EMBER: Efficient Multiple-Bits-Per-Cell Embedded RRAM Macro for High-Density Digital Storage

Akash Levy, *Graduate Student Member, IEEE*, Luke R. Upton, *Graduate Student Member, IEEE*, Michael D. Scott, Dennis Rich, *Graduate Student Member, IEEE*, Win-San Khwa, *Senior Member, IEEE*, Yu-Der Chih, *Member, IEEE*, Meng-Fan Chang, *Fellow, IEEE*, Subhasish Mitra, *Fellow, IEEE*, Boris Murmann, *Fellow, IEEE*, and Priyanka Raina

***Abstract*— Designing compact and energy-efficient resistive RAM (RRAM) macros is challenging due to: 1) large read/write circuits that decrease storage density; 2) low-conductance cells that increase read latency; and 3) the pronounced effects of routing parasitics on high-conductance cell read energy. Multiple-bits-per-cell RRAM can boost storage density but has further challenges resulting from reliability problems due to conductance relaxation and slow write due to narrow conductance levels. This work presents a multiple-bits-per-cell RRAM macro called Efficient Multiple-Bits-per-Cell Embedded RRAM (EMBER), which: 1) demonstrates read/write circuit compaction through constrained optimization of driver and pass gate transistor sizes; 2) introduces a common-mode bleed conductance at the sense amplifier inputs, reducing read settling time by 11.35× for low-conductance cells, and 3) cuts read path capacitance to further reduce read access time and energy. To address reliability and write speed, EMBER contains a configurable on-chip read/write controller. We present a level allocation scheme that uses array-level characterization data to find sufficiently reliable allocations, while simultaneously maximizing write bandwidth. EMBER is the first embedded RRAM storage macro to achieve fully integrated multiple-bits-per-cell readout and write-verification without any off-chip reference generation or sensing. The macro operates at 100 MHz with 64k × 48 = 3 M cells in TSMC 40-nm CMOS, achieving 1 b/cell read operation with 1.0 pJ/bit energy at 2.4 Gbps, and 2 b/cell read with 1.1 pJ/bit at 1.6 Gbps. 1 b/cell write-verify operates with 0.40 nJ/bit energy at 12.4 Mbps (BER < 6e-4), and 2 b/cell write-verify operates with 1.2 nJ/bit at 3.8 Mbps (BER < 3e-3). The array-level endurance is found to be 10 K for 1–2 b/cell. Normalizing for process scaling, the macro demonstrates the highest effective RRAM cell density to date of 5.6e-3 b/F^2 for 1 b/cell and 1.3e-2 b/F^2 for 2 b/cell, an improvement of 31% and 204%, respectively, over the best prior work.**

Manuscript received 1 December 2023; revised 25 March 2024; accepted 29 March 2024. Date of publication 17 April 2024; date of current version 28 June 2024. This article was approved by Associate Editor Danilo Manstretta. This work was supported in part by the AI Chip Center for Emerging Smart Systems (ACCESS), Hong Kong, SAR; in part by the Defense Advanced Research Projects Agency (DARPA) 3DSoC Project; and in part by NSF FuSe-TG under Award 2235462. The work of Akash Levy and Dennis Rich was supported by NSF Graduate Research Fellowship Program (GRFP). *(Akash Levy and Luke R. Upton contributed equally to this work.) (Corresponding authors: Akash Levy; Luke R. Upton.)*

Akash Levy, Luke R. Upton, Michael D. Scott, Dennis Rich, Subhasish Mitra, Boris Murmann, and Priyanka Raina are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: akashlevy@alumni.stanford.edu; lupton@stanford.edu).

Win-San Khwa, Yu-Der Chih, and Meng-Fan Chang are with TSMC, Hsinchu 300, Taiwan.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSSC.2024.3387566.

Digital Object Identifier 10.1109/JSSC.2024.3387566

***Index Terms*— Embedded memory, multilevel cell (MLC), multiple-bits-per-cell, nonvolatile memory, resistive RAM (RRAM).**

I. Introduction

RESISTIVE random access memory (RRAM) is a suitable on-chip memory technology for edge computing because of its back-end-of-line compatibility with CMOS, high unit cell density, multiple-bits-per-cell storage, nonvolatility, and low-energy operation versus embedded flash [1]. Multiple-bits-per-cell RRAM ideally increases storage density by a factor of n [where $n = \log_2(\#\text{ of levels per cell})$] over single-bit-per-cell [2] but faces several challenges, illustrated in Fig. 1: (1) large read/write peripheral circuits attenuate storage density improvements, (2) for multiple-bits-per-cell operation, the decreased low-conductance state slows down the read operation, (3) the increased high-conductance state increases read energy, (4) conductance relaxation [3] can lead to high bit error rates (BERs) as narrow conductance distributions broaden with time, and (5) narrow conductance levels require more write pulses to target and hence cost more write energy. Finally, (6) to our knowledge, no prior RRAM storage macro fully integrates the circuitry needed for both multiple-bits-per-cell readout and write-verify operation.[1]

In this article, we present the Efficient Multiple-Bits-per-Cell Embedded RRAM macro [5] (EMBER; Fig. 2) that addresses the above-mentioned limitations by: 1) decreasing total bitline/source line (BL/SL) driver and pass gate area by 6.07× compared to [6] by using thinner oxide high-voltage transistors and trading off BL/SL write pass gate area with BL/SL driver area; 2) reducing sense amplifier settling time for low-conductance RRAM cells by 11.35× with a common-mode bleed conductance; 3) reducing read energy by 37% through read access switch partitioning; 4) using a bandwidth-aware level allocation scheme to enable reliable read/write operation while; 5) maximizing write bandwidth; and 6) fully integrating multiple-bits-per-cell readout and write-verification on chip, without any off-chip reference generation or sensing. Contributions 4) and 5) were performed after our original conference publication [5].

[1]Fully integrated multiple-bits-per-cell operation with an RRAM array is demonstrated in [4], but the read/write circuits in that design are neither compact nor adjacent to the RRAM unit cells as is typical in a storage macro.

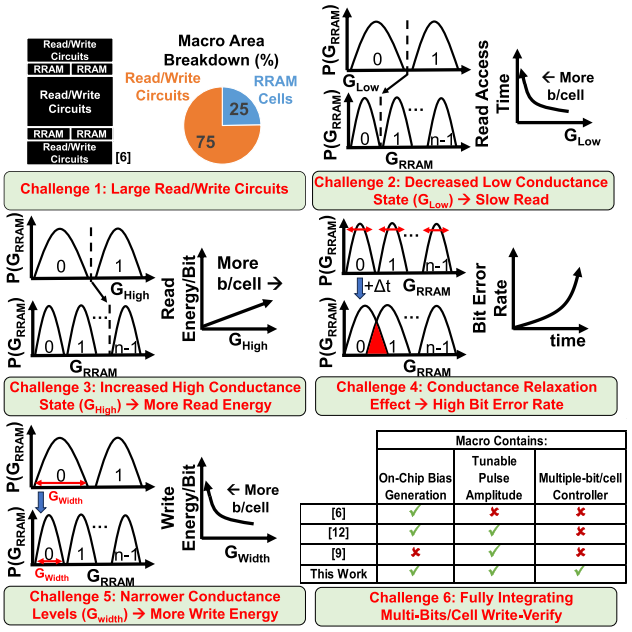
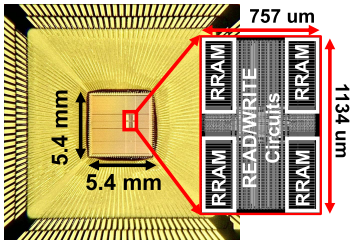


Fig. 1. Multiple-bits-per-cell RRAM macro design challenges.



(a)

Technology	40 nm CMOS
Max Freq	100 MHz
Cell Count	64k×48 = 3M
Read Time	12 ns (2 cycles)
Read Voltage	100-250 mV (BL) 2.5 V (WL)
Write Voltage	1-3.1 V (BL/SL) 0.5-2.5 V (WL)
Supply	I/O: 2.5 V (read) I/O: 3.3 V (write) Core: 0.9 V

(b)

Fig. 2. Summary of EMBER macro. (a) Die photograph of the EMBER chip and (b) overview of the macro specifications.

EMBER demonstrates reliable 1–2 bits per cell on-chip write-verify and readout. It achieves an average read energy of 1.0 pJ/bit at 2.4 Gbps (on par with the state-of-the-art) for 1 b/cell and 1.1 pJ/bit at 1.6 Gbps for 2 b/cell. For 1 b/cell write-verify, EMBER achieves 0.40 nJ/bit energy at 12.4 Mbps (BER < 6e-4), and for 2 b/cell write-verify, it achieves 1.2 nJ/bit at 3.8 Mbps (BER < 3e-3). The array-level endurance is 10 K for 1–2 b/cell. The RRAM array occupies 36.9% of the macro area with the remainder used for read and write circuits. EMBER's normalized array density is 5.6e-3 b/F^2 for 1 b/cell and 1.3e-2 b/F^2 for 2 b/cell, an improvement of 31% and 204%, respectively, over the state of the art [6]. EMBER's density, speed, and energy efficiency enable it to serve as an on-chip nonvolatile memory in edge devices.

