A Low-Power Embedded SRAM for Wireless Applications

Stefan Cosemans, *Student Member, IEEE*, Wim Dehaene, *Senior Member, IEEE*, and Francky Catthoor, *Fellow, IEEE*

***Abstract*—This paper introduces a novel ultra-low-power SRAM. A large power reduction is obtained by the use of four new techniques that allow for a wider and better trade-off between area, delay and active and passive energy consumption for low-power embedded SRAMs. The design targets wireless applications that require a moderate performance at an ultra-low-power consumption. The implemented design techniques consist of a more efficient memory databus, the exploitation of the dynamic read stability of SRAM cells, a new low-swing write technique and a distributed decoder. An 8-KB 5T SRAM was fabricated in a 0.18-μm technology. The measurement results confirm the feasibility and the usefulness of the proposed techniques. A reduction of active power consumption with a factor of 2 is reported as compared to the current state of the art. The results are generalized towards a 32-KB SRAM.**

***Index Terms*—Embedded memory, low power, SRAM.**

I. Introduction

EMBEDDED memories play a crucial role in contemporary electronic systems. They are used in different sizes ranging from a few kilobytes for local scratchpads to a few megabytes for on-chip caches. This paper focuses on embedded SRAMs smaller than 1 Mb for use in the lowest levels of the memory hierarchy. Memories at this level of the hierarchy are used very intensively, so their energy consumption has a significant impact on the energy consumption of the entire system.

The demand for more functionality in mobile applications is continuously increasing. New applications such as Digital Audio Broadcast receivers [1] and portable video applications [2] require not only a large amount of calculations, but they also have to handle huge amounts of data. For these data dominated applications, memory performance is crucial.

At the same time, technology scaling beyond 90 nm introduces a number of new challenges, most notably an increase in passive power consumption and increasing within-die variations, which will make it much more complicated to fully benefit from the advantages scaling could theoretically provide [3].

In this paper, a novel SRAM design is introduced. The design is based on four new techniques that allow for a wider and better trade-off between area, delay and active and passive energy consumption. Although the memory is designed in a 0.18-μm technology, the potential impact of the new techniques on passive power consumption and on the impact of within-die variations was considered during the design. A performance of 250 MHz was targeted.

The paper is structured as follows. Section II introduces the new techniques. First, it is shown how short bitlines with a buffer allow for a more efficient memory databus, reducing both energy consumption and delay. Second, it is shown that the static noise margin for read operations (SNM_R) used in SRAM design is too pessimistic when short, buffered bitlines are used. Cells with a very limited SNM_R are shown to operate correctly over a wide range of process and intra-die variations. This observation allows the use of a five-transistor (5T) SRAM cell without introducing more complicated matrix organizations or peripheral circuits. Third, a new, area-efficient technique to reduce the write cycle energy consumption is proposed based on a shared local write receiver. Fourth, a distributed decoder is presented that makes use of low-swing transmission of address bits.

Section III describes the fabricated design. Section IV discusses circuit level details of some of the communication links used. Section V presents the measurement results. Section VI addresses some of the issues associated with implementing the proposed techniques in more advanced technologies. Conclusions are drawn in Section VII. All examples in this paper assume a 0.18-μm technology and a supply voltage of 1.6 V, unless specified differently.

II. New Techniques

A. Short Buffered Bitlines and the Memory Databus

In traditional low-power memory designs, the cell read current $I_{read,cell}$ must create a large enough voltage difference on the bitlines. Since the bitline capacitance is very large, this step makes up a large part of the memory access delay. Therefore, $I_{read,cell}$ must be made as large as possible, which results in high cell leakage currents because a large $I_{read,cell}$ requires large transistor widths, low threshold voltages or a high supply voltage for the cell. The problem of optimizing the read speed could be decoupled to some extent from the problem of reducing cell leakage by introducing dynamic voltage control schemes [4], [5]. However, this does add some overhead. This technique is not further discussed in this paper.

Not only the nominal value of $I_{read,cell}$ is important though. Because the cells need to be very small, the intra-die variation of $I_{read,cell}$ will be large [6]. This requires a large safety margin on the memory delay. Since all cells will remain activated until the slowest cell is ready, this also causes an important increase in energy consumption.

These problems are remedied when the amount of charge that the cell must draw from the bitline is reduced. Therefore, the

Manuscript received November 17, 2006; revised February 6, 2007. This work was supported by IMEC under the Technology Aware Design (TAD) project.

S. Cosemans and W. Dehaene are with the ESAT-MICAS Laboratory, Katholieke Universiteit Leuven, B-3001 Leuven, Belgium (e-mail: stefan.cosemans@esat.kuleuven.be).

F. Catthoor is with IMEC-DESICS, B-3001 Leuven, Belgium. He is also with the Katholieke Universiteit Leuven, B-3001 Leuven, Belgium.

Digital Object Identifier 10.1109/JSSC.2007.896693

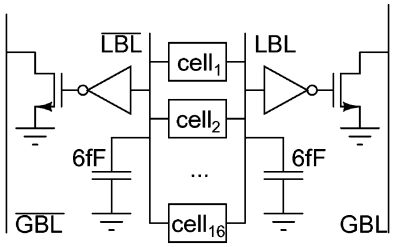


Fig. 1. A short, buffered local bitline.

bitline is divided into shorter local bitlines that connect to a global bitline (GBL) through a multiplexer. Often, this multiplexer consists of a simple pass transistor [7]. This adds relatively little area overhead. However, the cell still needs to sink all current from the global bitline. When an actual read buffer is inserted between the local and global bitline [8], [9], the cell requirements are relaxed even further. Fig. 1 depicts an implementation of the buffered bitline technique. The local bitline uses a large voltage swing, which allows the use of a simple inverter as a sensing element for the buffer. Because the capacitance on the local bitline is small, this larger voltage swing has only a limited impact on the energy consumption of the entire memory, in the range of one to three percent. The GBL driver in the buffer consists of a single scaled-up nMOS transistor, so $I_{read,buffer}$ can be much larger than $I_{read,cell}$ and suffers less from intra-die variation. If an additional low voltage power supply, for instance at 0.4 V, is available to the system, this supply can be used to precharge the GBL. This results in an important reduction in energy consumption as compared to the non-buffered situation, where a relatively large bitline precharge voltage (e.g., 1 V) is needed to allow for an acceptably large I_{read} without endangering the cell stability. This buffered bitline structure is the key enabler for most of the techniques introduced in this paper.

Impact on the Memory Databus: Fig. 2(a) shows the approach normally used in low-power memories to transfer data from the cell to the memory outputs. The capacitances in the figure are wire capacitances only for a 8-KB memory organized as 256 rows by 16 words of 16 bit. At the column level, sense amplifiers amplify the voltage difference on the bitlines to a full level signal. In traditional designs, this amplification at the column level is required to limit the impact of $I_{read,cell}$ on the memory speed. After this amplification, the data still needs to be transmitted to the memory output.

In the buffered bitline approach the buffer can easily deliver more current. This allows the global bitlines to be extended to the memory output. Fig. 2(b) illustrates this approach. Table I compares the two approaches for the 8-KB memory mentioned. The new solution is slightly faster and consumes 34% less energy if all charge has to be derived from a single 1.6-V power supply. If an additional power supply at the precharge voltage of the GBL is added to the system, the reduction in energy consumption is even larger. An additional advantage of the new approach is that it only requires one set of sense amplifiers for the entire memory instead of one set for each column. This can reduce the area overhead, or it could allow the use of more advanced sense amplifiers.

