A Loadless 6T SRAM Cell for Sub- & Near- Threshold Operation Implemented in 28 nm FD-SOI CMOS Technology

Even Låte[1,*], Trond Ytterdal[2], Snorre Aunet[3]

Department of Electronic Systems
Norwegian University of Science and Technology
O.S. Bragstads plass 2a, 7034 Trondheim

Abstract

Most ultra low power SRAM cells operating in the sub and near threshold region deploy 8 or more transistors per storage cell to ensure stability. In this paper we propose and design a low voltage, differential write, single ended read memory cell that consists of a total of 6 transistors. The innovative idea is to bring the loadless 4-transistor latch into the realm of low voltage memory cells by exploiting features of the 28 nm FDSOI Process and by adding a 2-transistor readbuffer with a footer line. Stand-alone and on a system level, the cell is stable during read, write and hold operations and it has great write-ability due to its differential write and loadless nature. The single NWELL option in 28 nm FD-SOI allows the loadless core to have minimal device widths while greatly improving the time it takes to evaluate the read bit-line. The cell has, in this paper, been used in a 128kb (2^{17}) SRAM in a 16 block configuration exploring 3 different types of logic libraries for the peripheral logic of the system. Depending on the application, the IO-peripheral logic may be implemented using either high threshold voltage transistors or low threshold voltage transistors in where the power consumption of the 128kb system was found to range from 1.31μW-71.09μW, the maximum operational frequency lies within 1.87 MHz and 14.97 MHz while the read energy varies from 13.08 to 75.21 fJ/operation/bit for a supply voltage of 350mV. The minimum retention voltage of the loadless SRAM cell is found to be 230mV covering 5σ of variation with Monte Carlo simulations.

Keywords: Loadless, Near-Threshold, Low-Power, SRAM, Memory-Cell.

1. Introduction

The number of devices connected to the Internet is predicted to increase from 12.5 billion in year 2010 to 50 billion within year 2020 by Cisco [1]. The explosive growth of on-line devices motivate advances in sensor node design as well as in data center efficiency. To enable this growth, an important area of research is energy efficiency for battery driven devices [2] and power efficiency in data center solutions to mitigate the dark silicon effect [3].

Static Random Access Memories (SRAMs) are heavily utilized in microprocessors as register files, instruction memories and data memories. Thus, SRAMs account for significant fractions of on-chip area and power & energy - consumption in microprocessors and therefore also in data centers and in sensor networks. To improve the energy & power efficiency in sensor networks and in data centers, innovations are needed within the field of SRAM design.

By designing digital circuits to operate in the sub- and near-threshold region one may trade operational speed for power consumption. Typically there is a sweet-spot for the supply voltage in terms of energy per operation that is normally located around the threshold voltage of the transistor, it is called the minimum energy point (MEP) [4]. For SRAM memories, the voltage where the MEP is located is typically higher than that of static CMOS logic [5], yet it is so low that operation on the MEP raises quite a few challenges where ensuring stability of all bit-cells in the memory arrays during hold, read and write operations are among the main concerns.

To lower the minimum operating voltage (V_{min}) of SRAM cells, architectural innovations are needed. Today most SRAM cells that are made for low supply voltages are made stable by adding read and write assists to the conventional 6T SRAM cells thereby increasing the number of transistors per cell to 8T [6], 9T [7], 10 T[8], 11T [9] 12T [10], 14T [11] and even higher numbers. In applications where large amounts of embedded memory is needed, having such large transistor count per memory cell and thus proportionally large area per memory bit leads to infeasible total area consumption. In this paper we therefore propose a 6T SRAM cell architecture that is suitable for low voltage applications. The cell has a retention voltage of 230mV, and for a 2^{17}bit (128kb), 16 block, SRAM configuration with I/O peripheral logic made from a custom,

*Corresponding author. Tel.: +47 92 89 33 21
URL: even.laate@iet.ntnu.no (Even Låte)
[1]PhD Candidate
[2]Professor & Supervisor
[3]Professor & Supervisor

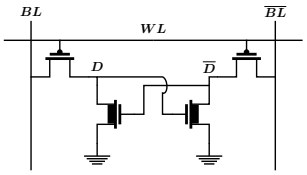
low leakage library, a leakage per bit of $9.99pW$ and energy per bit accessed in a 64bit word read-operation of $13.08fJ$.

In section 2, the proposed cell architecture is explained in detail, the cell sizing is elaborated with respect to the cell's hold and read stability of the cell together with the cell's write ability. The simulation setup is also listed in the same section. Following is the results in section 3 in terms of well established figures of merit for memory systems. The results are discussed in section 4 and conclusions are drawn in the same section.

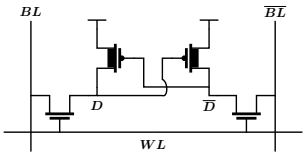
2. Methodology

2.1. Memory Cell Design, Stability & Reliability

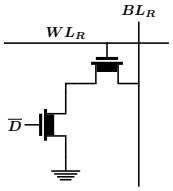
2.1.1. Memory Cell Design



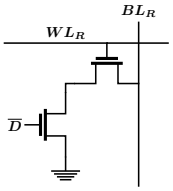
(a) Loadless SPW PDC [12].



(b) Loadless SNW PUC [13].



(c) 2T SPW RB.



(d) 2T SNW RB[14].

Figure 1: a) & b): Loadless single p-well (SPW) pull down (PD) and single n-well (SNW) pull up (PU) cores that are prone to read failures at low V_{DD}. c) & d): read buffers to mitigate the problem.

To achieve a low transistor count a loadless 4T cell core was created as shown in Fig. 1(a) by the Nippon Electric Company (NEC) for ultra-high density SRAM macros [12]. NEC based their work on the 4T NMOS memory cell with pull up resistors that has dominated the standalone SRAM market. By removing the resistive loads that requires a dedicated process step, NEC introduced the 4T loadless cell for embedded SRAM applications. The loadless pull down core structure in Fig. 1 (a) operates on a principle of high leakage PMOS pass-gates and low leakage cross coupled NMOS pull down transistors. Data retention is achieved by pre-charging the bitlines to VDD whenever the SRAM is idle. The high leakage PMOS pass gates then provides the cross coupled NMOS pair with supply current. Two design criteria for robust data retention in loadless pull down SRAM cells exist:

- The off current of the NMOS transistors should be significantly lower than the off current of the PMOS transistors in order to store a logic high.
- The on current for the NMOS should be significantly higher than the off current of the PMOS transistors in order to store a logic low.

