STAGGER: Enabling All-in-One Subarray Sensing for Efficient Module-level Processing in Open-Bitline ReRAM

Chengning Wang, Dan Feng*, Yuchong Hu, Wei Tong and Jingning Liu
Wuhan National Laboratory for Optoelectronics, Key Laboratory of Information Storage Systems, Engineering Research Center of Data Storage Systems and Technology, Ministry of Education of China (School of Computer Science and Technology, Huazhong University of Science and Technology), Wuhan, China.
{cnwang, dfeng, yuchonghu, tongwei, jnliu}@hust.edu.cn *Corresponding Author

ABSTRACT

Emerging resistive RAM (ReRAM) devices can in-situ execute vector-matrix-multiplication (VMM) and is able to achieve energy-efficient in-memory scientific computing. However, the peripheral separated S&Hs and ADCs for row buffering and sensing in conventional designs are the system bottleneck. We propose an ADC-less all-in-one processing-in-ReRAM design that enables the precharge once, readout multiple-bits (PORM) functionality for overlapping the tRCD latencies of different-significance result bits sensed out on the same bitline. Specifically, to support PORM sensing, we propose a cascaded-feedback bitline sensing architecture for VMM and a buffering-and-sensing-collocated sense amplifier elemental design with the bitline and the storage node fully decoupled for enabling conflict-free column accesses. We further propose cross-level interleaving mechanism for successive column VMM accesses to reduce the overall latency through improving the hardware spatiotemporal utilization. Experimental results show that our proposed design achieves 297% overall performance improvement and 85.8% energy reduction, compared with an aggressive baseline.

CCS CONCEPTS

• **Hardware → Analysis and design of emerging devices and systems**; **Memory and dense storage**.

KEYWORDS

interconnect parasitic effects, *RC* latency, resistive devices

ACM Reference Format:
Chengning Wang, Dan Feng*, Yuchong Hu, Wei Tong and Jingning Liu. 2024. STAGGER: Enabling All-in-One Subarray Sensing for Efficient Module-level Processing in Open-Bitline ReRAM. In *61st ACM/IEEE Design Automation Conference (DAC '24), June 23–27, 2024, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3649329.3655968

1 INTRODUCTION

Emerging resistive RAM (ReRAM) can in-situ perform vector-matrix multiplication (VMM) in cell-arrays and is promising to achieve energy-efficient iterative solution of linear systems derived from mathematical physics problems, including electromagnetics, power network, and semiconductor device and circuit simulation [1–4, 15, 18]. However, the bitline current accumulation imposes significant area and latency overheads of VMM result buffering and sensing when multiple rows are activated simultaneously [21], which has become the system bottleneck. Previous works [3, 14, 15, 18] used sample-and-hold (S&H) circuits for each bitline to buffer the output results, and set multiplexed analog-to-digital converters (ADC) to fetch the result from the S&Hs for sensing, as shown in Figure 1(a). This separated design enlarges the latency and energy overhead of bitline sensing. Besides, the ADCs occupy a large portion of bitline-side pitches, due to the non-custom-optimized general-purpose ADC used in previous works [3, 14, 15]. As a result, at least 8 or more bitlines have to share a ADC for sensing [12, 21], which significantly limits the parallelism and performance of column accesses.

