Energy-Efficient Reconfigurable SRAM: Reducing Read Power Through Data Statistics

Chuhong Duan, *Student Member, IEEE*, Andreas J. Gotterba, *Member, IEEE*, Mahmut E. Sinangil, *Member, IEEE*, and Anantha P. Chandrakasan, *Fellow, IEEE*

***Abstract*— This paper introduces a framework for designing data-dependent SRAMs taking advantage of statistical dependencies present in the binary values processed and stored in the intermediary stages of various algorithms. To demonstrate the framework, a reconfigurable conditional precharge (CP) SRAM is designed in a 28-nm fully-depleted silicon-on-insulator CMOS process. To reduce read power consumption, the SRAM reconfigures its prediction scheme for each column as the data statistics evolve. A 10T bit cell, a prediction-based CP circuit, and a compact column circuit implemented in a 16-kbit SRAM test chip demonstrate the power savings of 63%, 50%, and up to 69% for the applications sparse fast Fourier transform, object detection, and motion estimation, respectively, as compared with similar memories with naive prediction. Analysis tools for optimal prediction selection for the presented class of low-power memories are also provided.**

***Index Terms*— Correlation, low power, reconfigurable, SRAM, statistics.**

I. Introduction

Across a broad spectrum of emerging and established applications, embedded systems provide economically and technically sound solutions, which, in turn, drive need for their high density, high bandwidth, and low-power operation. Notable examples of such application spaces include Internet of Everything/wearables and mobile platforms. In such contexts, the efficiency of data storage is critical in that data traffic, workload, and capacity continues to grow. Among possible memory solutions, SRAM is the IC system bottleneck as it consumes increasingly disproportionate amounts of die area and total power. Read power constitutes the majority of dynamic power consumed by an SRAM in settings where data accesses outnumber data stores [1]. During read operations, over half of the total dynamic power dissipates during the switching of highly capacitive bitlines. This power, denoted as P, is specified according to

$$P = \alpha_{0\to 1} C_{\mathrm{BL}} V_{\mathrm{dd}} (\Delta V) f \tag{1}$$

Manuscript received February 15, 2017; revised May 10, 2017; accepted July 13, 2017. Date of publication August 11, 2017; date of current version September 21, 2017. This work was supported in part by DARPA and in part by NSF. This paper was approved by Guest Editor Jaeha Kim. *(Corresponding author: Chuhong Duan.)*

C. Duan was with the Massachusetts Institute of Technology, Cambridge, MA 02139 USA. She is now with Texas Instruments, Dallas, TX 75243 USA (e-mail: chuduan@alum.mit.edu).

A. J. Gotterba is with NVIDIA, Santa Clara, CA 95050 USA.

M. E. Sinangil was with NVIDIA, Santa Clara, CA 95050 USA. He is now with TSMC North America, San Jose, CA 95134 USA.

A. P. Chandrakasan is with the Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2017.2731814

where $\alpha_{0\to 1}$ is the so-called switching activity factor, C_{BL} is the bitline capacitance, V_{dd} is the supply voltage, ΔV is the voltage differential, and f is the operating frequency.

Scaling down V_{dd} reduces the dynamic power in (1) while introducing a tradeoff between energy efficiency and delay [2]. However, the robust operation of an SRAM demands higher supply voltages than surrounding digital blocks ensuring reliable performance under worst case process, voltage, and temperature conditions. This effect compounds with device scaling toward meeting high-density requirements. To address this, low-power SRAM systems typically use assist methods to incrementally lower V_{dd} [3]. However, aggressively scaled technologies tend to have high variability limiting the effectiveness of such assist methods, especially at low voltages. Straightforward techniques for reducing the contributions of C_{BL} and ΔV include architecting hierarchical bitlines [4] and using offset-compensated sense amplifiers [5], [6]. These methods incur large area overheads, in practice, thus are of limited use. Consequently, recent designs focus on reducing switching activity by leveraging statistical patterns present in SRAM data.

From system-level architectures to circuit topologies, novel designs often benefit from designers' in-depth understanding of the end-use application: properties describing the environment the system will operate in, models characterizing typical data sequences to be processed, and so on. On-chip memories, modules that deal with large amounts of data bandwidth, could benefit significantly by adapting to the statistics of the data being accessed or stored. The designs proposed in [7] and [8] recognized that 8T SRAMs are intrinsically data-dependent and used bit inverting and column-based data encoding, respectively, to convert the majority of bits in the array to either 1 or 0, which saved power by reducing bitline switching during read operations. These approaches only perform well when either the data are heavily biased to one bit value, i.e., mostly 1s or 0s, or when consecutive rows share similar values thus limiting the randomness of the access patterns. Furthermore, such bit encoding schemes introduce the risk of error propagation.

As opposed to the above-mentioned techniques relying upon data manipulation, novel architectural and bit-cell designs have also been used to leverage data correlation. For example, Noguchi *et al.* [9] eliminated the precharge phase in a 10T array using statically driven read bitlines. Consecutively accessing the same value results in reduced bitline switching yet the penalties associated with latency and area remain. In the context of real-time video applications, which are well

known for their heavily correlated data patterns, [10] introduced a priority-based 6T/8T hybrid SRAM, which allows for video quality to be traded for power reduction. The key insight to this approach involves recognizing the different quality indications of MSBs and LSBs of luminance pixels. A similar split-data-aware 8T/10T SRAM design was discussed in [11]. A prediction-based reduced bitline switching activity (PB-RBSA) 10T SRAM relying on high spatial and temporal data correlations was introduced in [1]. This method reduces read power using bitwise read-data predictions generated by a separate module that computes moving averages. A drawback of this technique is the significant power overhead incurred when changing predictions due to the highly capacitive prediction lines, meaning it only performs well with repeated data patterns across many consecutive cycles. Building from this key principle, [12] extended the usage of data prediction to the global bitlines of a 6T SRAM. From a statistical perspective, the previously mentioned designs target data sequences with very specific statistical properties and incur large penalties for data sets with heterogeneous features.

In this paper, the design challenges outlined earlier are addressed through a memory framework wherein reconfigurable, prediction-based SRAMs adapt their operation to the statistics of data being processed for improved energy efficiency. The focus on adapting the SRAM operation to data statistics generalizes the notion of application-specific design, since some applications are sufficiently characterized by a single statistical measure of their data while others require multifaceted measures. The reconfigurable memory framework proposed in this paper is discussed in Section II. Then, in Section III, three example applications are presented along with requisite tools for characterizing data statistics. Next, in Section IV, a prediction-based conditional precharge (CP) 10T array is presented, which is then used as a building block in the design of a columnwise reconfigurable SRAM that is the central focus of Section V. Finally, in Section VI, measurement results of a 16-kbit SRAM test chip fabricated in a 28-nm fully-depleted silicon-on-insulator (FD-SOI) CMOS process are provided.