Similarly for the loadless pull up core in Fig. 1 (b) [13], the bitlines are pre-charged to ground potential whenever the SRAM cell is in the hold state. To retain the stored data, the high leakage NMOS pass gates pulls sufficient supply current from the cross coupled PMOS pair. We have the following two design criteria for data retention:

- The off current of the PMOS transistors should be significantly lower than the off current of the NMOS transistors in order to store a logic high.
- The on current for the PMOS should be significantly higher than the off current of the NMOS transistors in order to store a logic low.

Both cell cores in Figs. 1 (a) and (b) were at first implemented on a schematic level. To fullfill the design criteria itemized above: Low threshold voltage PMOS devices (LVTPFETs) were used as access gates and regular threshold voltage devices (RVTNFETs) were used as pull down devices for the loadless pull-down cell. And for the pull-up cell, LVTNFETs were used as access gates and RVTPFETs were used as pull up devices. Exploiting the flip-well structure of the LVT devices in 28 nm FD-SOI allows the entire bitcell to be placed in a single P-well and a single N-well for the pull-down SRAM cell and for the pull-up cell respectively. Simulations proved that both loadless cores failed to retain their memory state during read operations for a supply voltage of 350mV. The differential read, differential write -cores were thus made into single ended read, differential write -cores by adding the single ended read buffers in Figs. 1 (c) and (d). The transistor count is increased from 4T to 6T to completely isolate the memory cells during read operations, equating the read margin to the hold margin for both cores.

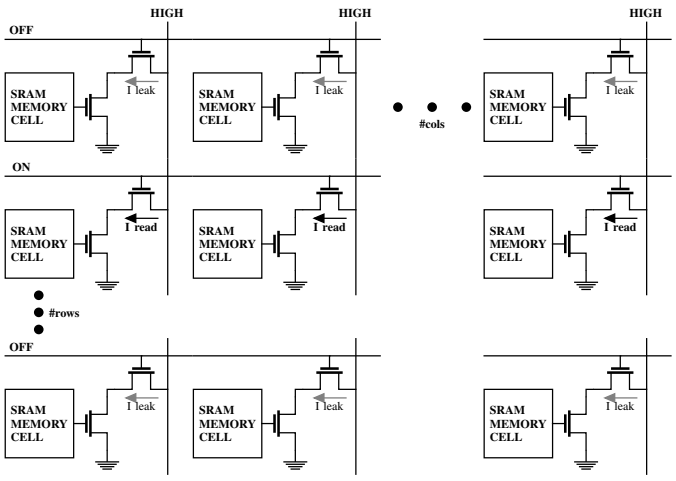
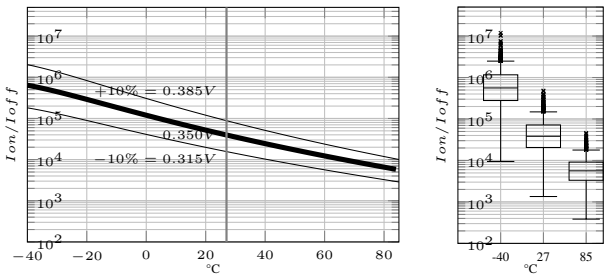


Figure 2: Read BL problem for conventional 2T read buffers.

The main issue with operating the conventional read buffer architectures in the near-threshold and subthreshold -regions, is the low I_{on}/I_{off} ratios of the transistors in the read buffers. The low on-current and relatively high off-current makes it difficult to distinguish the voltage differential on the bitlines caused by the non-selected rows

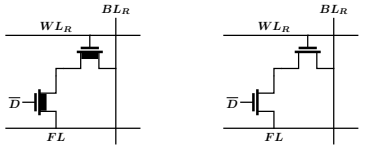


(a)I_{on}/I_{off}, T & V -variation. (b) I_{on}/I_{off}, P-variation

Figure 3: PVT variations of the I_{on}/I_{off} ratio for an LVTNFET. In a) V_{DS} & $V_{GS_{on}}$ varies ± 10% from the supply voltage of 350mV, the bold line in a) and for plot b).

in the same column from the read voltage differential generated by the accessed read buffer. Fig. 2 illustrates the bitline leakage problem. The higher the bitline is, in terms of column bitcells, the more cells are not accessed but still contribute to pulling down the pre-charged capacitance which again makes it more difficult to distinguish between logic 1s and 0s for the read-out logic. Since the readout buffer is single ended and, let us say that it is connected to the $\overline{D}$ node in the loadless SRAM core, the worst case scenario occurs when the accessed cell has $D = High$ with a corresponding $\overline{D} = Low$ and if all other cells in the column have $D = Low$, $\overline{D} = High$ ie. open buffer transistors and closed access transistors. The bitline is in this case supposed to retain its value since the accessed read buffer transistor is turned off, however BL_R ends up getting discharged by the sum of leakage currents from the unaccessed cells. This is also furthermore amplified by the impact on the I_{on}/I_{off} ratios from process voltage and temperature (PVT) variations. Fig. 3(a) depicts that the I_{on}/I_{off} ratio of an LVTNFET transistor varies by several orders of magnitude over the industrial temperature range from $T = -40°C$ to $T = 85°C$ with a supply voltage variation range of ± 10%. Process and mismatch variation adds even more variation to the I_{on}/I_{off} ratio which can be seen in (b).

To be able to perform read operations in the worst case



(a)SPW RB w/FL. (b)SNW RB w/FL.

	Read	Else
$\mathbf{BL_R}$	High → Value	High
$\mathbf{FL_{sel}}$	Low	High
$\mathbf{FL_{\overline{sel}}}$	High	High

(c)RB FL Functionality.

Figure 4: Used read buffer architecture and its functionality[15].

read situation mentioned above, a dynamic read footer line (FL) is used instead of the constant ground potential in the read buffer scheme to achieve a V_{DS} of 0 for all unaccessed read buffers. The footer-line read buffer scheme is depicted in Fig. 4. FL is, during the bitline evaluation phase of a read cycle, pulled to ground potential for the selected (sel) word-line. This eliminates the leakage from unaccessed readbuffers ($\overline{sel}$) and enables fast decision making in read operations [6].