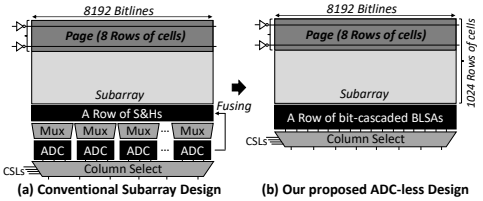

Figure 1: Fusing S&Hs and ADCs into Bit-Cascaded BLSAs

In this work, we try to leverage the elemental memory-native sense amplifier (SA) that can buffer a single bit while performing sensing, to construct a ADC-less buffering-sensing-collocated all-in-one sensing architecture for executing VMM and fully eliminate the separation of peripheral S&Hs and ADCs. However, the *challenges* of all-in-one subarray sensing are nontrivial. First, to only occupy a unit column pitch to guarantee the sensing parallelism, the layout of different-significance SAs dangling on the bitline should be folded rather than expanded. Second, during the sensing process, each SA should exactly buffer the result bits in their storage nodes (SN) for subsequent column accesses to fetch, while using the least number of SAs. Third, the interplay between the bitline and the storage node should be treated cautiously. For conventional SA design, the MSB SA restores and destructs the developed bitline voltage when the SA is enabled. This bitline voltage restoration makes it impossible for the subsequent CSB SA and LSB SA to sense the bitline in the same row cycle. In this scenario, three row cycles have to be issued to sense the MSB, CSB and LSB of bitline results, which significantly prolongs the overall latency. Also, precharging the bitlines destroys the data in the SA storage nodes, therefore the column accesses cannot be parallelized with the bitline precharging process.

To fundamentally solve the above challenges, we design a bit-cascaded bitline-SN-decoupled sensing architecture that fully enables the precharge once, readout multiple (PORM) functionality

DAC '24, June 23–27, 2024, San Francisco, CA, USA

ACM ISBN 979-8-4007-0601-1/24/06…$15.00
https://doi.org/10.1145/3649329.3655968

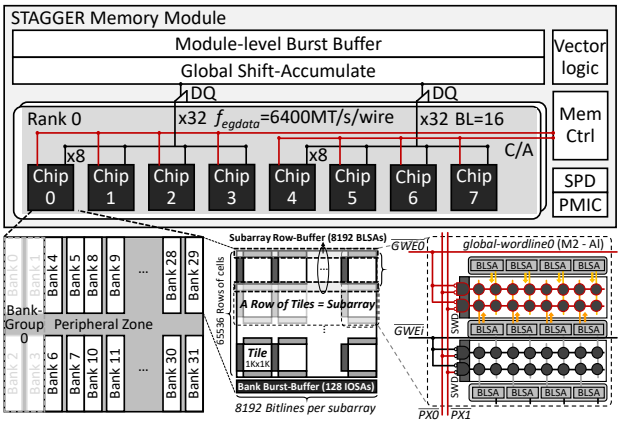

Figure 2: Overall STAGGER Memory Modular Hierarchy

for VMM, to overlap the tRCD latencies of MSB, CSB and LSB. With the bit buffer co-located in the SAs, the MSB column accesses can be interleaved with the bitline CSB and LSB sensing, which significantly improves the hardware spatial efficiency. Furthermore, the bitline precharging can be parallelized with column accesses by the decoupling, which further saves the row cycle time. Specifically, we make the following contributions.

- We propose a CF bitline sensing architecture that enables parallelization of column accesses with lower bits sensing.
- We propose a bitline-SN-decoupled SA elemental design that enables the PORM functionality in a single row cycle.
- We present SA-centric cross-level interleaving scheme to improve hardware utilization by exploiting open-bitline layout.
- We systematically evaluate our proposed design with performance improved by 297% and energy reduced by 85.8%.

2 BACKGROUND AND MOTIVATION

2.1 Open-Bitline ReRAM Organization

Generally, ReRAM refers to any memory technology that uses nonvolatile resistance to store information. Here we focus on a subcategory called metal-oxide ReRAM that has a MIM device structure [7]. Figure 2 shows the overall ReRAM modular hierarchy. ReRAM is composed of banks that can be independently accessed in parallel [10]. A bank is divided into tiles to reduce the RC latency of long wires [6, 17]. A tile comprises the periphery and a cell-array, where each cell uses the high-conductance state to store one and the low-conductance state to store zero. A row of tiles arranges in a subarray. Here, a row refers to a row of cells, and a row of bitline sense amplifier(BLSA) storage nodes is called a subarray row-buffer. Multiple rows (usually size of eight [13]) in a subarray that are simultaneously activated for executing VMM is called a page.