II. Reconfigurable Memory Framework

The proposed reconfigurable memory framework is shown in Fig. 1. Application-specific information, such as data statistics and addressing sequences, is extracted in the statistical analysis module, which can be implemented in software or hardware, and is then sent as metadata to the prediction module to dynamically configure the memory array. The prediction module, embedded in the memory system, provides functionality for multiple prediction modes supporting low-power data access in a variety of applications. The framework is reconfigurable as the prediction mode and associated parameters can be altered as data statistics evolve. In this section, we develop statistical tools for the analysis module and discuss their consequences on power consumption.

A. Low-Complexity Statistical Measures

In the sequel, we refer to the binary values stored in the column direction of an SRAM as a *column sequence* (CS)

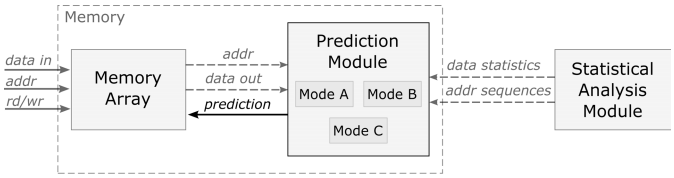


Fig. 1. Proposed memory framework consisting of a statistical analysis module, prediction module, and memory array. Configuration of the memory array is realized via the prediction module with inputs from the analysis module.

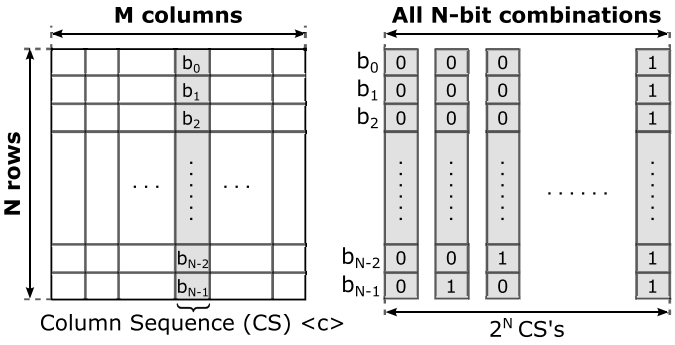


Fig. 2. Two SRAM arrays illustrating the definition of a CS and an example data set comprising 2^N N-bit CSs.

where the order of bit values reflects the read access pattern. To illustrate this, an example SRAM organized into N rows and M columns is presented in Fig. 2. Referring to Fig. 2, the highlighted N bit-cell values in column <c> constitute a CS. The array on the right depicts the enumeration of all possible N-bit CSs.

The *switching activity factor*, denoted by $\alpha_{0\to1}$ in (1), indicates the toggle density or ratio of $0 \to 1$ bit-level transitions to the maximum possible number of transitions within a given CS and can be computed as

$$\alpha_{0\to1} \triangleq \frac{\text{number of } 0 \to 1 \text{ transitions}}{\text{column sequence length} - 1}. \tag{2}$$

By definition, $\alpha_{0\to1}$ is restricted to the interval $[0, 0.5]$. The bias of a CS, denoted by ρ, indicates the relative frequency of the most frequently occurring bit and is computed as

$$\rho \triangleq \frac{\text{number of occurrences of the most frequent bit}}{\text{column sequence length}}. \tag{3}$$

By definition, ρ is restricted to the interval $[0.5, 1]$.

There are generally many statistical features that can be useful in characterizing the predictability of a given data set or CS. These include measures of statistical dispersion, such as standard deviations, and measures of distribution shape, such as skewness or kurtosis, or could be derived from more sophisticated stochastic models, such as Markov chains. In real-time processing settings, which are encompassed by the present framework, low-complexity measures are preferred so as to incur minimal power and area overhead. The test chip discussed in Section VI utilizes these low-complexity statistical measures computed by software so as not to limit our testing capability to a subset of measures implemented in hardware. Exploration of more advanced statistics is the subject of future work.

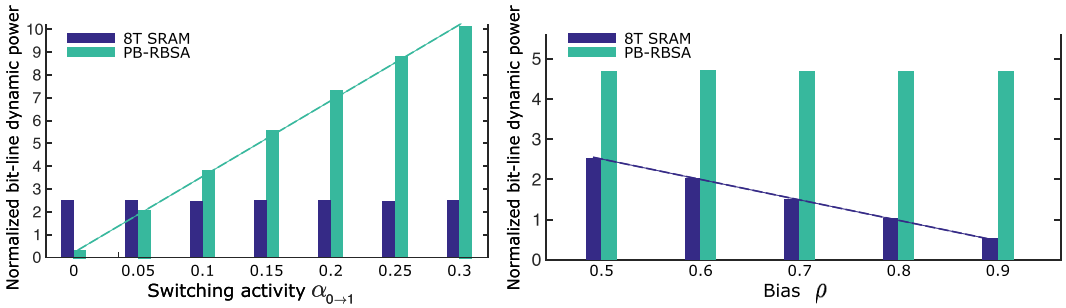


Fig. 3. Graphs portraying the power versus statistics for two main types of data-dependent memory models. The read bitline power consumption of 8T SRAM and PB-RBSA SRAM scales linearly with ρ and $\alpha_{0\to1}$, respectively. In each graph, the statistical measure on the abscissa is varied while the other measure is held constant.

B. Relationships Between $\alpha_{0\to1}$ and ρ and Read Power Consumption

The read power consumed by the existing memory models, such as 8T and PB-RBSA SRAMs, varies monotonically with the measures $\alpha_{0\to1}$ and ρ. This observation is emphasized by the plots in Fig. 3. In particular, Fig. 3 shows the normalized bitline power of the two data-dependent memories when accessing CSs with either varying $\alpha_{0\to1}$ or ρ. Observe that the read power consumption of the 8T SRAM scales linearly with ρ while the PB-RBSA SRAM's power consumption is an increasing function of $\alpha_{0\to1}$.

It is important to note that the statistics $\alpha_{0\to1}$ and ρ chosen in this paper are not independent of one another. As an extreme example, a CS with $\alpha_{0\to1} = 0$, such as the all-zero CS, unambiguously implies $\rho = 1$. As will be observed later, the complete characterization of CSs in a number of applications requires the use of both measures in deciding on optimal prediction modes. This observation motivates a need to combine multiple memory models to better leverage the discussed relationships and our access to both measures.

III. Dual-Mode Prediction Scheme

Data statistics can vary widely from application to application. Moreover, the statistics of individual CSs in a data set or SRAM array can also vary significantly from bit position to bit position, even for the same application. Motivated by this point, the SRAMs discussed in this paper allow prediction modes to be configured on a per-CS basis. In this section, we address pertinent questions involving the selection of hardware appropriate prediction modes and identify the values of the statistical measures under which those modes are optimal from a power consumption perspective.