Since the carrier mobility of N-channel devices is

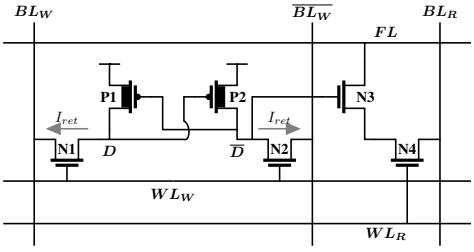


Figure 5: The proposed 6T loadless SRAM cell.

greater than that of P-channel devices, the loadless SPW pull-down core with LVTPFET passgates, would require very wide pmos access transistors, according to simulations 7.5 times the minimum width to satisfy the leakage design criteria introduced above. Instead, the SNW solution utilize the LVTNFET devices as the high leakage passgate allowing it to have minimum width, The SPW cell is therefore discarded. The remaining SNW loadless SRAM cell is not just smaller than the SPW, it also has the beneficial advantage of having LVTNFETs in the SNW readbuffer as well, greatly improving the bitline pulldown time. The complete schematic of the proposed 6T SNW loadless SRAM cell is shown in Fig. 5. It is realized with all minimum channel widths and with gate lengths equal to 48nm reducing cell leakage. With this configuration, the cell outperforms an implemented reference version of the 8T conventional subthreshold SRAM cell [6], both on the bitcell's minimum retention V_{DD} by 9% and on the bitcell leakage by 51.8% at $V_{DD} = 0.35V$. This is despite the reduction in transistor count from 8T to 6T. The reference cell was also made of all minimum width, 48nm length mixed V_{th} transistors in a SNW, and the comparison was done by running transient simulations with the write bitlines of both cells set to ground potential.

The memory cell is laid out in a flat, twisted and skewed layout scheme shown in Fig. 6. By twisted we mean that the area consuming cross coupling of the cell's core is avoided by twisting the right side 180 degrees around the horizontal axis, the halves of the SRAM cell are then skewed relative to each other to allow the gate of the right pull-up device to be connected to the drain of the left pull up device with a straight wire and vice versa. Also by making the layout in a flat manner with a horisontal row of transistors we avoid RVT to LVT spacing rules in the vertical direction, saving area directly. The height of the SRAM cell is limited by 3 minimum sized metal lines and 3 minimum sized line distances. To save one horizontal poly-poly spacing, the cell is flipped horizontally and the readbuffer pass gate poly is shared between neighboring columns in Fig. 7, in addition, the cell is flipped vertically to share supply rail connections between neighboring rows. By grouping 4 bitcells in one layout cell, memory

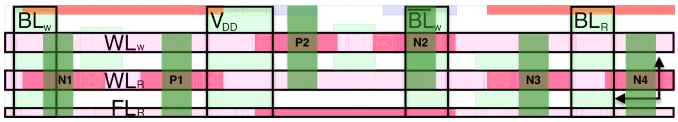


Figure 6: Layout of the proposed 6T loadless SRAM cell with shared RB footer line. Red = poly/metal gate, green = active area, light-blue = metal 1, light-green = metal 2, salmon = metal 3.

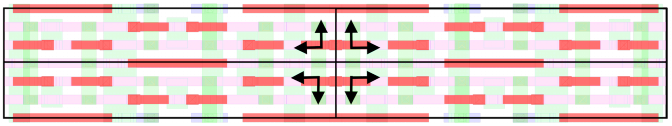
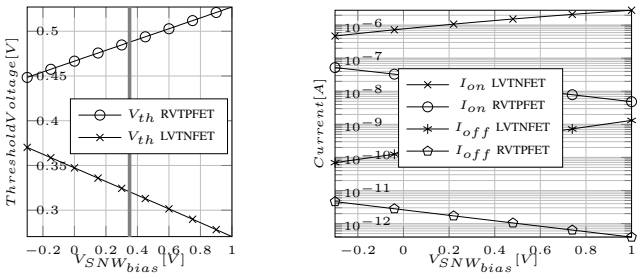


Figure 7: Layout of 4 Cells flipped horizontally and vertically to create mosaic-able cells.

	C_{BL_W} $[aF]$	$C_{\overline{BL}_W}$ $[aF]$	C_{WL_W} $[aF]$	C_{BL_R} $[aF]$	C_{WL_R} $[aF]$	C_{FL_R} $[aF]$
Pre-Layout	41.84	41.84	135.91	41.8366	67.95	41.84
Post-Layout nom	102.60	116.77	478.36	113.52	433.25	329.17
Post-Layout max	120.22	136.59	567.47	132.44	530.91	389.91
Post-Layout min	89.12	101.87	410.64	99.74	361.69	283.05

Table 1: Decoupled nodal capacitances on bitcell from extracted netlist.

matrices may be created with the mosaic functionality in the virtuoso layout editor achieving a SRAM cell area of $0.3\mu m$ x $1.824\mu m = 0.5472\mu m^2/bit$. Again, to provide a comparative metric, the layout of the conventional subthreshold 8T cell [6], was drawn using the single poly row approach. The area of this cell was found to be $0.3\mu m$ x $2.306\mu m = 0.6918\mu m^2/bit$, which reveals an area reduction for the loadless 6T core proposed in this paper of 20.90% in comparison. A benefit of choosing such a wide bitcell scheme is reduced bitline capacitance. Nodal capacitance for both pre and post-layout netlists are given in table 1, the post-layout capacitances are generated using the Cadence QRC layout extraction tool with the decoupled option activated for nominal, minimum and maximum corner scenarios. The above assumptions of low bitline capacitances for wide SRAM cells is confirmed by the difference between the post-layout BL capacitance and the post layout FL capacitance listed in the table. The corresponding penalty of choosing the flat layout scheme is increased WL capacitance which is also shown in the same table. It is however better to drive higher WL capacitances from a stronger row-decoder rather than increasing the size of the read buffer to pull down the higher BL capacitances through the area- and leakage- constrained memory cell. The post layout parasitic capacitance was also found for the 8T conventional subthreshold reference cell. The loadless 6T SNW cell has, with it's retention current to the write bitlines, no need for vertical ground lines, reducing the C_{FL} by 21.3%, C_{WL_R} by 17.6% and C_{WL_W} by 16.3% in comparison to the 8T cell. This again contributes to a lower dynamic power consumption since less parasitic capacitance is switched on memory operations.



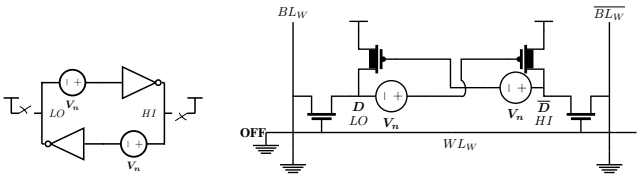
(a) Threshold Voltage. (b) Ion & Ioff.

Figure 8: Design knob, biasing the SNW from -0.3V to 1V at $V_{DD} = 350mV$.

