

# An 8-T ULV SRAM macro in 28nm FDSOI with 7.4 pW/bit retention power and back-biased-scalable speed/energy trade-off

Thomas Haine, Denis Flandre, David Bol

ICTEAM Institute, Electronic Circuits and Systems (ECS)

Université catholique de Louvain, Louvain-la-Neuve, Belgium

Email: thomas.haine@uclouvain.be, david.bol@uclouvain.be

**Abstract**—In this paper, we propose an ultra-low-voltage SRAM macro in 28nm FDSOI based on an 8-T ULP bitcell. Data retention in the proposed bitcell is obtained with two CMOS negative-differential resistance structures, which allow for using only low- $V_t$  transistors for density and speed without prohibitive leakage. The bitcell was sized to reach low failure rate for 32-kB macro with a gradient-based importance sampling methodology. Thanks to the exclusive use of low- $V_t$  transistors, adaptive back biasing allows for compensating PVT variations as well as scaling the power/performance trade-off. At 0.55V, the proposed SRAM can operate at up to 200 MHz and reaches an average access energy of 20.4 fJ/bit. In sleep mode with data retention, the leakage power goes down to 7.4 pW/bit.

**Index Terms**—SRAM, ULV, ULP, NDR, Importance Sampling.

## I. INTRODUCTION

Ultra-low power (ULP) Systems on chips (SoCs) for demanding Internet of things applications can rely on ultra-low-voltage (ULV) operation to maximize energy efficiency. ULV operation can be used for ULP SoCs while preserving descent logic speed performance in nanoscale CMOS [2]. However, these SoCs face a memory bottleneck due to the complex leakage/performance/area trade-off of ULV SRAMs. In [1], we proposed a 0.4V single-ended 7-T ULP bitcell with adaptive forward back biasing (ABB) in 28nm FDSOI based on two CMOS negative-differential resistance (NDR) structures to allow the use of only low- $V_t$  (LVT) transistors with the flipped-well architecture of 28nm FDSOI for density and speed without prohibitive leakage. However, this bitcell required the use of boosted-wordline (WL) assist technique with a limited density of 0.71 Mbit/mm<sup>2</sup> and no retentive sleep mode. In this paper, we show that the use of an extra transistor in an 8-T ULP bitcell combined with 0.55V supply allows for getting rid of the boosted-WL assist without frequency penalty and shrinking the bitcell transistor sizes, which result in a 40% density increase without access energy penalty. In sleep mode, the NMOS and PMOS back biasing voltages, noted BBN and BBP respectively, are tied at 0V to save leakage. We sized the bitcell to ensure data retention in this mode and reduce the total leakage by 4× compared to [1].

## II. PROPOSED ULV SRAM BITCELL

As shown in Fig. 1, the proposed bitcell based on [3] relies on two NDR structures forming a ULP latch with only LVT transistors and forward back bias applied to both NMOS and PMOS transistors. Despite these conditions, ultra-low-leakage is reached in the bitcell thanks to the self negative- $V_{gs}$  bias of the transistors in the cut-off NDR structure [3]. We use

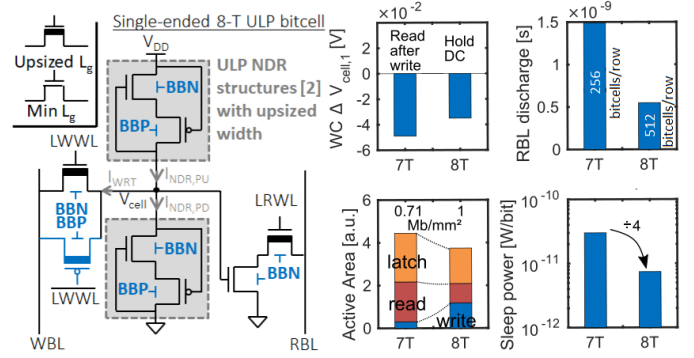


Fig. 1: (a) Proposed single-ended 8T ULP bitcell with the two ULP NDR structures, the 2-T read buffer and the write transmission gate. (b) Worst-case (WC)  $V_{cell}$  logic 1 level drop with respect to  $V_{DD}$ , (c) nominal RBL discharge time when performing a read after write for a logic 1, (d) active area normalized to a minimum-sized conventional 6-T bitcell and (e) sleep power (BBN=BBP=0V) (SPICE simulations in TT corner at 25°C at  $V_{DD}$ =0.55V).