To illustrate the above-mentioned point, we use data obtained from three applications as running examples throughout. The applications are: 1) the sparse fast Fourier transform (sFFT); 2) a support vector machine-based object detection (OD) system; and 3) a motion estimation (ME) engine for video coding. For these applications, data retrievals outnumber data stores. Previous hardware implementations in [13]–[15] suggest that read power is disproportionately large compared with total power consumption, which implies opportunity in improving overall system energy efficiency with low read power SRAMs.

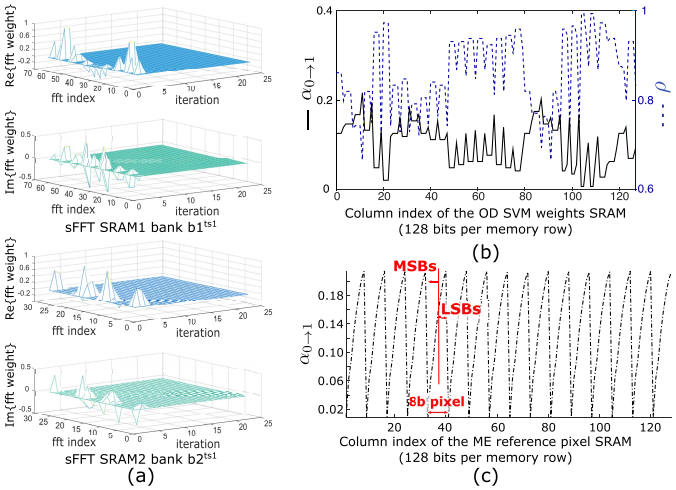


Fig. 4. Analysis showing the statistical features of the three example applications used throughout this paper. (a) sFFT. (b) OD. (c) ME. The data for OD and ME are organized into 128-bit-wide SRAMs, each containing 128 CSs.

The statistics associated with on-chip SRAMs for the three applications are portrayed in Fig. 4. Referring to Fig. 4(a), a typical progression of the data in two FFT coefficient SRAMs over the course of the sFFT's collision resolution procedure is depicted. The inherent sparsity of the sFFT manifests itself in that both SRAMs contain mostly 0s after two to three iterations, suggesting that low-power read operations for data with high bias ρ would be of merit. Referring to Fig. 4(b) and (c), the data processed by OD and ME systems possess small $\alpha_{0\to1}$ values, which can be understood from the well-known fact that frames in video coding and image processing exhibit large amounts of spatial correlation, since neighboring pixels typically have similar significant-bit values. The ME data in Fig. 4(c) specifically highlight the difference between $\alpha_{0\to1}$ values in MSB CSs and LSB CSs, motivating a need for their separate treatment. To elaborate on this observation, the distribution of $(\alpha_{0\to1}, \rho)$-pairs computed along each CS for each application is shown in Fig. 5 independently. The ME distribution, in particular, illustrates a bimodal outcome with MSBs having small $\alpha_{0\to1}$ and large ρ and LSBs having the opposite.

Responding to the opportunities afforded by the above-mentioned analysis, we define a columnwise dual-mode

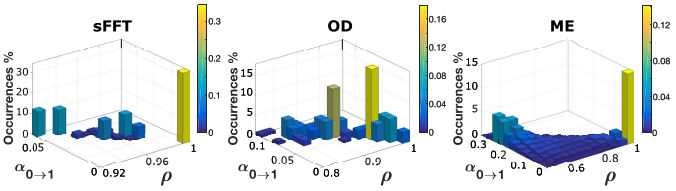


Fig. 5. Histograms depicting the distribution of $(\alpha_{0\to1}, \rho)$-pairs for 256-bit CSs from sFFT, OD, and ME SRAMs.

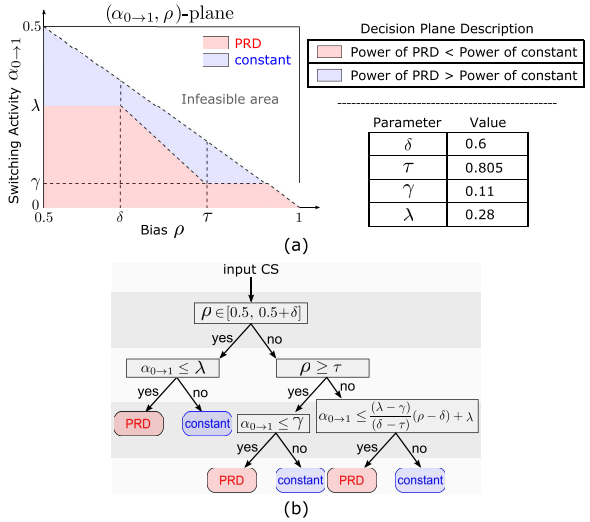


Fig. 6. (a) Optimal prediction modes for different $(\alpha_{0\to1}, \rho)$-pairs obtained by simulating read power consumption for each mode for all possible distinct 16-bit CSs. Decision boundaries are extrapolated from test-chip measurement results. (b) Example configuration of the decision tree.

prediction scheme where, for each CS, either a constant or previous-read-data (PRD) prediction mode is selected depending on the CS's $(\alpha_{0\to1}, \rho)$-pair. For operation in constant mode, the prediction is held constant at 0 or 1 depending on which constitutes the most frequent bit, i.e., the bit used to compute the numerator of (3). This is equivalent to equipping an 8T SRAM with majority logic and is asymptotically equivalent to arithmetic averaging over increasingly large window sizes. Constant mode is well suited for CSs with high bias. For operation in PRD mode, the current data value is used as the prediction for the next data value. This mode is well suited for CSs with low switching activities irrespective of their bias.

A summary of the optimal prediction mode as parameterized by $(\alpha_{0\to1}, \rho)$ values is provided in Fig. 6. The selected mode provides the lowest power consumption for any data sequence characterized by the corresponding statistics assuming CS lengths of 16 or more bits. In order to identify the decision boundary parameters listed in Fig. 6, the power consumed by the proposed SRAM operated in PRD and constant modes was measured for a data set containing all possible distinct 16-bit CSs, which corresponds to 45 distinct $(\alpha_{0\to1}, \rho)$-pairs, and then extrapolated to the entire $(\alpha_{0\to1}, \rho)$-plane [16]. As indicated by the depicted plane, the joint consideration of $\alpha_{0\to1}$ and ρ is required to identify the optimal mode for each CS. A decision tree is also provided in Fig. 6(b). Equipped with these tools, selecting the optimal mode for any data sequence, including entirely random CSs, is straightforward.

