

MIT Open Access Articles

Application-Specific SRAM Design Using Output Prediction to Reduce Bit-Line Switching Activity and Statistically Gated Sense Amplifiers for Up to 1.9x Lower Energy/Access

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Sinangil, Mahmut E., and Anantha P. Chandrakasan. "Application-Specific SRAM Design Using Output Prediction to Reduce Bit-Line Switching Activity and Statistically Gated Sense Amplifiers for Up to 1.9x Lower Energy/Access." IEEE Journal of Solid-State Circuits 49, no. 1 (January 2014): 107–117.

As Published: <http://dx.doi.org/10.1109/JSSC.2013.2280310>

Publisher: Institute of Electrical and Electronics Engineers (IEEE)

Persistent URL: <http://hdl.handle.net/1721.1/95890>

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-Noncommercial-Share Alike



An Application Specific SRAM Design Using Output Prediction to Reduce Bit-line Switching Activity and Statistically Gated Sense-Amplifiers for up to $1.9\times$ Lower Energy/Access

Mahmut E. Sinangil, *Member, IEEE*, Anantha P. Chandrakasan, *Fellow, IEEE*

Abstract—This paper presents an application-specific SRAM design targeted towards applications with highly correlated data (e.g. video and imaging applications). A prediction-based reduced bit-line switching activity scheme is proposed to reduce switching activity on the bit-lines based on the proposed bit-cell and array structure. A statistically gated sense-amplifier approach is used to exploit signal statistics on the bit-lines to reduce energy consumption of the sensing network. These techniques provide up to $1.9\times$ lower energy/access when compared to an 8T SRAM. These savings are in addition to the savings that are achieved through voltage scaling and demonstrate the advantages of an application-specific SRAM design.

Index Terms—10T-SRAM, application-specific, correlation of data, energy-efficient SRAM, signal statistics.

I. INTRODUCTION

CONTINUOUS scaling of process technologies driven by Moore's law has resulted in the integration of more transistors and more functionality on a single chip. This advancement has led to a wide range of new applications including mobile devices such as smartphones and tablet devices. These mobile devices pack increasingly more processing capabilities (e.g. quad-core processors) but can do little for extending the battery life and cooling due to their compact form factor. Hence, mobile applications require circuits to be extremely energy efficient and at the system level, this requires careful and joint selection of solutions not only at the algorithm and architecture levels but also at the circuits level [1]. Hence, circuits should also be thought within the context of its target application and designed by considering the application-specific properties.

SRAMs are the most common type of on-chip memories and a larger fraction of chip area has been allocated for SRAMs in modern integrated circuits as highly-parallelized designs often benefit from larger on-chip storage. Consequently, for system-level implementations, designing low-power, area-efficient and robust SRAMs can be detrimental for the overall power and cost of the design.

Voltage scaling is an effective way of reducing power consumption and low-voltage SRAM designs have been widely

investigated in the literature. As conventional 6T-based SRAM designs fail to operate at low voltage levels, recent work has developed new bit-cell topologies [2]–[6], new array architectures [7]–[9] and various assist techniques [10]–[16] to enable robust operation at low-voltage levels.

This work proposes an application-specific SRAM design that is targeted towards motion-estimation engine of a video encoder hardware. Correlation of storage data is used in the transistor-level design to reduce bit-line switching activity and consequently energy consumed during read operations. Signal statistics are considered in the design of the sensing network to provide further energy savings. Although this design is targeted for motion estimation, ideas can be generalized to other applications which possess similar or same properties with motion estimation. Lastly, it should be noted that energy savings achieved through the application-specific SRAM design are in addition to the savings from voltage scaling and help to maximize energy efficiency.

A. Application-Specific SRAM Designs

Application-specific designs can improve energy efficiency when compared to general-purpose designs because the former has the opportunity to optimize for the specific needs of a single scenario whereas the latter has to support a wider range of possible scenarios. Conventional SRAMs are designed without considering the data that will be stored in its array of memory cells. Moreover, SRAMs are generally tested with random data or worst-case access patterns for the specific design to characterize for the limits of the operational range. Although it is critical to test with worst-case patterns and ensure memories can work under these extreme cases, the properties of the data stored in the cell array can be utilized to improve energy efficiency without compromising the operation under the extreme cases.

Data stored in an SRAM can often have particular properties that can change from one application to the other and an application-specific SRAM design can be tailored for the target application by taking these properties into consideration. Moreover, access patterns to the SRAMs that are specific for the applications can also provide useful information for the design. These additional information can provide a new dimension for circuit designers to explore and can lead to designs that are optimized better for higher energy efficiency, smaller area or higher performance.

M. E. Sinangil is with NVIDIA in Bedford, MA 01730-1401 USA e-mail: msinangil@nvidia.com

A. P. Chandrakasan is with Electrical Engineering and Computer Science Department at Massachusetts Institute of Technology, Cambridge, MA 02139 USA e-mail: anantha@mit.edu

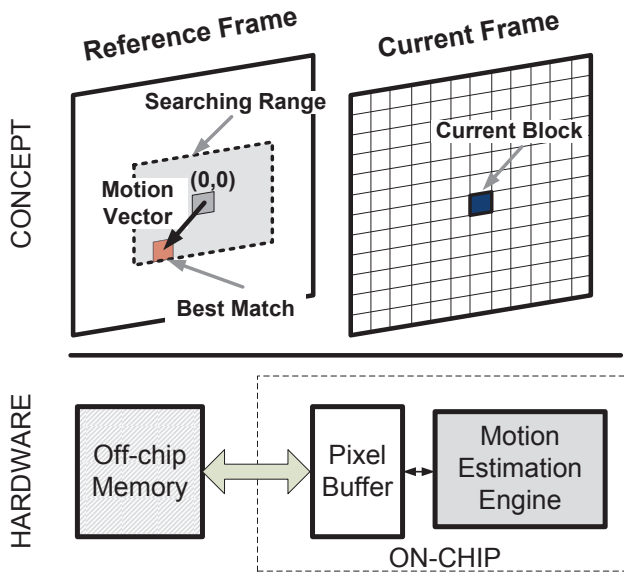


Fig. 1. Block based motion estimation concept and high level hardware implementation diagram.

The work in [17] is an example of an application-specific SRAM design. In this work, conventional 8T SRAM bit-cell is used to store the data. Because this bit-cell has a single-ended read port, read-bit-line is only discharged when a ‘1’ is held in the cell. Hence, to avoid the energy consumption due to the switching of the read-bit-lines, this design employs a scheme where an inversion bit is added for each word to store the data or its complement in the array such that the number of ‘1’s in every word is minimized.

A similar idea is proposed in [18] where least-significant-bits (LSB) of each word is stored in an error-prone but area-efficient 6T SRAM. Occasional bit-errors at low voltages are tolerable for its target application as these errors are limited to the LSB of each word.

Lastly, the work in [19] uses a bit-cell with a full inverter and a transmission gate as the read port such that read bit-lines can be driven by the bit-cell and pre-charging operation can be avoided. If the data stored in the array is correlated, read bit-lines can be driven to the same state across multiple cycles and power consumption is reduced.

B. Motion Estimation and Video Applications

Motion estimation is the process of finding the movement of objects within a sequence of image frames [20]. To find the movement, a motion search is performed on a previously encoded reference frame (Fig. 1). This motion search involves the pixel-by-pixel evaluation of a metric between the search range of the reference frame and the current block from the current frame. Hence, consecutive memory accesses to an on-chip reference pixel buffer are done during a motion search to access the necessary data. These on-chip buffers hold the pixels from reference frames and are generally implemented as SRAMs.

It should be noted here that motion estimation is a complex and computationally intensive process and is generally implemented with many engines performing the motion search in parallel. Hence, the duty cycle of the memories constituting the on-chip reference buffers are high and energy consumption of these memories is an important part of the overall motion estimation design [21].

This paper is structured as follows: Section II talks about the design decisions for the application-specific SRAM design targeted towards video and motion estimation. This section talks about specific features of the motion estimation and how these features can be used to reduce energy consumption. Then Section III explains the prediction-based reduced bit-line switching activity (PB-RBSA) SRAM design, starting with the bit-cell and array implementation and then focusing on prediction generation circuit and statistically gated sense amplifiers. Lastly, Section IV presents measurement results from a test chip fabricated in 65nm low-power CMOS process and Section V concludes the paper.