II. Background and Prior Work

An RRAM cell has a metal–insulator–metal (MIM) structure, enabling the creation or dissolution of conductive filaments within the cell's insulating oxide layer in response to applied voltage pulses. This switching mechanism can be unipolar, where the same electric field direction serves for both increasing and decreasing conductance, or bipolar, which involves distinct electric field directions for conductance increase and decrease. Bipolar switching is generally favored for its enhanced reproducibility and write endurance [1], [7], [8]. In the context of the bipolar HfO_x RRAM employed in this study, each RRAM unit cell is equipped with an access transistor to mitigate off-state current leakage, along with a MIM structure situated on the drain side of the access transistor, which stores information in the form of cell conductance. This 1-transistor 1-resistor (1T1R) unit cell configuration is commonly adopted for data storage. In the 1T1R configuration, the access line linked to the source of the access transistor is designated as the SL, the access line connected to the top of the MIM junction is referred to as the BL, and the access transistor gate is controlled by the word line (WL). The SET process is accomplished by applying a BL-to-SL voltage pulse, which increases the cell's conductance. Conversely, the RESET process involves an SL-to-BL voltage pulse, which reduces the cell's conductance. Reading the conductance of the cell involves applying a small voltage from BL to SL and measuring the resulting current response.

RRAM cell storage density can be improved through the use of intermediate conductance states. 2-3 bits/cell storage has been demonstrated in RRAM arrays with the assistance of off-chip source-measure units and pulse generation [9], [10], [11]. A prior RRAM-based compute-in-memory macro exhibit three-level write-verification using on-chip readout and write circuitry [4], but to our knowledge, a fully integrated macro has not been demonstrated with ≥2 bits/cell storage capability.

III. Array and Read/Write Circuit Architecture

The EMBER macro architecture (Fig. 3) consists of four 0.75 Mcell RRAM unit cell array quadrants, divided by shared peripherals used for read/write. Each quadrant has 24 subarrays, each with 512 WLs and 64 BLs (64 × 512 × 24 = 0.75 M cells). The RRAM arrays in EMBER employ a 1T1R architecture with a common SL for improved cell density [6]. Each half of the array contains 24 sense amplifiers and write drivers, so a word of up to 48 cells can be read or programed in parallel.

The array peripherals consist of WL and BL/SL components. The input address is used to select a WL and drive a DAC-generated voltage on that WL. Each WL spans two quadrants of the array. On the BL/SL, DACs adjust the applied write voltage and sense amplifiers enable read operations on the BL. Sense amplifiers compare a selected cell's conductance against a 6-bit user-controlled reference, at a BL voltage specified by the read DAC. The read/write analog interfaces in EMBER are operated by a digital controller, which receives commands over a serial peripheral interface (SPI).

A. Technique 1: Write Circuit Compaction Through Constrained Optimization of Driver and Pass Gate Sizes

The total peripheral area (dominated by BL/SL pass gates and write drivers) is comparable to the area of the RRAM cells [e.g., Fig. 4(a)], which significantly lowers the effective bit density. The write path transistors must be wide enough to minimize IR drop when resetting high-conductance RRAM

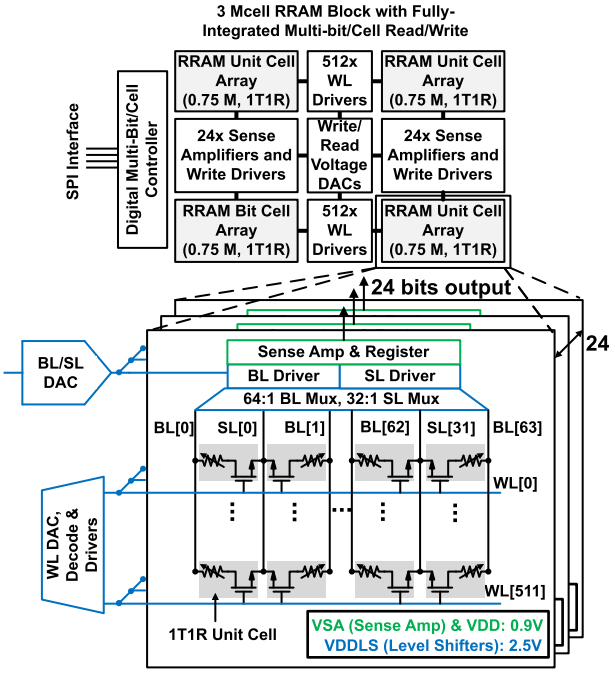


Fig. 3. EMBER top-level architecture and array design.

cells to a low conductance, but making these peripheral transistors too large lowers macro density and slows reading because of extra RC delay. We thus aim to minimize total peripheral area under the constraint that requisite cell SET/RESET voltages are produced. We perform a write transistor size sweep with relative sizing parameters [Fig. 4(b)]

$$\Gamma = \frac{V_{\text{on,SL}}}{V_{\text{on,BL}}} = \frac{W_{\text{Driver,BL}}}{W_{\text{Driver,SL}}} = \frac{W_{\text{Passgate,BL}}}{W_{\text{Passgate,SL}}} \tag{1}$$

$$\lambda = \frac{V_{\text{on,Drivers}}}{V_{\text{on,Passgates}}} = \frac{W_{\text{Passgate,BL}}}{W_{\text{Driver,BL}}} = \frac{W_{\text{Passgate,SL}}}{W_{\text{Driver,SL}}}. \tag{2}$$

$V_{\text{on,SL}}$ and $V_{\text{on,BL}}$ are the total SL-side and BL-side on-voltages in the write path and $V_{\text{on,Drivers}}$ and $V_{\text{on,Passgates}}$ are the total driver and pass gate on-voltages. $W_{\text{Driver,BL/SL}}$ and $W_{\text{Passgate,BL/SL}}$ correspond to the widths of pull-up/pull-down transistors in the BL/SL side of the write path. Note that only the width of pull-up (down) transistors should be compared to the width of other pull-up (down) transistors. All transistors in the pass gates and drivers use the process-defined minimum length, so parameters λ and Γ simplify total write path transistor width minimization to a two-parameter search.

We sweep (Γ, λ) with a minimum required V_{cell} at a desired maximum G_{cell} to find $\sum \text{Area}_{\text{SET}} + \sum \text{Area}_{\text{RESET}}$ [Fig. 4(c)]. The appendix contains the equations and assumptions used to calculate the SET/RESET path area. The global minimum sum of SET and RESET path area appears at $\lambda = 0.22$, $\Gamma = 0.65$. Using (2), (8) and $N_{\text{pg/drv}} = 16$, $\sum \text{Area}_{\text{SET/RESET}} \approx A_{\text{Driver}} + 16(A_{\text{Driver}}/4)$, meaning that the pass gate area is roughly 4× larger than the driver area, which can be verified in Fig. 4(a). Using this minimization technique and choosing thinner oxide write path transistors resulted in a 6.07× reduction in write circuit area compared to [6].

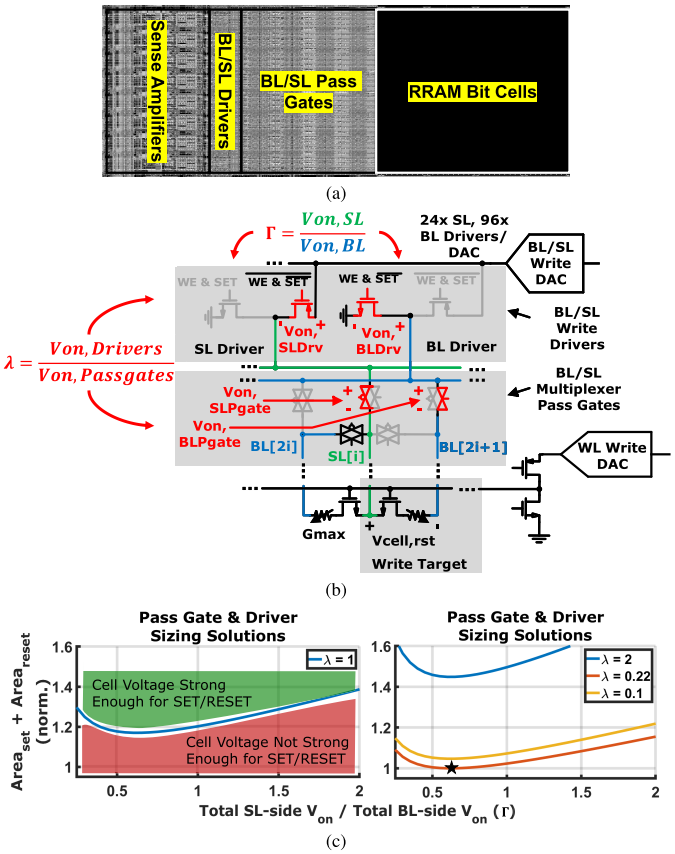
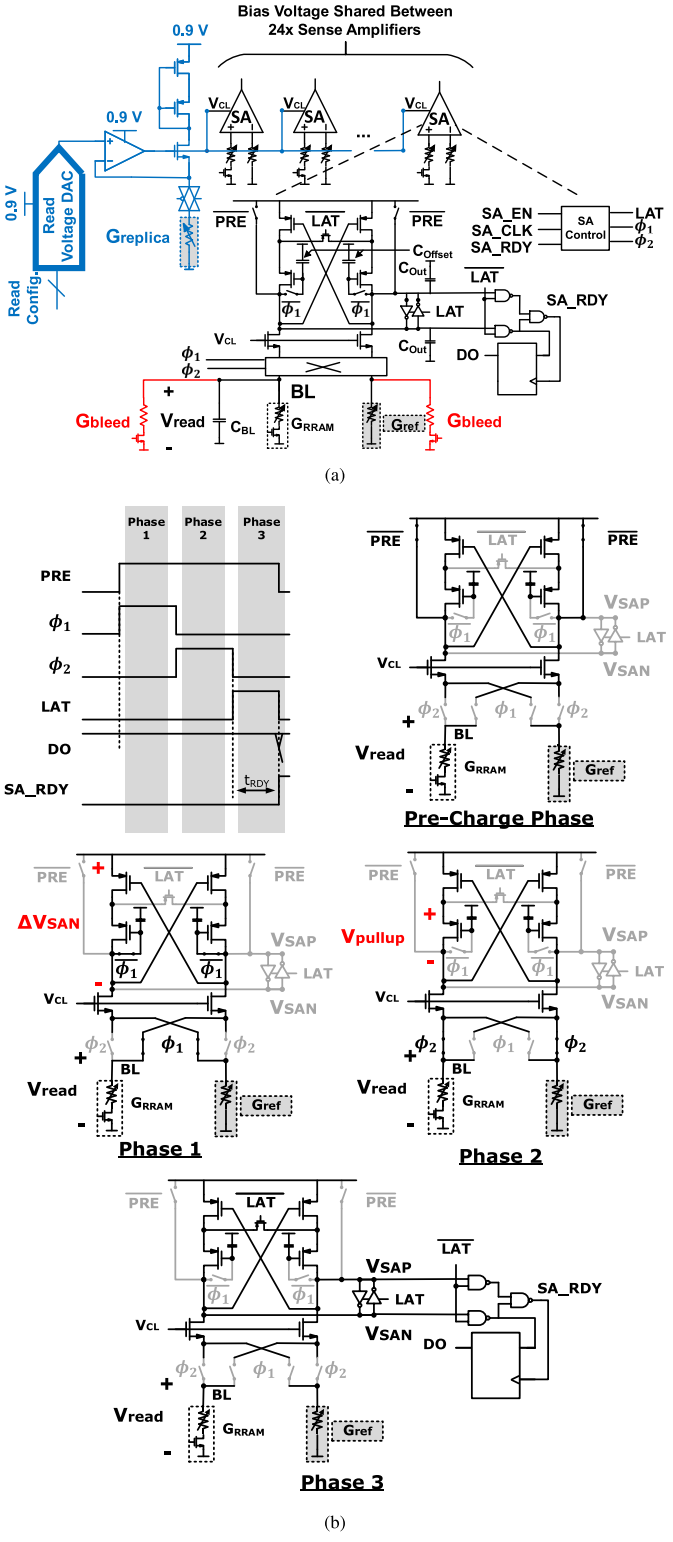