2.1.2. Biasing the SRAM Cell's SNW

Since all devices in the SRAM cell are placed in a single well, they all inherit the same back gate bias conditions. By biasing the back gate of the entire SRAM cell, the threshold voltage of the RVTPFETs increases with increasing back gate voltage while the threshold voltage of the LVTNFET decreases as Fig. 8 (a) shows. This again affects the on and off currents of the RVT and LVT devices in the cell in an inverse relationship as shown in (b). I_{on} LVTNFET & I_{off} RVTPFET increases with increasing $V_{SNW_{bias}}$, while I_{on} RVTPFET & I_{off} LVTNFET decreases with increasing $V_{SNW_{bias}}$. Since the LVT pass gates are used as SRAM load, and since the pull up devices react inversely to a change in the bias condition, the bias voltage therefore impact the logic high voltage level of the memory cell. A logic high voltage close to that of the pull up voltage is desired for data retention.

2.1.3. Hold & Read Stability



(a) General SNM. (b) Loadless Core SNM.

Figure 9: Hold and read SNM for the loadless 4T pull up core.

To maximize the stability of SRAM cells in low supply voltages, it is desirable to strengthen the static noise margin (SNM) of the internal latch, both while retaining and reading data. The hold SNM for a conventional SRAM cell is defined as the maximum value of V_n in Fig. 9a that the cell may tolerate without flipping it's stored state unintentionally[16]. The hold SNM definition of the loadless 6T SRAM cell is directly derived from that of the conventional SRAM latch and is shown as V_n in Fig. 9b, note that the latch is disconnected from the read bitline when retaining data. Since this also is the case during read operations on the SRAM cell, the same definition may be used for the SNM in the read state making the SRAM cell

implemented here as read stable as it is hold stable. We can model static noise margins of SRAMs analytically by finding the voltage transfer curve (VTC) of each of the SRAM's compromising inverters, and then, after inverting one of the VTCs, by finding the length of the side of the maximal square fitted inside each eye of the resulting butterfly plot.

For SRAMs operating below the threshold voltage the VTCs of the inverters may be derived using the widely adopted model for subthreshold drain current in MOSFETs and Kirchoff's current law [17].

$$I_{DS} = I_S e^{\frac{V_{GS}-V_{th}}{nV_T}} \left(1 - e^{\frac{-V_{DS}}{V_T}}\right)\left(1 + \lambda V_{DS}\right) \tag{2.1}$$

The exponential relationship between the drain current and node voltages for an NMOS transistor in the subthreshold region is given by Eq. 2.1. V_{th}, V_{GS} and V_{DS} is the threshold voltage, the gate to source and the drain to source voltage respectively. V_T is the thermal voltage given by $V_T = kT/q$ for an absolute temperature T, with the Boltzmann constant, k, and the elementary charge q. n is the subthreshold slope factor that is equal to $1+C_D/C_{ox}$. I_S is the current that flows through the transistor for a $V_{GS} = V_{th}$. We included a model for the Drain Induced Barrier Lowering (DIBL) effect to get a more realistic slope on the VTCs, the incorporated factor is $1 + \lambda V_{DS}$, where λ is the output impedance constant of the transistor.

$$I_{DSn} = I_{DSp} \tag{2.2}$$

For an arbitrary loadless inverter in the loadless SRAM cell, the current through the pullup PMOS is equal to the leakage current through the LVT NMOS in both the retention state and in the read state, Eq. 2.2. Inserting Eq. 2.1 into Eq. 2.2 yields Eq.2.3, it is further transformed into the closed-form expression in 2.4 to obtain the VTC for the loadless inverter.

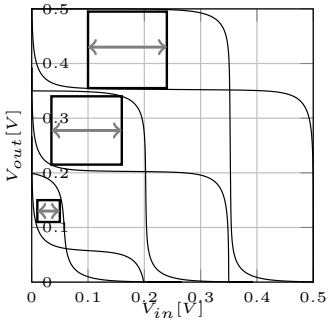
$$I_{Sn} e^{\frac{-V_{thn}}{n_n V_T}} \left(1 - e^{\frac{-V_{out}}{V_T}}\right)\left(1 + \lambda_n V_{out}\right) = I_{Sp} e^{\frac{V_{DD}-V_{in}-|V_{thp}|}{n_p V_T}} \left(1 - e^{\frac{-V_{DD}+V_{out}}{V_T}}\right)\left(1 + \lambda_p (V_{DD} - V_{out})\right) \tag{2.3}$$

$$V_{in} = V_{DD} - |V_{thp}| - ln\left(\frac{I_{Sn}}{I_{Sp}} e^{\frac{-V_{thn}}{n_n V_T}} \frac{\left(1 - e^{\frac{-V_{out}}{V_T}}\right)}{\left(1 - e^{\frac{-V_{DD}+V_{out}}{V_T}}\right)} \frac{\left(1 + \lambda_n V_{out}\right)}{\left(1 + \lambda_p (V_{DD} - V_{out})\right)}\right) n_p V_T \tag{2.4}$$

Parameters for the minimum sized LVTNFET and the minimum sized RVTPFET are extracted from the 28nm FDSOI kit and are listed in Fig. 10a. These are used to generate the butterfly plots in Fig. 10b. It can be observed from the figure that the plots have large eyes, for the targeted supply voltage og 350mV the SNM for read and hold is equal to 125mV. In adition to the analytical

	LVTNFET	RVTPFET
$V_t h$[V]	0.4113	0.4952
$\lambda\ [mV^{-1}]$	109.7	116.3
I_S [μA]	28.7	8.33
n	1.457	1.390
V_T [V]	0.026	0.026

(a)Model Parameters.

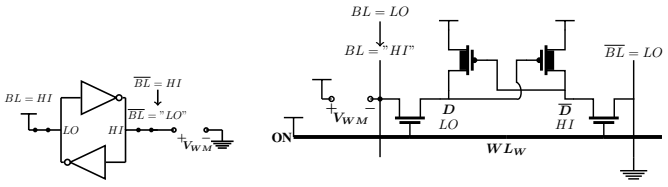


(b) Butterfly Plots varying V_{DD}.

Figure 10: Model Parameters from the targeted area of operation with 0V back-gate bias and butterfly plots for an analytical evaluation of read & hold SNM.

analysis of the read and hold SNM, transient simulations were performed with the data node D and $\overline{D}$ in Fig. 9b initialized to a logic low and a logic high respectively, this is given by the direction of the static noise sources in the figure. As transient static noise sources, verilog-A voltage ladders were used with long enough settling times for the circuit to settle to its steady state in between voltage increments. The read/hold-SNM is compared that of an implemented reference 6T cell in additional simulations.

2.1.4. Writability



(a) General WM. (b) Loadless Core WM.

Figure 11: Write margin for the loadless 6T SRAM cell.

