A 500 fW/bit 14 fJ/bit-access 4kb Standard-Cell Based Sub-V_T Memory in 65nm CMOS

Pascal Meinerzhagen*, Oskar Andersson†, Babak Mohammadi†, Yasser Sherazi†,
Andreas Burg*, and Joachim Neves Rodrigues†
*Institute of Electrical Engineering, EPFL, Lausanne, VD, 1015 Switzerland
Email: pascal.meinerzhagen@epfl.ch, andreas.burg@epfl.ch
†Department of Electrical and Information Technology, Lund University, Lund, 22100 Sweden
Email: oskar.andersson@eit.lth.se, babak.mohammadi@eit.lth.se, yasser.sherazi@eit.lth.se, joachim.rodrigues@eit.lth.se

***Abstract*— Ultra-low power (ULP) biomedical implants and sensor nodes typically require small memories of a few kb, while previous work on reliable subthreshold (sub-V_T) memories targets several hundreds of kb. Standard-cell based memories (SCMs) are a straightforward approach to realize robust sub-V_T storage arrays and fill the gap of missing sub-V_T memory compilers. This paper presents an ultra-low-leakage 4kb SCM manufactured in 65nm CMOS technology. To minimize leakage power during standby, a single custom-designed standard-cell (D-latch with 3-state output buffer) addressing all major leakage contributors of SCMs is seamlessly integrated into the fully automated SCM compilation flow. Silicon measurements of a 4kb SCM indicate a leakage power of 500 fW per stored bit (at a data-retention voltage of 220 mV) and a total energy of 14 fJ per accessed bit (at energy-minimum voltage of 500mV), corresponding to the lowest values in 65 nm CMOS reported to date.**

I. Introduction

Biomedical implants and sensor nodes, whose power and area budgets are often dominated by embedded memories, require ultra-low power consumption at low operating frequencies, and are therefore preferably operated in the sub-V_T domain. While logic circuits operate reliably in this domain, it is more difficult to build robust sub-V_T memories.

Commercial memory compilers are mainly oriented toward above-V_T operation and yield SRAM macros based on the conventional 6-transistor (6T) bitcell, which cannot be operated in the sub-V_T domain. Several research groups have proposed 8-transistor (8T) [1,2] or 10-transistor (10T) [3] SRAM bitcells reliably operating in the sub-V_T domain. However, such sub-V_T SRAM macros still have high leakage currents often dominating the leakage power of ultra-low-power (ULP) systems. To remedy excessive leakage currents, [4] has proposed a 14-transistor (14T) bitcell using high-threshold voltage (high-V_T) I/O transistors, stack forcing, and channel length stretching.

Designing dedicated 8T, 10T, or 14T SRAM macros for each new system and for each memory configuration is associated with a considerable design effort. Standard-cell based memories (SCMs) are an interesting alternative to full-custom sub-V_T SRAM macros in order to significantly reduce the design effort, ensure reliability, and even reduce the area cost for storage capacities smaller than a few kb [5]. However, as many SRAM macros [1–3], the SCMs presented in [5] suffer from high leakage currents, as they are implemented with latches from a commercial standard-cell library, which is primarily optimized for speed, but not for leakage.

Contributions: In this work, a new approach to the efficient design of embedded sub-V_T ULP memories is proposed. Relying on the fact that the bitcells (latches) together with the read multiplexers consume almost the totality of the leakage power of SCMs, a single custom-designed standard-cell, namely a low-leakage D-latch with 3-state output buffer, is integrated in the automated SCM compilation flow. As opposed to previous work [6], the proposed SCM design flow does not restrict the leakage minimization to the bitcells, but extends it to the peripheral circuits by using a 3-state read logic, accepting a speed degradation of the otherwise rather fast SCMs [5] for the benefit of lower leakage.