II. DESIGN DECISIONS FOR THE APPLICATION-SPECIFIC SRAM FOR VIDEO APPLICATIONS

This part of the paper will focus on the motion estimation specific features that are used in the application-specific SRAM design. These application specific features drive the design decisions and shape the transistor-level design of the SRAM.

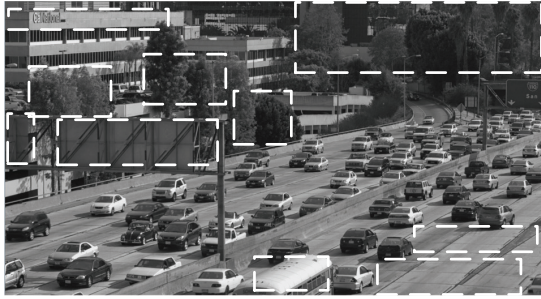
A. Motion Estimation Specific Features

As mentioned in Section I, SRAMs in motion estimation store pixels for the motion search. It should be noted that hardware implementation of the motion estimation engines generally use the luminance component of the pixel intensities (represented with 8-bit unsigned values) to perform a motion search.

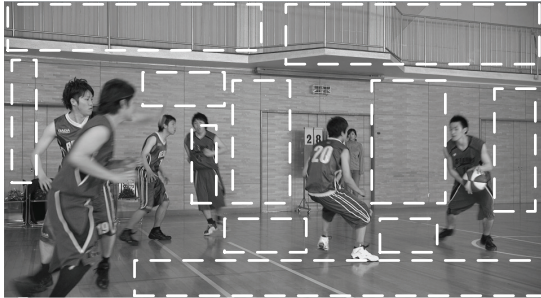
Two example frames are shown in Fig. 2 [22]. In this figure, areas of example frames with high correlation of pixel intensities are highlighted. Qualitatively, it can be observed that the pixels belonging to the same object or the same background are highly correlated in their intensities. Consequently, when reading a reference frame from the on-chip reference buffers during motion search, it is likely that pixel intensities will be correlated. The context of the frame is an important parameter here and frames with a lot of details will possess lower correlation when compared to frames with large and smooth objects or backgrounds.

It should be noted that as we go to higher resolutions, which is likely because of consumer demand in the past years, more pixels will be used to represent an object and the correlation of pixel intensities will be higher.

To quantify the correlation of pixels, a block average can be calculated for every 16×16 region and then the difference of each pixel value from the block average can be plotted. Fig. 3 shows the distribution of these differences for the video frames in Fig. 2. The distributions of differences from block average show that 58% and 76% of the pixels lie within ± 3 bits of the block average. In other words, out of 8 bits of a



(a)



(b)

Fig. 2. (a) Traffic (2560 × 1600) and (b) Basketball (1920 × 1200) example frames. Areas with high correlation of pixel intensities are highlighted.

pixel, more than half of the bits can be same with the block average.

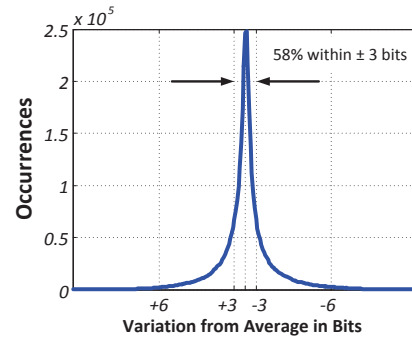
It should be noted that the binary representation of the pixels will result in the most-significant-bits (MSB) to switch at certain values. For example, from the binary representation of 127 (01111111) to 128 (10000000), all bits do change. However, this is a corner case. Moreover, a simple mapping can be done by shifting every pixel’s binary representation by a certain value [23] to reduce these effects.

Second feature that is specific to motion estimation and drives the design decisions for this work is that number of read accesses is significantly larger than the number of write accesses for motion estimation buffers. This is mainly because of the data reuse between consecutive blocks and the overlapping search ranges. On average, a pixel that is written to the reference pixel buffer is read more than three times [24] and energy consumption of read accesses is far more important than the energy consumption during write accesses.

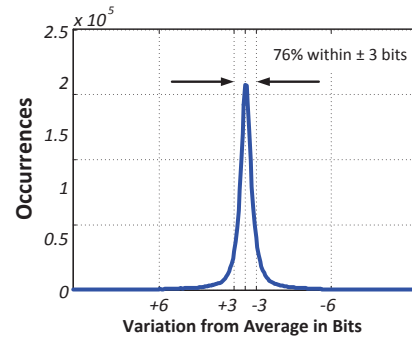
Based on these features, an SRAM design that provides lower energy/access in read accesses is suitable for SRAMs used in motion estimation engines and the correlation of pixel intensities can be utilized to achieve this.

B. Bit-line Switching Activity During Read Accesses

In high density arrays, bit-lines are shared by a large number of cells across a column (e.g. 256 or 512 bit-cells). The parasitic capacitance due to the devices connected to this signal as well as the metal parasitic capacitance due to routing of this signal contribute to the total capacitance of a bit-line and total bit-line capacitance is large in high-density arrays.



(a)



(b)

Fig. 3. Distribution of differences of each pixel in a 16 × 16 block from the block average for (a) Traffic and (b) BasketballDrive sequences. 58% and 76% of the differences lie within 3 bits of the block average.

**8T SRAM Read Power Breakdown
Measured at $V_{DD}=0.6V$**

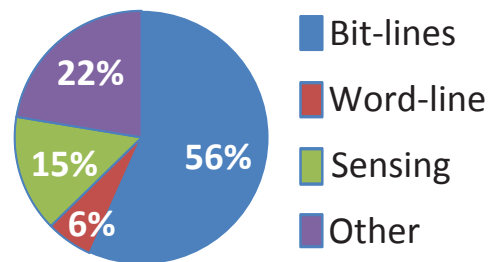


Fig. 4. Measured power consumption breakdown for a conventional 8T SRAM operating at 0.6V and at room temperature. Bit-line switchings account for more than half of the total power consumption of the SRAM.

Fig. 4 shows the measured power consumption breakdown of a conventional 8T SRAM operating at 0.6V with 64-bit words and a column multiplexing ratio of one. Because of the large bit-line capacitance, bit-lines account for more than half of the total power consumption during read accesses and a reduction in bit-line power can reduce overall power consumption during read accesses.

Analyzing the power consumption due to bit-line switching in SRAMs more closely, power consumed by a bit-line switching can be written as

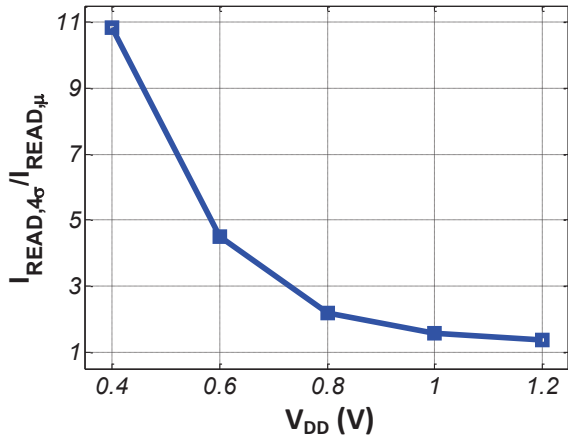


Fig. 5. Ratio of 4σ cell read current to average cell read current across different voltages. The ratio is close to unity at 1.2V but quickly increases as the operating voltage approaches sub-threshold region.

$$P_{BL,consumed} = \alpha_{0 \rightarrow 1} \times C_{BL} \times V_{DD} \times \Delta V_{min} \times f$$

where $\alpha_{0 \rightarrow 1}$ is the activity factor, C_{BL} is the total switching bit-line capacitance, V_{DD} is the supply voltage, ΔV_{min} is the minimum amount of voltage development on bit-lines that can be resolved correctly by sense-amplifiers and f is the frequency of operation. It should be noted that although minimum voltage swing on the bit-lines (ΔV_{min}) is intended to be small (e.g. 50-100mV), ΔV_{min} approaches V_{DD} at low voltages and the average development across all bit-lines will be closer to V_{DD} as described in [25]. This is because of the increasing ratio of the worst-case cell current to average cell current at low-voltages as shown in Fig. 5. This ratio is close to unity at 1.2V whereas it becomes nearly 5 at 0.6V. As the timing of the signals are adjusted to track the worst-case cell in the array, all the remaining cells that are faster than the worst-case cell in the array discharge the bit-lines at a faster rate and this results in the effective voltage swing on the bit-lines to be much larger than ΔV_{min} and approach to V_{DD} .

