1.03pW/b Ultra-low Leakage Voltage-Stacked SRAM for Intelligent Edge Processors

Jingcheng Wang, Hyochan An, Qirui Zhang, Hun Seok Kim, David Blaauw, Dennis Sylvester
University of Michigan, Ann Arbor (email: jiwang@umich.edu)

Abstract

A stacked voltage domain SRAM is proposed where arrays are split into two sets (top and bottom) with their supplies connected in series. System supply current is reused by top and bottom sets, and supply voltage is divided among the two sets of arrays, enabling seamless integration of very low voltage SRAM retention in a larger system with a nominal supply, without need for an efficiency-reducing LDO. An array swapping approach provides stable access to arbitrary banks within one system clock cycle. A comprehensive sizing strategy (W&L) is employed to optimally balance hold stability and bitcell size. Integrated in an IoT imaging system in 40nm CMOS, the proposed 8.9Mb SRAM achieves 1.03pW/bit leakage, a >100× reduction over conventional SRAM in the same technology.

Keywords: low leakage, sub/near threshold, SRAM, voltage stacking, charge recycling, array swapping

Introduction

Intelligent IoT devices seek to fit complete neural network models into on-chip memories to avoid costly off-chip DRAM accesses. As a result, SRAMs can consume >80% of chip area and 90% of standby power [1]. Prior work focused on reducing SRAM leakage via techniques such as HVT/thick-oxide devices [2-3], reverse body bias [4-5], floating bitline [6], raising VSS [7-8], and lowering VDD [9-10]. Apart from HVT usage, which enables 10× leakage reduction and is readily deployed, supply voltage lowering is one of the most effective approaches to reduce leakage due to the DIBL effect. However, it raises two issues: 1) Commercial bitcells are not sized to hold data at subthreshold supplies and require a careful hold margin / density tradeoff; 2) Voltage regulation is required to generate the voltage level for SRAM arrays. LDOs are conventionally used, incurring area and power overheads due to efficiency loss. Voltage stacking generates an intermediate voltage level by placing voltage domains in series and has been used in microprocessors [11] and high bandwidth data buses [12]. The major challenge in voltage stacking is balancing the active current between top and bottom levels, to maintain a stable mid-rail voltage level. This often requires an additional voltage regulator, reducing the stacking benefits. SRAM arrays, however, are dominated by near-constant leakage (writing a bit draws 10s of pA average current versus µA-level background leakage), making them ideal for voltage stacking.

Voltage Stacking and Array Swapping

To allow access to arbitrary arrays during operation while avoiding insertion of complex level converters, we propose a novel array voltage swapping mechanism. The SRAM peripherals are not stacked and therefore wordline and bitline voltages remain at VDDcore (~2VDDmid) for faster operation speed, inherent write/read noise margin enhancement (Fig. 1). This also removes the need for level converters but, as a result, only bottom arrays can be read/written. When a top array is accessed, the memory controller first swaps the voltage of a bottom array in the same quad-array SRAM bank with the desired top array (Fig. 3). This swap mechanism ensures the leakage current remains balanced and is completed in one system clock cycle due to the relatively low IoT processer clock frequency. In addition to leakage reduction from reduced supply voltage, the approach offers an additional 2× leakage reduction in top arrays due to their inherent reverse body bias and reduced bitline leakage effects. As a result, total leakage is minimized by increasing the fraction of top arrays to > 50% (e.g., 75%); this is supported by measurements while the optimal ratio can be set by a memory controller.

Memory Cell and Super-Cutoff Read

Fig. 2 shows the bitcell schematic and layout. The cross-coupled 4T uses HVT devices to minimize hold leakage while LVT devices in the read port provide faster sensing speed. The bitcell is upsized for improved hold noise margin (HNM). Channel length is increased to the point where leakage is minimum (18% less), also improving HNM by 10% while incurring 8% cell density loss. Channel width is increased, initially improving HNM faster than leakage power, providing a favorable tradeoff. Final sizes are chosen to balance among density, HNM, and leakage. To decouple read/write operation, we use a Z8T structure [13] instead of a traditional 8T, as differential sensing provides faster read speed and larger sensing margin. Since in our stacked SRAM the array voltage is ~½ the bitline voltage, it inherently avoids the clamping current problem in the original Z8T as unselected cells are super-cutoff with negative VGS. Further, the stacked configuration results in word-line overdrive, greatly increasing write margin.

