A Reconfigurable 65nm SRAM achieving Voltage Scalability from 0.25-1.2V and Performance Scalability from 20kHz-200MHz

Mahmut E. Sinangil*, Naveen Verma, Anantha P. Chandrakasan
Massachusetts Institute of Technology, Cambridge, MA
E-mail*: sinangil@mit.edu

***Abstract*— A 64kb SRAM array fabricated in 65nm low-power CMOS operates from 250mV to 1.2V. This wide supply range is enabled by a combination of circuits optimized for both sub-V_t and above-V_t regimes. Reconfigurable circuits are used extensively, as low voltage assist circuits are required for functionality, but they must not limit performance during high voltage operation. The SRAM operates at 20kHz with a 250mV supply and 200MHz with a 1.2V supply. Over this range the leakage power scales by more than 50X.**

I. INTRODUCTION

Highly energy constrained applications, such as wireless sensor nodes and biomedical implants, preferentially operate at low voltage levels and at low frequencies to be close to the minimum energy point [1]. However these applications typically need to elevate their performance levels for short bursts of time to meet a certain system constraint. Thus, SRAMs for dynamic performance applications should operate efficiently over a large voltage and performance range. Also, since SRAMs account for a significant portion of the total area and power consumption of modern digital systems, Ultra-Dynamic Voltage Scalable (U-DVS) design should minimize the area and power overheads of achieving the large operating range.

The designs in [2], [3] and [4] demonstrate sub-V_t SRAMs. Because of the increased effect of variation due to random dopant fluctuation (RDF) and severely-degraded I_{on}/I_{off} ratio in sub-V_t region, these designs use different topologies and peripheral assist circuits to overcome these effects and enable functionality. For example, [2] and [4] uses 10T SRAM cells and [3] proposes an 8T cell. Although these designs achieve very low energy consumption, they cannot operate efficiently at higher voltages and consequently at higher frequencies since their circuits are designed to mainly target sub-V_t functionality. On the contrary, the design demonstrated in [5] works at high voltages and over a larger range but only in above-V_t . By restricting the voltage range as such, this design can employ static topologies with no reconfigurability and still support a large range since the trade-offs do not vary much within one region. However, more aggressive leakage power and active energy savings are required in highly energy constrained applications and therefore lowering the supply voltage into the sub-V_t region is crucial. Designing SRAMs for both sub-V_t and above-V_t regions is very challenging because of the fundamentally different trade-offs governing the circuit operation between these two regimes. In order to operate efficiently in both regions, circuits must be able to adapt themselves to the varying trade-offs over the voltage range. This adaptability is enabled by designing peripheral circuits with hardware reconfigurability.

This paper presents an SRAM that is designed for both sub-V_t and above-V_t operation. The design is operational from 250mV which is in deep sub-V_t region to 1.2V which is the nominal-V_{DD} for the process. An 8T bitcell is used to construct a high density array. Assist circuitry which is necessary for low-voltage functionality is designed with low overhead reconfigurability in mind. One of three different write-assist schemes are activated depending on the supply voltage level to prevent excess power. Multiplexed sense-amplifiers are used in the sensing network to minimize sensing delay. Lastly, bitcell and peripheral circuits are designed for optimal operation over the large voltage range.

II. DESIGN CONSIDERATIONS FOR U-DVS SRAM

A. Bitcell Design and Sizing

The traditional 6T SRAM cell fails to operate at low voltages because of read and write failures due to the degradation of Read Static-Noise-Margin (RSNM) and Write Margin. To limit the area overhead, an 8T bitcell is used in the design (Fig. 1) [3]. Two NMOS devices constitute the read-buffer and decouple the read and write ports of the cell. A write operation is done through WL, BL and BLB ports and a single-sided read operation is done through RDWL and RDBL ports. Decoupling of write and read ports gives designer more freedom to optimize the sizing of individual devices inside the cell for low-voltage functionality. Specifically, the 6T part of the cell can be sized for writability and the two read-buffer transistors can be sized for better read performance.

BVSS node is shared on each row, and it gets pulled-up if the row is not accessed. As a result, the voltage drop across the read-buffers of all un-accessed rows is brought to 0V. This causes the leakage from RDBL through un-accessed rows to be greatly reduced.

MCHd node is also shared on each row and connected to a gated driver. Since MCHd is the virtual supply node for the cross-coupled inverters inside the cell, bringing its voltage

978-1-4244-2362-0/08/$25.00 ©2008 IEEE.