It can be seen from the above equation for power consumed by bit-line switchings that a reduction in the activity factor can provide proportionate savings in power consumption due to bit-line switchings. During a read access, all bit-lines go through pre-charging and signal development phases. For 6T SRAMs, differential nature of the cell results in one of the bit-lines to be actively pulled down during a read access. Hence, activity factor of one of the bit-lines in a 6T SRAM is 1. For a conventional 8T SRAM, discharging of RBLs depends on the data that is being read from the bit-cell. Hence, the activity factor of read bit-lines in a conventional 8T SRAM is data dependent and can range from 0 to 1 depending on how many '0's and how many '1's are present in the array. The data-dependent nature of the activity factor of bit-lines in a conventional 8T SRAM motivates for designing a cell and array structure that can reduce the activity factor of the bit-lines dynamically depending on the bits that are accessed

from the SRAM. This can provide significant savings for the power consumption of the SRAM.

III. PREDICTION-BASED REDUCED BIT-LINE SWITCHING ACTIVITY (PB-RBSA) SRAM DESIGN

A. PB-RBSA Bit-cell Design

To reduce bit-line switching activity of bit-lines using the correlation of video data, a new bit-cell topology is proposed that uses a bit-wise prediction.

Fig. 6 shows the PB-RBSA bit-cell topology. It consists of a cross-coupled inverter pair, two NMOS access devices connected to the storage nodes and two read-buffers. The footer node of the read-buffers are not connected to ground but connected to a predictor (pred) and its complement (predB). In other words, pred is a prediction of what is stored in the bit-cell.

Write operation is standard in the PB-RBSA bit-cell where BL and BLB overwrite the previous state of the cross-coupled inverters through the two NMOS pass transistors when WWL signal is asserted. A column multiplexing ratio of one is used in this work and as a result, bit-cells do not experience a half-select disturbance during write accesses.

Fig. 7 shows two different cases for read operation. For both cases, at the beginning of the access, RBL0 and RBL1 are pre-charged to V_{DD} and then RWL is asserted. Fig. 7-a shows a correct prediction case where a '1' is stored in the cell and pred is also '1'. In this case, both RBL0 and RBL1 stay at V_{DD} as the read buffer connected to RBL1 is turned off and there is no voltage difference across the read buffer connected to RBL0. Hence, with correct prediction, neither RBL0 nor RBL1 is discharged during a read access, preventing the activity on the read-bit-lines. In the case of an incorrect prediction (Fig. 7-b), on the other hand, read-buffer connected to RBL1 is turned off but RBL0 can be discharged to GND as there is a voltage difference from RBL0 to pred. Consequently, in the case of an incorrect prediction, one of the read-bit-lines is discharged to GND.

This work uses an architecture with 256 cells on a bit-line. However, to oppose the effect of leakage on bit-lines, it is possible to create a hierarchical bit-line structure and apply a similar prediction scheme to reduce bit-line switching activity.

During a read access, pred and predB signals are driven first and then RWL for the accessed row is asserted. This is ensured by delaying the RWL enable pulse with respect to the clock edge.

It should be noted that if the data stored in the array is correlated and it can be predicted with high accuracy, correct

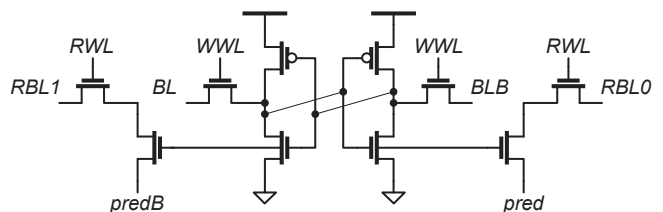


Fig. 6. PB-RBSA bit-cell.

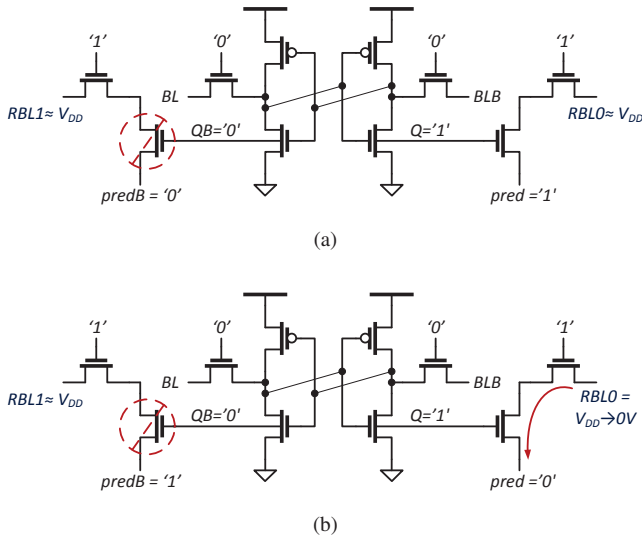


Fig. 7. (a) Correct and (b) incorrect prediction cases during a read operation with the PB-RBSA bit-cell. RBL0 and RBL1 stay at V_{DD} with correct predictions whereas RBL0 or RBL1 is discharged to GND with incorrect predictions.

predictions will be more frequent than incorrect predictions and the activity factor on the read-bit-lines can be significantly reduced.

B. Construction of the PB-RBSA Cell Array

Fig. 8 shows the cell array architecture for the PB-RBSA SRAMs. BL/BLB, RBL0/RBL1 and pred/predB pairs are routed in the column-wise direction and shared by the entire column of PB-RBSA bit-cells. WWL and RWL signals are routed in the horizontal direction and shared across a row of bit-cells. The row circuit for the PB-RBSA design is very similar to the row circuit for the conventional 8T design with separate drivers for the read and write ports of the cell and occupy roughly the same area in layout.

Sensing network resolves the RBL0 and RBL1 voltage levels during a read access and decides if the prediction was correct or not. Depending on the prediction being correct or not, pred or predB is driven to the output respectively. Drivers for BL/BLB and pred/predB are designed to be static inverter-based buffers and drive these nodes during the entirety of the accesses.

It should be noted that pred and predB lines span the entire height of the array and have a large capacitive load. Hence, switching these two lines on a cycle-by-cycle basis would introduce significant energy consumption. However, if the data in the array is correlated, the predictors can be updated at a much lower rate (e.g. every 16 or 32 cycles) and the additional energy consumption associated with the switching of pred/predB pair can be amortized across multiple cycles.

Fig. 9 depicts a layout sketch of the PB-RBSA bit-cell. Logic rules are used to implement the PB-RBSA and the 8T SRAM. Because of the additional devices, the width of the cell is longer and this allowed all column-wise signals to be routed in MET2 as shown in Fig. 9. WWL and RWL are routed in MET3 and MET5 respectively. PB-RBSA cell area is 20%

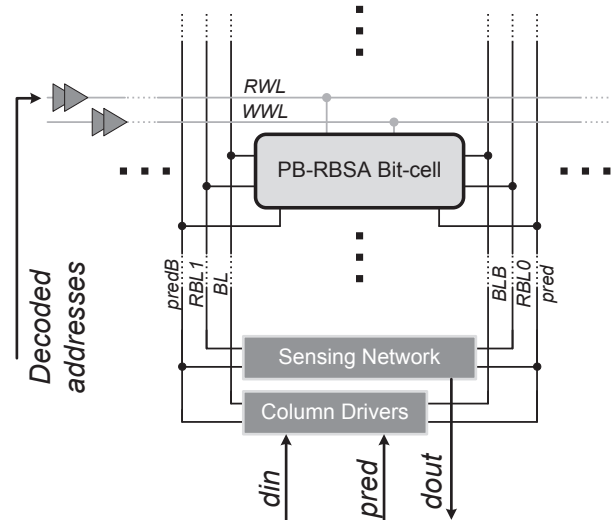


Fig. 8. PB-RBSA array architecture with drivers for RWL and WWL, column drivers for BL/BLB and pred/predB pairs and sensing network for RBL0/RBL1 voltage development.

larger than a conventional 8T bit-cell drawn with logic rules. It should be noted that the footer of read buffers (i.e. pred and predB) in the PB-RBSA cell are column-wise signals and the metal resistance along with pred/predB driver resistance can have a limiting effect on read performance especially in deep sub-micron CMOS technologies.

