18.2 An SRAM Using Output Prediction to Reduce BL-Switching Activity and Statistically-Gated SA for up to 1.9× Reduction in Energy/Access

Mahmut E Sinangil[1], Anantha P Chandrakasan[2]

[1]Nvidia, Bedford, MA, [2]Massachusetts Institute of Technology, Cambridge, MA

Mobile applications such as tablets pack increasingly more processing capability comparable to workstations or laptops but can do little for cooling or extending the battery life in their form factors. SRAMs account for a large fraction of chip area and are critical in this context. Recent work has focused on voltage scaling in SRAMs, which is an effective way of achieving energy efficiency [1,2]. These conventional SRAMs are mostly general-purpose in the sense that they are designed without considering the specific features of the data they will store. However, application-specific features such as statistics of storage data can be exploited and incorporated into the transistor-level design to provide a new dimension towards achieving the next level of energy savings in addition to the savings provided through voltage scaling. The work in [3] is an example where an inversion bit is added for each word to reduce read-bitline (RBL) transitions in an 8T-cell-based design with a single-ended read port. Similarly, the work in [4] stores only the LSBs of each word in 6T SRAMs where occasional bit-errors at low voltages are tolerable for its application. In this work, we focus on video; however, the ideas can be generalized to different applications. In video encoders, pixel processing is performed over large partitions of image frames (e.g., 192×192 pixels), which are stored in on-chip SRAMs and accessed frequently. Image frames generally consist of smooth backgrounds or large objects where the intensity of pixels is spatially correlated. For the video image frame in Fig. 18.2.1, the deviation of each pixel's intensity from its block average for a 16×16 block shows that 76% of pixels lie within 3 LSB of the average. This additional information can be used to design an SRAM where correlation of data is used to reduce bitline activity factor which, for an 8T SRAM in a 65nm low-power CMOS process, accounts for ~50% of total energy consumption during read accesses at 0.6V. In this work, we present a prediction-based reduced-bitline-switching-activity (PB-RBSA) scheme along with a hierarchical sensing network with statistical sense-amplifier gating to exploit the correlation of storage data. Reduction of switching activity on the bitlines and in the sensing network of the memory provide up to 1.9× reduction in energy/access.

PB-RBSA scheme and the bit-cell are shown in Fig. 18.2.2. Differential read buffers are controlled with RWL0 and RWL1 signals and the footer of these buffers are connected to the bit-wise prediction (pred) or its complement (predB), both of which are routed in column direction parallel to BL/BLB and RBL0/RBL1 pairs (Fig. 18.2.2). Second read buffer introduces an area overhead of 20% over the conventional 8T cell. During a read operation, RBL0 and RBL1 are pre-charged to V_{DD} and pred/predB settle to their final values prior to the assertion of RWL0 and RWL1. pred and predB are actively driven by drivers in column circuit at all times. When the prediction is correct, both RBL0 and RBL1 stay high, saving energy of pre-charging large RBL0/RBL1 capacitances after the read access (Fig. 18.2.3). In the case of an incorrect prediction, RBL0 or RBL1 is discharged by the bit-cell due to its differential nature. The activity factor of the RBLs can be greatly reduced if the prediction is correct most of the time. Note that an incorrect prediction does not lead to an erroneous output but results in the discharging of one of the RBLs. Figure 18.2.3 also shows the prediction-generation circuit where 64b (8×8b/pixel) memory outputs are accumulated to create a pixel average over 2^N cycles. This predictor is used for every 8 bits of an SRAM word corresponding to individual pixels. The value of N can be adjusted to minimize energy/access based on the following trade-off. With smaller N, the predictor is updated more frequently and the correct prediction percentage increases resulting in a reduction of the switching activity of RBL0/RBL1. However, with smaller N, the switching activity of pred/predB signals increase. In our measurements, we found N = 4 to 5 provides a good balance for most video frames minimizing the energy consumption. Also note that RWL0 and RWL1 signals are routed separately which provides an option to use PB-RBSA SRAMs in a high-throughput (HT) mode where pred and predB are driven to '0', RWL0 and RWL1 are asserted on different addresses and SRAMs can support 2R1W access.