For conventional 6T SRAM cells the bit-lines are precharged to V_{DD} in data retention mode and in read mode. To write to the bit-cell, one of the bit-lines is pulled low while the other one remains at V_{DD}. Fig. 11(a) depicts the write operation of the conventional SRAM latch, here the write margin, V_{WM}, is defined as the largest logic low bitline voltage that intentionally flips the state of the SRAM cell [18]. The higher the write margin, that is, the less one of the bitline has to be pulled down in order to write data to the cell, the greater the write-ability of the cell.

An analytical expression for the write margin of the cell can not be obtained with a similar methodology as in the previous section, this is due to the fact that there are no pull-down transistors in the latch active in the write phase, greatly improving the writability.The loadless cell proposed here has a default precharged bit-line voltage at ground potential to retain the data within the cells in hold

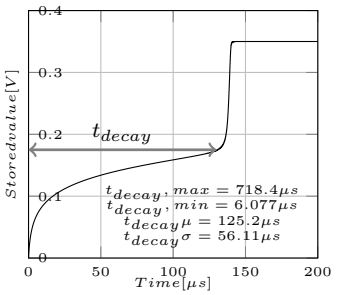
mode and in read mode. During a write operation, the desired bit-line should be pulled up and not down in order to flip the state of the cell, this implies a corresponding change in the write margin definition. The write margin of the loadless pull-up cell is equal to V_{DD} minus the required voltage increment on the bit-line to flip the cell. The less the bit-line has to be pulled up for a write operation, the greater the writability is for the cell.

To find the write margin with simulations for the loadless 6T SRAM cell, transient write simulations are performed with the data node D and $\overline{D}$ in Fig. 11(b) initialized to a logic low and a logic high respectively. The dc value of the bit-line BL is incremented in consecutive transient simulations until the cell succeeds to toggle and store the state of the cell after WL_W goes low again and closes the pass-gates. Also in the case of writability, the loadless core is compared to that of an implemented conventional 6T cell.

2.1.5. Write Stability for Unaccessed Cells in the Accessed Column

	C_D [aF]	$C_{\overline{D}}$ [aF]
Pre-Layout	146.74	214.69
Post-Layout nom	169.17	219.75
Post-Layout max	198.47	258.04
Post-Layout min	**147.44**	192.13

(a)Capacitance of D&$\overline{D}$.



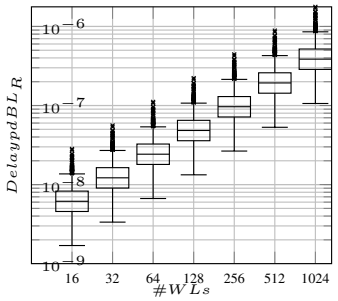
(b) Neighbour data-decay during write.

Figure 12: Maximum allowable time a bitline may be pulled up during a memory write operation for a full write-bitline swing.

The loadless core retains its data by holding the write bitlines at ground potential, forcing a voltage across the series RVTPFET and the LVTNFET of VDD, this gives origin to the retention current from VDD to the write bitlines. During a write operation, all memory nodes that are connected to the bitline that is pulled up, and that stores a logic 0 on the pulled up side, relies on their internal data node capacitance to retain the stored data during the column write access. This implies a maximum time that a bitline may be kept high, and therefore a timing requirement for write operations. Monte Carlo simulations of the nodal voltage-increase with a defined upper limit of the logic low node of 50% give rise to a maximum allowed bitline high-time of $6.077\mu s$ as Fig. 12(b) shows. The Monte Carlo run is here performed on the worst case internal nodal capacitance, that is the lowest value of the extraction corners listed in Fig. 12(a).

#WLs	μ	σ	min	max
16	6.9	3.36	1.7	28.2
32	13.6	6.7	3.4	55.9
64	27.1	13.2	6.7	111.2
128	54.0	26.4	13.3	221.8
256	107.9	52.7	26.6	443.1
512	215.7	105.3	53.1	885.6

(a) Delay Data in ns.



(b) Boxplot of the delay.

Figure 13: Read Delay vs # of WLs in 3σ MC simulations.

2.2. Memory Matrix, Peripheral logic & timing

2.2.1. Choosing the Memory Matrix Dimensions

The number of rows in an array of SRAM cells significantly impact the time it takes to evaluate the read bit lines. To illustrate this aspect of the dimensioning of the SRAM configuration, the time it takes to discharge the read bitlines for 16 to 512 words is given in Fig. 13. The worst case extracted capacitance corner is used in the Monte Carlo analysis.

For this work, to enable comparison to existing, state of the art, ultra low voltage SRAMs [19][20][21], a 128kb SRAM (2^{17}) is built up from 16 x 8kb SRAM sub-arrays of the loadless 6T memory cell. The dimensions of each sub-array or "block" is set to be 128 rows of 64-bit words to fit in modern 64 bit CPU architectures.

The 128x64x16 system configuration that is used as a benchmark for the 6T loadless SRAM cell is shown in Fig. 14. It is implemented with 3 different custom logic libraries designed for a supply voltage of 350mV. The first using minimum sized LVT transistors, the second using minimum sized RVT transistors and the third using RVT transistors with gate lengths of 48nm trading off speed for power consumption. All libraries are implemented with switching points of $V_{DD}/2$. The timing diagram for read and write operations on the 2^{17} SRAM system is shown in Fig. 15. The timing module is written as a 16-state state-machine in VHDL and is synthesized using Synopsys design compiler.

2.3. Simulation Setup

All simulations are performed with $gmin = 10^{-18}$, $reltol = 10^{-9}$, $vabstol = 10^{-9}$ & $iabstol = 10^{-15}$ to achieve accurate current simulation accuracy in the femto-amp range. Mismatch and process Monte Carlo simulations with specific sigma accuracy are performed using the Low-Discrepancy Sequence (LDS) sampling algorithm with yield targets set to the desired expected fraction of population inside range. All Monte Carlo simulations that imply functionality and yield are performed with 5σ accuracy. A sigma accuracy of 5 corresponds to a LDS yield target of 99.99994267%, that is, 1 in 1744278 samples lies outside of the simulation results which is a suitable number for a 2^{17}bits SRAM.