a 2-T read buffer as in [1], but with a NMOS/PMOS write transmission gate connected to a single write bitline (WBL) (instead of an NMOS-only pass gate in [1]). This ensures a proper write-1 voltage on  $V_{cell}$  without boosting the local write WL (LWWL). In the 7-T bitcell, the read-after-write-1 speed is significantly degraded by the capacitive drop on  $V_{cell}$  at the release of the LWWL. This is alleviated in the 8-T bitcell by the reduced LWWL swing (not boosted) and the opposite swing of the PMOS  $\overline{LWWL}$  (Fig. 1.b.). Moreover, in the proposed 8-T ULP bitcell,  $V_{DD}$  is increased to 0.55V instead of 0.4V in [1]. As a result of these two modifications, the 2-T read buffer widths can be downsized while reaching a 3 times lower read-bitline (RBL) discharge time with 512 bitcells/column instead of 256 in [1] (see Fig 1.c). The hold stability depends on the current equilibrium between the NDR structures ( $I_{NDR,PU}$  and  $I_{NDR,PD}$ ) and the write transistors leakage ( $I_{WRT}$ ). Write and NDR transistor sizes were optimized to ensure statistical hold stability in both forward back-biased active mode (BBN=1V, BBP=-2V in TT corner at 25°C) and reduced-back-biased sleep mode (see Fig 3.a.). For fast statistical calculation of hold failure rate for a 32-kB macro, we used Gradient Importance Sampling (GIS) methodology from [4]. Fig. 2.a. shows that hold stability is more difficult in sleep mode due to the low back-bias voltage to save leakage, as this results in a lower NDR peak current than in active mode. Fig. 2.b. shows that the size optimization preserves hold stability in sleep mode with

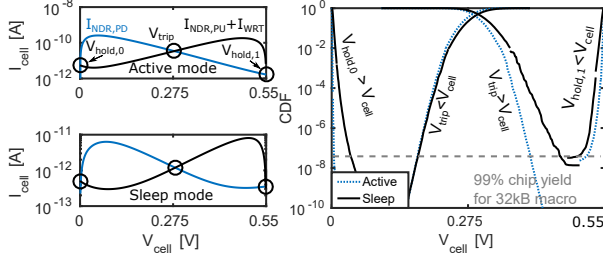


Fig. 2: (a) Retention of the bitcell in sleep and active modes (SPICE simulations with  $V_{DD}=0.55V$ , TT corner at  $25^{\circ}C$ ). (b) Assessment of the hold stability is done via GIS methodology. The curves represent the cumulative distribution function CDF of the failure rate.

99% chip yield. Compared to [1], this optimization allows for significantly reducing the latch area (Fig. 1.d.) while keeping write transistors with minimum width and upsized gate length to prevent loss of retention due to their leakage current thanks to the strong  $V_t$  roll-off in 28nm FDSOI. The total bitcell area is  $0.76\mu m^2$  without using the SRAM DRC pushed rules (25% reduction compared to [1] as shown in Fig 1.e.). Thanks to the optimized latch and read buffer, sleep power is divided by  $4\times$  (see Fig 1.e.).