II. Custom Low-Leakage Latch Design

Approximately 66% of the leakage power of SCMs are consumed by the latches, whereas the read multiplexers dominate the remaining power. This section addresses the most dominant leakage contributors by a custom low-leakage latch design. Latch topologies using 3-state buffers inherently have transistor stacks and consequently low leakage currents, while topologies using transmission-gates and static-CMOS gates suffer from higher leakage currents [7]. The best latch topology exhibiting the lowest leakage current has 1) the lowest number of paths from V_{DD} to ground, and 2) the highest resistance on each such paths, directly leading to a topology with 3-state buffers only. Having identified the best latch topology, transistor stacking (for parts of the latch which do not yet have transistor stacks) and channel length stretching are applied to further reduce leakage currents. The stacking factor is strictly limited to 2 since higher factors give diminishing returns in leakage reduction [7] and compromise reliability for sub-V_T operation. Moreover, the point of diminishing returns of channel length stretching is found to be 1.5-2X minimum channel length [7]. The right-hand side of Fig. 1 shows the transistor-level schematic of the final custom-designed standard-cell latch (with 3-state output buffer), while the left-hand side shows the SCM architecture.

III. Low-Leakage 3-State Read Logic

The read multiplexers, routing the selected word to the data output, are an integral part of the read logic and can be

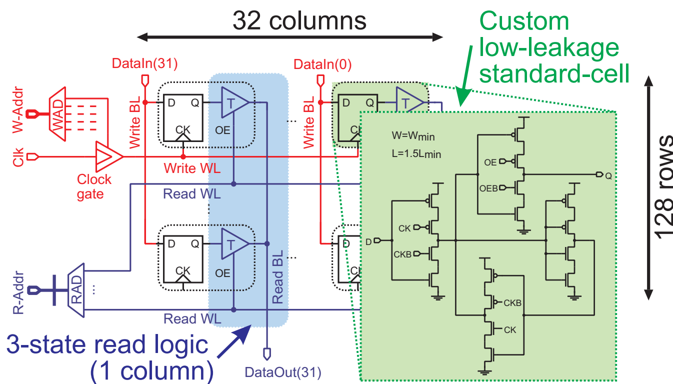

Fig. 1. Architecture of low-leakage 4kb standard-cell based memory (SCM): the write logic uses clock-gates [8], while the 3-state inverters used for the read functionality are integrated in the low-leakage latch design.

implemented with 3-state buffers [8], in order to address the dominant leakage contributor of SCM peripheral circuits. The already stacked output inverter of the custom-designed D latch is easily converted into a 3-state inverter, thereby addressing all major SCM leakage contributors by designing a single custom standard-cell.

The remainder of this section aims at finding the optimum transistor sizing of the 3-state drivers to simultaneously reduce overall leakage and improve speed, which is not contradictory in the sub-V_T regime, as expatiated on below. The presented 4kb SCM consists of 128 rows and 32 columns, as shown in Fig. 1. Thus, 128 3-state buffers are connected to the same read bit-line (RBL). During a read operation, the 3-state buffer in the selected word has to drive the RBL against 127 unselected, yet leaking 3-state buffers. To investigate the impact of the 3-state drive strength on the RBL (dis-)charge delay, a strong and a weak driver, defined in Table I, are considered. For a compact layout fitting nicely onto the standard-cell grid, and symmetric rise and fall times being only a secondary goal for the targeted low-speed ULP applications, the 3-state drivers are non-symmetric with equal NMOS and PMOS transistor sizes. As a result, RBL rise times are always longer than RBL fall times. Table I shows the 50%-to-50% rising-RBL propagation delay of the selected 3-state driver for the typical-typical (TT) process corner at 27 °C, for both above-V_T and sub-V_T supply voltages, and for both drive strengths. The considered low-power (LP) high threshold-voltage (HVT) 65nm CMOS technology has a nominal V_{DD} and a threshold-voltage of 1.2 V and 650 mV, respectively. Thus, a V_{DD} of 400 mV is already deep in the sub-V_T domain. Simulation results indicate that the stronger 3-state driver is faster for operation at nominal V_{DD} where on-to-off current ratios (I_{on}/I_{off}) are as high as 10^7 (for both NMOS and PMOS transistors), whereas the weaker 3-state driver is faster for sub-V_T operation, due to much lower I_{on}/I_{off} ratios of around 10^4 and the resulting non-negligible impact of the leakage current of unselected 3-state drivers.