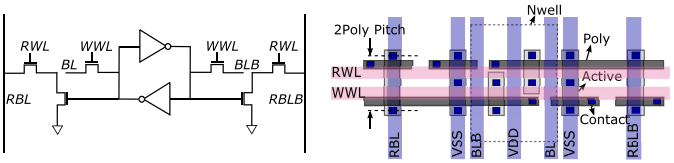


Fig. 7. Proposed 10T bit-cell design (left) and the custom layout (right, not to scale).

IV. Conditional Precharge 10T Array

This section discusses the design of a prediction-based SRAM array comprising a 10T bit cell and the associated CP circuitry. The SRAM is designed, such that the dynamic power incurred on the bitlines during a read operation is significantly reduced when the data value being accessed is correctly predicted. The predicted value can be obtained using either of the modes discussed in Section III.

A. Bit-Cell Design

The schematic provided in Fig. 7 on the left shows the proposed bit-cell topology and the diagram on the right indicates the general structure of the custom layout. Referring to the schematic, the 10T memory cell subsumes a standard 6T bit cell as well as a second differential read port. Read and write operations are decoupled from one another, since the word lines (RWL and WWL) and bitline pairs (RBL/RBLB and BL/BLB) separately handle read and write control and access, respectively. As indicated in Fig. 7, the separate access ports imply that the bit cell is exempt from read stability issues as is the case with conventional 8T memory cells. Furthermore, the bit cell can also be configured to construct two-port SRAMs wherein simultaneous read and write bitline accesses are permitted. The design as presented avoids the prediction line power overhead observed in the related 10T bit-cell design in [1] by grounding the sources of the two bottom read port devices. Alternatively, prediction signals are passed to the array via precharge circuitry as will be discussed in Section IV-B.

B. Conditional Precharge Circuit

The CP scheme is implemented by feeding prediction information to the precharge circuitry. As shown in Fig. 8(a), a pair of PMOS switches `P1` and `P2` are added in series to the conventional precharge devices `P3` and `P4`. Depending on the value being predicted, only the read bitline on the side predicted to store a 0 will be precharged. For instance, during the precharge phase, if the predicted value in the current cycle is 1, i.e., $\texttt{pred} = 1$ and $\texttt{predB} = 0$, RBLB is precharged to V_{dd} when `prechB` asserts and RBL is kept low through the cross-coupled NMOS devices `N1` and `N2`.

The two possible scenarios of read bitline activity are shown in Fig. 8(b). They specifically correspond to: 1) a correct prediction when the prediction matches the storage node `Q` and 2) an incorrect prediction when the prediction equals the complement of node `Q`. Suppose the bit-cell nodes take values $\texttt{Q} = 0$ and $\texttt{QB} = 1$. For a correct prediction $\texttt{pred} = \texttt{Q} = 0$, in the precharge phase, RBL is charged to V_{dd} and RBLB is

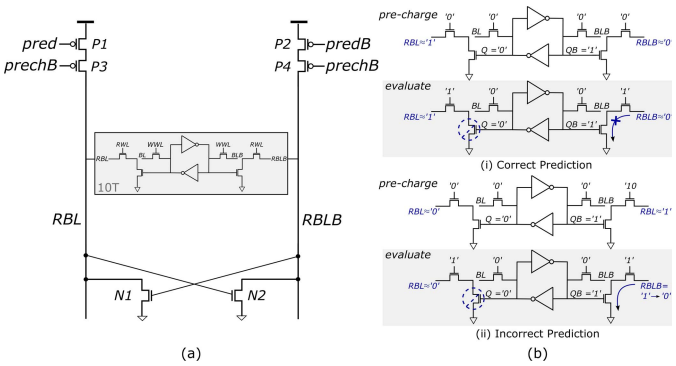


Fig. 8. (a) Prediction-based CP circuit and (b) read bitline activity during evaluation for (i) correct and (ii) incorrect predictions.

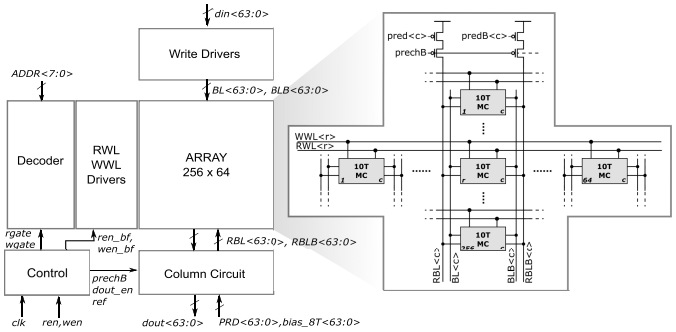


Fig. 9. Architecture of the proposed SRAM.

held low. During evaluation, RBL stays high for the next read cycle, since there is no conducting path through the pull-down NMOS devices. On the other hand, when the prediction is incorrect, i.e., `pred` = `QB` = 1, RBLB is charged high in the precharge phase, and subsequently discharges to ground during evaluation. Regarding data resolution, `Q` = `pred` if either read bitline remains high and `Q` = `predB` if both read bitlines are low at the cycle's end. As intended, a correct prediction led to a reduction in precharge activity and, therefore, decreased the total energy spent.

V. Columnwise Reconfigurable SRAM

In this section, we present a columnwise reconfigurable SRAM, which extends the CP array techniques discussed in Section IV by enabling the dual-mode prediction scheme from Section III. This is accomplished by equipping the array with a compact, reconfigurable column circuit.

A. Proposed SRAM Architecture

The architecture of the proposed reconfigurable CP SRAM is provided in Fig. 9 and is compatible with the chip implementation discussed in Section VI. The array contains 16 kbit of the 10T cells in Fig. 7 and is organized into 256 rows and 64 columns. Each column shares a CP circuit and two pairs of bitlines BL/BLB and RBL/RBLB for write and read accesses, respectively. For simplicity, write operations employ 64 write drivers to statically drive the bitlines BL/BLB to the desired bit values of input `din`. For read operation, a column circuit resolves the read bitlines RBL/RBLB values as well as generates predictions for the CP circuit. Extensions incorporating small-signal sensing, assist methods, or column

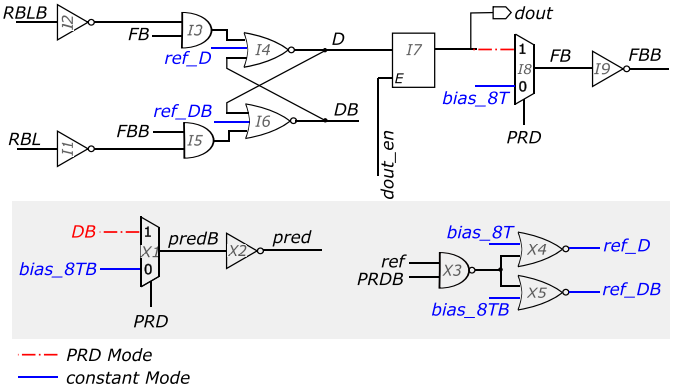


Fig. 10. Column circuit design resolving data output, generating data predictions, and allowing for prediction mode selection. The signal paths for PRD mode (dashed lines) and constant mode (solid lines) are highlighted for emphasis.

interleaving are straightforward to implement, and are outside the scope of this paper.