Fig. 4. (a) Read/write interface layout. (b) RRAM write path schematic. BL/SL write DAC current passes through (1) the BL/SL driver (SET/RESET), (2) a pass gate to (3) induce voltage drop across a unit cell, and (4) an SL/BL (SET/RESET) pass gate and (5) driver to ground. (c) Γ sweep for fixed λ. The blue line represents the minimum SET and RESET paths required to meet the needed write strength (left). (c) Two-parameter search to find the smallest write path design (right). Minimum area is achieved when $\lambda = 0.22$, $\Gamma = 0.65$.

B. Technique 2: Faster Read Times With Bleed Conductance

EMBER uses an offset-canceling current-mode sense amplifier design inspired by [12], with the addition of common-mode bleed conductances to limit sense amplifier settling time [shown in red in Fig. 5(a)]. The sense amplifier operates in three phases [Fig. 5(b)] dictated by the sense amplifier controller and remains in a precharge state when not actively reading. In Phase 1, the amplifier samples the current offset between both halves using diode-connected PMOS devices. In Phase 2, the amplifier samples $G_{\text{RRAM}} - G_{\text{ref}}$, and the offset current cancels itself out. In Phase 3, a cross-coupled inverter latch determines if $G_{\text{RRAM}} > G_{\text{ref}}$ (“1”) or $G_{\text{RRAM}} < G_{\text{ref}}$ (“0”). The SA_RDY signal marks the end of Phase 3, and the sense amplifier automatically reenters the precharge state.

The conductance reference G_{ref} is a 6b linear conductance DAC constructed with poly resistors and replica transistors that mimic the I/O transistors in the G_{RRAM} read path (Fig. 6). The two least significant bits of G_{ref} use a thermometer decoding scheme for conductance steps while the four most significant bits use a binary design. This segmented design saves layout area compared to a purely binary design.



(a)

(b)

Fig. 5. (a) Offset-canceling current-mode sense amplifier design (inspired by [12]), with modifications highlighted in red and blue. A shared circuit generates V_{CL} and sense amplifier control blocks generate ϕ_1 and ϕ_2 locally for each sense amplifier. (b) Sense amplifier operating phases. ΔV_{SAN} and V_{pullup} are dc operating voltages utilized in noise analysis.

The amplifier input settling time is described by [13]

$$\tau_{\text{settling}} = \frac{C_{\text{BL}} n V_t}{V_{\text{read}}(G_{\text{RRAM}} + G_{\text{bleed}})} \tag{3}$$

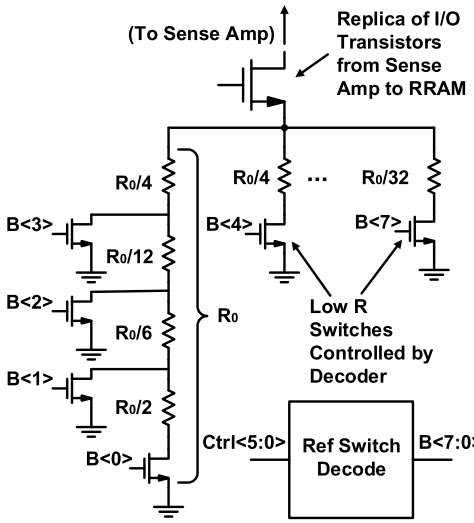


Fig. 6. G_{ref} schematic. The decoder converts the 6-bit conductance request to the switching signals necessary for the desired conductance in a segmented DAC design.

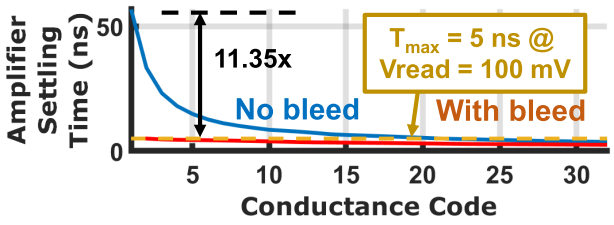


Fig. 7. Simulation of settling times. Minimum G_{bleed} that meets $T_{\text{max}} = 5$ ns (100 MHz) is used.

$$V_{\text{BL,settling}} = V_{\text{CL}} - nV_t \ln\left(\frac{(1+\Delta)V_{\text{read}}(G_{\text{RRAM}} + G_{\text{bleed}})}{I_{\text{DO}}}\right) \tag{4}$$

$$t_{\text{settling}} = -\tau_{\text{settling}} \times \ln\left(1 - \frac{(G_{\text{RRAM}} + G_{\text{bleed}})\exp\left(\frac{V_{\text{BL,settling}} - V_{\text{CL}}}{nV_t}\right)}{I_{\text{DO}}/V_{\text{read}}}\right). \tag{5}$$

Here, C_{BL} is the input-referred capacitance on the BL side of the amplifier, $n = 1.5$ (ratio of MOSFET subthreshold slope to 60 mV/decade), $V_t = (k_B T/q)$, V_{read} is the cell read voltage, G_{RRAM} is the cell conductance, I_{DO} is the subthreshold current constant for the transistors with $V_G = V_{\text{CL}}$, $V_{\text{BL,settling}}$ is the BL voltage needed to achieve transistor current settling within the margin Δ of the steady-state read current, and G_{bleed} is the common-mode bleed device conductance (Fig. 5(a), red). When using this design with low G_{RRAM} and no G_{bleed}, the 99% ($\Delta = 0.01$) input current settling time needed for comparisons increases due to the $1/G_{\text{RRAM}}$ dependence from (3). G_{bleed} forces 99% read current settling to occur by $T_{\text{max}} \approx 4\tau_{\text{settling}}$, and the sense amplifier clock frequency becomes $(1/2T_{\text{max}})$. G_{bleed} detaches the choice of RRAM conductance operating range from the read frequency specification, allowing for low G_{RRAM} in a multiple-bits-per-cell scheme (Fig. 7).

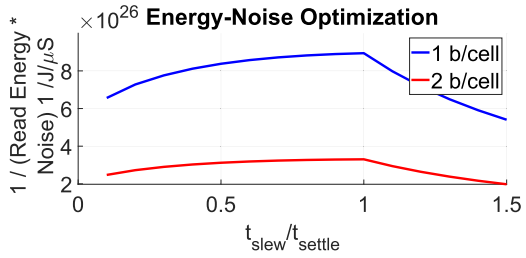


Fig. 8. Noise–energy efficiency plot for C_{offset} sizing, where $C_{\text{out}} = 12.5$ fF. The objective of the $t_{\text{slew}}/t_{\text{settle}}$ sweep is to minimize the product of per-bit read energy and input-referred noise, or maximize (energy × Noise)$^{-1}$ (shown in this plot).

We designed the sense amplifier such that: (1) G_{bleed} enforces a maximum settling time criterion; (2) the input SNR $\geq$ 10 dB; and (3) the noise-energy product of sensing was minimized. Such conditions guarantee stable operation while maximizing sensing efficiency with a provided read energy. The input-referred conductance noise of the sense amplifier architecture can be described using the following equation:

$$G_{\text{rms}}^2 = \frac{k_B T}{(C_{\text{offset}} + C_{\text{BL}})} \frac{(G_{\text{RRAM}} + G_{\text{bleed}})^2}{2V_{\text{pullup}}} \frac{g_{m,p}}{I_d} \times \left(\frac{C_{\text{offset}} + C_{\text{BL}}}{C_{\text{out}}} + \frac{g_{m,p}}{I_d} V_{\text{pullup}} \right) \tag{6}$$

where the capacitor is used for offset cancellation [Fig. 5(a)] can be sized with

$$C_{\text{offset}} = \frac{t_{\text{slew}}}{t_{\text{settle}}} \frac{\beta C_{\text{BL}} n V_t}{\Delta V_{\text{SAN}}} \times \ln\left(1 - \frac{(G_{\text{RRAM}} + G_{\text{bleed}}) \exp\left(\frac{V_{\text{BL,settling}} - V_{\text{CL}}}{n V_t} \right)}{I_{\text{DO}}/V_{\text{read}}} \right). \tag{7}$$

Above, $(g_{m,p}/I_D) = 20$ S/A for the PMOS devices in Phase 2, ΔV_{SAN} [Fig. 5(b)] is the expected voltage across C_{offset} at the end of Phase 1 of amplifier operation, and V_{pullup} is the expected voltage across the offset current sourcing transistors during Phase 2, $\beta = 0.75$ is the fraction of average current through $G_{\text{ref}} + G_{\text{bleed}}$ versus the final steady-state current in Phase 1, and $t_{\text{slew}}/t_{\text{settle}}$ is the ratio of C_{offset} slew time to RRAM read current settling time during Phase 1 of amplifier operation.