Figure 3 shows the hexagonal layout of open-bitline ReRAM where cylindrical ReRAMs are stacked on top of storage-element contacts over bitlines [22], with the top metal electrodes connected to the common ground plate. The open-bitline architecture features an interdigitated layout of odd and even bitlines to the top and bottom BLSAs at the edges of the subarray respectively, and each BLSA connects to two bitlines from two adjacent subarrays. The open-bitline architecture can achieve $6F^2$ cell area, which is 25% denser than the $8F^2$ folded-bitline architecture [20], where F is the

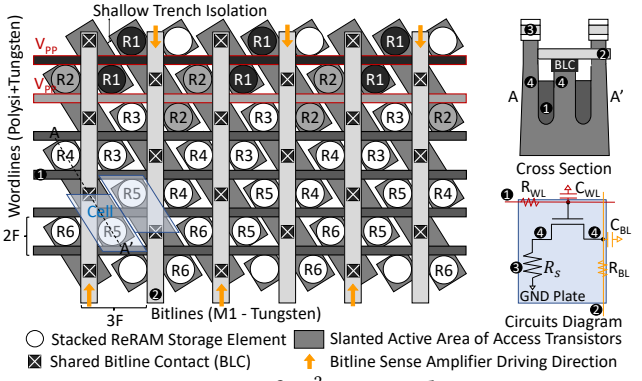

Figure 3: Layout of $6F^2$ Open-Bitline ReRAM

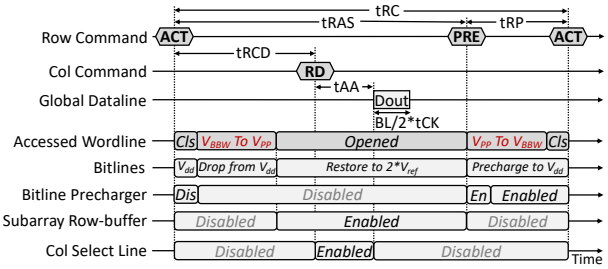

Figure 4: Read Access Timing of Open-Bitline ReRAM

feature size. For bipolar write accesses, the bitlines are biased at negative SET voltage and positive RESET voltage. For read accesses, the bitlines are first precharged to a positive read voltage (V_{dd_RD}) to prepare for the voltage development by floating the bitline.

2.2 Memory Access Commands and Timing

We briefly overview the controller-side memory row access and column access commands and timing parameters that guarantee the analog signal integrity [8, 10], as shown in Fig. 4.

1) Row ACTIVATE on Subarray (ACT). The ACT command with a row address first disables the precharger to let the bitlines floating and starts to open the wordline from V_{BBW} to V_{PP}, and the precharged bitline voltage starts to drop due to the discharging of bitline parasitic capacitance through the opened cell to the ground. Then the subarray row-buffer (BLSAs) enables, to sense and fetch the content of the whole row of cells into the row-buffer storage nodes. The latency from the issue of ACT command to the subarray row-buffer stabilization is called the $\overline{\text{RAS}}$-to-$\overline{\text{CAS}}$ Delay time (tRCD), and the time interval from the issue of ACT command to the issue of the PRE command is called the $\overline{\text{RAS}}$ time (tRAS).

2) Row PRECHARGE on Subarray (PRE). The PRE command with a row address starts to close the wordline already opened (V_{PP}) back to the completely closed state (V_{BBW}). PRE command also enables the precharger to bias the bitlines back to the V_{dd_RD} precharged state for preparation of accessing another row. The latency of the whole process is called the $\overline{\text{RAS}}$ Precharge time (tRP). The tRAS plus tRP together is called the Row Cycle time (tRC).

3) Column READ on Periphery (RD). When the row-buffer storage nodes and bitlines are sufficiently stabilized during activation, the RD command is issued with a column address to enable a specific column select line (CSL) and transfer a strip of data (typically 128 bits) *from the subarray row-buffer* [11] via local datalines (LDL) to

