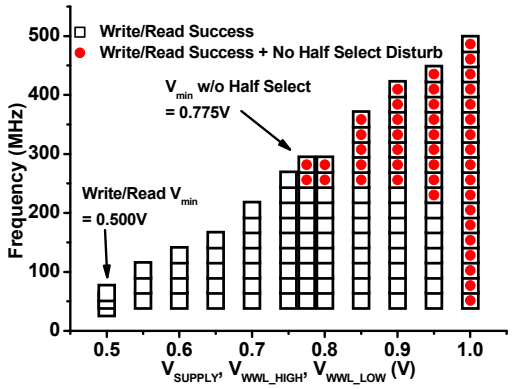

Figure 12. Measured shmoo plot to find V_{min}. V_{min} is 0.775V when all voltage levels are identical.

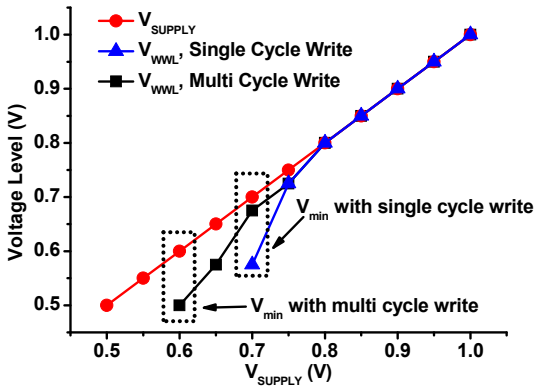

Figure 13. Measured V_{min} is reduced to 0.7V with conventional WWL underdrive, and to 0.6V with WWL underdrive and multi-cycle write. For further V_{min} reduction, V_{WWL_HIGH} and V_{WWL_LOW} must be separately optimized.

disturb dominates. Hence, at higher frequency there are fewer half select events (e.g., at V_{WWL} = 0.5V). As WWL voltage decreases, half select disturb is mitigated but write failures become prominent. At V_{WWL} < 0.425V, write failures begin dominating, as indicated by a higher number of failures at higher frequencies.

To concurrently fix both half select disturb and read/write failures, the proposed adaptive voltage level modulation scheme is applied at V_{SUPPLY} = 0.5V (Figure 15). If V_{WWL_LOW} is too high, half select may occur during the first cycle (see Figure 8, top right). To avoid this, V_{WWL_LOW} should be ≤ 390mV based on measured results. Also, V_{WWL_HIGH} must be high enough to write data to worst-case bitcells in the second cycle. All failures are fixed at V_{WWL_HIGH} = 475mV. Based on measured results, we can simplify the system by only underdriving V_{WWL_LOW} while setting V_{WWL_HIGH} = V_{SUPPLY} to maximize writability and half select disturb immunity. With V_{SUPPLY} = 500mV, V_{WWL_HIGH} = 500mV, and V_{WWL_LOW} = 390mV, the operating frequency with no failures is 64MHz and the average and maximum number of write cycles for the array is 2.25 and 3, respectively.

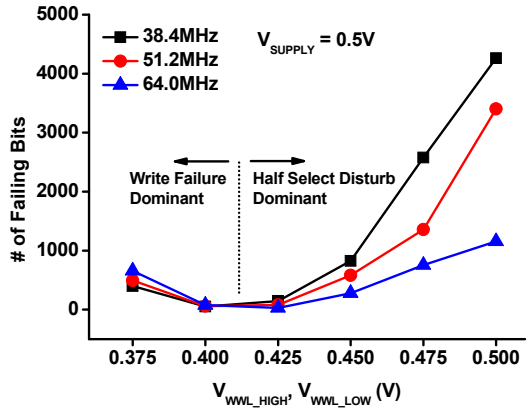

Figure 14. Measured number of failing bits with V_{SUPPLY} = 0.5V and WWL underdrive. As WWL voltage decreases, immunity to half select disturb is improved while writability worsens.