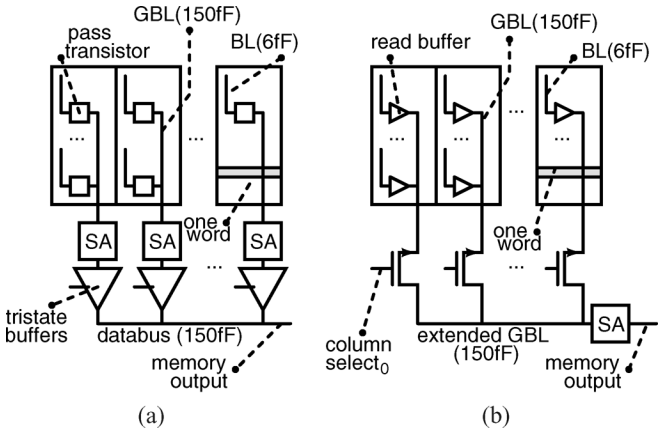


Fig. 2. Solutions to transfer data from the cell to the memory output. (a) Traditional solution. (b) Proposed solution.

TABLE I
COMPARISON OF SOLUTIONS FOR THE MEMORY DATABUS

	Delay [ps]	Energy [fJ/bit]	
		1.6V supply	one supply added
Traditional solution	860	500	480[a]
Proposed solution	740	330	180[b]

[a] BL and GBL are charged from a 1-V supply.

[b] GBL and extended GBL are charged from a 0.4-V supply.

B. Read Operation Using Dynamic Stability

A memory cell must be designed in such a way that the stored value is not corrupted during a read operation. for SRAMs, the SNM_R [10] is used during the design to ensure this. The SNM_R is based on DC simulations in which the wordline is fixed to its high voltage and the bitlines are fixed to the precharge voltage. Recently, a number of alternative methods to analyze cell stability have been proposed. In the N-curve approach [11], [12], both a current margin and a voltage margin are used, while specific problems related to PD/SOI technology are taken into account in [13]. These new methods are still based on DC calculations.

With further scaling, the device mismatch increases and the static margins decrease. Based on this observation, it is argued that it will be impossible to create stable 6T SRAM cells in future very deep-submicron technologies [14]. Therefore, one will be forced to add additional transistors to the cell, which increases cell cost.

Dynamic Read Stability: While the SNM_R is an acceptable read stability indicator for cells in a memory architecture with large bitline capacitances, it is much too pessimistic in memory architectures featuring short buffered bitlines. This is because during the actual read access, the bitline will already be discharged to a safe voltage before the cell can change its state. Fig. 3 shows a 6T cell without SNM_R that does not suffer from data corruption during read operations. Fig. 3(c) defines an ad hoc *read margin* as the minimal value of $V(n2) - V(n1)$. The only relevant property of this *read margin* is that as long as this *read margin* is positive (as long as the voltages on the cell nodes do not cross) the cell data is not destroyed. Fig. 3(d) shows the distribution of this *read margin*. Since the cell data is

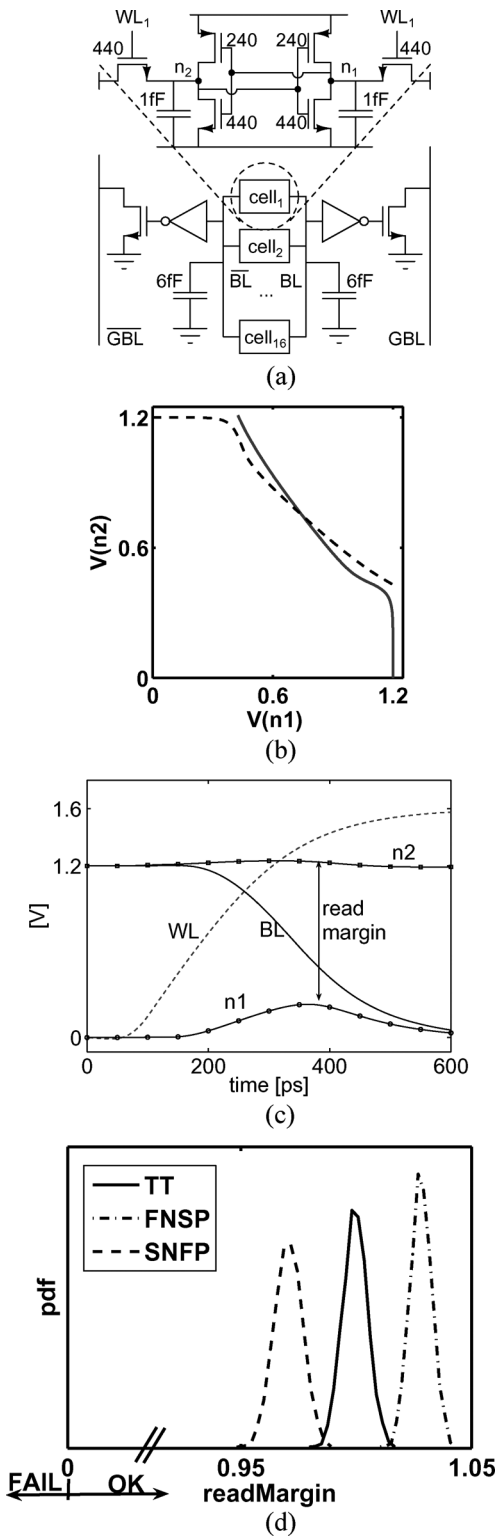


Fig. 3. With short, buffered bitlines, a 6T cell without SNM_R can still operate correctly under process and intra-die variability. All plots use $V_{DD,WL} = 1.6$ V, $V_{DD,Cell} = V_{BL,precharge} = 1.2$ V. (a) The simulation setup. (b) The cell does not have any SNM_R. (c) The read operation. No destructive read because of dynamic effects. (d) Monte Carlo simulations are performed to account for intra-die variations. This results in a distribution for the read margin. The simulation is repeated for each process corner.

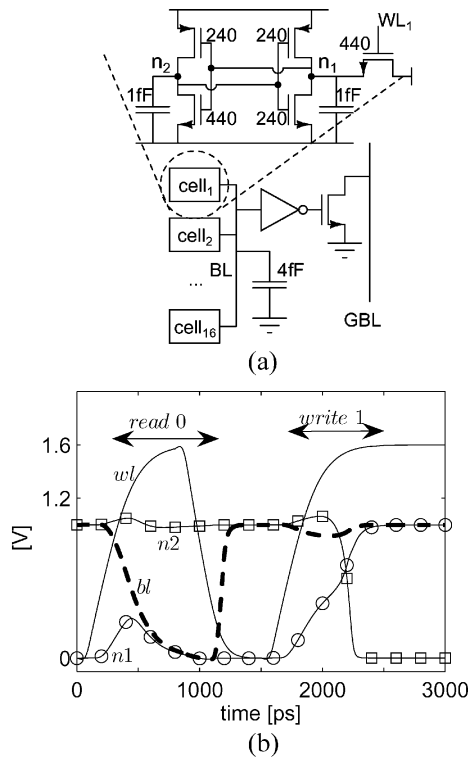


Fig. 4. Dynamically stable 5T cell ($V_{DD,Cell} = V_{BL,write} = V_{BL,precharge} = 1$ V, $V_{WL} = 1.6$ V). (a) The simulation setup. (b) The read and write operation for the nominal cell.

only overwritten when *read margin* becomes negative, the distributions show that the cell operates correctly under process and intra-die variations. Of course, the cell has a normal SNM when the wordline is disabled to ensure static data retention.

A 5T SRAM Cell Using Dynamic Read Stability: The 5T SRAM cell is no new concept. The advantages compared to a regular 6T SRAM cell are clear yet not huge: a slightly smaller area, a reduction in the number of leakage paths, a reduction in the number of bitlines and a potentially large reduction in active read energy consumption [15]. To operate correctly for both read and write, the 5T SRAM cell requires either a more complicated matrix organization or a more complex periphery. Possible solutions include the use of different cell, wordline or bitline voltages for the read operation as compared to the write operation or the introduction of additional control wires. This added overhead and the relatively limited benefits has prevented the widespread use of 5T cells.

When a buffered bitline is used, the 5T cell can be used without these complications by exploiting the dynamic stability of the cell during the read operation. Fig. 4 shows a 5T cell and the simulated voltage waveforms during the two critical operations for the cell: reading a 0 value and writing of 1 value. Fig. 5 indicates that the cell operates correctly under process and intra-die variations.