Bank Architecture and Swapping Control

Each SRAM bank has 4 arrays with power switches that connect an array to either top or bottom voltage domains (Fig. 3). Each bank can have 0−3 top arrays but at least one array must be in the bottom domain. When accessing a top array, the SRAM controller swaps this array with a bottom array in the same bank in two steps: First, the two arrays (target and swapping) are expanded to full voltage, after which they are collapsed to the appropriate half range. Since the two arrays are physically close, local charge sharing minimizes the disturbance to the mid-rail. All on-chip SRAM arrays in the system are connected to the same power/ground/mid-rail, resulting in a large amount of innate decoupling capacitance and background load current to suppress transient noise. To smooth transitions and reduce coupling noise, each power switch consists of both small and large headers/footers that are turned on in sequence. Each swap consumes ~8pJ, which is comparable to a single 128b read. To minimize the frequency of swaps, instruction memories (exhibiting mainly random accesses) are placed in the bottom domain, whereas neural engine memories with mostly sequential access patterns are primarily placed in the top domain. SRAM peripherals are power gated immediately after each access to reduce leakage.

Measurement Results

The proposed SRAM was implemented in a 40nm CMOS image processing IoT chip with 8.9Mb memories (Fig. 4). Fig. 5 shows measured leakage across voltage and temperature. As the number of top arrays increases, the mid-rail voltage rises while leakage continually drops (Fig. 6). Fig. 7 shows excellent mid-rail voltage stability; VDDmid varies only ±16mV across 100°C, drops ≤1.74mV when arrays swap every 11 cycles, and is unaffected with read/writes every cycle at full speed. The design operates at 438kHz at 0.7V (enabling 14fps in the supported image processing system) with 67fJ/bit access energy at 0.6V. Fig. 8 compares this work to other state-of-the-art low leakage SRAMs. The proposed work achieves 1.03pW/b at 0.58V without using large thick-oxide device, extra supply levels, or SOI with aggressive body biasing.

References

[1] S. Han et al., ISCA 2016.
[2] G. Chen et al., ISSCC 2010
[3] T. Fukuda et al., ISSCC 2014.
[4] R. Ranica et al., VLSI 2013
[5] M. Yabuuchi et al., VLSI 2017.
[6] Y. Wang et al., ISSCC 2009
[7] J. Chang et al., JSSC 2007.
[8] T. Kim et al., JSSC 2009
[9] F. Hamzaoglu et al., ISSCC 2008.
[10] N. Verma et al., JSSC 2008
[11] K. Blutman et al., JSSC 2017.
[12] J. Wilson et al., ISSCC 2016
[13] J. Wu et al., JSSC 2011.
[14] Q. Dong et al., VLSI 2017

978-1-7281-9942-9/20/$31.00 ©2020 IEEE

[15] A. Teman et al., JSSC 2011. [16] N. Maeda et al., JSSC 2013
[17] F. Tachibana et al., ISSCC 2013.
[18] F. Frustaci et al., ISSCC 2017. [19] H. Fujiwara et al., VLSI 2013

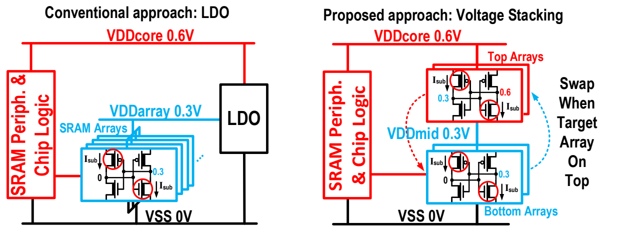

Conventional approach: LDO
VDDcore 0.6V
SRAM Periph. & Chip Logic
VDDarray 0.3V
SRAM Arrays
LDO
VSS 0V
Proposed approach: Voltage Stacking
VDDcore 0.6V
Top Arrays
SRAM Periph. & Chip Logic
VDDmid 0.3V
Swap When Target Array On Top
VSS 0V
Bottom Arrays

Fig. 1 Proposed array stacking and swapping technique.