### III. PROPOSED ULV SRAM MACRO

To minimize the access energy while avoiding the half-selected write pitfall, the SRAM uses divided-wordline (D-WL) architecture as in [1] but with 16 banks of 512 words of 32 bits for the 32-kB macro to preserve a good aspect ratio as shown in Fig. 3. Global wordline (GWL) have also been split in two (one for each half plane of the macro) to decrease further the dynamic energy. The row decoder features 1024 outputs, each of them attacks 8 words in 8 adjacent banks. To prevent a tangled RBL layout to the column mux, one sense amplifier (SA) is used per column as the columns of the same word bit are physically far away from each other with the D-WL architecture. A compact single-ended non-strobed SA based on a skewed inverter (see Fig. 3) precedes a low-capacitance minimum-sized second stage. This allows a fast SA output transition even at low RBL. The SA is followed by a compact tristate inverter used as bank multiplexer. This architecture allows to spare routing area compared to the standard transmission gate approach used in [1]. The read operation is the critical point of our design: the read-1 operation is relatively slow due to the large RBL capacitance with the 512 connected bitcells and in the read-0 operation, the leakage current of the 511 non-accessed cells in a column leads to a slow spurious RBL discharge. Fig. 4 shows the CDF of the bank delay. For a chip yield of 99% in a 32kB macro, the data have to be latched between 3ns and 14ns. Fig. 4 shows the evolution of the worst-case bank delay in function of the corner and the temperature. By adapting BBN/BBP voltages, we compensate PVT variations and prevent timing failures. Indeed, increasing the BBN/BBP voltages in slow corner (SS  $-40^{\circ}C$ ) preserves

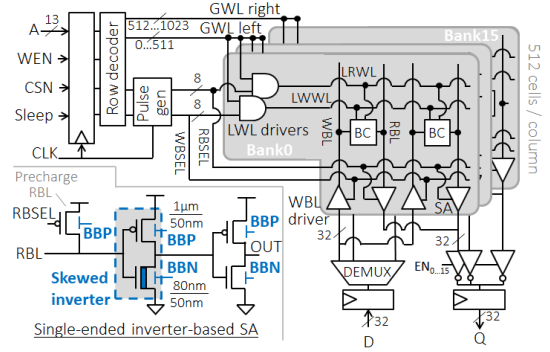


Fig. 3: Divided-wordline SRAM architecture and compact single-ended non-strobed SA schematic

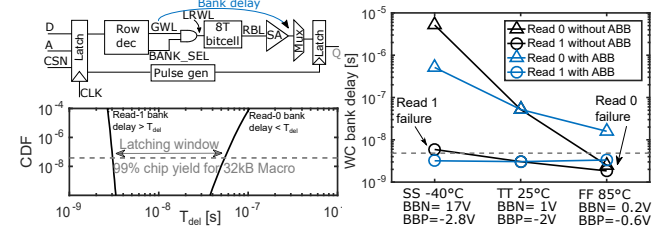


Fig. 4: (a) Bank delay schematic (b) Assessment of the nominal bank delay via GIS methodology (SPICE simulations with  $V_{DD}=0.55V$ , BBN=1V and BBP=-2V, TT,  $25^{\circ}C$ ). (c) Effect of the PVT and ABB on the bank delay (SPICE simulations with  $V_{DD}=0.55V$ ).

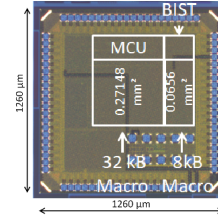


Fig. 5: Microphotograph of the  $1120 \times 1120 \mu m^2$  die in 28nm FDSOI.

short bank delay and lowering BBN/BBP voltages in fast corner (FF  $85^{\circ}C$ ) keeps the read buffer leakage under control to avoid prohibitively fast spurious RBL discharge. As the 32-kB macro is embedded in an MCU SoC, we use the adaptive BBN and BBP voltages generated by the ABB generator of the MCU to avoid extra power or area in the SRAM.