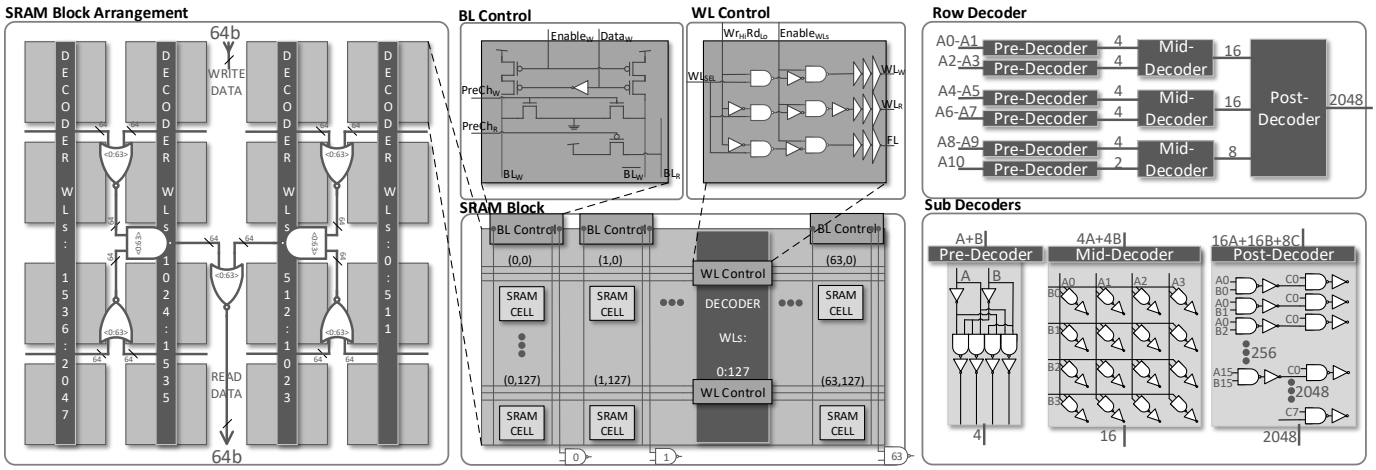


Figure 14: 2^{17}bit SRAM array structure for a configuration of 128 rows x 64 columns x 16 blocks.

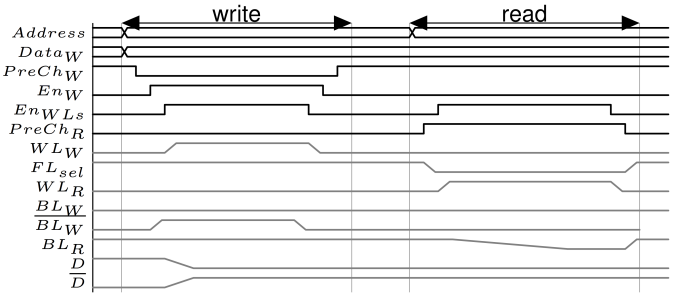


Figure 15: Timing diagram for a write and a read operation on the SRAM system in Fig. 14.

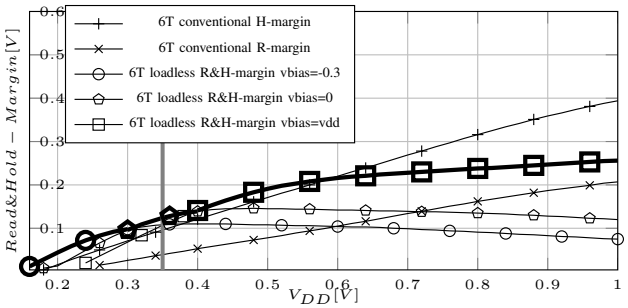


Figure 16: Read and hold margin for the loadless and conventional SRAM cells.

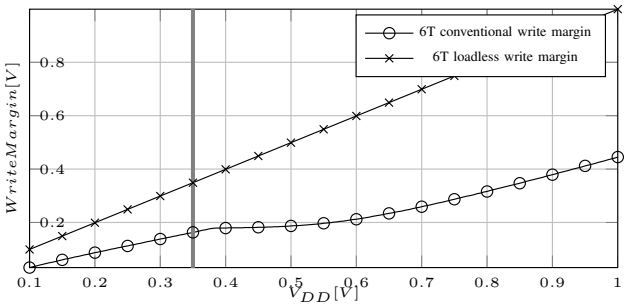


Figure 17: Write margin for the loadless and conventional SRAM cells, loadless is independent of vbias.

All experiments on the total 2^{17} SRAM configuration are performed with a reduced netlist where only the accessed row and column are instantiated with extracted models, this to reduce the simulation time to a feasible level. To achieve an accurate energy-per-operation metric, the read-time is multiplied by the 5 σ mean idle leakage power of all removed non-activated cells and the resulting energy is then added to the system's simulation result. This was done for the 15 non-selected SRAM blocks and for the SRAM cells that are not connected to the actived bit-line or the actived word/foot-lines of the selected block.

3. Simulation Results

3.1. Robustness of the Loadless 6T SRAM Cell

The hold and read stability of the SRAM cell is given in Fig. 16 as a function of supply voltage for different back gate bias potentials. In the same figure the read and hold SNM of a conventional 6T SRAM cell are used as a comparative metric. The bold line indicates the best biasing condition for the loadless SRAM cell at given values of V_{DD}. It can be noted that the bold line is superior in comparison to the read margin of the conventional SRAM cell for all values of the supply voltage. For a VDD of 350mV a bias voltage of 0V gives the best hold and read margin for the loadless core.

The write margin of the loadless core is quite interestingly found to be close to V_{DD} for all values of V_{DD} as Fig. 17 shows. For a 350mV voltage supply, the write margin is, thanks to the loadless nature of the cell, found to be approximately equal to 349mV.

3.2. Minimum Supply Voltage

Fig. 18 is made from multiple 5 σ monte carlo simulations revealing that the minimum retention voltage of the SRAM is 230mV for a single NWELL bias of -300mV.

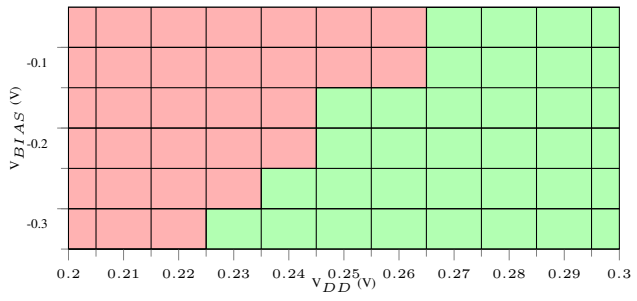


Figure 18: 5 σ shmoo plot of data retention for V_{BIAS} vs V_{DD}

	Idle Leakage				Active Speed		Active Energy	
I/O Logic Library	**Leakage SRAM Cells** [μW]	**Leakage I/O** [μW]	**Total leakage** [μW]	**Leakage** [pW/bit]	**Delay Read** [ns]	**Max. Rd. Freq** [MHz]	**Read Energy** [fJ/OP]	**Read Energy per bit** [fJ/OP/bit]
LVT30	1.07	70.02	71.09	542.37	66.79	14.97	4813.20	75.21
RVT30	1.07	6.57	7.64	58.29	109.75	9.11	908.56	14.19
RVT48	1.07	0.24	1.31	9.99	535.00	1.87	837.52	13.08

Table 2: Systems with different I/O implementations.