Authorized licensed use limited to: MIT Libraries. Downloaded on January 10, 2009 at 10:37 from IEEE Xplore. Restrictions apply.

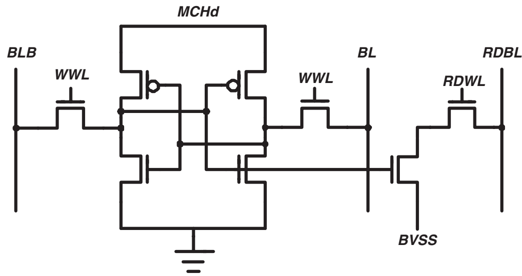

Fig. 1. 8T bitcell used in the design.

down is very effective to ease a write operation which will be discussed in the following sections.

Sizing of the read buffer devices inside the cell must be done by considering the entire voltage range. At voltage levels in or close to the sub-V_t region, using longer channel lengths increases the drive currents considerably. This is due to a decrease in the threshold voltage of the device with increasing gate length which is known as the reverse short channel effect (RSCE) [6]. However, in above-V_t regime, using longer channel length results in a decrease in the drive current.

Fig. 2 shows the 4σ drain current of a read-buffer with different gate lengths normalized to the minimum length read-buffer. There is a significant increase in 4σ current at low voltages which is due to RSCE. Larger channel area also improves 4σ current. The lengths of the read-buffer devices are chosen to be approximately 2 times the minimum length in order to increase performance at low voltages without degrading it too severely at high voltages.

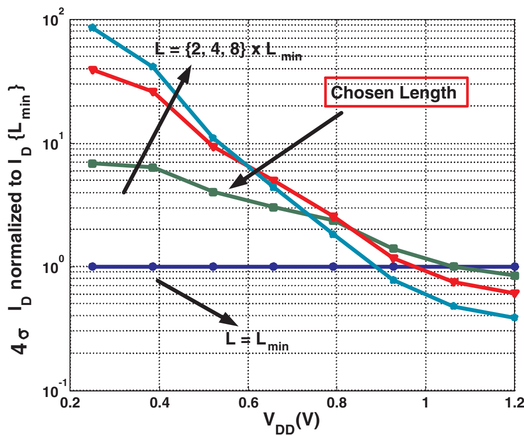

Fig. 2. 4σ drain current through a read-buffer for different gate lengths normalized to the minimum length read-buffer. Increasing the gate length results in a large current at low voltages but a smaller current at high voltages due to RSCE and a larger channel area.

B. BVSS Driver Design

As explained above, BVSS driver pulls the row-wise BVSS node up in order to mitigate the leakage through the RDBL when the row is not accessed. This structure employs a charge-pump circuit as shown in Fig. 3a. Doubling the gate drive of the NMOS pull-down device increases its drive strength by nearly 500X in sub-V_t so it can sink the aggregated current through all read-buffers in a row. The design in [3] exploits this exponential dependence of current to gate drive and uses a nearly minimum size pull-down NMOS at the output of the charge-pump.

For the U-DVS design, an interesting trade-off exists for the sizing of the pull-down transistor in the BVSS driver. The charge pump cannot be enabled beyond V_{DD} =0.6V due to reliability concerns. This causes a sudden drop in the performance vs. V_{DD} curve as shown in Fig. 3b. Up-sizing the NMOS transistor by nearly 10X makes the off-region smaller and ensures a nearly continuous performance improvement with increasing supply voltage.

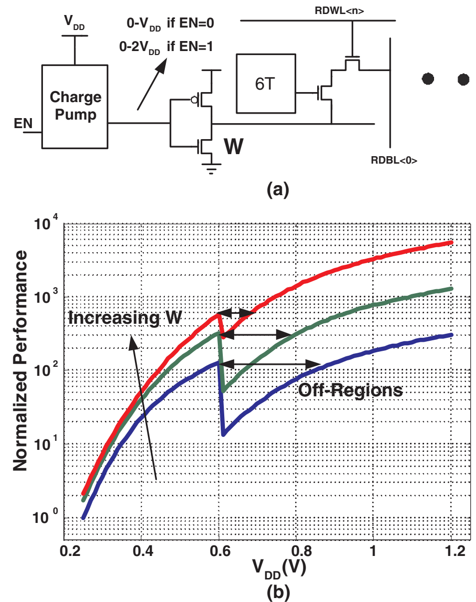

Fig. 3. BVSS driver (a) and normalized read performance with different widths for the pull-down device in BVSS driver (b). Increasing the width is necessary for a continuous performance vs. V_{DD} curve beyond the voltage at which charge-pump cannot be activated.