IV. Reliability Analysis

While bitcell read-failures and write-failures are avoided by using a read buffer and by disabling the bitcell-internal keeper, respectively, hold-failures limit V_{DD} down-scaling [5]. To assess the minimum V_{DD} required to hold data (V_{DDhold}), the minimum V_{DD} for which both static noise margin (SNM) values (corresponding to data '1' and '0', or, in other words, to top and bottom eye of the butterfly curve [9]) are still positive are extracted from a 1k-point Monte Carlo (MC) circuit simulation (accounting for within-die (WID) parametric variations, in the TT corner, at 27 °C). Fig. 2 shows the hold-failure probability as a function of V_{DD}, while the inset shows the corresponding distribution of V_{DDhold}. The first hold-failure occurs at 200 mV, corresponding to a worst (maximum) value of V_{DDhold} equal to 210 mV.

Due to the strong impact of parametric variations and low I_{on}/I_{off} ratios in the sub-V_T regime, the total leakage current from a large number of disabled 3-state buffers might become high enough, compared to the active drive-current of a single weak 3-state buffer, to compromise the reliability of the 3-state read logic. However, 1k MC runs accounting for WID parametric variations in the slow-slow (SS) process corner at 27 °C indicate that for up to 128 words per RBL, a single 3-state driver successfully drives the RBL at a V_{DD} as low as 400 mV.

V. Silicon Measurements

Fig. 3 shows the chip microphotograph and the layout of the 4kb SCM based on 3-state-enabled low-leakage latches and manufactured in 65nm CMOS with LP-HVT transistors. The silicon area of the 4kb SCM block is 315 x 165 µm^2, corresponding to 12.7 µm^2 per bit. Functionality is verified by writing and reading back checker-board and random data patterns using a scan-chain test interface. Unless stated differently, the temperature is carefully controlled to 27 °C for all silicon measurements.

TABLE I
READ BIT-LINE (RBL) DELAY, TT CORNER, 27 °C.

Drive strength	**Strong**	**Weak**
W/W_{min}, L/L_{min}	15, 1	1, 2
V_{DD}	**RBL delay**	
1.2 V	1.064 ns	2.126 ns
400 mV	3.336 µs	2.688 µs

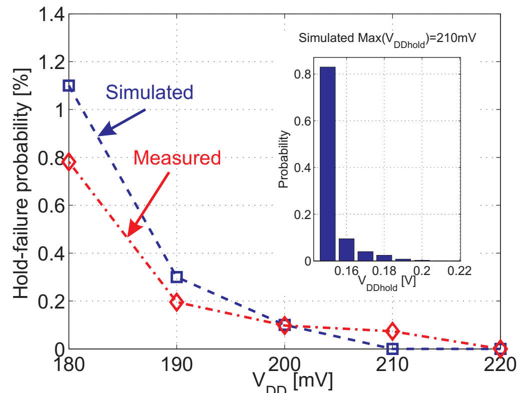

Fig. 2. Simulated and measured hold-failure probability versus V_{DD}. Inset: Simulated distribution of V_{DDhold}.

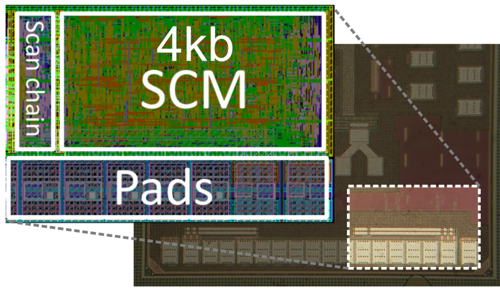

Fig. 3. Chip microphotograph and zoomed-in layout.