After assuming a range of reasonable C_{out} values, we sweep $t_{\text{slew}}/t_{\text{settle}}$, find the corresponding C_{offset}, fit G_{bleed} to meet 99% settling or slew under 5 ns across the input conductance range, use $G_{\text{RRAM,max}} + G_{\text{bleed}}$ to determine G_{rms} and enforce the 10-dB SNR cap, and then use the read current settling information given by (5) to estimate the read energy for 1 and 2 b/cell (Appendix). Fig. 8 shows that the energy-noise product $(G_{\text{rms}} * \Sigma_1^3 E_{\text{phase}})$ is minimized around $t_{\text{slew}}/t_{\text{settle}} = 1$ for this amplifier architecture. Thus, our design balances C_{offset} and C_{BL} current settling times.

In our sense amplifier noise and offset models, the RRAM cell behaves as an ideal resistor with a noise current density of $4k_B T G_{\text{RRAM}}$. The cell relaxation and random telegraph noise also impact the sense amplifier input conductance offset.

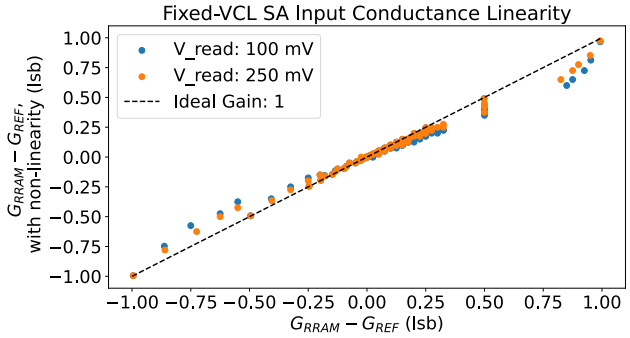


Fig. 9. Sense amplifier linearity analysis with fixed V_{CL}. G_{RRAM} is swept within 1 LSB of a G_{ref} value at fixed V_{CL} and the conductance on the G_{RRAM} input of the sense amplifier is referred to the G_{ref} side using the reference side read voltage. In all test configurations possible in EMBER, the slight nonlinearity does not cause a read result to flip or significantly reduce sense amplifier SNR.

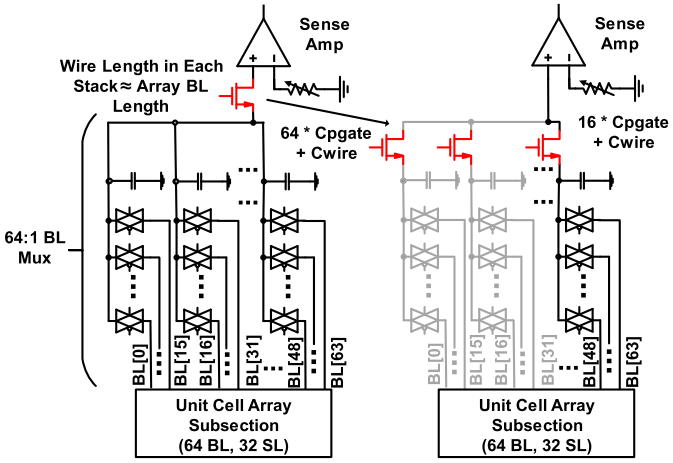


Fig. 10. Schematic of pass gate stacking from Fig. 4(a). Duplicating the read access switch allows cell reading without driving excess capacitance, decreasing read energy, and settling time.

To save energy in the read biasing circuits, the internally generated sense amplifier clamping voltage V_{CL} [Fig. 5(a), blue] is shared between 24 sense amplifiers present in each half of the BL/SL interface. This reduces sense amplifier area by 60% and reads energy by 30% compared to incorporating local feedback amplifiers [14]. Sharing V_{CL} has minimal impact on the readout result, even accounting for the nonuniformity in read voltage between an RRAM cell and a conductance reference (Fig. 9).

C. Technique 3: Reducing Sense Amplifier Load Capacitance via Smart Placement of Read Access Switches

Due to pass gate physical stacking at the array edge (Fig. 10), the read/write access routes to the pass gates branch from a shared sense amplifier. Rather than a single large read access switch, each branch has a dedicated smaller read access switch, limiting the wire and pass gate capacitance at the sense amplifier input. This change reduces settling time by 20% and read energy by 37% compared to increasing G_{bleed} to meet $T_{\text{max}} = 5$ ns with all branches loading the sense amplifier

input. Replicating the read access transistor leads to a 10% increase in the BL driver area and a 0.5% increase in the overall BL/SL interface area compared to using a common access transistor.

IV. Reliable Multiple-Bits-Per-Cell Storage

To reliably store multiple bits per cell with RRAM, a cell's measured conductance level must remain sufficiently close to its programed level. This proves to be challenging to enforce because RRAM suffers from *conductance relaxation* [3], [3], [15], where a cell's conductance evolves stochastically after the write process is complete. We show that when EMBER is used with an appropriate level allocation scheme, it achieves efficient and reliable multiple-bits-per-cell programming despite relaxation.

A. On-Chip Read/Write Controller Architecture

When storing information with cell conductance, boundaries between levels must be defined in such a way that a cell programed to one conductance level does not relax into a different level over time. To this end, distinct thresholds for read and write are used, with write windows being narrower than read windows. EMBER contains a configurable on-chip write-verify controller to generate the necessary waveforms to push RRAM cells into the desired conductance windows and read out their state. The controller operates at 100 MHz and contains a user-programmable register file with settings defining the read/write thresholds and pulsing strategy for each conductance level (up to 16 levels/cell), discussed further in Section IV-B. Commands (such as `SET`, `RESET`, `SENSE`, and `CYCLE`) execute pulsing/sensing operations. `READ` and `WRITE` commands enable multiple-bits-per-cell read and write-verify operations.[2]

During read, the sense amplifier's reference conductance is set to each level's upper readout threshold. As the level index is increased, it can be inferred which level a cell is in when the sense amplifier output changes from 1 (greater than reference) to 0 (less than reference). Once a cell's level is determined, that cell is "masked" to disable further sensing operations on it. During this "ramp read" process, our controller applies two more optimizations: 1) the sensing operation for the final level can be skipped since we know that any unmasked cells must belong in that final level; and 2) if all cells become masked before reaching the $(N-1)$th level, reading terminates early.

During write, we find all cells in a word that need to be programed to a particular level, then apply SET pulses to each one whose conductance is below the lower write threshold. After each SET pulse is applied, we verify which cells have achieved a conductance greater than this lower threshold. These cells are masked, and for the remaining cells, we apply another SET pulse with greater pulse strength (using either an increased WL voltage, BL voltage, or pulsewidth). This "SET-verify" sequence is repeated until all target cells are successfully pushed above the lower write threshold. Next, a "RESET-verify" pulse train is applied similarly to push the target cells below the upper write threshold. The SET-verify and RESET-verify operations are alternated until all cells are found to be stable within the write programming window (or the maximum pulse count is exceeded, causing the process to abort). This write-verify technique is performed for each target level to achieve multiple-bits-per-cell storage.

B. Conductance Level Allocation

To prevent reliability problems due to conductance relaxation, it is important to intelligently allocate conductance levels. A commonly used prior method to determine read/write windows is sigma-based allocation (SBA) [10], in which the width of each conductance window is set to be proportional to the measured standard deviation of conductance after programming. However, it was shown in [17] that SBA produces suboptimal allocations because it assumes that conductance values are normally distributed after write. In place of SBA, percentile-based allocation (PBA) [17] was proposed, in which measured conductance distributions are used for level allocation rather than normal distributions.[3] In both SBA and PBA, conductance levels are allocated greedily based on a user-specified cell error rate (CER) tolerance γ. PBA guarantees that the CER between neighboring levels will be bounded by γ. SBA does not provide such a guarantee unless the conductance distributions are normal. However, PBA does not necessarily produce allocations with the best write speed. We propose BandWidth-Aware Percentile-Based Allocation (BWA-PBA), which searches for allocations having Pareto-optimal write bandwidth and BER (under the same γ constraint as PBA).

C. Bandwidth-Aware PBA (BWA-PBA)

Like PBA, BWA-PBA requires a sufficiently large characterization dataset to be effective, which is collected as follows.

1) Program all cells in an array to a uniformly random distribution of conductance values between the minimum and maximum measurable conductance (0–256 μS).
2) Enumerate a comprehensive set of write conductance windows, (a, b) where $a < b$ and $\{a, b\} \in \mathbf{G}$, where $\mathbf{G}$ is the set of all possible conductance values resolvable by the read circuitry (e.g., 0–256 μS, in steps of 4 μS).
3) For each word in the array, program the cells in that word such that each cell's conductance g is $a < g \leq b$. Record the elapsed time to estimate the write bandwidth of that window.
4) Measure each cell's conductance immediately after programming and once more $\sim 10^4$ s later (to account for conductance relaxation).

After characterization, BWA-PBA is executed as follows.