Figs. 6 and 7 show two other distributions, which can be considered as alternative measures of stability which are much easier to measure accurately on chip. Fig. 6 shows the distribution of the highest cell supply voltage at which an individual cell suffers from destructive reads. The distribution has a mean of 767 mV and a standard deviation of 21 mV. The distribution was obtained by performing 1000 Monte Carlo simulations of a cell using the typical process. According to this simulation, a

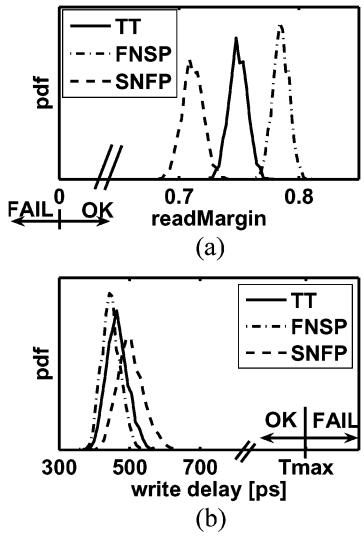


Fig. 5. Distributions for a 5T cell using dynamic read stability ($V_{DD,Cell} = V_{BL,write} = V_{BL,precharge} = 1$ V, $V_{WL} = 1.6$ V). (a) Read margin. (b) Write delay.

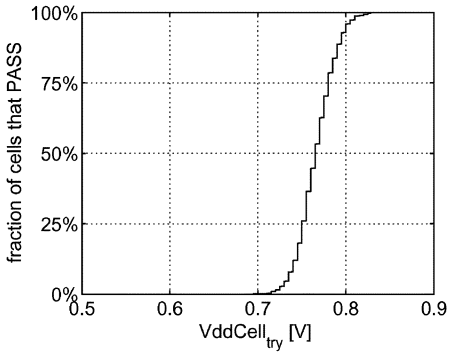


Fig. 6. Fraction of cells that do not suffer from destructive read when operated at a reduced cell supply voltage ($V_{BL,precharge} = 1.2$ V, $V_{WL} = 1.6$ V). (simulated for typical process).

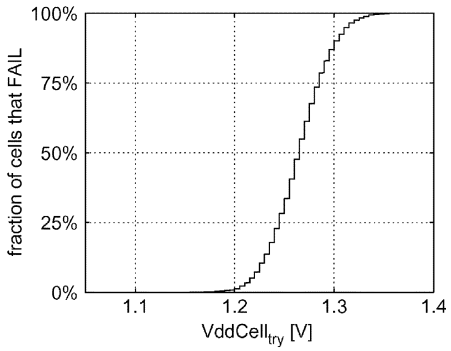


Fig. 7. Fraction of cells that are not written correctly at a given cell supply voltage ($V_{BL,write} = 1.2$ V, $V_{WL} = 1.6$ V). (simulated for typical process).

cell supply above 893 mV (6-sigma) allows virtually all cells on a typical die to be read correctly. Fig. 7 shows the distribution of the lowest cell supply voltage at which an individual cell can no longer be written. The distribution has a mean of 1264 mV and a standard deviation of 29 mV. The distribution was obtained by performing 1000 Monte Carlo simulations of a cell using the typical process. According to these simulations, a cell supply below 1089 mV (6-sigma) allows virtually all cells on a typical die to be overwritten correctly.

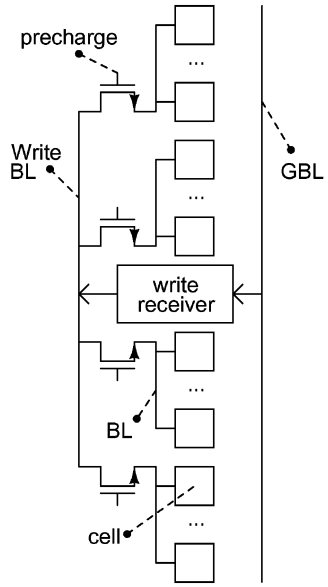


Fig. 8. Shared low-swing write receivers.

C. Low-Swing Write Using Shared Low-Swing Receivers

Several solutions have been proposed to reduce the write energy in memories. In [16], a hierarchical bitline is used. The local bitlines connect to the global bitline through pass transistors. A low-swing receiver is added to each local bitline to amplify the low-swing write data from the global bitline onto the local bitline. The main advantage of this approach is that there are no additional requirements imposed on the cells. The main disadvantage is the large area overhead. This is especially problematic when very short local bitlines are used, as is desirable for the read operation.

A new alternative is shown in Fig. 8. Essentially, a low-swing write receiver is shared between multiple local bitlines. This reduces the area overhead significantly, while the energy for a write operation increases only slightly. To share the receiver, an additional write bitline must be added. This write bitline only has one connection per local bitline, so it is possible to route this wire in one of the higher metal layers, which is not feasible for the local bitlines since they have to connect to each and every cell. Since the write bitlines are always high except when they are used in a write operation, the pass transistor connecting the local bitline with the write bitline can also be used to precharge the local bitline. Furthermore, the write bitlines can be used as shield wires for the global bitlines. This further reduces the cost of this technique.

D. Distributed Decoder

The use of low-swing techniques for the data transfers makes that the decoder is responsible for a large share of the energy consumption. A traditional low-power SRAM design uses divided wordlines and an X-Y decoder. At the crossing point of the selected global wordline (from the X-decoder) and the selected column-enable line (from the Y-decoder), a single local

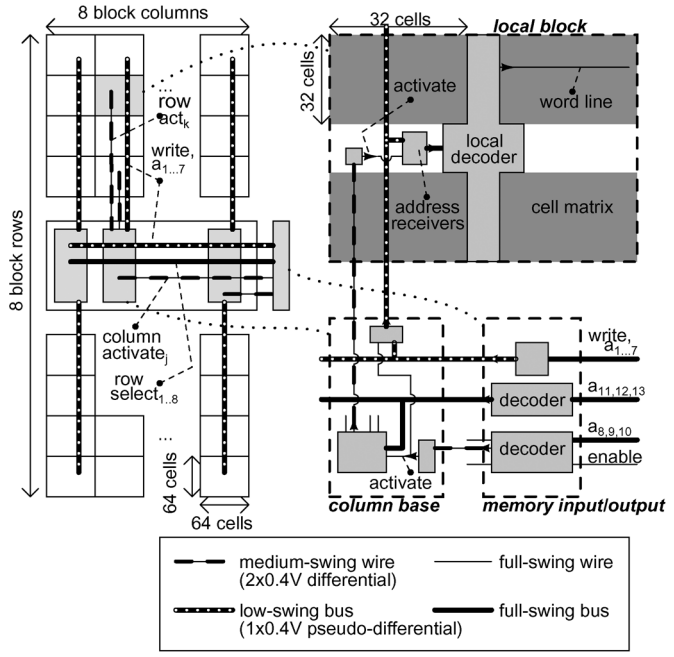


Fig. 9. A distributed decoder.

wordline is activated by the last decoder stage. The global wordlines have to be routed on a single-cell pitch, which results in a large capacitance per unit wire length.

For larger memories, it might be advantageous to use low-swing techniques in the decoder too. In [17], a half-swing pulse-mode gate family is proposed which allows the use of pulsed half-swing signals in the decoder, resulting in significant power savings without affecting performance. However, since the technique relies on a large number of half-swing pulsed gates with a significantly lower noise margin than normal gates, it might prove difficult to scale this solution beyond the 130-nm node.

Adding more complicated low-swing receivers just in front of the last decoder stage is not an economical option since this would introduce too much area overhead and passive power consumption. It also does not solve the cell-pitched routing of metal layers and it is further simplified when cells with a single bitline are the global wordlines.