With PB-RBSA scheme, the probability of RBLs being high at the end of a read access is higher provided that the correct prediction percentage is >50%. Hence, statistically, a sensing network consuming less energy when sensing a high RBL can provide savings. Figure 18.2.4 shows the sensing network consisting of a larger size sense-amp (M-SA) and a smaller one (S-SA). Only the RBL on the side of a low pred or predB can be discharged by the bit-cell so pred/predB is used to select the correct RBL to do sensing. S-SA is activated with the assertion of an earlier signal (snsEarly) and makes an initial evaluation of the RBL. If the output of the S-SA is high (i.e., correct prediction), M-SA is gated and does not turn on with the active-low snsB signal. However, because S-SA is sized smaller than M-SA, its offset distribution is wider and can have erroneous outputs when asserted early [5]. By skewing the offset distribution and by setting its reference voltage (REFH) to a larger voltage than M-SA's reference (REF), it can be ensured that erroneous outputs only occur when outputting a '0' when RBL is actually high. This causes the M-SA to be activated unnecessarily but RBL can be resolved correctly at the end. The offset distribution of S-SA in Fig. 18.2.4 shows that erroneous outputs occur for <5% of total S-SAs. Figure 18.2.5 shows the energy savings with statistically gated sense-amplifier and having only M-SA. For correct prediction percentage larger than 40%, energy savings are possible. At 100% correct prediction, up to 3× energy savings in the sensing network can be achieved.

A test-chip is fabricated in a 65nm low-power CMOS process with two blocks for each PB-RBSA SRAMs and conventional 8T SRAMs (Fig. 18.2.7). Pixel data from actual test video sequences as well as artificial patterns are provided to both SRAMs to make real-time energy/access comparisons. Read and write access patterns are based on the motion search algorithm given in [6]. Figure 18.2.5 shows normalized energy/access for PB-RBSA SRAM with varying correct prediction percentage and for 8T SRAMs. Because of the data-dependent nature of RBL activity in 8T SRAMs, input data is ensured to have 50% '0's and 50% '1's. With 100% correct prediction, PB-RBSA SRAMs can provide up to 1.75× reduction in energy/access.

Figure 18.2.6 shows distribution of average measured energy/access savings with PB-RBSA SRAMs with respect to the 8T SRAMs for 1100 video frames from 11 different video sequences with different contents and properties. Energy savings as large as 1.9× are measured. With even higher resolutions in the future (e.g., 4K×2K or 8K×4K), savings with PB-RBSA SRAMs can be larger. Finally, Fig. 18.2.6 shows a summary table for the 65nm test-chip and operating frequency measurement results. Low-voltage operation down to 0.52V is achieved. The savings with PB-RBSA scheme are built on top of the savings from voltage scaling providing a new dimension for circuit designers to achieve maximized energy efficiency.

Acknowledgements:
This project is funded by Texas Instruments. The authors thank the TSMC University Shuttle Program for chip fabrication.

References:

[1] E. Karl, et al., "A 4.6GHz 162Mb SRAM Design in 22nm Tri-Gate CMOS Technology with Integrated Active VMIN-Enhancing Assist Circuitry," *2012 IEEE ISSCC*, pp. 230-232, Feb. 2012.

[2] 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," *2011 IEEE ISSCC*, pp. 254-256, Feb. 2011.

[3] 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.

[4] I. J. Chang, et al., "A Priority-Based 6T/8T Hybrid SRAM Architecture for Aggressive Voltage Scaling in Video Applications*," IEEE TCSVT,* vol. 21, no. 2, pp. 101–112, Feb. 2011.

[5] B. Wicht, et al., "Yield and Speed Optimization of a Latch-Type Voltage Sense Amplifier," *IEEE JSSC,* vol. 39, no. 7, pp. 1148-1158, Jul. 2004.

[6] M. E. Sinangil, et al., "Hardware-Aware Motion Estimation Search Algorithm Development for High-Efficiency Video Coding (HEVC) Standard*," 2012 IEEE ICIP*, pp. 1529-1532, Sep. 2012.

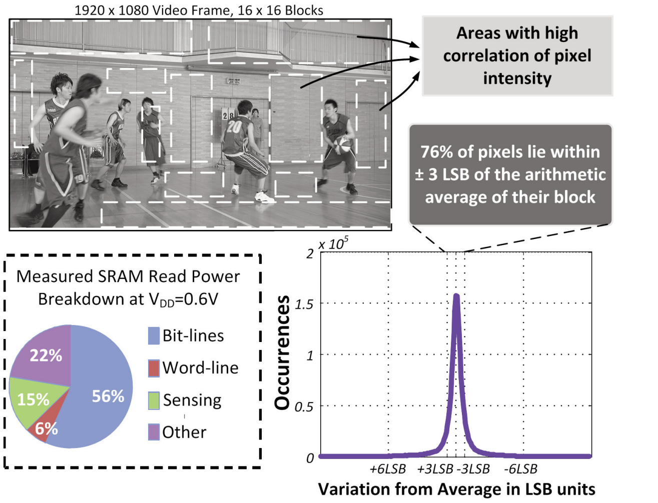


Figure 18.2.1: For pixels belonging to the same object or background, intensities are highly correlated. This additional information can be used to reduce BL switching activity.