C. Prediction Generation Circuit

Prediction generation is an important aspect of the PB-RBSA SRAM design as correct predictions lead to lower energy consumption. Various different prediction generation designs are possible and in this work we used arithmetic averaging of previous outputs of the SRAM to calculate the predictor because arithmetic averaging provides a good solution. Arithmetic averaging captures the common component of the correlated pixel intensities and it can be implemented

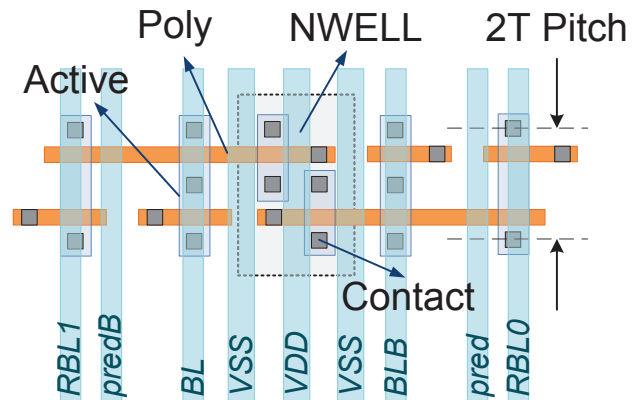


Fig. 9. Sketch of the PB-RBSA bit-cell layout. Vertical rectangles show the approximate placement of MET2 routing for various signals.

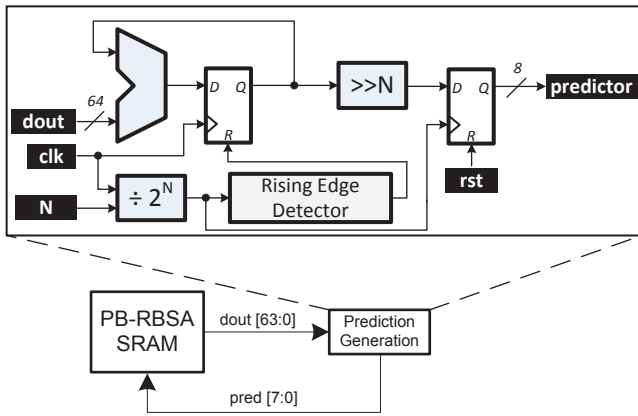


Fig. 10. Prediction generation circuit. 8 pixel wide outputs from the SRAM blocks are accumulated for 2^N cycles and their average is calculated and used as the predictor for the next 2^N cycles.

in circuits with simple adders and shifters. Fig. 10 shows the circuit calculating the predictor. 8 pixel wide (64-bit) output from the SRAM banks are accumulated for 2^N cycles where N is an input to the circuitry. Then, average of the accumulated sum is calculated and latched to be used as the predictor for the next 2^N cycles. It should be noted here that the output of the prediction generation circuit is an 8-bit pixel-predictor.

Selection of N introduces an interesting trade-off for the energy consumption of the PB-RBSA SRAMs (Table I). As we access pixels from the SRAMs continuously, there will be changes in the pixel intensities based on the context of the image frame. However, these changes can be a part of an object boundary in the image frame which will affect many pixels for many cycles or it can be a small detail or noise affecting only a couple of pixels. With smaller N and a shorter latency in prediction generation circuit, predictor is allowed to adjust to changes in the pixel intensities more rapidly so that the prediction can be correct more often. With larger N , on the other hand, a longer averaging period results in the predictor to be updated after a longer latency. However, smaller N and updating the predictor more frequently also results in pred/predB pair in the SRAM array to be switching more often and results in additional switching activity and energy consumption.

TABLE I
TRADE-OFFS FOR THE SELECTION OF N

| Selection of N | Prediction Accuracy (Benefit) | pred/predB Activity (Cost) |
|------------------|-------------------------------|----------------------------|
| Smaller N | Higher | Higher |
| Larger N | Lower | Lower |

To demonstrate this effect, a special case is considered in Fig. 11 where an image is constructed with a gradient of pixel intensities going from white to black and access pattern is adjusted to cycle between these gradients of intensities continuously. In this scenario, the output of the SRAM will change from all ‘0’s to all ‘1’s and the selection of prediction update period, 2^N , will be important.

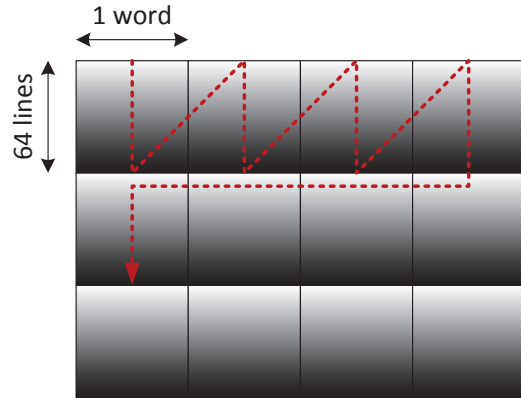


Fig. 11. An example frame constructed with a gradient of pixel intensities from light to dark. Red dashed arrow shows the access pattern which cycles between these gradients of intensities.

Fig. 12 shows normalized power consumption of the PB-RBSA SRAM with varying 2^N for the case explained in Fig. 11. For smaller values of 2^N , power consumption is dominated by the switching activity of the pred/predB pairs in the array. For larger values of 2^N , SRAM power is mostly due to bit-line switching activity as the predictors start to be incorrect after a couple of cycles. These two conflicting trends result in $2^N = 16$ to provide a minimum point in the curve.

In this work, for most of the sequences used for measurements [22], selection of 2^N as 16 or 32 is found to provide the lowest power consumption.

D. Statistically-Gated Sense-Amplifiers

PB-RBSA SRAMs reduce bit-line switching activity by using predictors in the design and provide savings when the data in the storage array is correlated. This leads to a natural consequence for the sensing network that the sense-amplifiers

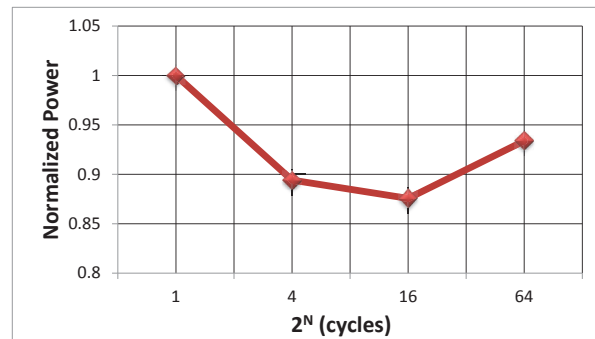


Fig. 12. Normalized power consumption of the PB-RBSA SRAM with varying 2^N for prediction generation circuit under the case explained in Fig. 11.

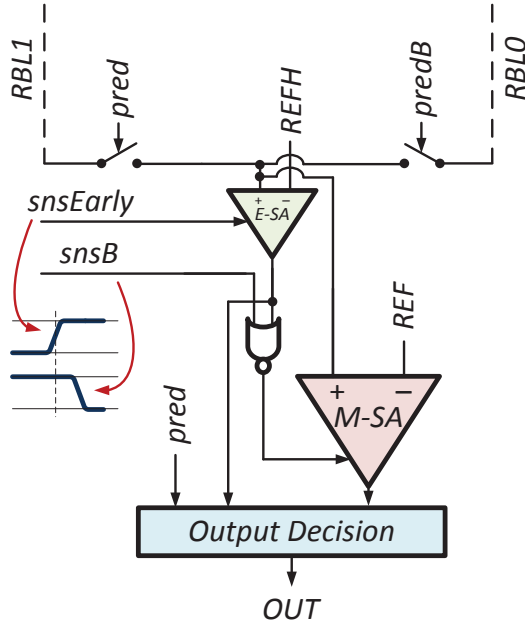


Fig. 13. Sensing network used in this work employing statistically gated sense amplifiers.

will be sensing a correct prediction (i.e. both RBL0 and RBL1 staying high) more often than an incorrect prediction (i.e. one of the RBLs discharged to GND). Hence, designing a sensing network that consumes less energy than a conventional design when sensing a correct prediction and more energy when sensing an incorrect prediction can be more energy efficient if the predictions are correct most of the time.