1) Pick a γ value as the desired upper bound for the error rate between neighboring levels.
2) Compute the cumulative distribution function (cdf) of cell conductance after 10^4 s for each write window

[2]Pseudo-code for the `READ` and `WRITE` commands can be found in [16].

[3]These measured distributions are typically collected on a small set of samples, processed with off-chip software, and the resulting level allocations are applied to all chips. This is the same approach described in [17].

(a, b) based on the characterization data. Using data after 10^4 s enables the resulting conductance allocation to be resilient to short-term relaxation.

3) For each window (a, b), calculate the conductance values $g_{\min}(a, b)$ and $g_{\max}(a, b)$, where the cdf of (a, b) intersects $\gamma/2$ and $1 - \gamma/2$, respectively.
4) Construct a directed acyclic graph, where each write window (a, b) is a node and a directed edge exists between (a_1, b_1) and (a_2, b_2) iff $g_{\max}(a_1, b_1) < g_{\min}(a_2, b_2)$.
5) Traverse all possible paths between nodes $(0, n)$ and $(m, G_{\max})$, where $n < m$ and $G_{\max}$ is the maximum reference conductance. Each such path represents a possible level allocation to consider.
6) For each level allocation, compute the optimal read boundary between each neighboring pair of levels by minimizing the expected error rate between them, using the CDFs computed earlier.
7) Compute the expected write bandwidth and expected BER. The write bandwidth can be estimated by averaging the characterized bandwidth for each write window in the allocation. The expected BER can be calculated by finding the rates of confusion between levels and computing the expected rate of bit flips under a Gray coding scheme.
8) Compute the Pareto-optimal set of allocations for write bandwidth and BER. Write bandwidth should be maximized and BER should be minimized. Choose one or more "allocations of interest" along the Pareto front.
9) Validate the chosen allocations by performing test writes across a large number of cells, in a checkerboard (CB) pattern or pseudorandom pattern generated from a linear feedback shift register (LFSR).

Fig. 11 shows the tradeoff between write bandwidth and BER. We choose $\gamma = 1.2 \times 10^{-2}$ and enumerate allocations where the BER $< 3 \times 10^{-3}$, which is sufficient for BCH error correction [10]. PBA allocations for γ swept between $[0, 1.2 \times 10^{-2}]$ in steps of 4×10^{-4} are shown for comparison. BWA-PBA produces a set of allocations that is a strict superset of PBA. It can be seen that PBA's greedy approach produces allocations with low BER but slow write speeds, while BWA-PBA discovers several ways to trade off bandwidth and BER.

In Fig. 11, we circle a few 1 and 2 b/cell "allocations of interest" in purple. The distributions for these allocations are plotted in Fig. 12. The points were picked to illustrate low BER (allocations 1.1, 1.3, and 2.1), high bandwidth (allocations 1.2 and 2.3), and mid-BER mid-bandwidth (allocation 2.2). Since the BWA-PBA Pareto front is relatively flat for 1 b/cell with increasing BER, we choose two Pareto-optimal allocation points that lie at the low-BER end of the spectrum. For 2 b/cell, the lowest BER point happens to be a PBA allocation as well. Evaluations of these allocations across 10 000 words on EMBER are given in Table I. BWA-PBA correctly predicts which allocations will yield the lowest BER and which will yield the highest bandwidth. High-bandwidth allocations tend to require wider write windows, while low-BER allocations tend to require narrower write windows. PBA allocations ignore the size of the write windows, which usually ends up yielding low BER with smaller write windows and low bandwidth.

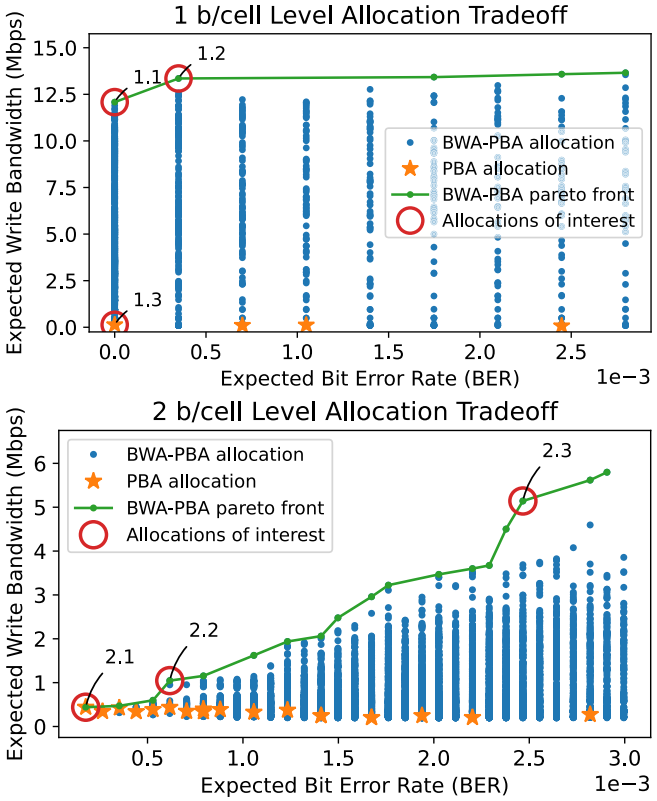


Fig. 11. Write bandwidth vs. BER for 1 b/cell (top) and 2 b/cell level (bottom) allocations. For both 1–2 b/cell, BWA-PBA produces a Pareto curve tradeoff between BER and write bandwidth, while PBA's best allocation minimizes BER at the cost of low write bandwidth. Circled allocations are evaluated on EMBER.

TABLE I

CONDUCTANCE ALLOCATIONS TESTED ON EMBER. THE BEST RESULTS FOR EACH BITS-PER-CELL/WRITE DATA COMBINATION ARE BOLDED. ESTIMATES OF THE RESULTS FROM THE CHARACTERIZATION DATA ARE PROVIDED IN PARENTHESES. SLIGHT DIFFERENCES IN BANDWIDTH/BER BETWEEN CB/LFSR DATA PATTERNS ARE MOST LIKELY DUE TO NOISE FROM SAMPLING DIFFERENT SETS OF RRAM CELLS. BW = WRITE BANDWIDTH, BER = BIT ERROR RATE, AND CER = CELL ERROR RATE

	Type	WData	BW (Mbps)	nJ/b	CER	BER
1.1	BWA	CB	10.94 (12.08)	0.49	**1.42E-4** (0)	**1.42E-4** (0)
1.1	BWA	LFSR	9.79 (12.08)	0.51	4.61E-4 (0)	4.61E-4 (0)
1.2	BWA	CB	**11.53** (13.35)	**0.46**	5.43E-4 (3.5E-5)	5.43E-4 (3.5E-5)
1.2	BWA	LFSR	**12.37** (13.35)	**0.40**	**3.08E-4** (3.5E-5)	**3.08E-4** (3.5E-5)
1.3	PBA	CB	1.26 (0.13)	3.77	2.05E-3 (0)	2.05E-3 (0)
1.3	PBA	LFSR	1.28 (0.13)	3.64	2.64E-3 (0)	2.64E-3 (0)
2.1	PBA	CB	1.50 (0.45)	3.12	**1.17E-3** (3.52E-4)	**5.85E-4** (1.76E-4)
2.1	PBA	LFSR	1.51 (0.45)	3.07	**1.37E-3** (3.52E-4)	**6.88E-4** (1.76E-4)
2.2	BWA	CB	2.41 (1.05)	1.94	2.54E-3 (1.23E-3)	1.27E-3 (6.17E-4)
2.2	BWA	LFSR	2.21 (1.05)	2.08	2.75E-3 (1.23E-3)	1.37E-3 (6.17E-4)
2.3	BWA	CB	**4.57** (5.14)	**1.04**	5.21E-3 (4.93E-3)	2.60E-3 (2.47E-3)
2.3	BWA	LFSR	**3.83** (5.14)	**1.20**	4.71E-3 (4.93E-3)	2.36E-3 (2.47E-3)

D. Array-Level Endurance Characterization

We test the endurance of a level allocation at room temperature by repeatedly writing synthetic data (e.g., CB and LFSR) across many words while monitoring the write bandwidth and

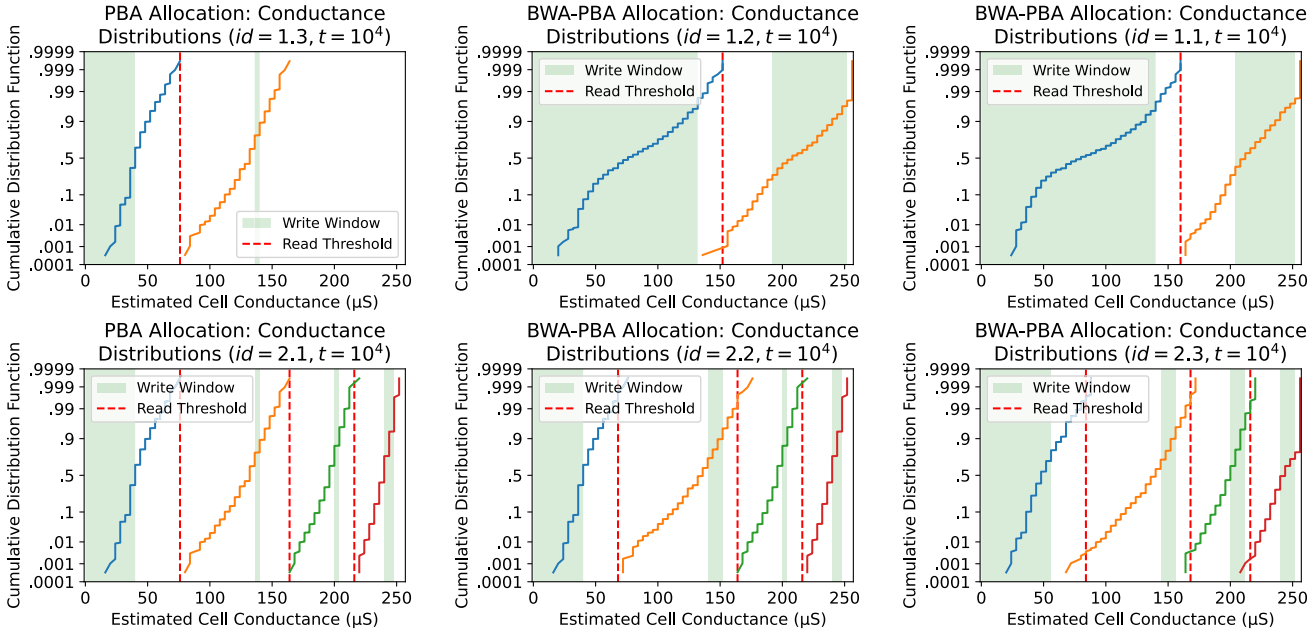


Fig. 12. Example conductance level allocations showing write windows and read thresholds chosen by BWA-PBA and PBA based on the measured conductance distributions 10^4 s after programming. The `id` value links these figures to Fig. 11.