C. MCHd Driver Design

Although the 8T cell can be sized for better writabilty, the effect of sizing is easily overcome by the variation as the supply voltage scales down. This causes the degradation of write margin of the cell. Fig. 4b shows the write margin distributions for the memory cell used in the design. A positive value shows a write failure. As the supply voltage goes down, the mean of the distribution becomes smaller. Additionally the distribution becomes wider which is due to the increased effect of variation at lower voltages. These two effects add up causing the tail of the distribution to get closer and closer to the point of failure as the supply voltage scales down. The distribution at 1.2V shows that there is enough write margin at this voltage. At 700mV, only the tail of write margin distribution fails whereas at 250mV, a significant portion of the memory cells are facing write failure. In order

Authorized licensed use limited to: MIT Libraries. Downloaded on January 10, 2009 at 10:37 from IEEE Xplore. Restrictions apply.

to maintain functionality at low voltages, peripheral assists should be employed to improve the write margin. However this peripheral assist circuits must be designed such that they introduce minimal power overhead to the design.

This motivates the concept of reconfigurability for write assists at different supply voltage levels. As mentioned above, MCHd node is connected to a gated driver as shown in Fig. 4a. MCHd driver is designed such that depending on the operating V_{DD} , MCHd voltage can be actively pulled-down, left floating or kept at V_{DD} during a write operation as shown in Fig. 4c. Reducing the supply node of the cell helps writability by degrading the strength of the internal feedback between the cross-coupled inverters. In *Keep at* V_{DD} mode, MCHd node is at V_{DD} all the times since no peripheral assist is required for functionality. In the *Float Header* mode, MCHd node is kept floating during a write access and the residual charge on this node is shared between the memory cells on the same row. This causes the MCHd node to droop to a slightly lower voltage during the write access. In the *Pull Header Down* mode, MCHd is pulled-down actively during the write cycle resulting in a better improvement in the write margin. Before the WL voltage goes down, MCHd node is actively pulled-up to V_{DD} in both modes. This programmable scheme prevents significant power overhead that would stem from keeping the write assists active at higher operating voltages.

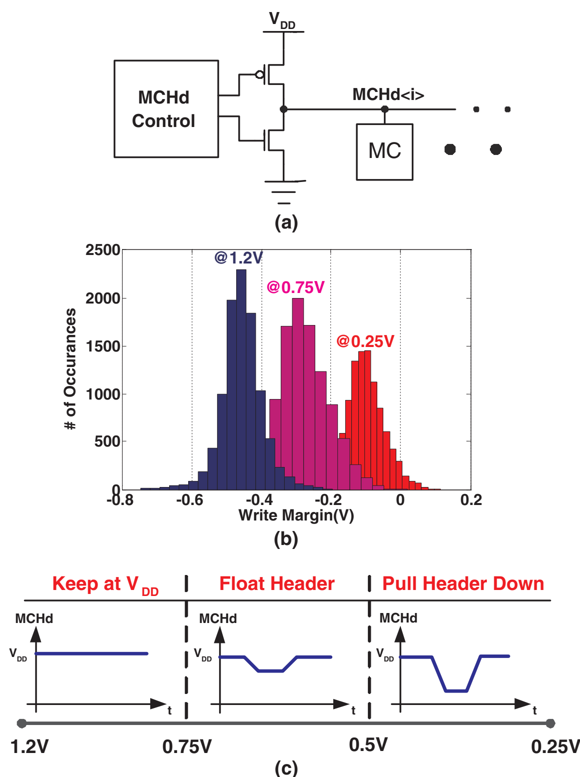

Fig. 4. MCHd driver circuit (a), write margin distributions at 1.2V, 0.7V and 0.25V (b) and three different writing schemes implemented in the design (c). One of the writing schemes is activated depending on the supply voltage level to prevent power overhead.

D. Sensing Network Design

Sense amplifier is in the critical path of a read access so the delay of this structure should also be considered very carefully. Since the U-DVS SRAM is intended to be operational in a very large voltage range, more than one sensing scheme should be employed to minimize the sensing delay.