### IV. MEASUREMENTS RESULTS

Two macros with 32-bit words have been prototyped in a 28nm FDSOI test chip (Fig. 5): a 32-kB macro with byte-wise write mask ( $0.271 \text{ mm}^2$ ) connected to a ULV MCU and an 8-kB macro ( $0.065 \text{ mm}^2$ ) connected to a built-in self-test (BIST) featuring various test patterns. Most of the present measurements were performed on the 8-kB as the BIST offers faster testability. Fig. 6.a. shows a pass/fail map of a typical die in function of BBN and BBP voltages. The use of too low BBN/BBP voltages lead to timing errors but too high voltages lead to hold errors due to excessive leakage. Fig 6.b. gives the cumulative number of errors measured on 7 dies

TABLE I: State-Of-The-Art comparison

|                                              | [5]        | [6]       | [7]       | [8]        | [1]                        | This work                   |
|----------------------------------------------|------------|-----------|-----------|------------|----------------------------|-----------------------------|
| Publication                                  | VLSIC      | JSSC      | ESSCIRC   | TCAS-I     | ESSCIRC                    | S3S                         |
| Year                                         | 2013       | 2014      | 2016      | 2018       | 2017                       | 2018                        |
| Technology                                   | 28nm       | 28nm      | 28nm      | 28nm       | 28nm                       | 28nm                        |
|                                              | FDSOI      | FDSOI     | FDSOI     | FDSOI      | FDSOI                      | FDSOI                       |
| Size [kbits]                                 | 128        | 64        | 128       | 32         | 64                         | 64                          |
| I/O word width [bits]                        | 32         | 32        | 64        | 32         | 32                         | 32                          |
| Column mux                                   | Yes        | Yes       | No        | Yes        | Yes                        | Yes                         |
| Bit cell                                     | 6T         | 10T       | 7T        | 10T        | 7T                         | 8T                          |
| Area efficiency [Mbit/mm <sup>2</sup> ]      | -          | -         | -         | 0.3        | 0.71 †                     | 1.0 †                       |
| Supply voltages [V]                          | 0.6-0.9    | 0.35-1.2  | 0.24-0.36 | 0.25       | 0.4+0.7                    | 0.55                        |
| Max frequency [MHz]                          | -          | 13-1020   | 15-37.5   | 0.3        | 80                         | 80-200                      |
| Access energy [fJ/bit] (50% read, 50% write) | 67 @200MHz | 36 @13MHz | 36 @15MHz | 2.9 @30kHz | 20(SRAM)<br>+8(ABB) @80MHz | 20.4 @80MHz<br>33.9 @200MHz |
| Leakage power [pW/bit]                       | 34         | -         | 10-100    | 1.65       | 320 (active)<br>30 (sleep) | 40.7(active)<br>7.4 (sleep) |

† Not using SRAM DRC pushed rules.

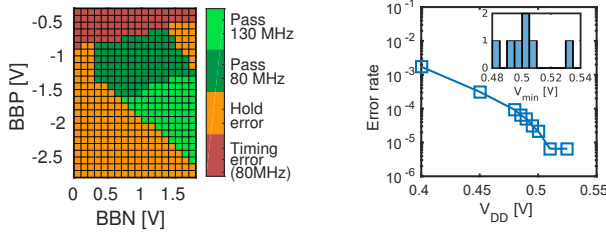
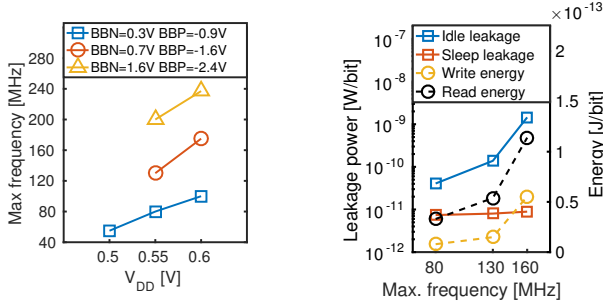


Fig. 6: (a) Error map of the SRAM in function of BBN and BBP. (b) Cumulative error rate on the tested 7 dies, with distribution of the minimum voltage without error.