A better option is presented in Fig. 9. In this example, the 32-KB matrix is divided in 8 by 8 blocks, each containing 128 words of 32 bits. In the input block of the memory, the address bits used to select the block are decoded. In this design, it is assumed that the highest address bits a_{11}, a_{12} and a_{13} have a low activity, as is the case for applications with good locality of reference. Therefore, they are statically decoded into the row select signal which is transmitted in full swing to all column bases. Because of the low activity, this has only a very limited impact on the total energy consumption. If there are no address bits for which low activity can be assumed, the energy consumption in this stage would still be less than 5% of the energy consumption of the entire access. In parallel, address bits a_8, a_9 and a_{10} are decoded into the column select signal. This signal is combined with the activate signal and is then sent to the base of the appropriate column using a medium swing transmitter and receiver. Also in parallel, the remaining address bits a_1 to a_7 and the write enable signal *write* are put on the horizontal low-swing address bus, which is at this time isolated from all vertical address busses by means of pass transistors.

Next, the column activate signal is received at the base of the column. The pass transistors between the horizontal address bus and the appropriate vertical address bus are enabled. Based on the row select that is already available, a medium swing block select signal is sent to the appropriate block.

By the time the block select signal is restored to a full level signal by the medium swing receiver in the local block, the data on the vertical address bus is available. The activate signal enables the low-swing receivers, which are implemented as sense amplifiers. The address bits a_1 to a_7 are then decoded within the local block. All needed control signals, such as the precharge signals for the local bitlines, are generated locally.

The area overhead of this scheme is limited. The local decoder is repeated in each local block and eight sense amplifiers are added to each block. In the given example, this results in an area overhead of only 3.5% as compared to a cell matrix in which only the most basic subdivided wordline scheme is implemented.

III. The Design

A low-power 8-KB 5T SRAM with a word length of 16 bits has been designed and fabricated in a 0.18-μm technology. The memory is organized as 64 blocks of 32 cells by 32 cells and implements a distributed decoder as in Fig. 9. The local bitline is implemented as in Fig. 4, with 16 5T SRAM cells connected to one buffered bitline. A low-swing write receiver is shared between four bitlines according to Fig. 8. The precharge transistor and the write transistor are combined into a single pass transistor. The extended global bitline scheme from Fig. 2(b) is used.

The memory layout and the chip photograph are shown in Fig. 10. The memory measures 850 μm by 810 μm. The area is 0.69 mm^2. A large part of the remainder of the die is used for on-chip signal monitoring circuits, which are not discussed in this paper.

IV. Communication Links

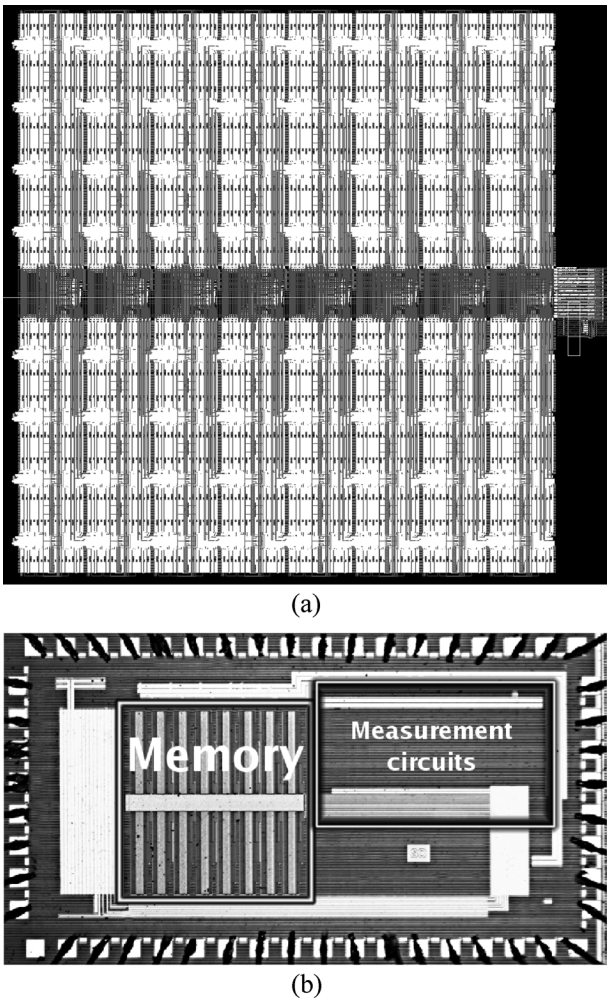
A. Data Transfer From LBL to Read Sense Amplifier

Each read cycle, the 16 bits from a word have to be transmitted from the LBL to the memory outputs as described in Section II-A. This transmission is part of the critical access path, and due to the fact that 16 bits have to be transmitted, the energy consumption of this communication link is crucial.

In this communication link, the read buffers are the transmitters, the GBLs and extended GBLs are tristate busses and the read sense amplifiers are the receivers. This setup requires the read buffers to have a high impedance output in standby. Since one read buffer is required per LBL, the area of the read buffer should be as small as possible. The size of the sense amplifiers is much less critical.

We considered a number of signaling techniques to implement this link, three of which are discussed in this section.

Pseudo-Differential Signaling [Fig. 11(a)]: For pseudo-differential signaling, only one GBL is used per bit. It is precharged to $V_{high} = (V_{offset} + 2 \cdot V_{sense})$. V_{sense} is the required input



(a)

(b)

Fig. 10. The fabricated memory chip. (a) Layout. (b) Chip microphotograph.

voltage difference for the sense amplifier. V_{offset} is the voltage headroom available to the driver [see Fig. 12(b)], which will have an impact on speed. The other input of the sense amplifier is connected to a reference line that is shared between all bits. This reference line is fixed at $V_{ref} = V_{offset} + V_{sense}$.

When a zero has to be transmitted, the GBL is pulled low. When a one has to be transmitted, the GBL is left at its precharged level, thus consuming no energy.

To ensure signal integrity, a shield wire has to be provided between different GBLs. This means two wires have to be routed per cell pitch. The resulting wire capacitance for the GBL will be called C_2 in this discussion. Assuming that a specific power source is provided to precharge the GBLs, the average energy consumption per bit for pseudo-differential signaling can be approximated as

$$E_{pseudo} = P_0 \cdot (V_{offset} + 2 \cdot V_{sense})^2 \cdot C_2.$$

In this formula, P_0 is the probability that a cell stores a 0.

Complementary Signaling [Fig. 11(b)]: For complementary signaling, two signal wires GBL and GBLc are provided for each cell column. They are both precharged to $V_{high} = (V_{offset} + V_{sense})$. V_{sense} is the required input voltage difference for the sense amplifier. GBL and GBLc connect to the inputs of the sense amplifier. During a read access, either GBL or GBLc is discharged. A shield wire is needed between all signal wires, which results in four wires routed per cell pitch.

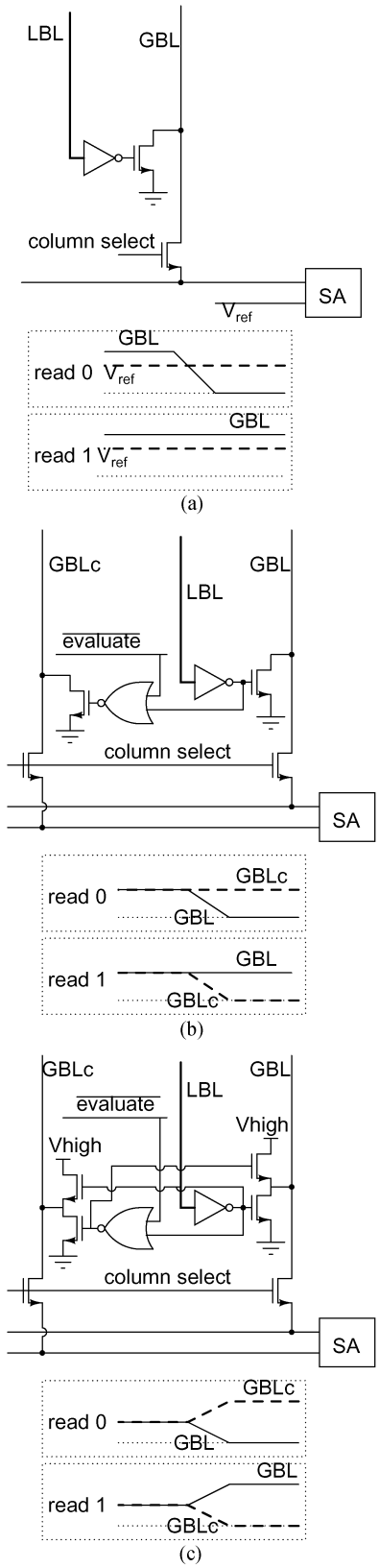


Fig. 11. Transmitters for the different signaling schemes. (a) Pseudo-differential scheme. (b) Complementary scheme. (c) Differential scheme.