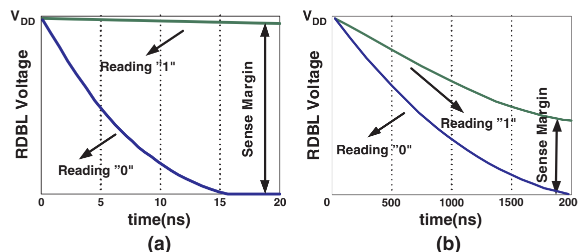

Fig. 5. Voltage waveforms in small-signal (a) and large-signal (b) sensing scenarios. Common-mode of the signals are close to V_{DD} for small-signal sensing whereas it is closer to ground level for large-signal sensing.

Small-signal and large-signal sensing schemes are widely used in SRAMs and both of these schemes have different design advantages over the other as explained in [7]. Fig. 5a and Fig. 5b show the RDBL voltage vs. time plots for two different scenarios. If the read access time is much faster than the parasitic droop on RDBL due to leakage (Fig. 5a), small-signal sensing scheme can be employed. However in the case of very long access times, which occurs at low voltage-performance modes, the increased effect of leakage results in severe degradation of the sensing margins (Fig. 5b). As a result, here, a large-signal sensing scheme must be employed.

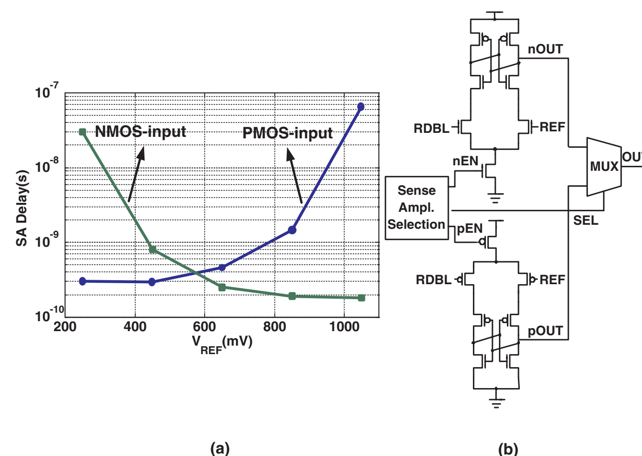

Fig. 6. Delay vs. REF for a sense-amplifier (a) and sensing network design (b). Devices for pre-charge/pre-discharge are omitted in the schematic for simplicity.

The sense-amplifier input voltages (i.e. RDBL and REF) are close to V_{DD} for small-signal sensing, whereas they are close to ground for large-signal sensing. Fig. 6a shows the delay vs. REF plots for NMOS-input and PMOS-input sense-amplifiers. The delay of this structure is shown to be highly dependent on the common-mode of the input voltages in [8]. Below a certain

Authorized licensed use limited to: MIT Libraries. Downloaded on January 10, 2009 at 10:37 from IEEE Xplore. Restrictions apply.

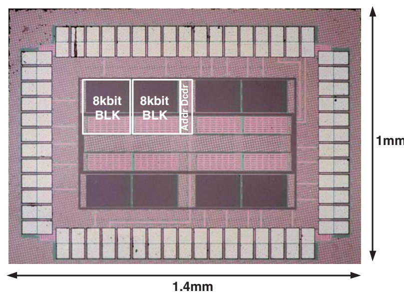

Fig. 7. Chip micrograph for the 64kbit SRAM implemented in 65nm CMOS process. The design consists of eight 8kbit sub-arrays.

level, as the common-mode input voltage decreases the delay of the sense amplifier increases. In order to keep the sensing delay low, two sense-amplifiers (one with NMOS-input and one with PMOS-input) are implemented along with a simple selection logic (Fig. 6b). At high voltage levels, the small-signal sensing scheme with the NMOS-input sense amplifier is activated. At low voltage levels large-signal sensing scheme is employed. Since the common-mode of the input signals is closer to ground level for large-signal sensing, PMOS-input sense amplifier is selected with this scheme.

III. Measurement Results

A 64kbit SRAM is fabricated in a 65nm low-power process. The array consists of eight 8kbit sub-arrays and address decoder (Fig. 7). The memory performance scales from 20kHz to 200MHz over the 250mV to 1.2V operating range (Fig. 8). Due to test setup limitations, the performance measurements for 0.8V-1.2V are done by measuring access time. Leakage power scales down by more than 50X over the voltage range resulting in very significant power savings. Leakage, active and total energy curves are shown in Fig. 9. Active energy decreases quadratically as supply voltage decreases. Leakage energy, on the contrary, increases as the supply voltage decreases because of integrating leakage power over a larger access period. Total energy reaches a minimum around 400mV.