The sensing network used in this work is shown in Fig. 13. It consists of two sense-amplifiers: E-SA and M-SA. E-SA is strobed earlier than M-SA and if the output of E-SA is ‘1’, then M-SA is gated and its energy consumption is avoided. Output prediction logic consists of a couple of static CMOS logic gates and decides if the prediction was correct based

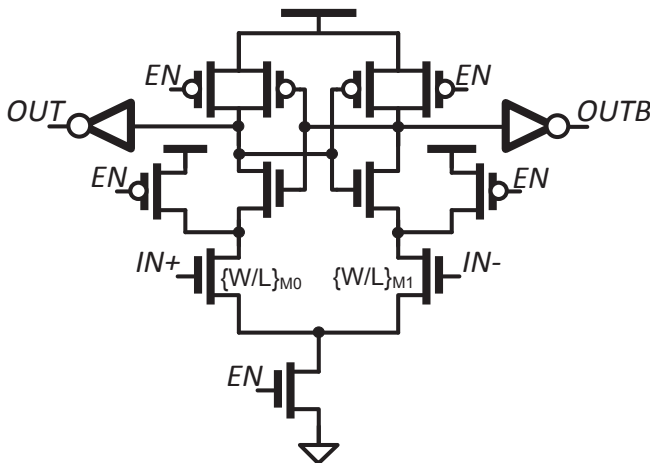
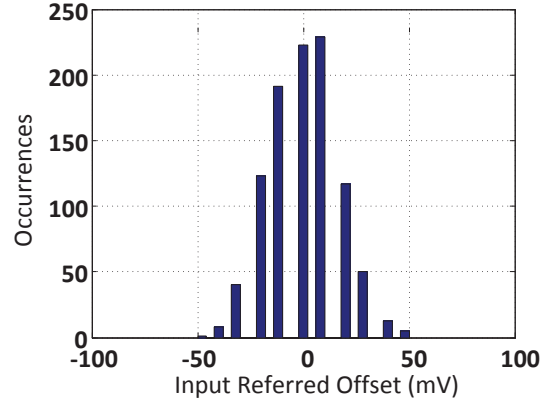
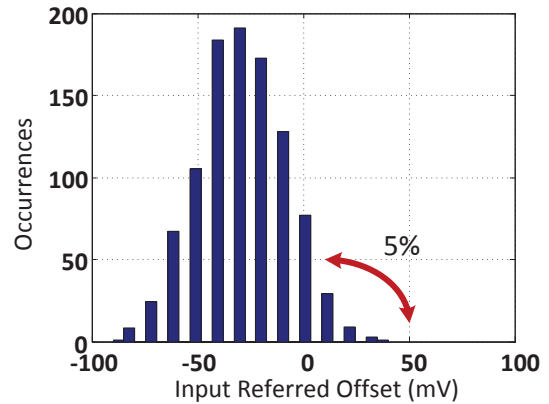


Fig. 14. Sense-amplifier design used in this work.



(a)



(b)

Fig. 15. Input offset distributions for (a) M-SA and (b) E-SA. Input offset of the E-SA is shifted towards negative offset voltages such that 5% of the sense-amplifiers have a positive offset.

on the output of E-SA and M-SA and finally drives pred or predB to the output. The switches connecting RBL0/RBL1 to the inputs of the sense-amplifiers are selected by the value of pred/predB as only the read buffer with a ‘0’ at its footer can potentially discharge its read-bit-line.

Both E-SA and M-SA are designed as latch-type sense-amplifiers with separately controlled reference voltages as shown in Fig. 14. The reference voltages are connected to the *IN*- terminal of the sense-amplifiers. M-SA is sized to be larger which results in a tighter offset distribution (Fig. 15-a). The offset span of the M-SA is around 100mV from simulations and from measurement results. This dictates a minimum of 100mV of separation between a ‘high’ and ‘low’ read-bit-line and determines the minimum read word-line pulse-width.

E-SA is sized to be smaller for lower energy but this also results into a wider (roughly 150mV span) offset distribution as shown in Fig. 15-b. Because of this larger offset span, E-SA requires a large voltage development on the read-bit-lines. Alternatively, operating E-SA in a scenario with a separation of 100mV for its inputs results in erroneous outputs as E-SA cannot resolve some of the cases correctly. It should be noted

that in the cases where the output of the E-SA is ‘1’, M-SA is gated but for the cases of E-SA outputting a ‘0’, M-SA is activated. If the erroneous outputs of E-SA are intentionally limited to the cases where E-SA outputs a ‘0’, M-SA can resolve the inputs correctly and fix the errors E-SA can make. This can be done by first shifting the offset distribution of E-SA towards negative offset voltages by properly adjusting the driving strengths of its input devices through sizing and setting its reference voltage (REFH) higher than M-SA’s reference voltage (REF). This will ensure that when E-SA outputs a ‘1’, it will be correct and when E-SA outputs a ‘0’, it will either be a correct ‘0’ or an erroneous ‘0’ which can be fixed by the M-SA. For this purpose, the output of M-SA is given a higher precedence in the cases when E-SA outputs a ‘0’. These modifications in the design ensure that the errors in E-SA output do not propagate to the final output and the sensing network operate correctly in all cases.

An example showing two cases to demonstrate the operation of the statistically gated sense-amplifiers is given in Fig. 16. For these cases, *pred* is assumed to be ‘1’ and *RBL0* is selected for the sensing. Reference voltage for E-SA is connected to V_{DD} and the reference voltage for M-SA is connected to $V_{DD} - 50mV$, in the middle of the offset voltage span of the sense-amplifiers. The effect of leakage on the read-bit-lines are not considered for simplicity and the reference voltage levels can be adjusted properly to mitigate the effects of leakage.

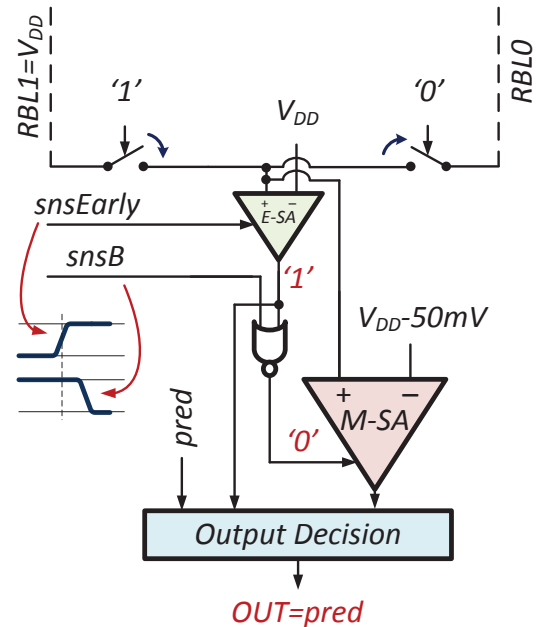
Fig. 16-a shows a correct prediction case and a negative input offset for the E-SA. Because of the correct prediction, *RBL0* is not discharged but stays at V_{DD} . All E-SAs with a negative input offset voltage can resolve this case correctly and output a ‘1’ which will gate M-SAs. In this case, output is resolved by only strobing E-SA.

Fig. 16-b also shows a correct prediction case but with a positive input offset for the E-SA. In this case, E-SA cannot resolve *RBL0* at V_{DD} correctly and outputs a ‘0’ erroneously. M-SA, in this case, is strobed and correctly output a ‘1’ and this output will be used in the output prediction logic as explained above.