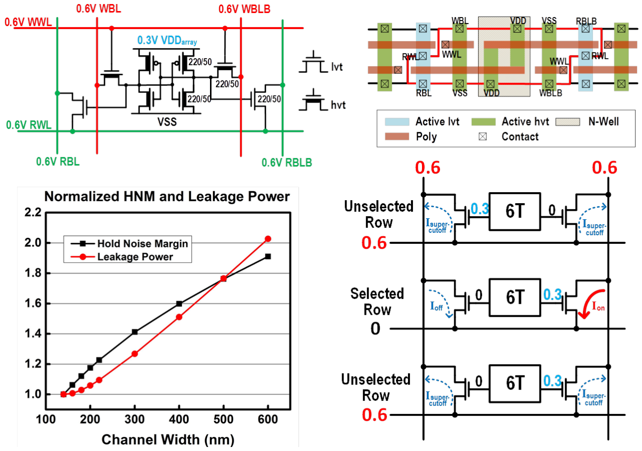


Fig. 2 Bitcell schematic and layout (top), hold noise margin (HNM) and leakage versus bitcell sizing (bottom-left), currents on the bitline during read operation (bottom-right).

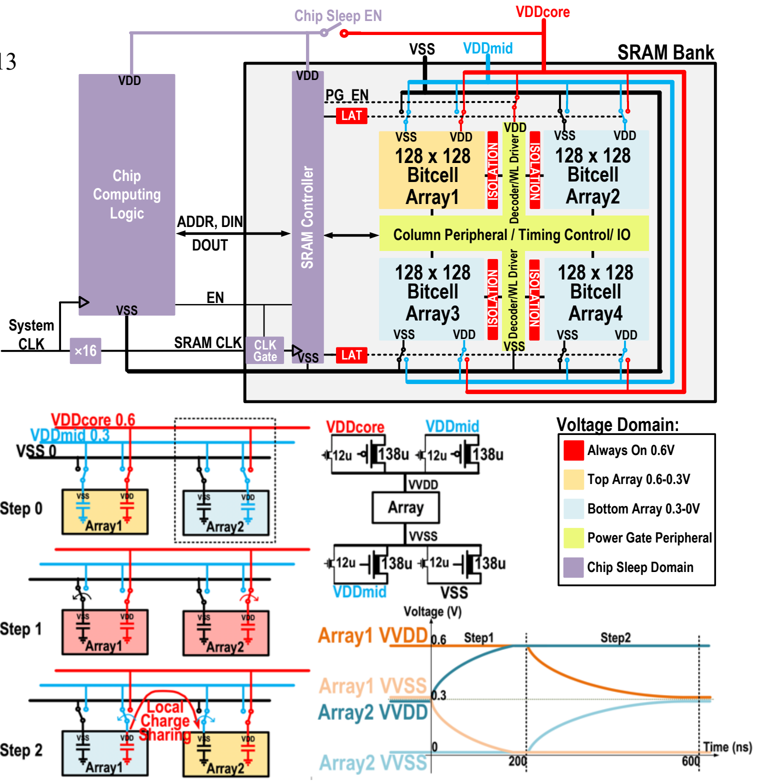


Fig. 3 SRAM bank architecture (top), array swapping algorithm and power switches (bottom).

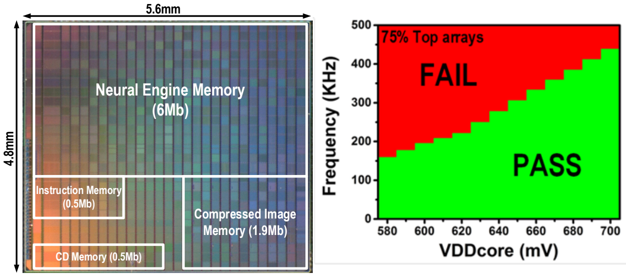


Fig. 4 Die photograph and shmoo plot.

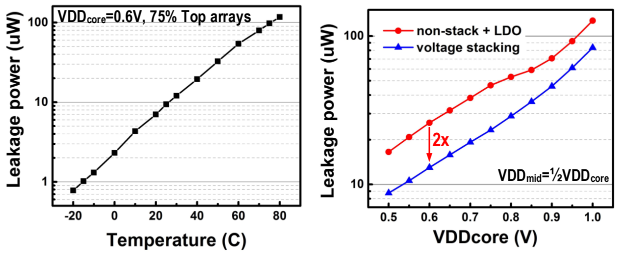


Fig. 5 Leakage across temperature and voltage.