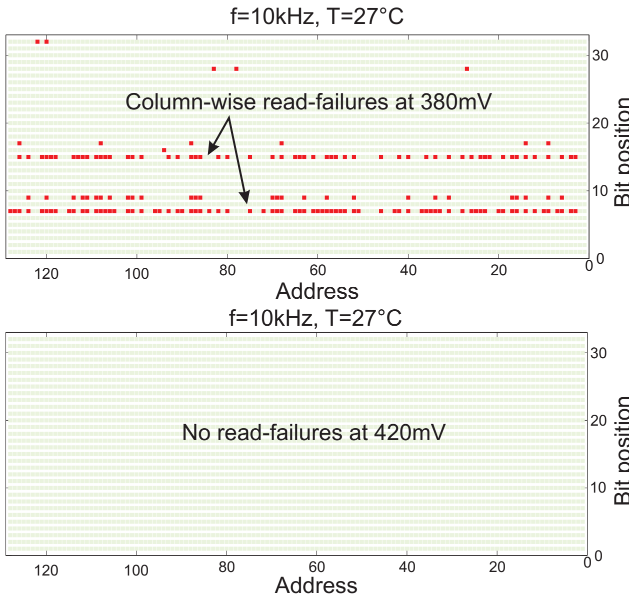

Fig. 4. Measured error maps for V_{DD} of 380 mV (top) and 420 mV (bottom).

A. Minimum V_{DD} for Data Retention and Memory Access

The measured minimum required supply voltages to guarantee correct hold, write, and read functionality are 220, 300, and 420 mV, respectively. The measured value of V_{DDhold} (220 mV) is in good agreement with the aforementioned simulated value (210 mV), as shown in Fig. 2. It is apparent that the low-leakage 3-state read logic limits the minimum voltage for read/write access (V_{DDmin}). For a closer inspection of the onset of read failures, Fig. 4 shows error maps: a green (bright) marker indicates correct access to a bitcell, while a red (dark) marker indicates an access failure. For $V_{\mathrm{DD}} = 380\,\mathrm{mV}$, it is apparent that failures occur column-wise, confirming that the 3-stated RBLs are the first point of failure under V_{DD} scaling. Completely error-free access is measured at $V_{\mathrm{DDmin}} = 420\,\mathrm{mV}$. Fig. 5 shows the the number of inoperative columns, i.e., columns containing at least one bitcell with access failure, as a function of V_{DD}, while the inset shows the total number of bitcell read-failures versus V_{DD}.

B. Access Energy, Frequency, and Leakage Power

Fig. 6 shows the measured energy per bit-access performed at maximum speed versus V_{DD}. The measured energy-minimum voltage is located at 500 mV, while the minimum

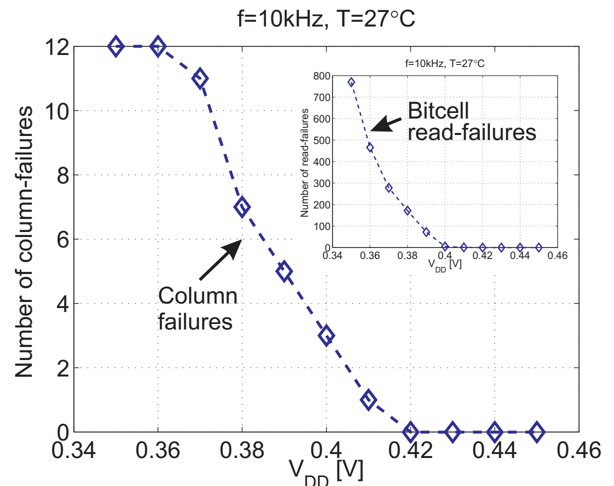

Fig. 5. Measured number of inoperative columns versus V_{DD}. Inset: Total number of read-failures versus V_{DD}.

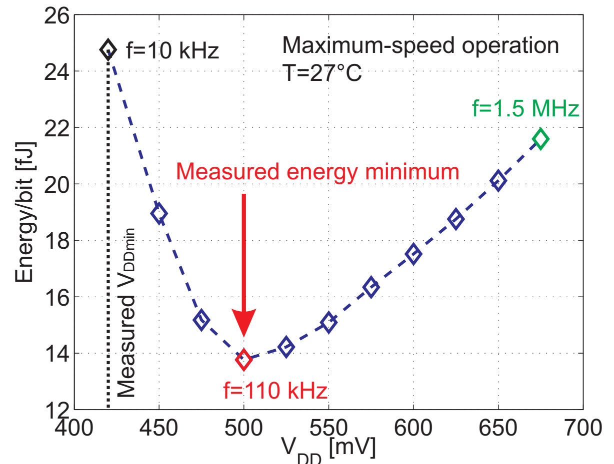

Fig. 6. Measured energy per bit-access.

energy dissipation per bit access is 14 fJ. At 675, 500, and 420 mV (V_{DDmin}), the maximum measured operating frequencies are 1.5 MHz, 110 kHz, and 10 kHz, respectively. The 3-state read logic limits V_{DDmin} and the read-access time, but satisfies the ambition of ultra-low leakage power and access energy, while the energy-minimum voltage is still higher than V_{DDmin}. At $V_{\mathrm{DDhold}} = 220\,\mathrm{mV}$, data is correctly held with a leakage power of 425-500 fW per bit (best and worst die), as shown in Fig. 7.