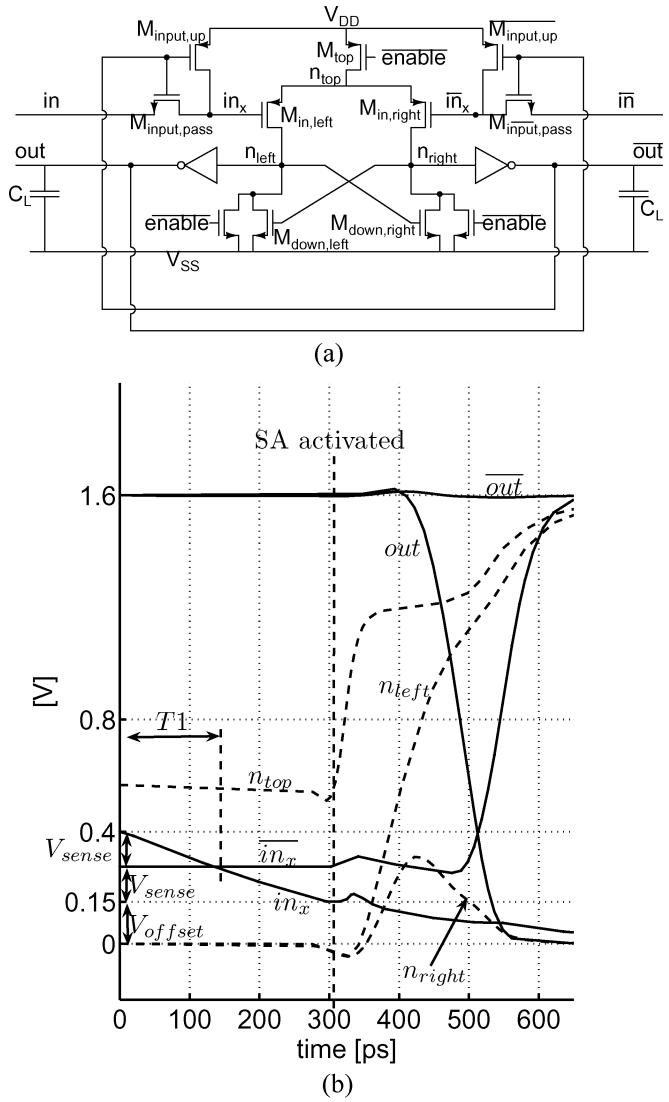


Fig. 12. The read sense amplifier. (a) Circuit schematic. (b) Waveforms during the sense operation.

The resulting GBL wire capacitance will be referred to as C_4. C_4 will be larger than C_2 since the available spacing between the different is smaller. The exact relation between C_2 and C_4 depends on the actual layout. If the wires must be routed on a single metal layer, the capacitance between wires on this layer dominates over capacitances to other metal layers, and since the available spacing between the wires is more than halved when four instead of two wires must be routed, C_4 will be more than $2 \cdot C_2$. Since the lateral capacitance varies nonlinearly with distance, the real capacitance C_4 will most certainly be larger than this. If plenty of metal layers are available to route the wires, the increase in capacitance will be somewhat less pronounced. However, the first layout sketches pointed out that these wires would have to be routed on a single metal layer, so in this analysis, we assume $C_4 = 2 \cdot C_2$, which is known to be too optimistic. The average energy consumption per bit can be approximated as

$$E_{complementary} = 1 \cdot (V_{offset} + 1 \cdot V_{sense})^2 \cdot C_4.$$

When V_{offset} is chosen equal to V_{sense} and P_0 is assumed to be 0.5, the pseudo-differential approach uses 44% less energy than the complementary approach. With the assumptions made, the speed of the two schemes will be comparable. The complementary approach avoids the delay $T1$ on Fig. 12(b), but this advantage is completely lost due to the larger GBL capacitance involved.

Differential Signaling [Fig. 11(c)]: For differential signaling, two signal wires GBL and GBLc are provided for each cell column. They are precharged to $V_{offset} + (V_{sense}/2)$. During a read operation, one line is pulled up to $2 \cdot V_{offset} + V_{sense}$, the other line is discharged to ground. GBL and GBLc connect to the sense amplifier. In principle, shield wires are not needed if bitline folding is applied correctly. In this analysis, the GBL capacitance is assumed to be C_2, which is the most optimistic case possible. If charge recycling is used, the average energy consumption per bit can be approximated as

$$E_{differential} = 1 \cdot (2 \cdot V_{offset} + V_{sense}) \cdot \left(V_{offset} + \frac{V_{sense}}{2} \right) \cdot C_2.$$

With $V_{offset} = V_{swing}$ and $P_0 = 0.5$, this results in exactly the same energy consumption as for the pseudo-differential solution.

Selection of Signaling Technique: If $P_0 \approx 0.5$, the energy and delay performance of pseudo-differential and differential signaling are very similar. Based on this observation, we selected the pseudo-differential signaling scheme based on area considerations, even though it requires an additional reference voltage. Another consideration in favor of this choice is the fact that when P_0 is known to be much smaller than 0.5 (as in [15]), the advantages of pseudo-differential signaling are very pronounced.

For this design, we selected $V_{offset} = 150$ mV and $V_{sense} = 125$ mV. This results in a precharge voltage of 0.4 V and a reference voltage V_{ref} of 0.275 V.

Read Sense Amplifier: We designed the sense amplifier shown in Fig. 12(a). As long as $\overline{enable}$ is high, n_{left} and n_{right} are forced to 0 V, in is connected to in_x through $M_{input,pass}$, $\overline{in} = V_{ref}$ is connected to $\overline{in_x}$ through $M_{\overline{input,pass}}$ and no DC currents are flowing. The waveforms associated with a read operation for this sense amplifier are shown in Fig. 12(b). When $\overline{enable}$ becomes low, the voltage on both n_{left} and n_{right} increases, but the voltage at the side which has the lowest input voltage connected to M_{in} increases much faster. This turns on the M_{down} transistor at the opposite side, which effectively stops the voltage increase on the other node. After the sensing operation is completed, the current through the input transistor which had the highest input voltage on its gate is stopped. This is done by isolating in_x or $\overline{in_x}$ from the input and pulling it up to V_{DD}.

This sense amplifier avoids the stack of three pMOS transistors which occurs in more traditional voltage-based sense amplifiers. This way, the sensing speed is improved, and only the mismatch between $M_{in,left}$ and $M_{in,right}$ has a significant impact on the offset voltage of the sense amplifier. In this sense amplifier, the inputs remain connected to the gates of the input transistors until the sensing operation is completed, so this sense

TABLE II
TRANSISTOR SIZES USED IN SENSE AMPLIFIER

transistor names	Width [nm]	Length [nm]
$M_{in,left}$, $M_{in,right}$	960	180
$M_{down,left}$, $M_{down,right}$	330	180
M_{top}	620	180
$M_{input,up}$, $M_{\overline{input},up}$	240	180
$M_{input,pass}$, $M_{\overline{input},pass}$	240	180

amplifier does not rely on dynamic storage, which would require a very careful layout.

The transistor sizes used are summarized in Table II.