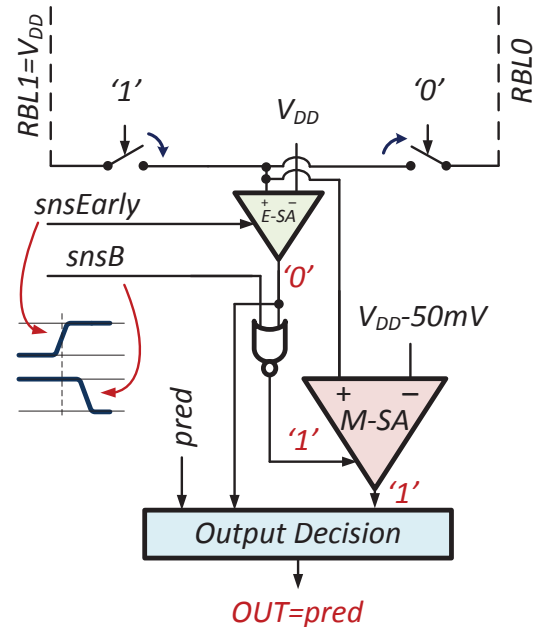
In the remaining two cases of incorrect prediction with a positive or negative input offset for E-SA, E-SA outputs a ‘0’ correctly. Although unnecessary, M-SA is strobed in these cases as we cannot distinguish an erroneous ‘0’ from a valid ‘0’.

Table II summarizes the four cases. It should be noted that correct predictions are expected to be more frequent than incorrect predictions and only 5% of the E-SAs have a positive offset. Hence, statistically, it is much more likely that only E-SA is activated during a read operation. Fig. 17 shows energy consumed in the sensing network with different correct prediction percentage numbers for statistically gated sense-amplifiers and a conventional design employing an M-SA alone. Statistically gated sense-amplifiers provide energy savings when the correct prediction percentage is larger than 40%.

Statistically gated sense-amplifier network results in roughly $2\times$ the area of the M-SA. However, this area overhead is amortized by the height of the memory cell array (i.e. number of cells on a column).



(a)



(b)

Fig. 16. Statistical sensing network operation with correct prediction and (a) negative input offset and (b) positive input offset for the E-SA.

TABLE II
IS M-SA GATED?

| | Correct Pred. | Incorrect Pred. |
|-----------------------|---------------|-----------------|
| E-SA with Neg. Offset | Yes | No |
| E-SA with Pos. Offset | No | No |

IV. TEST CHIP AND MEASUREMENT RESULTS

To demonstrate the ideas described in this work, a 65nm test chip is fabricated on a low-power CMOS process. A die

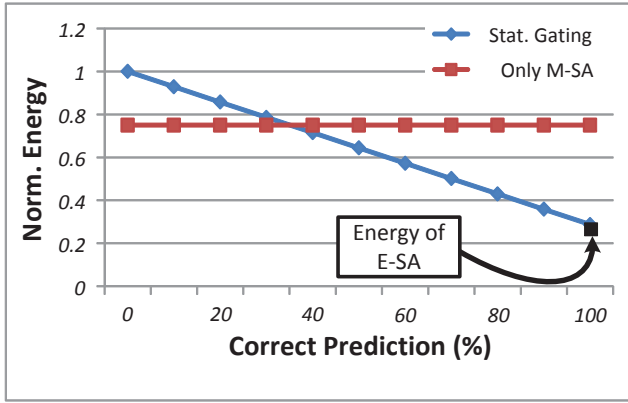


Fig. 17. Energy consumed in the sensing network with varying correct prediction percentage for statistically gated sense-amplifiers and a single conventional sense-amplifier.

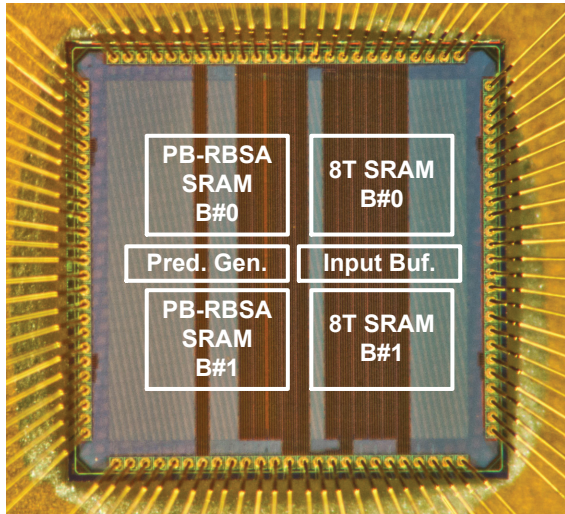


Fig. 18. Die photo of the 65nm test chip fabricated in low-power CMOS process. Two blocks of the PB-RBSA SRAM are placed side-by-side with two blocks of the conventional 8T SRAM to enable on-fly energy consumption comparison.

TABLE III
CHIP SPECIFICATIONS

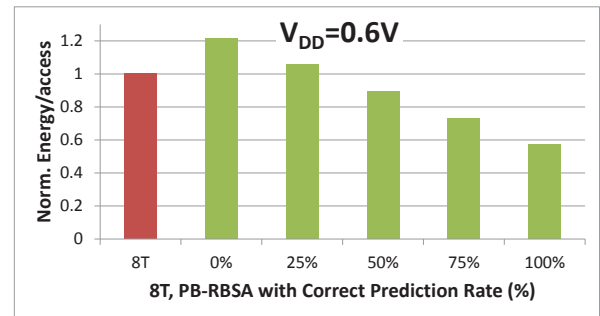
| | |
|---------------------------|--|
| Technology | 65nm Low-Power CMOS |
| Die Size | 2.3mm × 2.3mm |
| Operating Voltage Range | 0.52-1.2V |
| PB-RBSA SRAM Organization | 32Kbit (256 Rows x 64 Cols x 2 Blocks) |
| 8T SRAM Organization | 32Kbit (256 Rows x 64 Cols x 2 Blocks) |

photograph of the test chip is shown in Fig. 18 and Table III provides test chip specifications. PB-RBSA SRAM design works from 1.2V down to 0.52V. It should be noted that PB-RBSA SRAM uses voltage scaling and its application-specific design at the same time to maximize energy efficiency. The large voltage range enables this design to be used for a motion estimation engine that can scale its throughput to cover different resolutions of video data or different quality

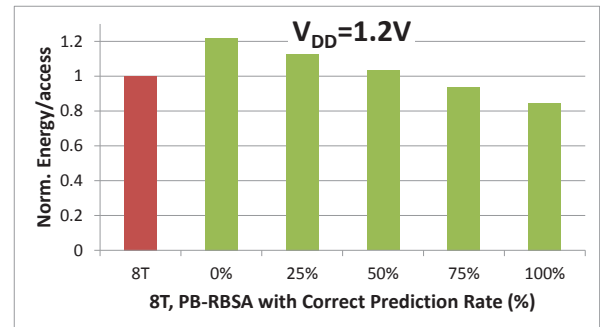
requirements of the encoding process.

To make real-time comparison of energy consumption, two blocks of the conventional 8T SRAM with conventional sense-amplifiers are placed on the test chip along with two blocks of the PB-RBSA SRAM and same input data is provided to both designs. Separate pins are used to power up 8T and PB-RBSA SRAMs and to do on-fly energy measurements. Reference voltages for the statistically gated sense-amplifiers (REF and REFH) are generated off-chip and input to the chip through dedicated pins. Prediction generation circuit is also implemented on the chip and occupies (0.012 μm^2) and introduces 3% area overhead when compared to the area of two PB-RBSA SRAM blocks. This overhead can be even smaller if the prediction generation circuit is shared across a larger number of PB-RBSA SRAM blocks.

Fig. 19 shows the measured read energy/access numbers for both the conventional 8T SRAM and the PB-RBSA SRAM at 0.6V and at 1.2V. For this experiment, 8T SRAM is



(a)



(b)

Fig. 19. Measured read energy/access numbers for the 8T SRAM and the PB-RBSA SRAM with varying correct prediction rates at (a) 0.6V and (b) 1.2V.