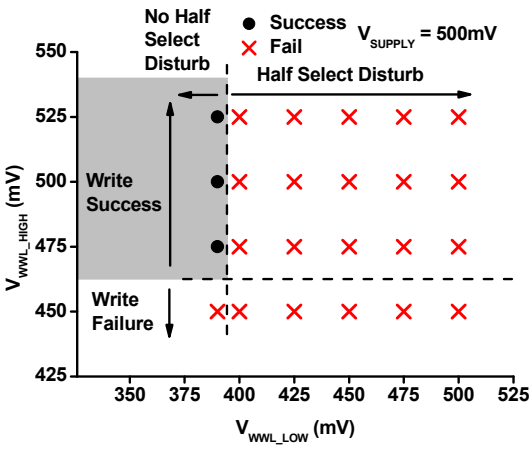

Figure 15. V_{WWL_HIGH} and V_{WWL_LOW} are separately optimized at V_{SUPPLY} = 0.5V. Measured results show that all failures are fixed at V_{WWL_LOW} = 0.39V

Overall, V_{min} using the adaptive WWL modulation scheme is measured to be 500mV, offering significant reductions beyond conventional WWL underdrive (700mV) and multi-cycle writes alone (600mV) (Figure 16). Note that this V_{min} (500mV) is equal to the V_{min} when neglecting half select completely. Measured power consumption at 64MHz is summarized in Table 1. The V_{min} reduction from the proposed technique enables 2.55× leakage power reduction and 2.4× active power reduction for the SRAM array, and enables the use of bit-interleaving in low voltage robust 8T SRAMs.

6. CONCLUSIONS

An adaptive write word-line pulse width and voltage modulation architecture is proposed for low voltage bit-interleaved 8-T SRAMs. By minimizing excessive margins and applying regenerative half select bit-lines scheme, V_{min} is lowered from 700mV to 500mV with 2.55× leakage power reduction and 2.4× active power reduction in a 65nm CMOS prototype chip.

7. ACKNOWLEDGEMENTS

The authors at University of Michigan acknowledge the support of NSF, GSRC, and Army Research Laboratory.

8. REFERENCES

[1] D. Kim *et al.*, "Variation-Aware Static and Dynamic Writability Analysis for Voltage-Scaled Bit-Interleaved 8-T SRAMs," *International Symposium on Low-Power Electronics and Design (ISLPED)*, pp. 145-150, Aug. 2011

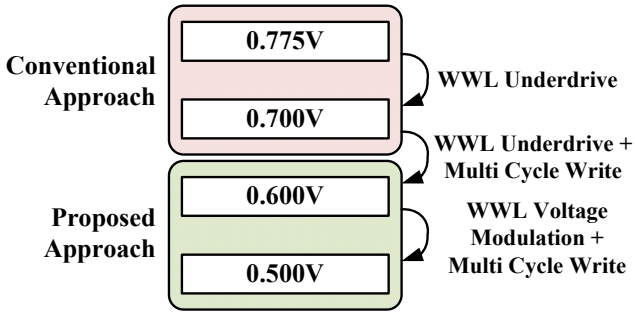

Figure 16. Conventional WWL underdrive can reduce V_{min} to 0.7V. With the proposed methods, V_{min} is further lowered to 0.5V.

Table 1. Measured power/energy improvements (at 64MHz). The conventional method is WWL underdrive. The proposed method includes multi cycle write and WWL voltage modulation. Using the proposed methods, V_{min} is lowered with large gains in leakage power and active power.

Cases	Supply (mV)			Power (µW)	
	V_{SUPPLY}	V_{WWL_HIGH}	V_{WWL_LOW}	Leakage	Active
Conventional	700	575	575	204.7	307.2
Proposed	500	500	390	80.2	127.9

[2] L. Chang *et al.*, "Stable SRAM cell design for the 32nm down and beyond," *IEEE Symposium on VLSI Circuits (VLSI Circuits)*, pp. 128-129, June 2005

[3] R. Joshi *et al.*, "6.6+ GHz Low V_{min}, read and half select disturb-free 1.2 Mb SRAM," *IEEE Symposium on VLSI Circuits (VLSI Circuits)*, pp. 250-251, June 2007

[4] S. Hanson *et al.*, "A Low-Voltage Processor for Sensing Applications with Pico-watt Standby Mode", *Journal of Solid State Circuits (JSSC)*, Vol. 44, pp. 1145-1155, Apr. 2009.

[5] S. O. Toh, Z. Guo, B. Nikolić, "Dynamic SRAM Stability Characterization in 45nm CMOS," *IEEE Symposium on VLSI Circuits (VLSI Circuits)*, pp. 35-36, June 2010

[6] H. Nho *et al.*, "A 32nm High-κ Metal Gate SRAM with Adaptive Dynamic Stability Enhancement for Low-Voltage Operation," *IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 346-347, Feb. 2010

[7] A. Raychowdhury *et al.*, "PVT-and-Aging Adaptive Wordline Boosting for 8T SRAM Power Reduction," *IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 352-353, Feb. 2010

[8] M. Sinangil, H. Mair, A. Chandrakasan, "A 28nm High-Density 6T SRAM wth Optimized Peripheral-Assst Circuits for Operation Down to 0.6V," *IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 260-261, Feb. 2011.