3.3. Power and energy consumption on a system level

Table 2 shows the idle leakage power for the 128kb SRAM and the maximum frequency & energy consumed for read operations. Results for the entire SRAM structure with I/O peripheral circuitry implemented with the LVT30 library, the RVT30 library and the RVT48 library are given.

4. Discussion & Conclusion

Ref.	**Tech. Node**	**T.-count** [T/bit]	**Cell Area** [$\mu m^2/bit$]	**BL Height** [Cells/BL]	**Min. V_{DD}** [mV]	**Max. R-Freq.** [Hz]	**Lk. Pwr.** [pW/b]	**Rd. Energy** [fJ/OP/b]
***@ LVT30**	28nm FD-SOI	6	0.547	128	230	14.97M (0.35V)	542.37 (0.35V)	75.2 (0.35V)
***@ RVT30**	28nm FD-SOI	6	0.547	128	230	9.11M (0.35V)	58.29 (0.35V)	14.2 (0.35V)
***@ RVT48**	28nm FD-SOI	6	0.547	128	230	1.87M (0.35V)	9.99 (0.35V)	13.1 (0.35V)
[19]	28nm FD-SOI	10	0.384	64	350	13M (0.35V)	N/A	71.8 (0.35V)
[20]	28nm FD-SOI	6	0.232	32	360	9M (0.36V)	N/A	52.5 (0.45V)
[21]	28nm FD-SOI	7	0.261	64	200	90M (0.30V)	0.9 (0.30V)	8.4 (0.30V)

Table 3: Comparison table to state of the art SRAMs. *= simulated results. @ = lack of access to SRAM transistors.

In this paper we have proposed a 6T SRAM cell for low voltage applications. The construction of the cell is reasoned for in the methodology section, it consumes 6 transistors and is built up from a loadless 4T core and a 2T read buffer. Fig. 16 reveals that the loadless structure dominates the read and hold margin of the conventional 6T cell for low voltages. For higher voltages the opposite is true unless a positive bias voltage is applied equal to that of the supply voltage. This is to correct for the change in the leakage ratios of the RVT pull up devices and the LVT access devices. Fig. 18 confirms the effect of reducing the SNW bias voltage on the hold margin. The cell retains its data at supply voltages as low as 230mV over 5 σ variation with process & mismatch variation taken into account.

It can furthermore be observed from Fig. 17 that the proposed cell inherits close to perfect write margins, this is due to its loadless nature. Since the loading of the loadless SRAM cell occurs as off-state leakage through the LVT-NFET pass gates, once the pass gates are opened, the content of the cell is forced to take the differential value of the differential bitlines without any load working against the write operation. The loadless nature of the loadless core acts as a write-assist, and allows very small differential voltages to be used for write operations, potentially saving power and write-delay since a full swing bit-line scheme may be relaxed.

3 different SRAM implementations were made using 3 classes of logic libraries varying in terms of speed and leakage. From table 2, one can see that the choice of logic libraries for the I/O peripherals greatly impacts all major SRAM performance metrics. The library that yields the best energy per operation per bit has a low operational speed. Choosing an I/O library is therefore highly application specific. Slow applications that infrequently performs memory operations, may benefit from the RVT48 library while fast applications that writes and reads in a high frequent manner may be forced to choose the implementation with the LVT peripheral circuitry.

A comparison to state of the art low voltage SRAMs is provided in table 3. Finding the best SRAM out of a few is ambiguous, especially when one compares conservative simulated results to results obtained through physical measurements. Even though such a comparison is not considered fair, it does provide enough information to decide whether or not one should pursue the idea further.

Slight variations in system implementations may result in great differences in well known metrics. For this implementation conventional RVT and LVT transistors are used, in the 28nm FD-SOI technology a set of more restricted dedicated SRAM transistors are also available that have relaxed layout rules and thicker oxide. The authors of [21] and the authors of [20] collaborate with ST Microelectronics while the author of [19] worked for STM at the time of writing and it is unclear from their papers whether they have access to the dedicated SRAM transistors or not. This may be a reason for their low reported cell-area. The purpose of bringing this up here is that greater leakage and area improvements is still possible for the proposed memory cell by porting the design to the dedicated SRAM transistors. Another reason for the difference in area consumption is the poly strip configuration, in this paper a flat layout scheme was chosen where only a single horizontal row of 6 poly strips makes up the memory cell, this was to minimize the bitline capacitance in order to reduce the read out time which is limited by the read buffer. It is possible to create a two-row configuration with 3 polystrips in each so that the following pins are shared with the neighbour cells above and below: top: BL_W, $\overline{BL_W}$, BL_R bot: V_{DD}, V_{DD}, FL. Such a physical configuration should have less area at the cost of a greater read out delay. As we saw in section 2, the area of the proposed cell is 20.7% smaller than the implemented 8T reference cell.

The 2^{17}-bit array of loadless 6T SRAM Cells has per-

formance and energy metrics according to the rightmost columns of Table 3. Using the LVT30 library for the peripheral logic results in a maximum frequency of 14.97Mhz at 0.35V and a read energy of 75.20 fJ/OP/bit, however the LVT30 implementation has a very high retention leakage and the peripheral logic should be power gated. Compared to the LVT30 design, the RVT48 memory system consumes roughly 1 54th of the retention power per bit with a speed reduction of only one 8th which results in the highest energy efficiency of the designs proposed here. If a charge pump scheme such as the one in [21] was applied to the read word-line of the SRAM proposed here, a great reduction of the read out delay and thus a higher energy efficiency on read operations would be expected at the cost of higher complexity.

The true yield of the SRAM remains to be found in a physical prototype of the system, but according to statistical simulations with 5 σ accuracy, the proposed loadless 6T cell performs on par with state of the art low voltage SRAMs using conventional figures of merit.

References

[1] D. Evans, The internet of things, Whitepaper, Cisco Internet Business Solutions Group (IBSG) 1 (2011) 1–12.

[2] L. Atzori, A. Iera, G. Morabito, The internet of things: A survey, Computer networks 54 (15) (2010) 2787–2805.