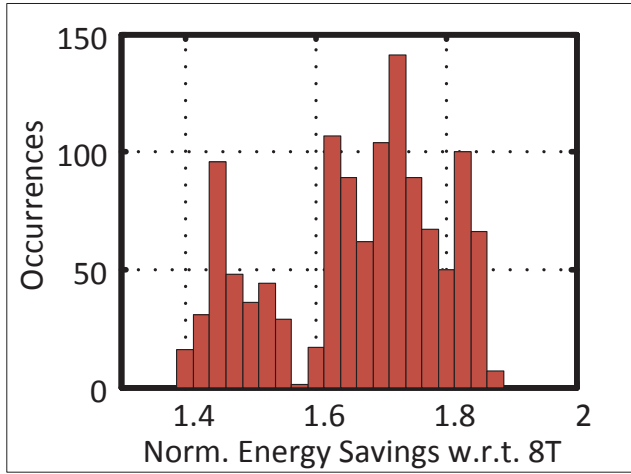
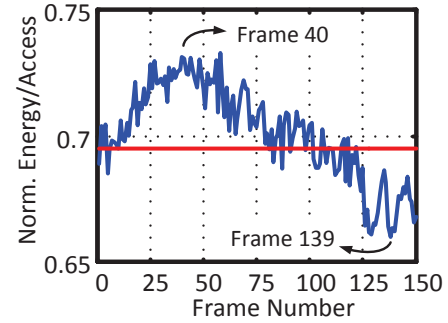


Fig. 20. Distribution of energy savings with respect to the conventional 8T SRAM with real video sequences. The test set consists of 11 different sequences and 1100 video frames with resolutions ranging from 1280×720 to 2560×1600 .

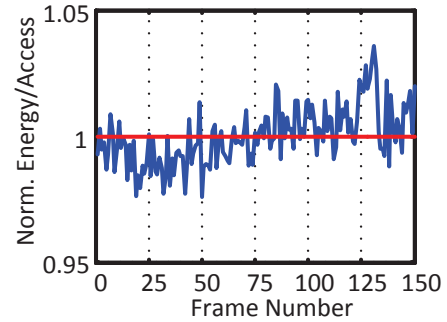
programmed to have 50% ‘0’s and 50% ‘1’s and the activity factor on the read-bit-lines of the 8T SRAM is 0.5. For the PB-RBSA SRAM, predictor and array data are programmed to provide varying correct prediction rates. At 0.6V (Fig. 19-a), PB-RBSA SRAM provides up to $1.75\times$ lower energy/access compared to the 8T SRAM with 100% correct prediction rate. The savings are smaller with lower correct prediction rate and at the 50% correct prediction rate, PB-RBSA SRAM provides $1.1\times$ lower energy/access mainly due to the statistically gated sense-amplifiers employed in the PB-RBSA design. Below 40% correct prediction rate, 8T energy/access is lower than PB-RBSA design. At 1.2V (Fig. 19-b), energy savings from the PB-RBSA design are smaller since the contribution of bit-line switching to overall SRAM energy is lower at higher voltages. At 1.2V, PB-RBSA SRAM provides up to $1.2\times$ lower energy/access compared to the 8T SRAM. These results are in agreement with the simulation results with post-layout extraction.

To get a distribution of energy savings with the PB-RBSA design, experiments with real video data are performed on the test chip and results are plotted in Fig. 20. 11 different sequences and 1100 image frames are used in the experiments with resolutions ranging from 1280×720 to 2560×1600 . Savings are plotted with respect to the energy consumption of the 8T design. Results showed that savings with the PB-RBSA SRAMs are higher with larger video resolutions and savings up to $1.9\times$ are reported. It should be noted that with the continuous increase of video resolutions and the introduction of $4K \times 2K$ and $8K \times 4K$ in the future, the savings with the PB-RBSA SRAMs can be expected to be even higher.

As mentioned in Section II, the context of the image frame is an important parameter for the correlation of pixel intensities. Fig. 21-a and Fig. 21-b shows frame-by-frame energy/access numbers for the PB-RBSA and the conventional 8T SRAMs for a sequence of 150 image frames. Average energy/access numbers for both designs are shown with the



(a)



(b)



(c)



(d)

Fig. 21. Frame-by-frame energy/access numbers for the (a) PB-RBSA and (b) 8T SRAM for a 150 frame long sequence. Frame (c) 40 and (d) 139 of the sequence are also provided to show the change in the contents.

red solid lines and all numbers are normalized to the average energy/access for the 8T SRAM. Fig. 21-c and Fig. 21-d shows the 40th and 139th frames from the sequence. When the contents of the frame are dominated by the grassy background

which is harder to predict because of many details (Fig. 21-c), energy/access for the PB-RBSA SRAM increases. On the contrary, when the image frame is dominated with the riders and horses (Fig. 21-d) which provides smooth surfaces that are easier to predict, energy/access for the PB-RBSA SRAM decreases. Lastly, for the 8T SRAM, energy/access numbers depend on the number of '0's and '1's in the pixel intensities and seem to be independent of the contents of the image frames.

V. CONCLUSION

Designing circuits to be application-specific can provide significant improvements in terms of energy efficiency by

- optimizing the design for a specific target and
- exploiting the specific features of the application.

However, these application-specific optimizations should not be limited to algorithm and architecture level but should be extended to cover the circuits level as well. This work proposes an application specific SRAM design targeted towards motion estimation and video applications where techniques are developed for utilizing the correlation of input data and signal statistics.

First, prediction-based reduced bit-line switching activity (PB-RBSA) scheme is proposed to exploit the correlation of data in the memories. Specifically, PB-RBSA scheme introduces a predictor for the read data of the SRAM and bit-line transitions are avoided when the predictor is correct. To complement this idea, a statistically gated sense-amplifier approach is developed to take advantage of the biased transition probabilities on the bit-lines of the PB-RBSA SRAMs. A smaller sense-amplifier that is intentionally designed with a non-symmetric input offset distribution is used to evaluate the most-likely case correctly and to gate the larger main sense-amplifier.

Proposed techniques are implemented in a 65nm prototype which is tested for functionality down to 0.52V. PB-RBSA scheme with sense-amplifier gating provides up to $1.9\times$ lower energy/access with respect to a conventional 8T design that is also implemented on the same test chip. It should be noted that energy savings achieved through application-specific SRAM design is the next level of savings on top of the savings from voltage scaling. In other words, PB-RBSA scheme does not prevent SRAMs to do voltage scaling but this scheme enables a completely new dimension for savings that can be achieved by using application-specific features.

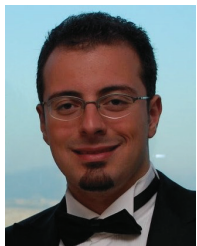
ACKNOWLEDGMENT

The authors would like to thank Texas Instruments for funding and TSMC University Shuttle Program for chip fabrication.