C. Measurements at Human-Body Temperature

Biomedical implants encounter a typical working temperature of 37 °C. At 37 °C, the first completely error-free read access to the entire array is measured at already 400 mV. As a desirable effect of higher temperatures, the maximum operating frequency doubles when heating the chips from 27 to 37 °C (measured at $V_{\mathrm{DD}} = 420\,\mathrm{mV}$). Unfortunately, the leakage power increases as well with increasing temperature, as shown in Fig. 7.

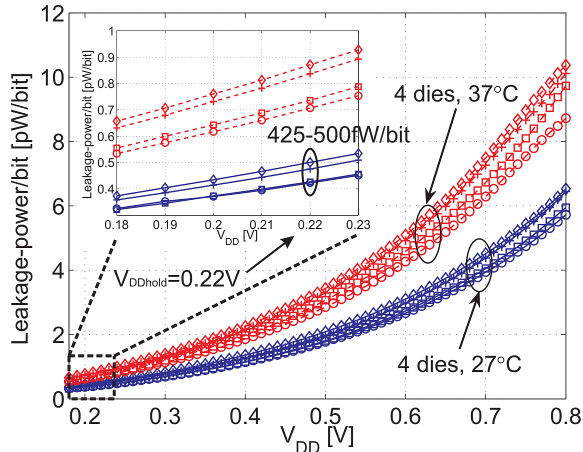

Fig. 7. Measured leakage power per bit, including overhead of peripheral circuits, measured for 4 dies, at 27 and 37 °C. Inset: Zoom around V_{DDhold}.

VI. Comparison with Prior-Art Sub-V_{T} Memories

Compared to a previous study on SCMs considering only commercially available standard-cell libraries [5], designing merely one custom standard-cell (3-state-enabled low-leakage latch) cuts the leakage power into half while maintaining the same silicon area.

Table II shows the best (in terms of access energy and leakage power) memories in 65nm CMOS reported to date. The energy figures ($E_{\mathrm{tot/bit}}$) correspond to the total (active and leakage) energy per memory access performed at maximum speed, normalized to the size of the data I/O bus. Unless stated in parentheses, $E_{\mathrm{tot/bit}}$ is given for V_{DDmin}. The power figures ($P_{\mathrm{leak/bit}}$) correspond to the leakage power of the memory macro (including peripheral circuits) during standby, normalized to the macro's storage capacity. Unless stated in parentheses, $P_{\mathrm{leak/bit}}$ is given for V_{DDhold}.

In [6], the standby leakage of the SRAM macro is dominated by the leakage of peripheral circuits, due to the aggressive reduction of array leakage. In this work, not only the bitcell (latch), but also the leakage-dominant peripheral circuits (read multiplexers) are leakage-optimized, which clearly pays off compared to [6] (see Table II).

With a total energy dissipation of 14 fJ per accessed bit and a leakage power of 500 fW per stored bit, the presented work outperforms all previous work in 65nm CMOS nodes. The reported clock frequencies are suitable for a wide range of biomedical applications, while most previously reported sub-V_{T} SRAMs are overdesigned. Even the silicon area of SCMs is smaller compared to sub-V_{T} SRAM hardmacros for storage capacities of up to several kb, due to less area for peripheral circuits [5]. For several tens of kb, an area-increase of roughly 4X [5], stemming from the larger bitcell, is acceptable for the benefit of the clearly lower leakage power and access energy.