B. Reconfigurable Column Circuit

The column circuit portion of the reconfigurable CP SRAM is shown in Fig. 10. By adding a trivial number of basic logic gates, e.g., multiplexers and inverters, two prediction paths are introduced at the circuit level. The signal paths for PRD and constant prediction modes are, respectively, highlighted with dashed and solid lines. The column circuit realizes three basic functions in a compact manner, namely, data resolution, prediction generation, and mode selection. Each of these functions is described next.

Data Resolution: The inverters `I1` and `I2` are designed with skewed sizing to speed up bitline sensing. An AND-OR-invert logic-based set-reset (SR) latch takes the outputs of `I1` and `I2` along with the current prediction and prediction-inverse `FB` and `FBB` as inputs to resolve the read value. If one of the bitlines stays high, meaning the prediction is correct, the SR latch holds the current values of `D` and `DB`; otherwise, the SR latch inverts the values of `D` and `DB`, reflecting that the read value is complementary to the prediction. Once the latch outputs settle and one of the bitlines precharges high according to the next prediction, an enable signal `dout_en` asserts and a second latch, `I7`, latches `D` as the read output `dout`. Clock-to-output delay is traded here for improved area and energy efficiency.

Prediction Generation: Depending on the selected prediction mode, the column circuit passes either the last cycle's output values `D`/`DB` (PRD mode) or preset values `bias_8T`/`bias_8TB` (constant mode) to the control signals `pred`/`predB`, which then drive the gates of `P1` and `P2` in the CP circuit in Fig. 8(a). Additionally, feedback signal `FB` takes the next cycle's predicted value, i.e., the current output `dout` for PRD mode and the bias `bias_8T` for constant mode. An additional pair of signals `ref_D`/`ref_DB` resets the SR latch at the read cycle's end in constant mode, so that `D` and `DB` reflect the preset prediction by the beginning of the following evaluation phase.

Mode Selection: The control signal `PRD` selects either constant mode (`PRD` = 0) or PRD mode (`PRD` = 1) using the MUXes `I8` and `X1`. The bias `bias_8T` determines whether

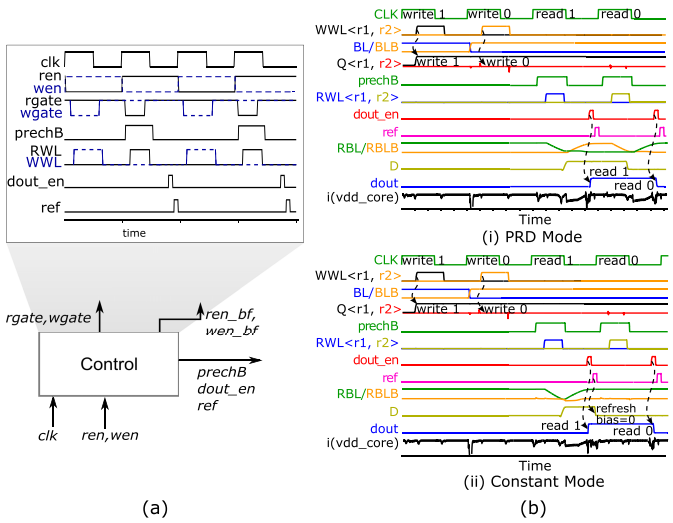


Fig. 11. (a) Timing control waveforms with write control signals highlighted with dashed lines. (b) Simulated waveforms of the reconfigurable CP SRAM operating in (i) PRD mode and (ii) constant mode for the operation sequence: write 1 to address `r1`, write 0 to address `r2`, read 1 from address `r1`, and read 0 from address `r2`. The initial prediction is set to be 0, and the constant mode bias is set to be 0.

to constantly precharge RBL (`bias_8T` $= 0$) or RBLB (`bias_8T` $= 1$) when operating in constant mode.

C. Fixed-Width Control

The off-state leakage current is required to be small compared with the total on-state current for the CP scheme to function correctly. This is generally true for regular-VT devices in technologies, such as FD-SOI CMOS [17]. Control signals are designed to have fixed-width pulses to further ensure the correct operation of the CP scheme across clock frequency domains. The primary timing control signals are shown in Fig. 11(a) for an access sequence of write-then-read. Specifically, in the read cycle, `prechB` has a fixed-width positive pulse meaning bitline evaluations happen in fixed amounts of time, thereby avoiding scenarios where leakage current erroneously discharges the precharged bitline over time, even when the prediction is correct. It is ensured that `dout_en` and `ref` have sufficient pulse widths and happen before the next rising edge of the clock, so that read output is correctly latched, and in constant mode, the SR latch's state is reset before starting the next read cycle. Standard delay elements are used to generate these timing signals for which the maximum operating frequency is determined by the read bitlines precharge and discharge delays.

Simulated waveforms displaying the reconfigurable CP SRAM operating in: 1) PRD and 2) constant modes are presented in Fig. 11(b). The operating sequence corresponds to: 1) writing 1 to address `r1`; 2) writing 0 to address `r2`; 3) reading 1 from address `r1`; and 4) reading 0 from address `r2`. Addresses `r1` and `r2` identify two memory cells in the same column but from different rows, and therefore, they share a column circuit. Referring to Fig. 11(b)-(i), in the first two cycles, values 1 and 0 are successfully written to the bit cells at `r1` and `r2`. In the third cycle, RBL precharges high, since the initial prediction is set to 0 in PRD mode. In the evaluation phase, RBL discharges to ground reading 1 at `r1`. As `dout_en` asserts, `dout` latches the correct output value of 1. In the fourth cycle, RBLB precharges high, since the new prediction is 1. It discharges to ground from reading 0 at `r2`, and, consequently, `dout` transitions to 0. Incorrect predictions in both read cycles result in the two observable current surges in `i(vdd_core)`. Similarly in Fig. 11(b)-(ii), where the array is operating in constant mode with `bias_8T` set to 0, RBL precharges high in both read cycles and only discharges in the third cycle when reading 1 from `r1`. By making a correct prediction, constant mode saves charging current and, therefore, power consumed in cycle four.

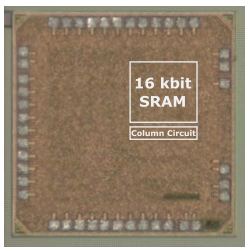


Fig. 12. Die photography of the 16-kbit reconfigurable CP SRAM in 28-nm FD-SOI CMOS.