REFERENCES

- [1] D. Markovic, B. Nikolic, and R. Brodersen, "Power and Area Minimization for Multidimensional Signal Processing," *Solid-State Circuits, IEEE Journal of*, vol. 42, no. 4, pp. 922-934, Apr. 2007.
- [2] L. Chang and et al., Stable SRAM Cell Design for the 32nm Node and Beyond, in *Symp. on VLSI Circuits (VLSI) Dig. Tech. Papers*, Jun. 2005, pp. 128129.
- [3] B. Zhai, D. Blaauw, D. Sylvester, and S. Hanson, A Sub-200mV 6T SRAM in 0.13m CMOS, in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2007, pp. 332333.
- [4] 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- V_{DD} and High-Speed Applications, in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2005, pp. 478479.
- [5] B. Calhoun and A. Chandrakasan, A 256-kbit Sub-threshold SRAM in 65nm CMOS, in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2006, pp. 628629.
- [6] I. J. Chang, J. Kim, S. P. Park, and K. Roy, A 32kb 10T Subthreshold SRAM Array with Bit-interleaving and Differential Read Scheme in 90nm CMOS, in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2008, pp. 388389.
- [7] L. Chang, Y. Nakamura, R. K. Montoyo, J. Sawada, A. K. Martin, K. Kinoshita, F. Gebara, K. Agarwal, D. Acharyya, W. Haensch, K. Hosokawa, and D. Janssek, A 5.3GHz 8T-SRAM with Operation Down to 0.41V in 65nm CMOS, in *Symp. on VLSI Circuits (VLSI) Dig. Tech. Papers*, Jun. 2007, pp. 252253.
- [8] S. Ishikura, M. Kurumada, T. Terano, Y. Yamagami, N. Kotani, K. Satomi, K. Nii, M. Yabuuchi, Y. Tsukamoto, S. Ohbayashi, T. Oashi, H. Makino, H. Shi-nohara, and H. Akamatsu, A 45nm 2port 8T-SRAM using hierarchical replica bitline technique with immunity from simultaneous R/Waccess issues, in *Symp. on VLSI Circuits (VLSI) Dig. Tech. Papers*, Jun. 2007, pp. 254255.
- [9] M. Sinangil, N. Verma, and A. Chandrakasan, A 45nm 0.5V 8T column-interleaved SRAM with on-chip reference selection loop for sense-amplifier, in *Solid-State Circuits Conference, 2009. A-SSCC 2009. IEEE Asian*, Nov. 2009, pp. 225-228.
- [10] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, and M. Bohr, A 3-GHz 70Mb SRAM in 65nm CMOS Technology with Integrated Column-Based Dynamic Power Supply, in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2005, pp. 474475.
- [11] M. Yamaoka, N. Maeda, Y. Shinozaki, Y. Shimazaki, K. Nii, S. Shimada, K. Yanagisawa, and T. Kawahara, Low-Power Embedded SRAM Modules with Expanded Margins for Writing, in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2005, pp. 480481.
- [12] Y. Fujimura, O. Hirabayashi, T. Sasaki, A. Suzuki, A. Kawasumi, Y. Takeyama, K. Kushida, G. Fukano, A. Katayama, Y. Niki, and T. Yabe, A configurable SRAM with constant-negative-level write buffer for low-voltage operation with $0.149\mu\text{m}^2$ cell in 32nm high-k metal-gate CMOS, in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2010, pp. 348349.
- [13] K. Takeda, H. Ikeda, Y. Hagihara, M. Nomura, and H. Kobatake, Redefinition of Write Margin for Next-Generation SRAM and Write-Margin Monitoring Circuit, in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2006.
- [14] M. Sinangil, H. Mair, and A. Chandrakasan, A 28nm high-density 6T SRAM with optimized peripheral-assist circuits for operation down to 0.6V, in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2011, pp. 260-262.
- [15] E. Karl, et al., A 4.6GHz 162Mb SRAM Design in 22nm Tri-Gate CMOS Technology with Integrated Active VMIN-Enhancing Assist Circuitry, *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, pp. 230-232, Feb. 2012.
- [16] H. Pilo, et al., A 64Mb SRAM in 32nm High-k metal-gate SOI technology with 0.7V operation enabled by stability, write-ability and read-ability enhancements, *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, pp. 254-256, Feb. 2011.
- [17] H. Fujiwara, et al., Novel Video Memory Reduces 45% of Bitline power Using Majority Logic and Data-Bit Reordering, *IEEE TVLSI*, vol. 16, no. 6, pp.620-627, Jun. 2008.
- [18] I. J. Chang, et al., A Priority-Based 6T/8T Hybrid SRAM Architecture for Aggressive Voltage Scaling in Video Applications, *IEEE TCSTV*, vol. 21, no. 2, pp. 101112, Feb. 2011.
- [19] H. Noguchi, Y. Iguchi, H. Fujiwara, Y. Morita, K. Nii, H. Kawaguchi, M. Yoshimoto, A 10T Non-Precharge Two-Port SRAM for 74% Power Reduction in Video Processing, in *Proc. IEEE Computer Society Annual Symp. VLSI (ISVLSI)*, pp. 107-112, March 2007.
- [20] J. S. Lim, Two-Dimensional Signal and Image Processing, Prentice-Hall, 1989.
- [21] H.-C. Chang, J.-W. Chen, B.-T. Wu, C.-L. Su, J.-S. Wang, J.-I. Guo, "A Dynamic Quality-Adjustable H.264 Video Encoder for Power-Aware Video Applications," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol.19, no.12, pp.1739,1754, Dec. 2009

- [22] Joint Call for Proposals on Video Compression Technology, *ITU-T SG16/Q6, 39th VCEG Meeting: Kyoto*, 17-22 Jan. 2010, Doc. VCEG-AM91.
- [23] R. Rithe, C.-C. Cheng, and A. Chandrakasan, Quad Full-HD transform engine for dual-standard low-power video coding, in *Solid State Circuits Conference (A-SSCC), 2011 IEEE Asian*, Nov. 2011, pp. 401-404.
- [24] M. E. Sinangil, V. Sze, M. Zhou, and A. P. Chandrakasan, Hardware-Aware Motion Estimation Search Algorithm Development for High-Efficiency Video Coding (HEVC) Standard, in *Image Processing (ICIP), 2012 19th IEEE International Conference on*, Sept. 2012, pp. 1529-1532.
- [25] A. Kawasumi, T. Suzuki, S. Moriwaki, S. Miyano, "Energy efficiency degradation caused by random variation in low-voltage SRAM and 26% energy reduction by Bitline Amplitude Limiting (BAL) scheme," in *Solid State Circuits Conference (A-SSCC), 2011 IEEE Asian*, Nov. 2011, pp.165.168.



Mahmut E. Sinangil (S06M12) received the B.Sc. degree in electrical and electronics engineering from Bogazici University, Istanbul, Turkey, in 2006, and the S.M. and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, in 2008 and 2012 respectively.

Since July 2012, he has been a Research Scientist in the Circuits Research Group at NVIDIA. His research interests include low-power and application specific on-chip memories targeted towards graphics

applications.

Dr. Sinangil was the recipient of the Ernst A. Guillemin Thesis Award at MIT for his Masters thesis in 2008, co-recipient of 2008 A-SSCC Outstanding Design Award and recipient of the 2006 Bogazici University Faculty of Engineering Special Student Award.



Anantha P. Chandrakasan (M95SM01-F'04) received the B.S., M.S. and Ph.D. degrees in Electrical Engineering and Computer Sciences from the University of California, Berkeley, in 1989, 1990, and 1994 respectively. Since September 1994, he has been with the Massachusetts Institute of Technology, Cambridge, where he is currently the Joseph F. and Nancy P. Keithley Professor of Electrical Engineering.

He was a co-recipient of several awards including the 1993 IEEE Communications Society's Best Tutorial Paper Award, the IEEE Electron Devices Society's 1997 Paul Rappaport Award for the Best Paper in an EDS publication during 1997, the 1999 DAC Design Contest Award, the 2004 DAC/ISSCC Student Design Contest Award, the 2007 ISSCC Beatrice Winner Award for Editorial Excellence and the ISSCC Jack Kilby Award for Outstanding Student Paper (2007, 2008, 2009). He received the 2009 Semiconductor Industry Association (SIA) University Researcher Award. He is the recipient of the 2013 IEEE Donald O. Pederson Award in Solid-State Circuits.

His research interests include micro-power digital and mixed-signal integrated circuit design, wireless microsensor system design, portable multimedia devices, energy efficient radios and emerging technologies. He is a co-author of *Low Power Digital CMOS Design* (Kluwer Academic Publishers, 1995), *Digital Integrated Circuits* (Pearson Prentice-Hall, 2003, 2nd edition), and *Sub-threshold Design for Ultra-Low Power Systems* (Springer 2006). He is also a co-editor of *Low Power CMOS Design* (IEEE Press, 1998), *Design of High-Performance Microprocessor Circuits* (IEEE Press, 2000), and *Leakage in Nanometer CMOS Technologies* (Springer, 2005).

He has served as a technical program co-chair for the 1997 International Symposium on Low Power Electronics and Design (ISLPED), VLSI Design '98, and the 1998 IEEE Workshop on Signal Processing Systems. He was the Signal Processing Sub-committee Chair for ISSCC 1999-2001, the Program Vice-Chair for ISSCC 2002, the Program Chair for ISSCC 2003, the Technology Directions Sub-committee Chair for ISSCC 2004-2009, and the Conference Chair for ISSCC 2010-2012. He is the Conference Chair for ISSCC 2013. He was an Associate Editor for the IEEE Journal of Solid-State Circuits from 1998 to 2001. He served on SSCS AdCom from 2000 to 2007 and he was the meetings committee chair from 2004 to 2007. He was the Director of the MIT Microsystems Technology Laboratories from 2006 to 2011. Since July 2011, he is the Head of the MIT EECS Department.