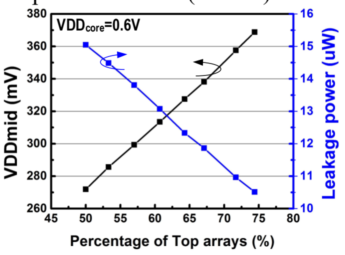
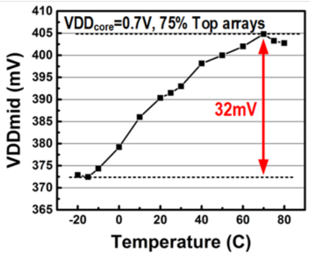


Fig. 6 Mid-rail voltage and leakage with number of top arrays.



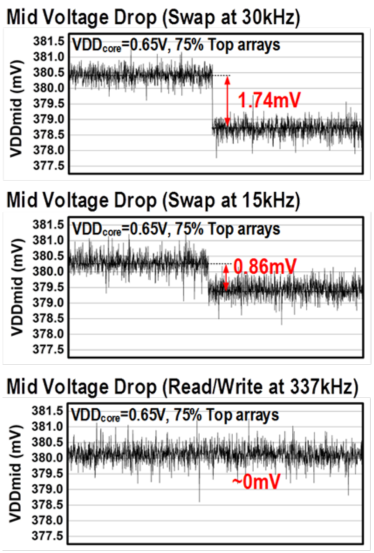


Fig. 7 Mid-rail variation with temperature (left), voltage drop due to various memory activities (right).

Acknowledgement

We thank Sony Semiconductor Solutions Corp./Sony Electronics Inc. for supporting this work

	This work	ISSCC10 [2]	ISSCC14 [3]	VLSI13 [4]	VLSI17 [5]	JSSC09 [8]	JSSC08 [10]
Process	40nm	180nm	65nm Thick Oxide	28nm FDSOI	65nm SOTB	130nm	65nm
Cell type	8T	10T	6T	6T	6T	8T	8T
Cell Area (um^2)	0.82*	17.48	2.159	0.12	0.5408	0.61	-
On-chip SRAM Capacity	8.91Mb	24kb	128kb	1Mb	8Mb	64kb	256kb
Leakage	1.03pW/bit (0.58V)	3.3fW/bit (0.4V)	32.4fW/bit (1.2V)	300fW/bit (0.6V)	13.7fW/bit (0.5V)	70pW/bit (0.23V)	6.45pW/bit (0.3V)
Extra Supply Level Required	No	Yes	No	Yes	No	Yes	No
Body Bias Voltage Required	No	No	No	-1.5V	-2V	No	No
Access time	143ns (0.7V)	13700ns (0.4V)	7ns (1.2V)	-	31.4ns (0.75V)	66.7ns (0.6V)	1250ns (0.6V)
Access Energy	67fJ/bit (0.6V)	-	195fJ/bit (1.2V)	-	224fJ/bit (0.75V)	25fJ/bit (0.4V)	-

*Logic design rule

	JSSCC11 [13]	VLSI17 [14]	JSSC11 [15]	JSSC13 [16]	ISSCC13 [17]	ISSCC14 [18]	VLSI13 [19]
Process	90nm	55nm	40nm	28nm	28nm	28nm	20nm
Cell type	8T	6T	9T	6T	6T	6T	6T
Cell Area (um^2)	-	0.803	1.058	-	0.12	-	-
On-chip SRAM Capacity	64kb	16kb	8kb	512kb	2Mb	32kb	128kb
Leakage	305pW/bit (0.23V)	97.6pW/bit (0.2V)	6.25pW/bit (0.4V)	10.6pW/bit (0.8V)	3.5pW/bit (0.7V)	109pW/bit (0.33V)	33.6pW/bit (0.6V)
Extra Supply Level Required	No	Yes	No	No	Yes	No	No
Body Bias Voltage Required	No	No	No	No	No	No	No
Access time	111ns (0.3V)	250ns (0.25V)	1000ns (0.4V)	0.42ns (1V)	-	-	1.16ns (0.9V)
Access Energy	1.15pJ/bit (0.5V)	5fJ/bit (0.25V)	11.3fJ/bit (0.4V)	-	-	93fJ/bit (0.6V)	69fJ/bit (0.6V)

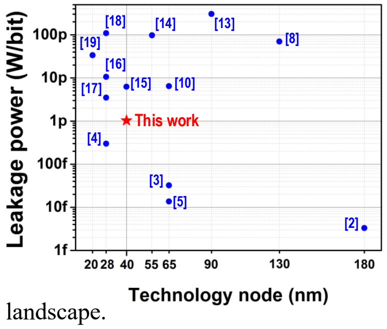


Fig. 8 Comparison table and design space landscape.