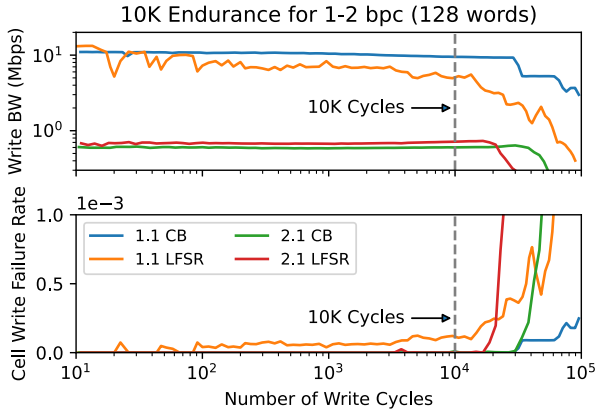


Fig. 13. Endurance characterization for 1–2 b/cell (low-BER allocations), showing write bandwidth degradation and write failure rate increase as cells wear.

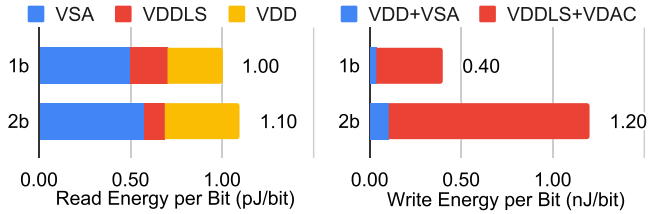


Fig. 14. Left: Measured average read energy breakdown at 100 MHz operation with $V_{\text{read}} = 100$ mV. Read was repeatedly performed on CB patterns with word address scrolling. VDD: core digital supply (0.9 V), VDDLS: level shifter supply (2.5 V), VSA: sense amplifier supply (0.9 V). Right: Measured average write energy breakdown at 100 MHz. Write was repeatedly performed with LFSR data patterns and word address scrolling. VDAC: write DAC supply, VDD + VSA both at 0.9 V, VDAC+VDDLS both at 3.3 V.

write error rate. The results of array endurance measurements for low-BER 1 and 2 b/cell allocations are shown in Fig. 13. The write bandwidth fluctuates with time due to the inherent stochasticity of the SET and RESET processes. The results indicate that 10 K write cycles can be achieved for both 1–2 b/cell with a relatively low write failure rate.

V. Measured Results and Discussion

EMBER was fabricated in TSMC 40-nm CMOS technology. Fig. 2(a) shows the die photo, and Fig. 2(b) lists the macro specifications. All measured results were collected on a single macro.

A. Read Energy and Bandwidth

EMBER's read bandwidth is 2.4 Gbps for 1 b/cell (48 bits per word, 2 cycles per word, 10 ns per cycle) and 1.6 Gbps for 2 b/cell (96 bits per word, 6 cycles per word, 10 ns per cycle). The average read energy using the full reference conductance range (4–256 μS) varies from 1.0 pJ/bit for 1 bit/cell to 1.1 pJ/bit for 2 bit/cell (Fig. 14). The sense amplifier supply (VSA) consumes 50% of the total per-bit read energy at the 100 MHz operating frequency. The average read energy per bit scales non-linearly with unit cell storage resolution since on average $(2^{n-1}/n)$ sense amplifier reads are performed per bit of information in an n-bit scheme. The macro quiescent power is 5 μW, excluding the digital controller and using off-chip power gating for the write DAC supply.

B. Area, Write Energy, Bandwidth, and BER

The total macro area is 0.86 mm^2 and the breakdown is presented in Fig. 15(a). Write energy, bandwidth, and BER are strongly dependent on the choice of conductance level allocation (Section IV). When evaluating the nominal performance, we consider the LFSR data pattern, which is representative of random bits being continuously programed. For 1 b/cell, we achieve 12.4 Mbps write speed at 0.40 nJ/bit, with BER of $<$ 6e-4 for cells before any cycling. For 2 b/cell,

TABLE II
COMPARISON WITH STATE-OF-THE-ART RRAM MACROS

	This Work	[9]	[12]	[19]	[6]
Integrated Read/Write	Yes (1-2 b/cell)	No	Yes (1 b/cell)	Yes (1 b/cell)	Yes (1 b/cell)
Process Node	40 nm	130 nm	22 nm	22 nm	40 nm
Capacity (Mcells)	3	0.016	3.6	13.5	11
Multi-Bits/Cell	1-4 b/cell	1-2.3 b/cell	None	None	None
Read Energy (pJ/bit)	1.0 (1 b/cell) 1.1 (2 b/cell)	1.8 (1 b/cell)	1.0* (1 b/cell)	1.5 (1 b/cell)	2.2 (1 b/cell)
Read Time (1 b/cell)	12 ns	23 ns	10 ns	10 ns	9 ns
Write Energy (nJ/bit)	0.4 (1 b/cell) 1.2 (2 b/cell)	N/A (off-chip)	N/R	N/R	N/R
Write Speed (Mbps)	12.4 (1 b/cell) 3.8 (2 b/cell)	N/A (off-chip)	1.8 (1 b/cell)	N/R	N/R
Density (bit/F^2) (with peripherals)	5.6e-3 (1 b/cell) 1.3e-2 (2 b/cell)	6.8e-5† (1 b/cell) 1.4e-4† (2.3 b/cell)	6.3e-4†	2.5e-3†	4.27e-3†

*Pre-silicon simulation, †Layout-estimated, N/R: not reported

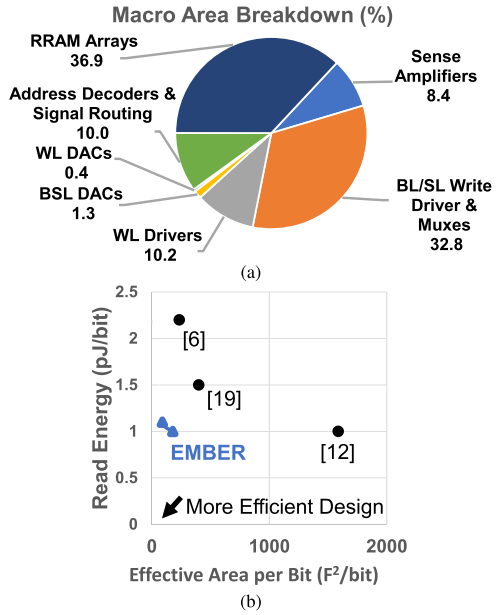


Fig. 15. (a) Macro area breakdown by percentage. (b) Read energy versus unit cell area for EMBER and prior RRAM macros.

we achieve 3.8 Mbps write speed at 1.2 nJ/bit, with BER of < 3e-3 for cells before any cycling. The write energy breakdown is given in Fig. 14. Other level allocations exist that trade off bandwidth for BER, shown in Table I.

C. Comparison With Other Macro Designs

A comparison between EMBER and the prior state-of-the-art is in Table II. EMBER achieves the highest effective density (normalized for process scaling), with an improvement of 31% over the state-of-the-art for 1 b/cell and 204% for 2 b/cell. At the same time, the macro matches the state-of-the-art in read energy per bit [Fig. 15(b)]. Compared with the foundry-provided RRAM macro in [18] in the same technology and with the same array size (64k × 48) that achieves a write bandwidth of 6.5 Mbps for 1 b/cell, EMBER shows a 90% improvement for 1 b/cell, and a 41% penalty for 2 b/cell. Compared with foundry-provided macro's 5.0 nJ/bit write energy for 1 b/cell, EMBER shows a ∼12× improvement for 1 b/cell and ∼4× improvement for 2 b/cell. EMBER achieves BERs that are sufficient for BCH error correction (<3e-3) for both 1–2 b/cell (the same error correction scheme supported by the foundry-provided macros). The write bandwidth and energy improvements in EMBER are mainly a result of the improved peripheral circuitry and optimized conductance level allocation.

VI. CONCLUSION

The EMBER macro showcases several circuit techniques to reduce RRAM energy and latency while increasing the read/write bandwidth and effective array density. On top of these analog design techniques, the macro's digital controller provides on-chip waveform generation to enable efficient multiple-bits-per-cell level allocation using BWA-PBA. Future work could focus on improving read/write speed and reliability further by 1) targeting higher clock speeds; 2) adjusting write-verify programming settings; and 3) increasing endurance by tuning conductance level allocations further.