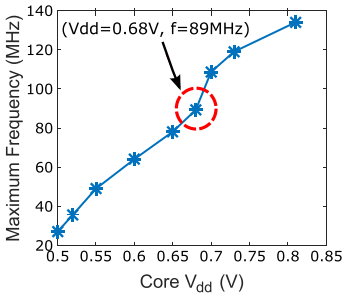


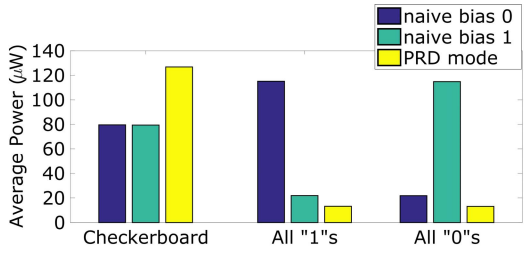
Fig. 13. Recorded performance of the test chip. The upper frequency bound of 134 MHz is a limitation of the testing equipment.

VI. Measurement Results

In this section, we report the results of a 16 kbit, reconfigurable CP SRAM fabricated in a 28-nm FD-SOI CMOS process [18]. Fig. 12 displays the die photography. The die size is 1.0×1.0 mm^2 from which the array occupies 0.015 mm^2. The SRAM was validated to perform with no read errors down to 0.5V, corresponding to a clock frequency of 27 MHz. Under the operating conditions of 0.81-V core supply voltage, the chip achieved 134-MHz maximum frequency, which was a limitation of the test equipment. Sampled operating points are provided in Fig. 13. The power measurements reported in this section used $V_{\text{dd}} = 0.68$ V and $f = 89$ MHz.

A. Power Measurements of Special Data Sequences

In addition to PRD and constant prediction modes, we introduce a baseline power comparison referred to as *naive prediction*. In this mode, the SRAM invariably precharges a user-selected read bitline irrespective of data statistics. This mode loosely approximates the operation of a conventional 8T SRAM with its single-ended read port hardwired to one



Data Set	$\alpha_{0\to1}$	ρ	PRD	naive (bias 0)	naive (bias 1)
Checkerboard	0.5	0.5	0% correct	50% correct	50% correct
All "1"s	0	1	100% correct	0% correct	100% correct
All "0"s	0	1	100% correct	100% correct	0% correct

Fig. 14. Power measurement results of the test chip applied to special data sequences.

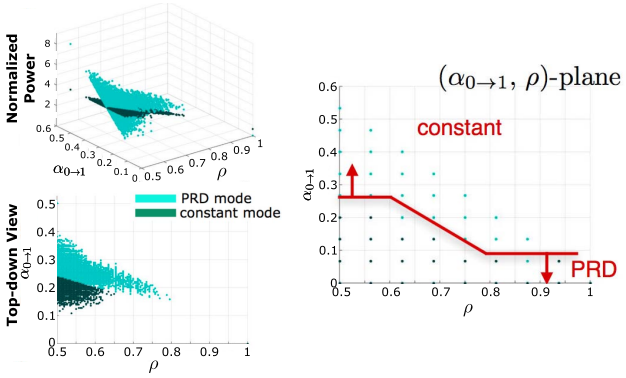


Fig. 15. Simulated $(\alpha_{0\to1}, \rho)$-plane with measured power numbers. Left: methodology behind populating the plane. Right: possible rule for determining optimal prediction modes.

of the storage nodes. To characterize power consumption across a broad range of data patterns, power was measured for data sequences with disparate statistics in each mode. More concretely, we consider the following three artificial data sets: 1) checkerboard, i.e., alternating 0 and 1 values; 2) all 0s; and 3) all 1s. The checkerboard pattern in particular corresponds to the worst case scenario for both PRD and constant modes, since $(\alpha_{0\to1}, \rho) = (0.5, 0.5)$, respectively, maximizes and minimizes the measures $\alpha_{0\to1}$ and ρ leading to prediction accuracies of 0% and 50%. The naive prediction mode achieves the same prediction accuracy as constant mode. When reading the data sets with all 0s or 1s, both PRD and constant modes achieve 100% accuracy, since $(\alpha_{0\to1}, \rho) = (0, 1)$, respectively, minimizes and maximizes the measures $\alpha_{0\to1}$ and ρ. The naive prediction mode achieves the same prediction accuracy as constant mode when the bias is correct and precharges the preselected bitline during every read cycle if the bias is incorrect. The statistical implications of these special patterns are summarized by Fig. 14.

Fig. 14 also presents the power consumed by processing each data set. As demonstrated by naive prediction, the average read power consumption approximately satisfies a linear relationship with the bitline switching activity where the deviations can be attributed to crowbar current in the column circuit during certain switchings. Naive prediction with a correctly chosen bias, equivalent to constant mode,

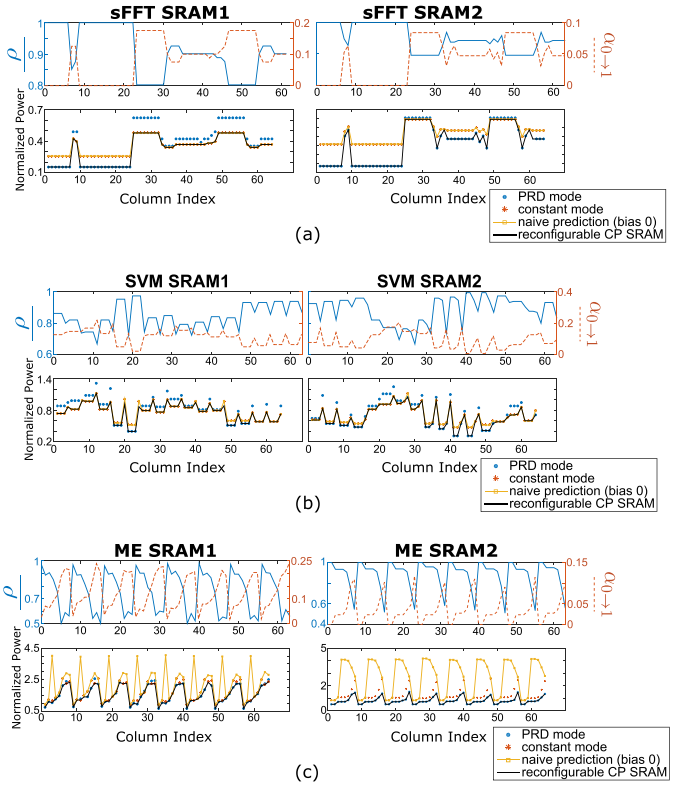


Fig. 16. Optimal column configurations for each of the application SRAMs based on test-chip power numbers (a) sFFT, (b) SVM, and (c) ME SRAMs.