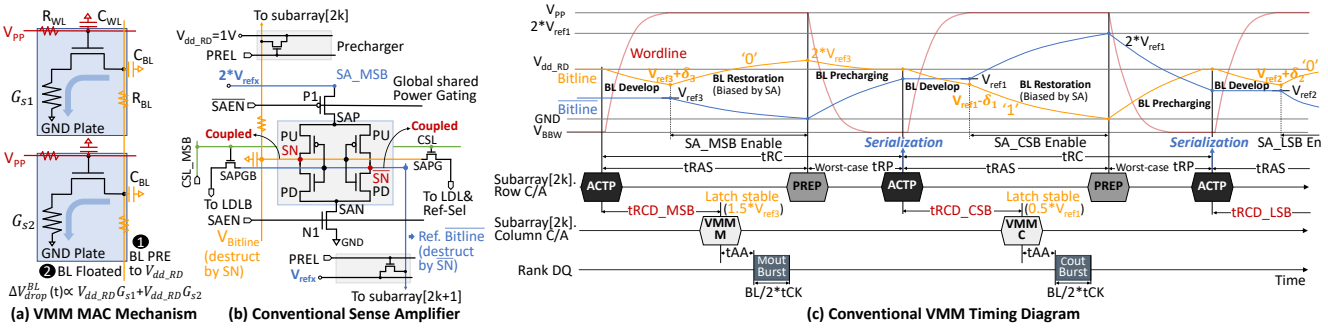

Figure 5: Conventional Design. (a) Sensing Mechanism, (b) SA Structure, and (c) Subarray VMM Access Timing.

write into the bank burst buffer (BBB) made up of 128 I/O Sense Amps (IOSA). Then, the BBB outputs the data with eight global I/O datalines (GDL) in the burst mode with a burst length (BL) of 16, at both the rising and falling edges of the clock. Thus the data rate of the external GDL is 32 times of that of the internal core clock frequency. The latency from issuing the RD command until the BBB stabilized and is going to output the first data bit into GDL is called the read $\overline{\text{CAS}}$ Latency time (tAA [8]), and the burst latency of GDL is calculated as BL/2*tCK, where tCK is the GDL clock cycle.

When we take binary voltage as wordline input to the gates of access transistors [12] and one-bit-per-cell storage configuration, VMM operation is de facto a read access. We slightly modify row ACT and PRE commands as follows and introduce VMM as a column command to support VMM access.

1) Row ACTP on Subarray. ACTP is the same as ACT except for wordlines with non-zero input in accessed page are opened to V_{PP}.

2) Row PREP on Subarray. PREP is the same as PRE except for wordlines in the accessed page are all closed to the V_{BBW} state.

3) Column VMM on Periphery. The column VMM command is analogous to column RD command that fetches the sensed results of 128 bitlines from subarray row-buffer storage nodes to the BBB and bursts out for global shift-accumulation. Here we introduce three versions of VMM command. VMMM fetches the MSB result, VMMC fetches the CSB result, and VMML fetches the LSB result.

2.3 VMM Operation Sensing Principle

For VMM operations, each bitline executes a column-wise dot product computation. The bitlines are first precharged to the full read voltage. When the wordlines in the page are opened, the charge in the bitline parasitic capacitance leaks through the opened cells, shown in Figure 5(a). Using wordline gate binary input [12], the decay of the bitline voltage (ΔV_{drop}^{BL}) within a given time is determined by the dot product $\sum_i V_i G_{ij}$, where V_i=V_{dd_RD} if $V_{wordline\_i}$=V_{PP}, and V_i=0 if $V_{wordline\_i}$=V_{BBW}, as shown in Figure 5(a). Here, G_{ij} is the cell conductance at the ith wordline and the jth bitline. The bitline voltage decays (i.e. discharging) faster if the bitline dot product result current is larger, yielding a lower post-dropped bitline voltage. If we activate eight adjacent wordlines at a time, we need a three-bit BLSA for each bitline and seven reference voltages in total. For the binary-search-based tiered sensing, the most-significant bit (MSB) is sensed out first, given the 3rd reference level. Then, the central-significant bit (CSB) is sensed out, given the 1st and the 5th reference levels. Finally, the least-significant bit (LSB) is sensed out, given the 0th, 2nd, 4th and 6th reference levels.