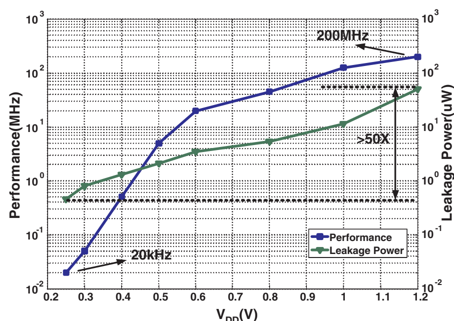

Fig. 8. Measured performance and power vs. V_{DD} plots. SRAM performance scales from 20kHz to 200MHz over the voltage range.

IV. Conclusion

An SRAM functional between 0.25V to 1.2V is presented in this paper. Sub-V_t operation requires additional circuitry

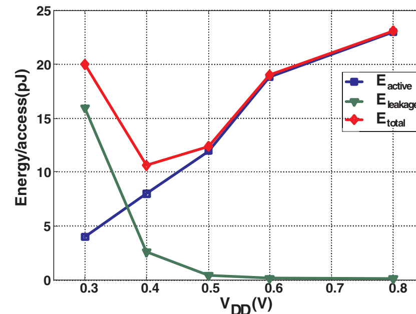

Fig. 9. Measured Energy/access vs. V_{DD} plot. Leakage and active components of the total energy is shown. Minimum energy point occurs around 400mV.

for correct functionality whereas standard techniques can be implemented for nominal V_{DD} levels. In order to maintain efficiency and prevent performance and power overheads, the SRAM features reconfigurable circuits to support this large voltage range and to manage different trade-offs associated with sub-V_t and above-V_t regions.

8T cell is sized for easy writability and good performance. One of the three write assist schemes are enabled depending on the supply voltage level to prevent power overhead. Multiplexed sense-amplifiers are used to minimize the sensing delay over the entire operating range. The design achieves 20kHz to 200MHz performance and more than 50X leakage power scaling.

Acknowledgment

This work is funded by DARPA and chip fabrication is provided by Texas Instruments. The authors thank Joyce Kwong and Masood Qazi for valuable discussions and support.

References

[1] A. Wang and A. Chandrakasan, “A 180mV FFT Processor Using Subthreshold Circuit Techniques,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2004, pp. 292–293.
[2] B. Calhoun and A. Chandrakasan, “A 256-kbit Sub-threshold SRAM in 65nm CMOS,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2006, pp. 628–629.
[3] N. Verma and A. Chandrakasan, “A 65nm 8T Sub-Vt SRAM Employing Sense-Amplifier Redundancy,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2006, pp. 328–329.
[4] T.-H. Kim, J. Liu, J. Keane, and C. H. Kim, “A High-Density Subthreshold SRAM with Data-Independent Bitline Leakage and Virtual Ground Replica Scheme,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2007, pp. 330–331.
[5] L. Chang, Y. Nakamura, R. K. Montoye, J. Sawada, A. K. Martin, K. Kinoshita, F. Gebara, K. Agarwal, D. Acharyya, W. Haensch, K. Hosokawa, and D. Jamsek, “A 5.3GHz 8T-SRAM with Operation Down to 0.41V in 65nm CMOS,” in *Symp. on VLSI Circuits (VLSI) Dig. Tech. Papers*, June 2007, pp. 252–253.
[6] B. Yu, E. Nowak, K. Noda, and C. Hu, “Reverse Short-Channel Effects and Channel-Engineering in Deep-Submicron MOSFETs: Modeling and Optimization,” in *Symp. on VLSI Technology (VLSI) Dig. Tech. Papers*, June 1996, pp. 162–163.
[7] K. Zhang, K. Hose, V. De, and B. Senyk, “The Scaling of Data Sensing Schemes for High Speed Cache Design in Sub-0.18μm,” in *Symp. on VLSI Circuits (VLSI) Dig. Tech. Papers*, June 2000, pp. 226–227.
[8] B. Wicht, T. Nirschl, and D. S-Landsiedel, “Yield and Speed Optimization of a Latch-Type Voltage Sense Amplifier,” *IEEE Journal of Solid-State Circuits*, vol. 39, pp. 1148–1158, July 2004.

Authorized licensed use limited to: MIT Libraries. Downloaded on January 10, 2009 at 10:37 from IEEE Xplore. Restrictions apply.