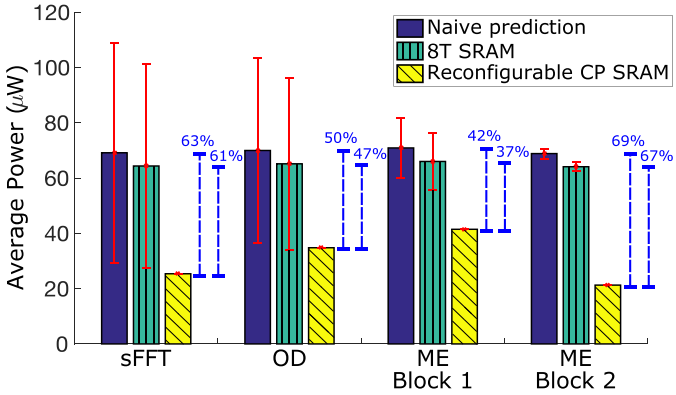


Fig. 17. Power measurement results of the test chip applied to application data sets. The power numbers of both naive prediction and reconfigurable CP SRAM were measured from the 28-nm test chip, whereas 8T SRAM power is estimated by using the measurements of 10T SRAM with naive prediction in combination with postlayout simulation results of critical SRAM blocks.

incurs slightly more power than PRD for the all 1s and all 0s cases due to the additional gate switchings of `ref_D` and `ref_DB`. The measurements from the above-mentioned experiments were used to construct a power model through which an experimental $(\alpha_{0\to1}, \rho)$-plane was generated for optimal mode selection. The result of this model is shown in Fig. 15. Referring to Fig. 15, the plots on the left show the power consumption of randomly generated data sequences and demonstrate the procedure by which the plane was populated. A comprehensive experiment where the data set contains all possible distinct 16-bit CSs is shown on the right. Through this analysis, the parameters in Fig. 6 were extrapolated to provide the rules followed for configuring each CS in an SRAM.

Work	Design Idea	Targeted Application(s)	Energy Savings*
Noguchi '07 [9]	A non-precharge 10T SRAM for reduction in bit-line switching activity	video applications	up to 58%**
Fujiwara '08 [7]	An 8T SRAM using majority logic and data-bit reordering for reduction in bit-line switching activity	video applications	28%
Chang '11 [10]	A priority-based 6T/8T hybrid SRAM for aggressive voltage scaling	video applications	32%**
Gong '12 [11]	An 8T/10T Split-Data-Aware SRAM for aggressive voltage scaling	video applications	N/A***
Sinangil '14 [1]	A 10T PB-RBSA SRAM using output prediction for reduction in bit-line switching activity	video applications	up to 1.75x
Biswas '16 [12]	A 6T SRAM using array-level data prediction for reduction in bit-line switching activity	video applications	up to 36%
Do '16 [8]	An 8T SRAM using column-based data encoding scheme for reduction in bit-line switching activity	image processing	up to 40%
This work	A conditional pre-charge 10T SRAM using reconfigurable prediction schemes for reduction in bit-line switching activity	sFFT, object detection, video coding	up to 67%

* Data-dependency enabled energy savings w.r.t conventional 8T/6T SRAMs with the same technology, operating condition, and array size
** Simulation result ***Energy savings were only reported with voltage scaling

Fig. 18. Comparison of data-dependent SRAM designs.

B. Power Measurements of Application Data

The switching activity and bias values as well as optimal prediction mode for each CS in the SRAMs of the: 1) sFFT; 2) OD; and 3) ME applications are provided in Fig. 16. As demonstrated by the power plots, the power consumed by constant mode is equal to or less than the power consumed by naive prediction. Reconfigurable CP SRAM utilizes either PRD or constant mode to minimize the power consumed by each memory column; therefore, it achieves the maximum read power efficiency possible for any given data set.

The average power consumed by reading test data from the three example applications is given in Fig. 17. ME blocks 1 and 2 represent different ME data sets selected from two histogram clusters in Fig. 5. The power consumed by the naive prediction mode with bias values of 1 and 0 is represented by the ends of the solid lines, and the average values of the two are used to generate the leftmost bars for comparison. This comparison has been extended to 8T SRAMs by compensating for the power overhead of additional switching in column circuitry and additional wiring capacitance due to increased bit-cell size. The read power of an 8T SRAM is estimated to be 0.931 times lower than that of a 10T SRAM operating in naive prediction mode and is shown in Fig. 17 through the middle bars for each application. The overhead of using 10T SRAMs decreases with increasing SRAM array size as would typically be found in applications including the three discussed in this paper. Finally, Fig. 18 compares our approach against all previously mentioned data-dependent SRAM designs. As emphasized by Fig. 18, this paper achieves significant energy savings for a broader range of applications and statistical features.

VII. Conclusion

This paper introduced a reconfigurable memory framework wherein data statistics can be leveraged to achieve power savings for energy-constrained systems. The methods discussed can readily be combined with traditional low-power techniques. As a proof-of-concept, a 16-kbit columnwise, reconfigurable SRAM was implemented in a 28-nm FD-SOI CMOS process. This test chip demonstrated the functionality of the proposed 10T bit cell, CP scheme, and column circuit. Furthermore, this paper showed that different types of data statistics are necessary for achieving low-power performance in many applications, and SRAMs, which support reconfigurability, achieve higher flexibility and better prediction accuracy, and, importantly, incur lower power penalties as compared with their existing counterparts. Power savings were demonstrated for sFFT, OD, and ME applications and are 63%, 50%, and up to 69%, respectively. In addition to circuit-level solutions, this paper provides low-complexity data analysis tools, which can be adapted to statistical learning environments at various levels and in various forms. Possible extensions include on-chip, real-time learning realized with simple computation and comparison modules.

Acknowledgment

The authors would like to thank STMicroelectronics for chip fabrication.

References

[1] M. E. Sinangil and A. P. Chandrakasan, "Application-specific SRAM design using output prediction to reduce bit-line switching activity and statistically gated sense amplifiers for up to 1.9× lower energy/access," *IEEE J. Solid-State Circuits*, vol. 49, no. 1, pp. 107–117, Jan. 2014.

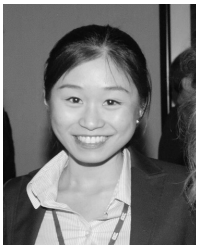
[2] G. K. Chen, D. Blaauw, T. Mudge, D. Sylvester, and N. S. Kim, "Yield-driven near-threshold SRAM design," in *Proc. Int. Conf. Comput.-Aided Design*, Nov. 2007, pp. 660–666.