B. Transmission of the Activation Signal

Each access, one local block is activated by the transmission of an activation signal to this block. The activation is part of the critical path of the memory access, so the delay for this transmission is critical. Since only one such signal is transmitted each cycle, the energy consumption for this transmission is slightly less crucial than was the case for transmitting the databits. As explained in Section II-D, block activation is implemented using a differential medium swing signal. Each block has its own differential pair of activation wires, so there is only one driver and one receiver on each wire pair.

Fig. 13(a) shows the circuit schematic for this communication link. Fig. 13(b) shows the associated waveforms. As long as the block is not activated, $activate$ is at $V_{low} = 0$ V and $\overline{activate}$ is at $V_{medium} = 0.4$ V. In this situation, the zero threshold (ZVT) transistor $M_{in,D}$ has a gate-source voltage of $V_{GS} = V_{medium}$ and keeps node $\overline{disable}$ low, which in turn makes sure node $\overline{enable}$ stays high. Transistor $M_{in,E}$ is disabled, since it has a $V_{GS} = -V_{medium}$. When the block has to be activated, $activate$ is pulled up to V_{medium} and $\overline{activate}$ is pulled to V_{low}. This quickly pulls $\overline{enable}$ to V_{low}. This activates $M_{up,p}$, which will slowly pull up $\overline{disable}$. All events that have to be triggered quickly after block enabling are derived from $\overline{enable}$, while all events that have to be triggered quickly after block disabling are derived from $\overline{disable}$. Table III summarizes the transistor sizes used in the design. The minimal gate length allowed for ZVT transistors in this technology is 300 nm.

A similar receiver could also be conceived without the use of ZVT transistors, but this would require a more complicated circuit, including a capacitive coupling from $activate$ and $\overline{activate}$ to the gates of the input transistors.

V. MEASUREMENT RESULTS

A. Performance Measurements

Table IV shows the measurement results for the fabricated 8-KB memory. It operates at 250 MHz and consumes 9.5 pJ per access (2.4 mW at 250 MHz). With an additional power supply at 0.4 V, only 6.8 pJ per access is consumed (1.7 mW at 250 MHz). For the energy consumption, a random data pattern and a row select signal activity of 1/16 is assumed. The table also contains performance estimates for a 32-KB memory with a word length of 32 bits implementing the techniques. As the comparison with the current state of the art in Table V indicates, the proposed design results in a large reduction in energy per access.

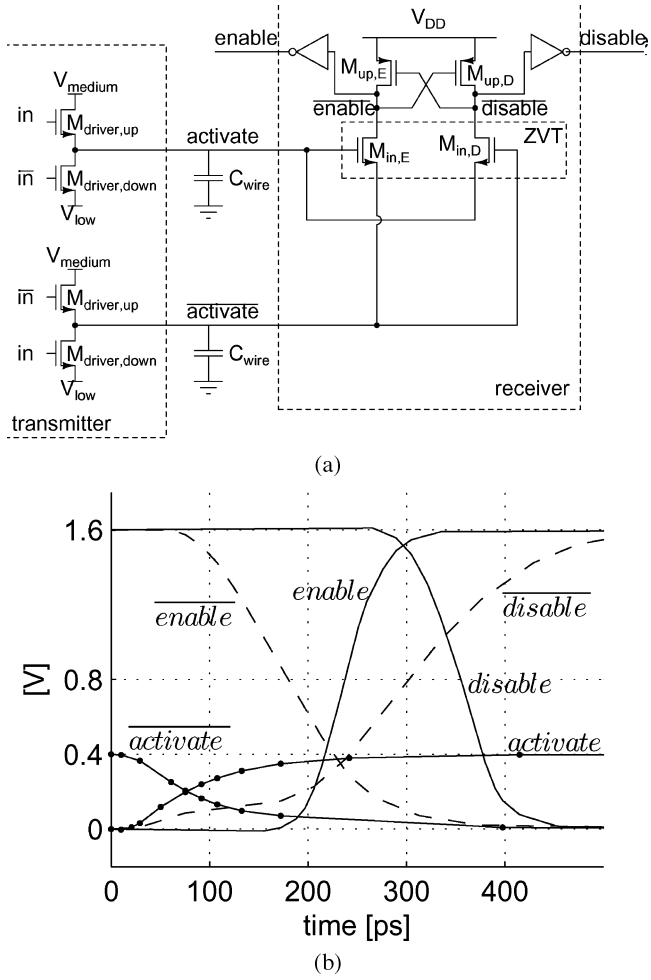


Fig. 13. The medium-swing receiver. (a) Circuit schematic. (b) Waveforms.

TABLE III
TRANSISTOR SIZES USED IN THE TRANSMISSION OF THE ACTIVATION SIGNAL

transistor names	transistor type	Width [nm]	Length [nm]
$M_{in,E}$, $M_{in,D}$	zero threshold	1440	300
$M_{up,E}$, $M_{up,D}$	regular	280	180
$M_{out,up}$ [a]	regular	770	180
$M_{out,down}$ [b]	regular	480	180
$M_{driver,up}$	regular	1440	180
$M_{driver,down}$	regular	1200	180

[a] The pMOS transistor in the output inventor.
[b] The nMOS transistor in the output inventor.

The energy per access is reduced to 54% of the value reported in [16]. At the same time, a significant area reduction is obtained, while speed is only degraded by 24%. With the additional supply voltage at 0.4 V, the energy consumption per operation is only 33% of the energy consumption reported in [16] with the same speed degradation.

B. Cell Measurements: Functional Operation Under Intra-Die Variations

To verify the functional operation of our cells under intra-die variations, we performed four measurements on all cells of a die.

TABLE IV
MEASUREMENT RESULTS FOR THE 8-KB MEMORY AND ESTIMATED PERFORMANCE FOR A 32-KB MEMORY ($Vdd_{Cell} = 1$ V, $V_{LocalBitline} = 1.2$ V, $Vdd_{decoder} = V_{WL} = Vdd_{SA} = 1.6$ V, $V_{ref} = 0.275$ mV, $V_{GlobalBitline} = 0.4$ V)

memory size	8KByte	32KByte
wordlength	16 bits	32 bits
technology	$0.18\mu m$	$0.18\mu m$
	measured	estimated
Access time	$4ns$	$4.3ns$
cell size	$6.06\mu m^2$	
memory core size	$0.69mm^2$	$2.4mm^2$
energy/access (1 supply) [a]	$9.5pJ$	$19pJ$
energy/access (2 supplies) [b]	$6.8pJ$	$12pJ$

[a] The charge for GBL, extended GBL and low-swing address wires is taken from 1.6-V supply.
[b] The charge for GBL, extended GBL and low-swing address wires is taken from a 0.4-V supply.

TABLE V
COMPARISON

memory size	8KByte	8KByte	32KByte	32KByte
wordlength	16 bits	16 bits	32 bits	32 bits
technology	$0.18\mu m$	$90nm$ $[0.18\mu m]$	$0.18\mu m$	$0.25\mu m$ $[0.18\mu m]$
	measured	literature[a]	estimated	literature[b]
Access time	$4ns$	$2ns$ $[4ns]$	$4.3ns$	$4.55ns$ $[3.27ns]$
memory core size	$0.69mm^2$	$0.28mm^2$ $[1.1mm^2]$	$2.4mm^2$	$6.1mm^2$ $[3.2mm^2]$
energy/access (1 supply)	$9.5pJ$	$6.3pJ$ $[50pJ]$	$19pJ$	$95pJ$ $[35.4pJ]$
energy/access (2 supplies) [c]	$6.8pJ$	/	$12pJ$	/

[a] The design described in [8] was implemented in 90 mm. The design operates over a wide V_{dd} operation range. Only one operating point is tabulated. The provided optimistically scaled values should be used with care.
[b] The design described in [16] was implemented in 0.25 μm. Optimistic scaling rules were used to ensure a fair comparison.
[c] GBL, extended GBL and low-swing address wires are charged from a 0.4-V supply.