Fig. 7: (a) Maximum frequency vs.  $V_{DD}$  for different pairs of BBN and BBP. (b) Evolution of leakage power and the access energy vs. BBN/BBP at  $V_{DD}=0.55V$ .

tested at 80MHz. The  $V_{min}$  of these dies ranges from 0.48 to 0.53V. Fig. 7.a. shows the maximum frequency without error. At 0.55V the SRAM can operate up to 200MHz. As shown in Fig. 7.b, the average access energy (50% write, 50% read) at 0.55V and 80MHz is 20.4 fJ/bit. The write energy (7.8 fJ/bit) is lower than the read energy (32.9 fJ/bit), which is largely dominated by the read-1 operation (58.4 fJ/bit) due to the RBL discharge. As was done for the LWL, splitting the RBL into local RBL might be a good way to improve this [7]. Finally, as shown in Fig 7.b., the leakage power reaches 40.7 pW/bit in active mode and goes down to 7.4 pW/bit in sleep mode, while preserving data retention.

## V. STATE-OF-THE-ART COMPARISON AND CONCLUSIONS

Measured performance is compared with recent low-energy SRAMs in Table I. Some previous SRAMs achieve lower access energy either at a lower frequency or by avoiding the

use of column muxes, which amortizes the access energy over more I/O word bits. For large macros, this implies a wide 64-bit I/O bus, which can lead to wasted energy at the system level in 32-bit ULP SoCs. Moreover, the proposed SRAM is fully tolerant to PVT variations, which is not mentioned for the other SRAMs from the research state-of-the-art. To summarize, we presented in this paper an ULV SRAM in 28nm FDSOI. The 8-T bitcell is based on CMOS NDR structures that allows using only LVT transistors for speed and density target without prohibitive leakage. Low failure rate of the bitcell is reached by sizing the transistors of the bitcell with GIS. Last but not least, PVT variations are compensated via ABB. At 0.55V, the proposed SRAM can operate at up to 200 MHz and reaches access energy of 20.4 fJ/bit. In sleep mode, data retention is achieved at 7.4 pW/bit. Compared to the ULV SRAM from [1], the density has been increased by 40% and the maximum frequency by 2.5x without increasing the energy, while reducing the sleep power by 4x.

## ACKNOWLEDGEMENTS

This work has been funded by the MSP-Project of the European Union, by the SAVE project of the Walloon region as well as the IDEES project co-funded by the European program FEDER and the Walloon region.

## REFERENCES

- [1] T. Haine et al., "An 80-MHz 0.4V ULV SRAM macro in 28nm FDSOI achieving 28-fJ/bit access energy with a ULP bitcell and on-chip adaptive back bias generation", in Proc. IEEE ESSCIRC, 2017.
- [2] D. Bol et al., "Can we connect trillions of IoT sensors in a sustainable way? A technology/circuit perspective (invited)", in Proc. IEEE S3S Conf., 2015.
- [3] D. Levaq et al., "Low Leakage SOI CMOS Static Memory Cell with Ultra-Low Power Diode", in IEEE JSSC, 2007.
- [4] T. Haine et al., "Gradient Importance Sampling: an Efficient Statistical Extraction Methodology of High-Sigma SRAM Dynamic Characteristics", in Proc. IEEE DATE, 2018.
- [5] H. Fujiwara et al., "A 20nm 0.6V 2.1μW/MHz 128kb SRAM with no half select issue by interleave wordline and hierarchical bitline scheme", in IEEE VLSIT, 2013.
- [6] F. Abouzeid et al., "Scalable 0.35V to 1.2 V SRAM bitcell design from 65nm CMOS to 28 nmFDSOI", in IEEE JSSC, 2014.
- [7] B. Mohammadi et al., "A 128 kb 7T SRAM Using a Single-Cycle Boosting Mechanism in 28-nm FD-SOI", in IEEE TCAS-I, 2018.
- [8] Y.C. Chien et al., "A 0.2 V 32-Kb 10T SRAM With 41 nW Standby Power for IoT Applications", in IEEE TCAS-I, 2018.