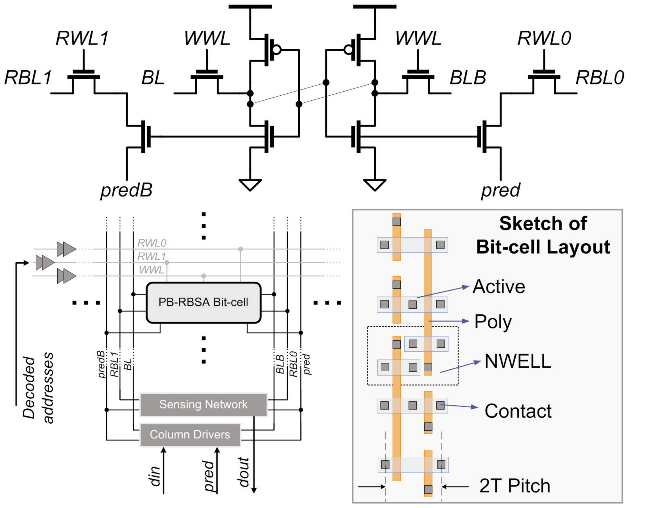


Figure 18.2.2: Bit-cell design and array organization for PB-RBSA scheme. BLs, RBLs and pred/predB are routed in MET2, WWL in MET3 and RWLs in MET5. Layout sketch does not reflect transistor sizing.

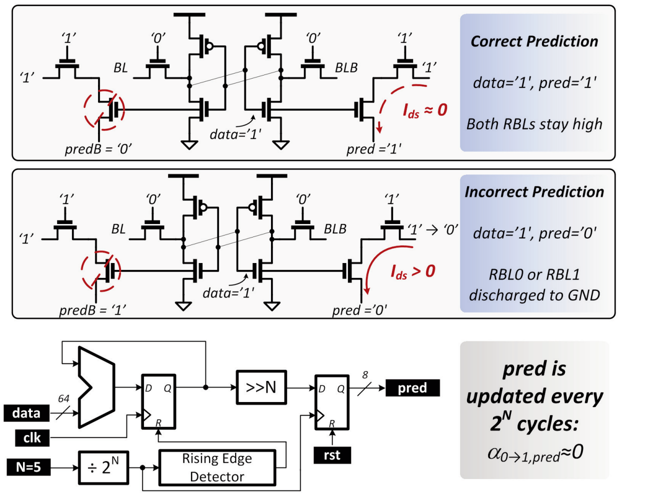
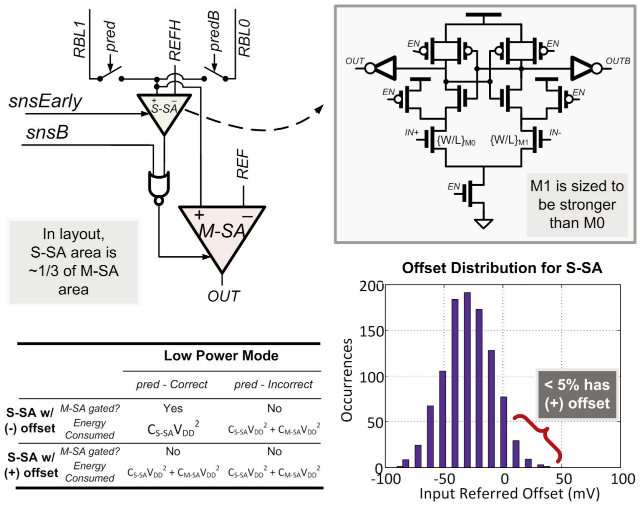
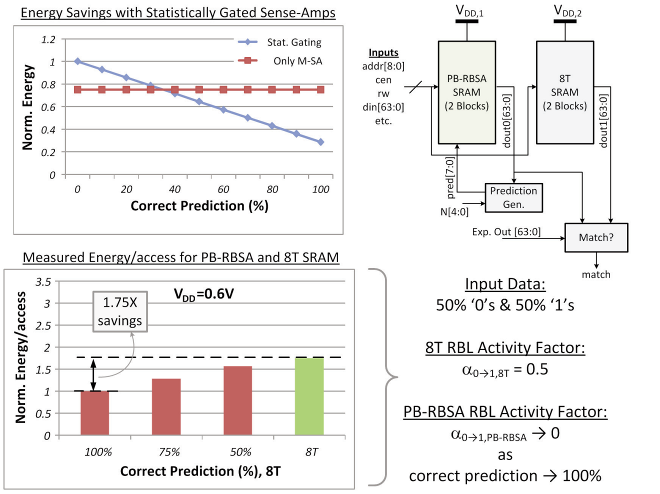


Figure 18.2.3: Correct prediction prevents RBLs from being discharged by the bit-cell. Prediction generation can be implemented with an average calculation based on previous outputs.