For the first measurement, a zero is written to all cells using a known good supply voltage for the cell. Next, the cell supply voltage is temporarily reduced to $VddCell_{try}$. Finally, the cell contents is read out using a known good supply voltage for the cell. This is repeated for decreasing values of $VddCell_{try}$. Fig. 14 shows the measurement results for all 65 536 cells on a single die.

The second measurement is identical to the first measurement, but now a high value is written to the cells. Fig. 15 shows the measurement results for all cells on a single die. A relatively large amount of cells have a high value as preferential state, which could have been expected given the asymmetric cell sizing and the high voltage on the local bitline.

These first two measurements indicate that these cells should not be operated below a cell supply voltage of 0.5 V. Since this

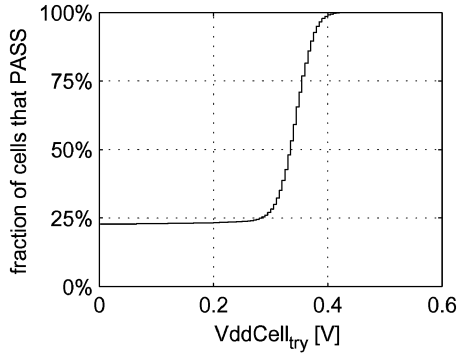


Fig. 14. Fraction of cells that correctly retain their 0-state when the cell supply voltage is temporarily reduced ($V_{BL} = 1.2$ V) (measured).

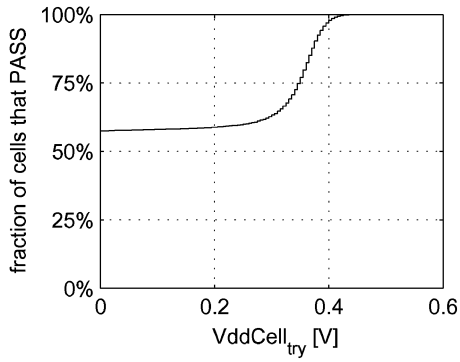


Fig. 15. Fraction of cells that correctly retain their 1-state when the cell supply voltage is temporarily reduced ($V_{BL} = 1.2$ V) (measured).

design never uses such low supply voltages for the cells, this does not pose any problems.

The third measurement identifies for each cell the highest cell supply voltage at which destructive read occurs. For this measurement, a zero is written to all cells using safe supply voltages. The cell supply voltage is reduced to $VddCell_{try}$ and a read access is performed using this reduced cell supply voltage. This is repeated for decreasing values of $VddCell_{try}$. Fig. 16 shows the measurement results for all cells on a single die. The distribution of $VddCell_{try}$ at which cells fail has a mean of 650 mV and a spread of 32 mV. These measurements show that cell supplies larger than 850 mV (6-sigma) do not result in destructive reads.

The fourth measurement identifies for each cell the lowest cell supply voltage at which it can no longer be written correctly. A zero is written to all cells. The cell supply voltage is increased to $VddCell_{try}$, and a write access is performed, attempting to write a high value to the cell. Afterwards, the cell state is examined using safe voltages. This is repeated for increasing values of $VddCell_{try}$. Fig. 17 shows the measurement results for all cells on a single die. The distribution of $VddCell_{try}$ at which cells are no longer writeable has a mean of 1.192 V and a spread of 23 mV. These measurement show that cell supplies smaller than 1.050 V (6-sigma) allow all cells to be written. These results are in sufficient accordance with the simulations.

The results from the last two measurements show that in this design, the application of dynamic read stability on a cell which does not have any SNM_R, created sufficient stability margin. This margin is likely to disappear in more advanced technology

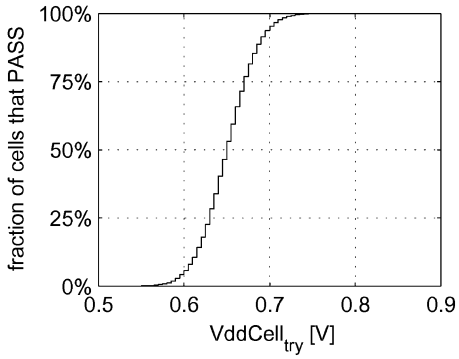


Fig. 16. Fraction of cells that do not suffer from destructive read when operated at a reduced cell supply voltage ($V_{BL,precharge} = 1.2$ V, $V_{WL} = 1.6$ V) (measured).

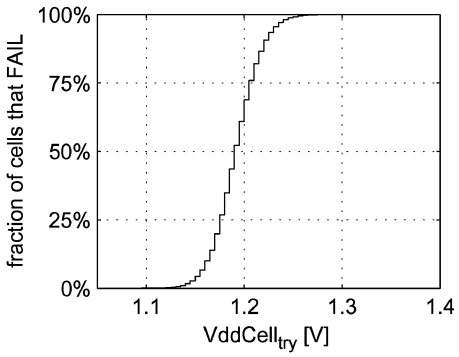


Fig. 17. Fraction of cells that are not written correctly at a given cell supply voltage ($V_{BL,write} = 1.2$ V, $V_{WL} = 1.6$ V) (measured).

nodes which suffer more from variability. However, dynamic read stability applied to somewhat more stable SRAM cells, such as regular 6T cells or single-ended cells with additional transistors to support the write operation [8], will allow to design sufficiently robust SRAMs even under the increased variability of these technologies.

VI. Scalability of the Proposed Techniques

The 5T cell used in this design does not scale well to future technology nodes. However, the proposed techniques do not depend on the use of this 5T cell, and most of these techniques do scale very well to 90 nm and beyond. The use of a short buffered bitline becomes more beneficial since it reduces the impact of the cell current on the memory delay. This reduced impact can be employed in two ways. On the one hand, it allows for the use of high-threshold transistors in the cells at a small delay penalty. This greatly reduces the static power consumption of the memory. On the other hand this reduced impact of the cell current reduces the impact of the large variations in cell current, which is most welcome. The fact that the proposed memory databus requires only one set of read sense amplifiers becomes very interesting as sense amplifier performance becomes more and more dominated by mismatch. With increasing variations and decreasing power supply voltages, the additional margin provided to the cell design by the ability to rely on the dynamic read stability becomes a large advantage, not just for 5T-cells but also for more stable cells such as the normal 6T-cell or modified single-ended cells. The shared low-swing write receivers do not pose any specific challenges for scaling. The advantages of using medium swing signals in the decoder might become smaller with scaling, since the difference between the nominal supply voltage and the voltage swing that is needed for these medium swing signals becomes smaller.

When applying these techniques to smaller technologies which suffer more from leakage, special attention is required to reduce the leakage currents in the read buffers, the write receivers and the local decoder, peripheral circuits that are best implemented using low threshold transistors. During sleep states, the peripheral circuits can be disconnected from the power supply. During the active operation, it is rather easy to provide power gating on an access-by-access basis for the read buffers, since there is quite some time available between block activation and the moment they start operating. Power gating on an access-by-access basis is also possible for the write receivers, but at a slight increase in write delay. Reducing the passive power consumption in the local decoder is more complicated. This power is dominated by the wordline drivers. These wordline drivers are required in all designs that use subdivided wordlines, which means that this is a generic problem for memories aimed at low active energy consumption.

VII. Conclusion

A novel SRAM design based on four new techniques has been presented in this paper. An 8-KB memory aimed at low-power wireless applications has been fabricated. The measurement results show a reduction of the energy consumption per access with a factor of 2, with a limited impact on performance. At the same time, the impact of within-die variations on the performance of this low-power memory was mitigated by the introduction of short buffered bitlines. Measurements performed on this memory prove that when these short buffered bitlines are used, the traditional static noise margins are overly pessimistic. By relying in part on dynamic read stability, the daunting task of designing stable SRAM cells in future technology nodes can be greatly simplified. Still, much more work remains to be done before we will really be able to cope with the increasing uncertainties associated with smaller technology nodes. In our future work, we will focus on techniques to reduce the impact of these variations on system performance and reliability.

Acknowledgment

The authors would like to thank the members of the Technology Aware Design (TAD) project at IMEC for the many fruitful discussions.