VII. Conclusions

This paper addresses the lack of ultra-low-power (ULP) sub-V_{T} memory compilers by utilizing a fully automated standard-cell based memory (SCM) compilation flow, especially interesting for ULP biomedical systems requiring only small storage capacities of several kb. A single custom-designed standard-cell (D-latch with 3-state output buffer) is designed, addressing all dominant SCM leakage contributors at once, and integrated into the SCM compilation flow, cutting leakage power into half compared to using only commercial standard-cell libraries.

Silicon measurements show that a 3-state read logic with up to 128 words per bit-line operates reliably in the sub-V_{T} regime down to 420 mV. Counter to intuition, weaker 3-state buffers not only reduce leakage, but also shorten the bit-line delay compared to stronger 3-state buffers. The 4kb SCM manufactured in 65nm CMOS consumes a leakage power of 500 fW per stored bit (at data-retention voltage of 220 mV) and dissipates a total energy of 14 fJ per accessed bit (at energy-minimum voltage of 500 mV).

TABLE II
Comparison with Prior-Art Sub-V_{T} Memories in 65nm CMOS

	[3]	[2]	[6]	**This work**
V_{DDmin} [mV]	380	250	700	**420**
V_{DDhold} [mV]	230	250	500	**220**
$E_{\mathrm{tot/bit}}$ [fJ/bit]	54 (0.4V)	86 (0.4V)	-	**14 (0.5V)**
$P_{\mathrm{leak/bit}}$ [pW/bit]	7.6 (0.3V)	6.1	6.0, 1.0[a]	**0.5**

[a] Leakage-power of bitcell only

Acknowledgment

This work was kindly supported by the Swiss National Science Foundation (PP002-119057), Swedish Vetenskapsrådet (621-2011-4540), and Swedish VINNOVA Industrial Excellence Centre (SOS).

References

[1] Y.-W. Chiu, J.-Y. Lin, M.-H. Tu, S.-J. Jou, and C.-T. Chuang, "8T single-ended sub-threshold SRAM with cross-point data-aware write operation," in *Proc. IEEE ISLPED*, Aug. 2011.

[2] M. E. Sinangil, N. Verma, and A. P. Chandrakasan, "A reconfigurable 8T ultra-dynamic voltage scalable (U-DVS) SRAM in 65 nm CMOS," *in IEEE JSSC*, Nov. 2009.

[3] B. H. Calhoun and A. P. Chandrakasan, "A 256-kb 65-nm sub-threshold SRAM design for ultra-low-voltage operation," *in IEEE JSSC*, March 2007.

[4] S. Hanson, M. Seok, Y.-S. Lin, Z. Y. Foo, D. Kim, Y. Lee, N. Liu, D. Sylvester, and D. Blaauw, "A low-voltage processor for sensing applications with picowatt standby mode," *in IEEE JSSC*, April 2009.

[5] P. Meinerzhagen, S. M. Y. Sherazi, A. Burg, and J. N. Rodrigues, "Benchmarking of standard-cell based memories in the sub-VT domain in 65-nm CMOS technology," *in IEEE JETCAS*, Aug. 2011.

[6] Y. Wang, H. J. Ahn, U. Bhattacharya, Z. Chen, T. Coan, F. Hamzaoglu, W. Hafez, C.-H. Jan, P. Kolar, S. Kulkarni, J.-F. Lin, Y.-G. Ng, I. Post, L. Wei, Y. Zhang, K. Zhang, and M. Bohr, "A 1.1 GHz 12 uA/Mb-leakage SRAM design in 65 nm ultra-low-power CMOS technology with integrated leakage reduction for mobile applications," *in IEEE JSSC*, 2008.

[7] B. Mohammadi, P. Meinerzhagen, O. Andersson, Y. Sherazi, A. Burg, and J. Rodrigues, "A 0.28-0.8V 320 fW D-latch for sub-VT memories in 65-nm CMOS," in *Proc. IEEE CICC, under review*, Sept. 2012.

[8] P. Meinerzhagen, C. Roth, and A. Burg, "Towards generic low-power area-efficient standard cell based memory architectures," in *Proc. IEEE MWSCAS*, Aug. 2010.

[9] B. Calhoun and A. Chandrakasan, "Static noise margin variation for sub-threshold SRAM in 65-nm CMOS," *in IEEE JSSC*, July 2006.