[3] H. Esmaeilzadeh, E. Blem, R. S. Amant, K. Sankaralingam, D. Burger, Dark silicon and the end of multicore scaling, in: ISCA 2011, IEEE, 2011, pp. 365–376.

[4] B. H. Calhoun, A. Wang, A. Chandrakasan, Modeling and sizing for minimum energy operation in subthreshold circuits, IEEE JSSC 40 (9) (2005) 1778–1786.

[5] G. Chen, D. Sylvester, D. Blaauw, T. Mudge, Yield-driven near-threshold sram design, IEEE transactions on very large scale integration (VLSI) systems 18 (11) (2010) 1590–1598.

[6] N. Verma, A. P. Chandrakasan, A 256 kb 65 nm 8t subthreshold sram employing sense-amplifier redundancy, IEEE JSSC 43 (1) (2008) 141–149.

[7] M.-H. Tu, J.-Y. Lin, M.-C. Tsai, C.-Y. Lu, Y.-J. Lin, M.-H. Wang, H.-S. Huang, K.-D. Lee, W.-C. Shih, S.-J. Jou, et al., A single-ended disturb-free 9t subthreshold sram with cross-point data-aware write word-line structure, negative bit-line, and adaptive read operation timing tracing, IEEE JSSC 47 (6) (2012) 1469–1482.

[8] J. P. Kulkarni, K. Kim, K. Roy, A 160 mv robust schmitt trigger based subthreshold sram, IEEE JSSC 42 (10) (2007) 2303–2313.

[9] J. Chen, L. T. Clark, T.-H. Chen, An ultra-low-power memory with a subthreshold power supply voltage, IEEE JSSC 41 (10) (2006) 2344–2353.

[10] Y.-W. Chiu, Y.-H. Hu, J.-K. Zhao, S.-J. Jou, C.-T. Chuang, A subthreshold sram with embedded data-aware write-assist and adaptive data-aware keeper, in: ISCAS 2016, IEEE, 2016, pp. 1014–1017.

[11] S. Hanson, M. Seok, Y.-S. Lin, Z. Foo, D. Kim, Y. Lee, N. Liu, D. Sylvester, D. Blaauw, A low-voltage processor for sensing applications with picowatt standby mode, IEEE JSSC 44 (4) (2009) 1145–1155.

[12] K. Noda, K. Matsui, K. Imai, K. Inoue, K. Tokashiki, H. Kawamoto, K. Yoshida, K. Takeda, N. Nakamura, T. Kimura, et al., A 1.9 μm^2 loadless cmos four-transistor sram cell in a 0.18 μm logic technology, in: 1998. IEDM 1998, IEEE, 1998, pp. 643–646.

[13] X. Deng, T. W. Houston, Loadless 4t sram cell with pmos drivers, uS Patent 6,731,533 (May 4 2004).

[14] L. Chang, D. M. Fried, J. Hergenrother, J. W. Sleight, R. H. Dennard, R. K. Montoye, L. Sekaric, S. J. McNab, A. W. Topol, C. D. Adams, et al., Stable sram cell design for the 32 nm node and beyond, in: VLSI Technology, 2005. Digest of Technical Papers. 2005 Symposium on, IEEE, 2005, pp. 128–129.

[15] J. Kwong, Y. K. Ramadass, N. Verma, A. P. Chandrakasan, A 65 nm sub-microcontroller with integrated sram and switched capacitor dc-dc converter, IEEE JSSC 44 (1) (2009) 115–126.

[16] E. Seevinck, F. J. List, J. Lohstroh, Static-noise margin analysis of mos sram cells, IEEE JSSC 22 (5) (1987) 748–754.

[17] R. M. Swanson, J. D. Meindl, Ion-implanted complementary mos transistors in low-voltage circuits, IEEE Journal of Solid-State Circuits 7 (2) (1972) 146–153.

[18] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, M. Bohr, A 3-ghz 70-mb sram in 65-nm cmos technology with integrated column-based dynamic power supply, IEEE JSSC 41 (1) (2006) 146–151.

[19] F. Abouzeid, A. Bienfait, K. C. Akyel, A. Feki, S. Clerc, L. Ciampolini, F. Giner, R. Wilson, P. Roche, Scalable 0.35 v to 1.2 v sram bitcell design from 65 nm cmos to 28 nm fdsoi, IEEE JSSC 49 (7) (2014) 1499–1505.

[20] A. Biswas, A. P. Chandrakasan, A 0.36 v 128kb 6t sram with energy-efficient dynamic body-biasing and output data prediction in 28nm fdsoi, in: ESSCIRC Conference 2016: 42nd, IEEE, 2016, pp. 433–436.

[21] B. Mohammadi, O. Andersson, J. Nguyen, L. Ciampolini, A. Cathelin, J. N. Rodrigues, A 128 kb single-bitline 8.4 fj/bit 90mhz at 0.3 v 7t sense-amplifierless sram in 28 nm fd-soi, in: ESSCIRC Conference 2016: 42nd, IEEE, 2016, pp. 429–432.

Even Låte graduated from the Norwegian University of Science and Technology (NTNU) with a M.Sc. degree in Electrical Engineering in 2014. His work on "Transaction Level Modeling of a PCI Express Root Complex" was awarded with the best master's thesis prize within the field of microelectronics at NTNU. Since the fall of 2014, Even has been employed as a PhD candidate at the Circuits and Systems Group. His research focus is mitigation of the dark silicon effect by designing ultra low power CMOS memory for use in the "Single-ISA Heterogeneous Many-Core Computer" project at NTNU.

Trond Ytterdal received his M.Sc. and Ph.D. degrees in electrical engineering from the Norwegian Institute of Technology in 1990 and 1995, respectively. He is a Professor at the Department of Electronics and Telecommunications, Norwegian University of Science and Technology. Current main research interests include design of analog integrated circuits and modeling of novel nanoscale transistors. He has authored and co-authored more than 180 scientific papers in international journals and conference proceedings. He is a co-author of three books. Prof. Ytterdal is a member of The Norwegian Academy of Technological Sciences and a Senior Member of IEEE.

Snorre Aunet Snorre Aunet received the Cand. Scient. degree in informatics, from the University of Oslo (UiO) in 1993, and the Dr.Ing. degree in physical electronics from the Norwegian University of Science and Technology (NTNU), in 2002. He worked with ASIC design at Nordic VLSI from 1994 to 1997. Later, he worked at NTNU and UiO, and is currently a professor at NTNU. He has been a visiting researcher at the University of Paderborn, Germany, for shorter periods, between 2004 and 2012. Between July 2015 and July 2016 he had a sabbatical at UCSD. Research interests include ultra low-power mixed-signal circuits.