References

[1] P. Op de Beeck, C. Ghez, E. Brockmeyer, M. Miranda, F. Catthoor, and G. Deconinck, "Background data organisation for the low-power implementation in real-time of a digital audio broadcast receiver on a SIMD processor," in *Proc. Conf. Design, Automation and Test in Europe (DATE)*, 2003, pp. 1144–1145.

[2] E. Brockmeyer, L. Nachtergaele, F. Catthoor, J. Bormans, and H. De Man, "Low power memory storage and transfer organization for the MPEG-4 full PEL motion estimation on a multimedia processor," *IEEE Trans. Multimedia*, vol. 1, no. 2, pp. 202–216, Jun. 1999.

[3] G. Gielen and W. Dehaene, "Analog and digital circuit design in 65 nm CMOS: End of the road?," in *Proc. Design, Automation and Test in Europe (DATE)*, 2005, pp. 37–42.

[4] F. R. Saliba, H. Kawaguchi, and T. Sakurai, "Experimental verification of row-by-row variable VDD scheme reducing 95% active leakage power of SRAMs," in *Symp. VLSI Circuits Dig. Tech. Papers*, 2005, pp. 162–165.

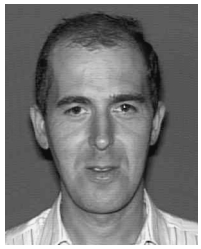
[5] P. Geens and W. Dehaene, "A small granular controlled leakage reductions system for SRAMs," *J. Solid-State Electron.*, vol. 49, pp. 1776–1782, Nov. 2005.
[6] M. J. M. Pelgrom, A. C. J Duinmaijer, and A. P. G. Welbers, "Matching properties of MOS transistors," *IEEE J. Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, Oct. 1989.
[7] A. Karandikar and K. K. Parhi, "Low power SRAM design using hierarchical divided bit-line approach," in *Proc. Int. Conf. Computer Design: VLSI in Computers and Processors*, 1998, pp. 82–88.
[8] K. Takeda, Y. Hagihara, Y. Aimoto, M. Nomura, Y. Nakazawa, T. Ishii, and H. Kobatake, "A read-static-noise-margin-free SRAM cell for low-VDD and high-speed applications," *IEEE J. Solid-State Circuits*, vol. 41, no. 1, pp. 113–121, Jan. 2006.
[9] J. Silberman, N. Aoki, D. Boerstler, J. L. Burns, S. Dhong, A. Essbaum, U. Ghoshal, D. Heidel, P. Hofstee, K. T. Lee, D. Meltzer, H. Ngo, K. Nowka, S. Posluszny, O. Takahashi, I. Vo, and B. Zoric, "A 1.0-GHz single-issue 64-bit PowerPC integer processor," *IEEE J. Solid-State Circuits*, vol. 33, no. 11, pp. 1600–1608, Nov. 1998.
[10] E. Seevinck, F. J. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *IEEE J. Solid-State Circuits*, vol. SC-22, no. 5, pp. 748–754, Oct. 1987.
[11] C. Wann, R. Wong, D. J. Frank, R. Mann, K. Shang-Bin, P. Croce, D. Lea, D. Hoyniak, L. Yoo-Mi, J. Toomey, M. Weybright, and J. Sudijono, "SRAM cell design for stability methodology," in *Proc. IEEE Int. Symp. VLSI-TSA*, Apr. 2005, pp. 21–22.
[12] E. Grossar, M. Stucchi, K. Maex, and W. Dehaene, "Read stability and write-ability analysis of SRAM cells for nanometer technologies," *IEEE J. Solid-State Circuits*, vol. 41, no. 11, pp. 2577–2588, Nov. 2006.
[13] R. V. Joshi, S. Mukhopadhyay, D. W. Plass, Y. H. Chan, C. Ching-Te, and A. Devgan, "Variability analysis for sub-100 nm PD/SOI CMOS SRAM cell," in *Proc. European Solid-State Circuits Conf. (ESSCIRC)*, Sep. 2004, pp. 211–214.
[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, K. W. Guarini, and W. Haensch, "Stable SRAM cell design for the 32 nm node and beyond," in *Symp. VLSI Technology Dig. Tech. Papers*, 2005, pp. 128–129.
[15] Y. Chang and F. Lai, "Dynamic zero-sensitivity scheme for low-power cache memories," *IEEE Micro*, vol. 25, no. 4, pp. 20–32, Jul.–Aug. 2005.
[16] B. D. Yang and L. S. Kim, "A low-power SRAM using hierarchical bit line and local sense amplifiers," *IEEE J. Solid-State Circuits*, vol. 40, no. 6, pp. 1366–1376, Jun. 2005.
[17] K. W. Mai, T. Mori, B. S. Amrutur, R. Ho, B. Wilburn, M. A. Horowitz, I. Fukushi, T. Izawa, and S. Mitarai, "Low-power SRAM design using half-swing pulse-mode techniques," *IEEE J. Solid-State Circuits*, vol. 33, no. 11, pp. 1366–1376, Nov. 1998.

Stefan Cosemans (S'04) received the M.S. degree in electrical engineering from the Katholieke Universiteit Leuven (K.U.Leuven), Heverlee, Belgium, in 2004. The subject of his M.S. thesis was the design of an operating system extension for runtime reconfigurable platforms.

Currently, he is a Research Assistant at the ESAT-MICAS Laboratory of the Katholieke Universiteit Leuven. He is working towards a Ph.D. degree on the variability-aware design of embedded memories.

Wim Dehaene (S'88–M'97–SM'04) received the M. Sc. degree in electrical and mechanical engineering in 1991 from the Katholieke Universiteit Leuven. In November 1996 he received the Ph. D degree at the Katholieke Universiteit Leuven. His thesis is entitled CMOS integrated circuits for analog signal processing in hard disk systems.

After receiving the M. Sc. Degree Wim Dehaene was a research assistant at the ESAT-MICAS Laboratory of the Katholieke Universiteit Leuven. His research involved the design of novel CMOS building blocks for hard disk systems. The research was first sponsored by the IWONL (Belgian Institute for Science and Research in Industry and agriculture) and later by the IWT (the Flemish institute for Scientific Research in the Industry). In November 1996 Wim Dehaene joined Alcatel Microelectronics, Belgium. There he was a senior project leader for the feasibility, design and development of mixed mode Systems on Chip. The application domains were telephony, xDSL and high speed wireless LAN. In July 2002 Wim Dehaene joined the staff of the ESAT-MICAS laboratory of the Katholieke Universiteit Leuven where he is now a professor. His research domain is circuit level design of digital circuits. The current focus is on ultra-low-power signal processing and memories. Wim Dehaene is teaching several classes on digital circuit and system design. Wim Dehaene is a senior member of the IEEE.



Francky Catthoor (S'86–M'87–SM'98–F'05) received the engineering degree and the Ph.D. degree in electrical engineering from the Katholieke Universiteit Leuven (K.U.Leuven), Belgium, in 1982 and 1987, respectively.

Since 1987, he has headed several research domains in the area of high-level and system synthesis techniques and architectural methodologies, all within the Design Technology for Integrated Information and Telecom Systems (DESICS—formerly VSDM) division at IMEC, Leuven, Belgium. He was an Assistant Professor at the Electrical Engineering Department of the K.U.Leuven since 1989, and full Professor (part-time) since 2000. His current research activities belong to the field of architecture design methods and system-level exploration for power and memory footprint within real-time constraints, oriented towards data storage management, global data transfer optimization and concurrency exploitation. Platforms that contain both customized architectures and (parallel) programmable instruction-set processors are targeted. He also has an extensive scientific curriculum, including over 500 international papers (with three best paper nominations), TPC membership of many major conferences, and editorship in IEEE TRANSACTIONS ON VLSI SYSTEMS, TRANSACTIONS ON MULTI-MEDIA and *ACM Transactions on Design Automation for Embedded Systems*. In 1986, he received the Young Scientist Award from the Marconi International Fellowship Council. He was elected an IEEE Fellow in 2005.