APPENDIX

A. Write Path Area Optimization

The total driver and pass gate transistor area for the SET operation is expressed as

$$\sum \text{Area}_{\text{SET}} = \left(A_{\text{NMOS,Drv,SL}} + A_{\text{PMOS,Drv,BL}}\right) + \frac{3A_{\text{shifter}}}{2} + N_{\text{pg/drv}} \times \left(A_{\text{NMOS,Pgate,SL}} + A_{\text{PMOS,Pgate,BL}} + \frac{A_{\text{shifter}}}{2}\right). \tag{8}$$

Substituting in the optimization parameters from (1) and (2) results in the following:

$$\sum \text{Area}_{\text{SET}} = \frac{(1+\Gamma)}{\alpha\lambda/(1+\lambda)}\left(\frac{X_n}{\Gamma} + X_p\right) + \frac{3A_{\text{shifter}}}{2} + N_{\text{pg/drv}}\left(\frac{(1+\Gamma)}{\alpha/(1+\lambda)}\left(\frac{X_n}{\Gamma N_{\text{BL/SL}}} + X_p\right) + \frac{A_{\text{shifter}}}{2}\right) \tag{9}$$

$$X_{n/p} = \frac{L_{n/p}^2 V_{\text{cell}} G_{\text{cell}}}{V_{\text{out,DAC}}\ \mu_{n/p} C_{\text{ox}} (V_{\text{DD,IO}} - |V_{\text{Tn/p}}|)} \tag{10}$$

$$\alpha = 1 - \frac{\sum R_{\text{wire}}}{\sum R_{\text{path,write}}} - \frac{V_{\text{cell}}}{V_{\text{out,DAC}}}. \tag{11}$$

The expression for the transistor area of the RESET operation is the same, except X_n and X_p are swapped (different pull-up/down transistors in BL/SL drivers and pass gates are toggled for SET/RESET). $N_{\text{pg/drv}}$ describes the BL muxing ratio for each BL driver, $N_{\text{BL/SL}}$ describes the ratio of BLs to SLs in a given unit cell array (2 for this design), and A_{shifter} is the area of level shifters used in the write drivers and pass gates. The minimum design rule spacing between driver and pass gate transistors is not accounted for in the SET/RESET path area estimates. Equation (10) relates parameters X_p and X_n to I/O FET minimum channel length $L_{n/p}$, the cell

voltage drop necessary for SET/RESET (V_{cell}) at maximum conductance G_{cell}, simulated conductivity parameters $\mu_{p/n}C_{\text{ox}}$, and switch overdrive voltage assuming triode operation with small V_{DS}. The parameter α accounts for the portion of BL/SL DAC voltage allotted to the pass gate and driver switches in the write path, with $(\sum R_{\text{wire}}/\sum R_{\text{path,write}})$ assigned to 15% for this work [see (11)]. The post-layout wire resistance aligns with the 15% assumption.

B. Sense Amplifier Energy

The BL voltage transient is approximated to a form similar to [13] by treating the current through G_{RRAM} and G_{bleed} like a fixed current source

$$V_{\text{BL}}(t) \approx V_{\text{CL}} + n\ V_t \ln\left(\frac{I_{\text{DO}}}{V_{\text{read}}(G_{\text{RRAM}}+G_{\text{bleed}})}\left(1-\exp\left(\frac{-t}{\tau_{\text{settling}}}\right)\right)\right). \tag{12}$$

The energy spent in phase 1 and phase 2 of sense amplifier operation can be found by summing the energy spent during current settling and steady-state for each phase

$$\begin{aligned} E_{\text{phase}}(T_{\max}) \approx & V_{\text{SA}} \int_0^{t_{\text{settling}}} I_{\text{DO}} \exp\left(\frac{V_{\text{CL}} - V_{\text{BL}}(t)}{nV_t}\right) dt \\ & + \max(T_{\max} - t_{\text{settling}}, 0) \\ & \times V_{\text{SA}} V_{\text{read}}(G_{\text{cell}} + G_{\text{bleed}}) \\ & + T_{\max} V_{\text{SA}} V_{\text{read}}(G_{\text{ref}} + G_{\text{bleed}}). \end{aligned} \tag{13}$$

The latching energy in Phase 3 can be simplified to

$$E_{\text{latch}} = C_{\text{out}} V_{\text{SA}}^2. \tag{14}$$

ACKNOWLEDGMENT

The authors would like to thank TSMC for chip fabrication.

REFERENCES

[1] H.-S. P. Wong et al., "Metal-oxide RRAM," *Proc. IEEE*, vol. 100, no. 6, pp. 1951–1970, Jun. 2012.

[2] B. Q. Le et al., "RADAR: A fast and energy-efficient programming technique for multiple bits-per-cell RRAM arrays," *IEEE Trans. Electron Devices*, vol. 68, no. 9, pp. 4397–4403, Sep. 2021.

[3] C. Wang et al., "Relaxation effect in RRAM arrays: Demonstration and characteristics," *IEEE Electron Device Lett.*, vol. 37, no. 2, pp. 182–185, Feb. 2016.

[4] J.-H. Yoon, M. Chang, W.-S. Khwa, Y.-D. Chih, M.-F. Chang, and A. Raychowdhury, "A 40-nm 118.44-TOPS/W voltage-sensing compute-in-memory RRAM macro with write verification and multi-bit encoding," *IEEE J. Solid-State Circuits*, vol. 57, no. 3, pp. 845–857, Mar. 2022.

[5] L. R. Upton et al., "EMBER: A 100 MHz, 0.86 mm^2, multiple-bits-per-cell RRAM macro in 40 nm CMOS with compact peripherals and 1.0 pJ/bit read circuitry," in *Proc. IEEE 49th Eur. Solid State Circuits Conf. (ESSCIRC)*, Sep. 2023, pp. 469–472.

[6] C.-C. Chou et al., "An N40 256KX44 embedded RRAM macro with SL-precharge SA and low-voltage current limiter to improve read and write performance," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 478–480.

[7] L. Goux et al., "On the gradual unipolar and bipolar resistive switching of TiN/HfO_2/Pt memory systems," *Electrochemical Solid-State Lett.*, vol. 13, no. 6, p. G54, Apr. 2010, doi: 10.1149/1.3373529.

[8] F. Nardi, S. Larentis, S. Balatti, D. C. Gilmer, and D. Ielmini, "Resistive switching by voltage-driven ion migration in bipolar RRAM—Part I: Experimental study," *IEEE Trans. Electron Devices*, vol. 59, no. 9, pp. 2461–2467, Sep. 2012.

[9] T. F. Wu et al., "A 43pJ/cycle non-volatile microcontroller with 4.7 μs shutdown/wake-up integrating 2.3-bit/cell resistive RAM and resilience techniques," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 226–228.

[10] B. Q. Le et al., "Resistive RAM with multiple bits per cell: Array-level demonstration of 3 bits per cell," *IEEE Trans. Electron Devices*, vol. 66, no. 1, pp. 641–646, Jan. 2019.

[11] E. R. Hsieh et al., "High-density multiple bits-per-cell 1T4R RRAM array with gradual SET/RESET and its effectiveness for deep learning," in *IEDM Tech. Dig.*, Dec. 2019, p. 35, doi: 10.1109/IEDM19573.2019.8993514.

[12] E. R. Hsieh et al., "A 3.6 Mb 10.1 Mb/mm^2 embedded non-volatile ReRAM macro in 22 nm FinFET technology with adaptive forming/set/reset schemes yielding down to 0.5 V with sensing time of 5 ns at 0.7 V," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*. IEEE, 2019, pp. 212–214.

[13] H. Rapakko and J. Kostamovaara, "On the performance and use of an improved source-follower buffer," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 54, no. 3, pp. 504–517, Mar. 2007.

[14] M.-F. Chang et al., "Challenges and circuit techniques for energy-efficient on-chip nonvolatile memory using memristive devices," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 5, no. 2, pp. 183–193, Jun. 2015.

[15] S. Ambrogio, S. Balatti, V. McCaffrey, D. C. Wang, and D. Ielmini, "Noise-induced resistance broadening in resistive switching memory—Part I: Intrinsic cell behavior," *IEEE Trans. Electron Devices*, vol. 62, no. 11, pp. 3805–3811, Nov. 2015.

[16] A. Levy, "Multi-pole NEM relays and multiple-bits-per-cell RRAM for efficient 3-D ICs," Ph.D. thesis, Dept. Elect. Eng., Stanford Univ., Stanford, CA, USA, Mar. 2024. [Online]. Available: https://searchworks.stanford.edu/view/in00000063124

[17] A. Wei et al., "PBA: Percentile-based level allocation for multiple-bits-per-cell RRAM," in *Proc. IEEE/ACM Int. Conf. Comput. Aided Design (ICCAD)*, Oct. 2023, pp. 1–9.

[18] K. Prabhu et al., "CHIMERA: A 0.92-TOPS, 2.2-TOPS/W edge AI accelerator with 2-MByte on-chip foundry resistive RAM for efficient training and inference," *IEEE J. Solid-State Circuits*, vol. 57, no. 4, pp. 1013–1026, Apr. 2022.

[19] C.-C. Chou et al., "A 22 nm 96KX144 RRAM macro with a self-tracking reference and a low ripple charge pump to achieve a configurable read window and a wide operating voltage range," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2020, pp. 1–2.

Akash Levy (Graduate Student Member, IEEE) received the B.S.E. degree from Princeton University, Princeton, NJ, USA, in 2018, and the M.S. and Ph.D. degrees from Stanford University, Stanford, CA, USA, in 2020 and 2023, respectively, all in electrical engineering.

His research has focused on integrating emerging nanotechnologies, such as resistive RAM and nanoelectromechanical relays, into digital ICs for improved efficiency.

Dr. Levy received the NSF Graduate Research Fellowship in 2018.

Luke R. Upton (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering from Purdue University, West Lafayette, IN, USA, in 2018, and the M.S. degree from Stanford University, Stanford, CA, USA, in 2020, where he is currently pursuing the Ph.D. degree in electrical engineering, under the supervision of Boris Murmann and H.-S. Philip Wong.