		Low Power Mode	
		pred - Correct	*pred - Incorrect*
S-SA w/ (-) offset	*M-SA gated?*	Yes	No
	Energy Consumed	$C_{S-SA}V_{DD}^2$	$C_{S-SA}V_{DD}^2 + C_{M-SA}V_{DD}^2$
S-SA w/ (+) offset	*M-SA gated?*	No	No
	Energy Consumed	$C_{S-SA}V_{DD}^2 + C_{M-SA}V_{DD}^2$	$C_{S-SA}V_{DD}^2 + C_{M-SA}V_{DD}^2$

Figure 18.2.4: Hierarchical sensing with S-SA. S-SA is sized to have a systematic negative offset. Second M-SA (necessary for 2R1W) is not shown in the figure.

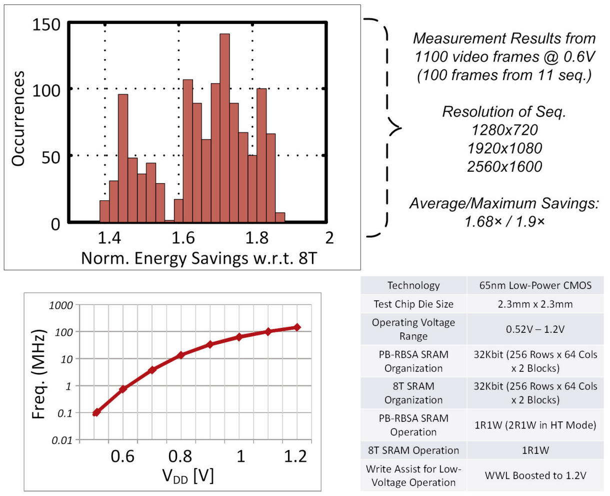


Input Data:
50% '0's & 50% '1's

8T RBL Activity Factor:
$\alpha_{0\to1,8T} = 0.5$

PB-RBSA RBL Activity Factor:
$\alpha_{0\to1,PB\text{-}RBSA} \to 0$
as
correct prediction → 100%

Figure 18.2.5: Gated sense-amps provide energy savings when correct prediction percentage is >40%. Test-chip features PB-RBSA and 8T SRAMs to make real time energy/access comparison.



Measurement Results from 1100 video frames @ 0.6V (100 frames from 11 seq.)

Resolution of Seq.
1280x720
1920x1080
2560x1600

Average/Maximum Savings:
1.68× / 1.9×

Technology	65nm Low-Power CMOS
Test Chip Die Size	2.3mm x 2.3mm
Operating Voltage Range	0.52V – 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)
PB-RBSA SRAM Operation	1R1W (2R1W in HT Mode)
8T SRAM Operation	1R1W
Write Assist for Low-Voltage Operation	WWL Boosted to 1.2V

Figure 18.2.6: Measured energy savings with PB-RBSA SRAMs with respect to the 8T SRAMs. Across 1100 different video frames, savings up to 1.9× are reported with the PB-RBSA scheme.

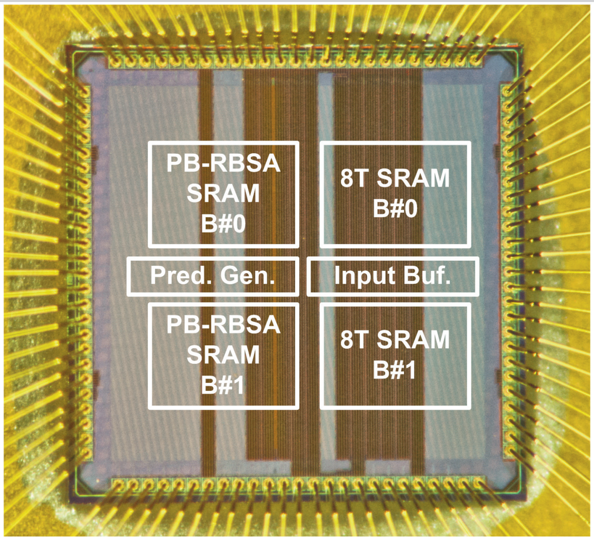


Figure 18.2.7: Die micrograph of the test-chip fabricated in 65nm low-power CMOS process. Die size is 2.3×2.3mm^2.