[3] B. Zimmer *et al.*, "SRAM assist techniques for operation in a wide voltage range in 28-nm CMOS," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 59, no. 12, pp. 853–857, Dec. 2012.

[4] 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.

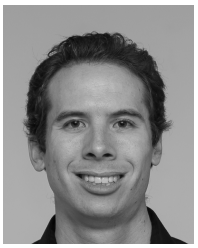
[5] B. Giridhar, N. Pinckney, D. Sylvester, and D. Blaauw, "A reconfigurable sense amplifier with auto-zero calibration and pre-amplification in 28 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 242–243.

[6] Y. Sinangil and A. P. Chandrakasan, "A 128 kbit SRAM with an embedded energy monitoring circuit and sense-amplifier offset compensation using body biasing," *IEEE J. Solid-State Circuits*, vol. 49, no. 11, pp. 2730–2739, Nov. 2014.
[7] H. Fujiwara *et al.*, "Novel video memory reduces 45% of bitline power using majority logic and data-bit reordering," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 16, no. 6, pp. 620–627, Jun. 2008.
[8] A. T. Do, S. M. A. Zeinolabedin, and T. T. Kim, "A 0.3 PJ/access 8T data-aware SRAM utilizing column-based data encoding for ultra-low power applications," in *Proc. Asian Solid-State Circuits Conf.*, Nov. 2016, pp. 173–176.
[9] H. Noguchi *et al.*, "A 10T non-precharge two-port SRAM for 74% power reduction in video processing," in *Proc. Comput. Soc. Annu. Symp. VLSI*, Mar. 2007, pp. 107–112.
[10] I. J. Chang, D. Mohapatra, and K. Roy, "A priority-based 6T/8T hybrid SRAM architecture for aggressive voltage scaling in video applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 2, pp. 101–112, Feb. 2011.
[11] N. Gong, S. Jiang, A. Challapalli, S. Fernandes, and R. Sridhar, "Ultra-low voltage split-data-aware embedded SRAM for mobile video applications," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 59, no. 12, pp. 883–887, Dec. 2012.
[12] A. Biswas and A. P. Chandrakasan, "A 0.36 V 128 Kb 6T SRAM with energy-efficient dynamic body-biasing and output data prediction in 28 nm FDSOI," in *Proc. 42nd Eur. Solid-State Circuits Conf.*, Sep. 2016, pp. 433–436.
[13] O. Abari *et al.*, "A 0.75-million-point Fourier-transform chip for frequency-sparse signals," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 458–459.
[14] A. Suleiman and V. Sze, "Energy-efficient HOG-based object detection at 1080HD 60 fps with multi-scale support," in *Proc. IEEE Workshop Signal Process. Syst. (SiPS)*, Oct. 2014, pp. 1–6.
[15] H. C. Chang, J. W. Chen, B. T. Wu, C. L. Su, J. S. Wang, and J. I. Guo, "A dynamic quality-adjustable H.264 video encoder for power-aware video applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 12, pp. 1739–1754, Dec. 2009.
[16] C. Duan, "Energy efficient reconfigurable SRAM using data-dependency," M.S. thesis, Dept. Electr. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2015.
[17] P. Magarshack, P. Flatresse, and G. Cesana, "UTBB FD-SOI: A process/design symbiosis for breakthrough energy-efficiency," in *Proc. Design, Autom. Test Eur. Conf. Exhibit.*, Mar. 2013, pp. 952–957.
[18] C. Duan, A. Gotterba, M. E. Sinangil, and A. P. Chandrakasan, "Reconfigurable, conditional pre-charge SRAM: Lowering read power by leveraging data statistics," in *Proc. Asian Solid-State Circuits Conf.*, Nov. 2016, pp. 177–180.



Chuhong Duan (S'13) received the B.S. degree in electrical engineering from the University of Virginia, Charlottesville, VA, USA, in 2013, and the M.S. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2015.

She is currently with the Analog Technology Development Team, Texas Instruments, Dallas, TX, USA. Her current research interests include on-chip volatile and nonvolatile memories and low-power systems design.



Andreas J. Gotterba (S'02–M'05) received the B.S. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 2003, and the M.Eng.Sc. degree in photovoltaics from the University of New South Wales, Sydney, NSW, Australia, in 2004.

From 2005 to 2009, he designed embedded memories with Novelics LLC, Aliso Viejo, CA, USA. In 2009, he joined Nvidia, Santa Clara, CA, USA, where he has been involved in developing high-speed and low-power caches. His current research interests include SRAMs and other custom circuits, particularly low-power and handshaking designs.

Mahmut E. Sinangil (S'06–M'12) received the B.Sc. degree in electrical and electronics engineering from Boğaziçi University, Istanbul, Turkey, in 2006, and the M.S. and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 2008 and 2012, respectively.

From 2012 to 2015, he was a Senior Research Scientist with the Circuits Research Group, NVIDIA, Santa Clara, CA, USA. He is currently a Technical Manager with TSMC North America, San Jose, CA, USA, where his responsibilities include the development of on-chip memory compilers and assessing next generation memory bottlenecks and developing solutions for them. His current research interests include low-power and high-density memory circuit design with a focus on low voltage operation and application specific circuit optimizations.

Dr. Sinangil was a recipient of the Ernst A. Guillemin Thesis Award at MIT for his master's thesis in 2008 and the 2006 Boğaziçi University Faculty of Engineering Special Student Award. He was a co-recipient of the 2008 A-SSCC Outstanding Design Award.

Anantha P. Chandrakasan (M'95–SM'01–F'04) received the B.S., M.S., and Ph.D. degrees from the University of California at Berkeley, Berkeley, CA, USA, in 1989, 1990, and 1994, respectively, all in electrical engineering and computer sciences.

He was the Director of Microsystems Technology Laboratories, Cambridge, MA, USA, from 2006 to 2011. Since 2011, he has been the Head of the EECS Department, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA. Since 1994, he has been with MIT, where he is currently the Vannevar Bush Professor of electrical engineering and computer science. He has co-authored *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). His current research interests include ultra-low-power circuit and system design, energy harvesting, energy efficient RF circuits, and hardware security.

Dr. Chandrakasan was a co-recipient of several awards, including the 2007 International Solid-State Circuits Conference (ISSCC) Beatrice Winner Award for Editorial Excellence and the ISSCC Jack Kilby Award for Outstanding Student Paper in 2007, 2008, and 2009. He received the 2009 Semiconductor Industry Association University Researcher Award and the 2013 IEEE Donald O. Pederson Award in Solid-State Circuits. In 2015, he was elected to the National Academy of Engineering. He has served in various roles for the IEEE ISSCC, including the Program Chair, the Signal Processing Sub-Committee Chair, and the Technology Directions Sub-Committee Chair. He has been the Conference Chair of ISSCC since 2010.