His research focuses on designing efficient read/write circuitry and controllers for resistive memory macros as well as modeling how factors, such as cell relaxation and impact macro performance.

Michael D. Scott received the B.S. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 1999, and the M.S. degree in electrical engineering from the University of California, Berkeley, CA, USA, in 2002.

He is the Founder of a private consulting firm delivering custom circuit design and support for a variety of client companies, with particular emphasis on high performance data converters.

Dennis Rich (Graduate Student Member, IEEE) received a dual B.S. degree in electrical engineering and physics from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 2019, and the M.S. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 2021. He is currently pursuing the Ph.D. degree in electrical engineering with the Robust Systems Group, Stanford University.

He researches methods for cooling 3-D ICs, combining new materials and physical design to enable new scaled architectures. He has also researched cleanroom processes for controlling stress in wafers to create thin films, carbon nanotube fabrication, and the nanophysics of water evaporating from porous surfaces. He has also worked at Cerebras Systems investigating cooling approaches for wafer-scale systems.

Mr. Rich is an NSF Graduate Research Fellow, an Edward J. McCluskey Graduate Fellow, and a Goldwater Scholar.

Win-San Khwa (Senior Member, IEEE) received the B.S. degree from the University of California at Los Angeles, Los Angeles, CA, USA, in 2007, the M.S. degree from the University of Michigan, Ann Arbor, MI, USA, in 2010, and the Ph.D. degree in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 2017.

In 2012, he joined Macronix International (MXIC), Hsinchu, while pursuing his Ph.D. degree. He is currently the Technical Manager with the Corporate Research Design Solution Department, Taiwan Semiconductor Manufacturing Company (TSMC), Hsinchu, on emerging memory path finding and IP development. His research interests include circuit-device optimization designs of emerging memories for artificial intelligence applications.

Dr. Khwa serves as a TPC member for CICC and DAC in 2021 and 2024, respectively.

Yu-Der Chih (Member, IEEE) received the B.S. degree in physics from National Taiwan University, Taipei, Taiwan, in 1988, and the M.S. degree in electronics engineering from National Tsing Hua University, Hsinchu, Taiwan, in 1992.

From 1992 to 1997, he was a Design Engineer of an Ethernet transceiver circuit for data communication and a Circuit Design Engineer for SDRAM with Taiwan Semiconductor Manufacturing Company (TSMC), Hsinchu. In 1997, he joined TSMC, for the development of embedded nonvolatile memory IP, including embedded flash, one-time programmable (OTP), multiple-times-programmable (MTP), and emerging memory. He is currently a TSMC Academician and the Director of the Embedded Nonvolatile Memory Library Department, Memory Solution.

Meng-Fan Chang (Fellow, IEEE) received the M.S. degree from Pennsylvania State University, State College, PA, and the Ph.D. degree from National Chiao Tung University, Hsinchu, Taiwan.

He is currently a Distinguished Professor with National Tsing Hua University, Hsinchu, and the Director of Corporate Research, TSMC, Hsinchu. His research interests include circuit design for volatile and nonvolatile memory, ultralow-voltage systems, 3-D memory, circuit–device interactions, spintronic circuits, memristor logics for neuromorphic computing, and computing-in-memory for artificial intelligence.

Dr. Chang was a recipient of several prestigious National-Level Awards in Taiwan, including the Outstanding Research Award from MOST-Taiwan, the Outstanding Electrical Engineering Professor Award, the Academia Sinica Junior Research Investigator Award, and the Ta-You Wu Memorial Award. He received recognition as a top-10 contributor of papers to ISSCC over the past 70 years in 2023. He has been serving as an Associate Editor for the IEEE JOURNAL OF SOLID-STATE CIRCUITS, IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (TVLSI) System, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS, and IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS and a Guest Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: EXPRESS BRIEFS and IEEE JOURNAL ON EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS. He has also been serving on the Executive Committee for IEDM, as well as the Subcommittee Chair for ISSCC, IEDM, DAC, ISCAS, VLSI-DAT, and ASP-DAC. He was a Distinguished Lecturer for the IEEE Solid-State Circuits Society (SSCS) and Circuits and Systems Society (CASS) as well as the Chair of the Nano-Giga Technical Committee of CASS, SSCS Taipei Chapter, and IEEE Taipei Section and an Administrative Committee Member of the IEEE Nanotechnology Council.

Subhasish Mitra (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA.

He is William E. Ayer Professor with the Departments of Electrical Engineering and Computer Science at Stanford University. He is also the Associate Chair (Faculty Affairs) of Computer Science. He directs the Stanford Robust Systems Group, leads the Computation Focus Area of the Stanford SystemX Alliance, and is a member of the Wu Tsai Neurosciences Institute. His research ranges across Robust Computing, Nanosystems, Electronic Design Automation (EDA), and Neurosciences. Results from his research group have influenced almost every contemporary electronic system, and have inspired significant government and research initiatives in multiple countries. He has held several international academic appointments—the Carnot Chair of Excellence in Nanosystems at CEA-LETI, France, an Invited Professor at EPFL in Switzerland, and a Visiting Professor at the University of Tokyo in Japan. He also has consulted for major technology companies including Cisco, Google, Intel, Samsung, and Xilinx. In the field of Robust Computing, he has created many key approaches for circuit failure prediction, online diagnostics, QED system validation, soft error resilience, and X-compact test compression. Their adoption by industry is growing rapidly, in markets ranging from cloud computing to automotive systems. His X-compact approach has proven essential for cost-effective manufacturing and high-quality testing of almost all 21st-century systems, enabling billions of dollars in cost savings.

Dr. Mitra is an ACM Fellow and a Distinguished Alumnus of the Indian Institute of Technology at Kharagpur, Kharagpur. With his students and collaborators, he demonstrated the first carbon nanotube computer. They also demonstrated the first 3-D Nanosystem with computation immersed in data storage and these received wide recognition: cover of NATURE, Research highlighted to the U.S. Congress by the NSF and highlighted as "important scientific breakthrough" by global news organizations. His honors include the Harry H. Goode Memorial Award (by the IEEE Computer Society for outstanding contributions in the information processing field), Newton Technical Impact Award in EDA (test-of-time honor by ACM SIGDA and IEEE CEDA), the University Researcher Award (by the Semiconductor Industry Association and Semiconductor Research Corporation to recognize lifetime research contributions), the Intel Achievement Award (Intel's highest honor), and the U.S. Presidential Early Career Award. He and his students

have published over ten award-winning papers across five topic areas (technology, circuits, EDA, test, and verification) at major venues including the Design Automation Conference, International Solid-State Circuits Conference, International Test Conference, Symposium on VLSI Technology, Symposium on VLSI Circuits, and Formal Methods in Computer-Aided Design. Stanford undergraduates have honored him several times "for being important to them."

Boris Murmann (Fellow, IEEE) received the Dipl.-Ing. (FH) degree in communications engineering from Fachhochschule Dieburg, Dieburg, Germany, in 1994, the M.S. degree in electrical engineering from Santa Clara University, Santa Clara, CA, USA, in 1999, and the Ph.D. degree in electrical engineering from the University of California at Berkeley, Berkeley, CA, USA, in 2003.

From 1994 to 1997, he was with Neutron Mikrolektronik GmbH, Hanau, Germany, where he was involved in the development of low-power and smart-power application-specific integrated circuits (ASICs) in automotive CMOS technology. From 2004 to 2023, he was with the Department of Electrical Engineering, Stanford University, Stanford, CA, USA, where he served as an Assistant Professor, an Associate Professor, and a Full Professor. He is currently with the Department of Electrical and Computer Engineering, University of Hawai'i at Manoa, Honolulu, HI, USA. His research interests include the area of mixed-signal integrated circuit design, with an emphasis on data converters, sensor interfaces, and circuits for embedded machine learning.

Dr. Murmann was a co-recipient of the Best Student Paper Award at the Very Large-Scale Integration Circuits Symposium in 2008 and 2021, the Best Invited Paper Award at the IEEE Custom Integrated Circuits Conference (CICC) in 2008, the Agilent Early Career Professor Award in 2009, the Friedrich Wilhelm Bessel Research Award in 2012, and the SIA-SRC University Researcher Award for lifetime research contributions to the U.S. semiconductor industry in 2021. He was the 2017 Program Chair of the IEEE International Solid-State Circuits Conference (ISSCC) and the 2023 General Co-Chair of the IEEE International Symposium on Circuits and Systems (ISCAS).

Priyanka Raina received the B.Tech. degree in electrical engineering from Indian Institute of Technology at Delhi, New Delhi, India, in 2011, and the M.S. and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2013 and 2018, respectively.

She was a Visiting Research Scientist with NVIDIA Corporation, Santa Clara, CA, USA, in 2018. She is currently an Assistant Professor of electrical engineering with Stanford University, Stanford, CA, USA, where she works on domain-specific hardware architectures and agile hardware–software codesign methodology.

Dr. Raina is a 2018 Terman Faculty Fellow. She was a co-recipient of the Best Demo Paper Award at VLSI 2022, the Best Student Paper Award at VLSI 2021, the IEEE JOURNAL OF SOLID-STATE CIRCUITS (JSSC) Best Paper Award in 2020, the Best Paper Award at MICRO 2019, and the Best Young Scientist Paper Award at ESSCIRC 2016. She has won the Sloan Research Fellowship in 2024, the National Science Foundation (NSF) CAREER Award in 2023, the Intel Rising Star Faculty Award in 2021, and the Hellman Faculty Scholar Award in 2019. She was the Program Chair of the IEEE Hot Chips in 2020. She serves as an Associate Editor for the IEEE JOURNAL OF SOLID-STATE CIRCUITS and IEEE SOLID-STATE CIRCUITS LETTERS.