2.4 Analysis of Conventional Sensing Procedure

Figure 5 shows the conventional BLSA structure [9, 19, 20]. The bitline and the reference-line are directly connected to the SN and $\overline{\text{SN}}$ of the BLSA, respectively. When the MSB SA is enabled, the MSB SA flips. Unfortunately however, the MSB SA injects charges into the bitline and restores the bitline voltage to either twice of the reference-line voltage or the ground, depending on the comparison result of bitline voltage and the reference-line voltage. Meanwhile, the reference-line voltage will also be restored to the reverse side by the MSB SA. Thus, both the developed bitline voltage and the constant reference-line voltage are destructed by this flipping-based restoration process. This means that when the MSB SA is sufficiently stabilized and is ready for reading out from its storage node, *the subsequent CSB SA and LSB SA cannot be enabled for result sensing in the same voltage development period*, as the developed bitline voltage is already destructed by the MSB SA. This leads to the "precharge once, readout single" limitation for VMM. In this scenario, the bitline voltage should be initialized by the precharge operation after reading out the MSB SA result, then the next activation command could be issued for bitline voltage developing again and then CSB SA sensing. The tRCD latencies of MSB, CSB and LSB have to be separated in multiple row cycles in a serial manner, which makes the hardware underutilized and prolongs the overall sensing latency.

3 STAGGER DESIGN

To solve the above bitline VMM sensing efficiency problem, we propose STAGGER design that encompasses the PORM mechanism and the cross-level interleaving scheme for VMM accesses.

3.1 Precharge Once, Readout Multiple (PORM)

3.1.1 Cascaded-Feedback (CF) Bitline Sensing Architecture. To construct duality bit-exact row buffer and directly store the sensed result bits in the SAs, we propose a cascaded-feedback bitline sensing architecture, shown in Figure 6. The interleaved BLSA is cascaded by three SAs, and each SA latches a single result bit. The latch output from each SA is fed back to select the reference voltage input of lower level SAs. Therefore, we only need three SAs to sense a three-bit result. The MSB SA, CSB SA and LSB SA are enabled one after another. The most important benefit is that by using the cascaded-feedback bit-by-bit buffering-in-SA structure, the MSB VMM column access for fetching the result can be parallelized with the CSB SA sensing, and the busy periods of MSB SA, CSB SA and LSB SA are overlapped, which saves the overall latency, as shown

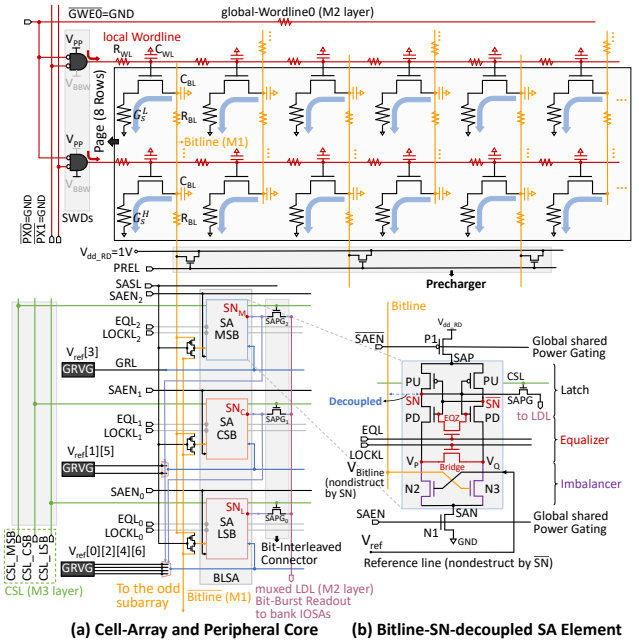

Figure 6: Design of Cell-Array Periphery for VMM

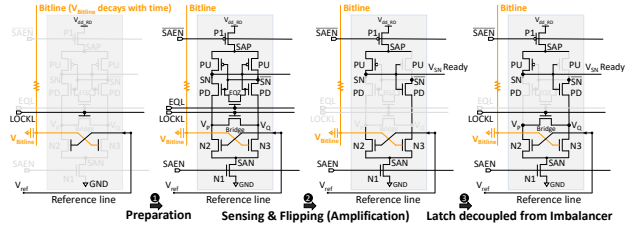

Figure 7: BLSA Multi-Phase Operation and Conformation

in Figure 8. We call this feature as *sense-amplifier-level parallelism*. By comparison, without the CF architecture, all the result bits can be output only after all the bits are sensed.

3.1.2 Bitline-SN-Decoupled SA Elemental Design. To completely isolate the bitline voltage developing process from the SA flipping, we propose a novel decoupled SA for tilt-based VMM sensing, which is inspired by the Wheatstone bridge, as shown in Figure 6(b). The SA is composed of three parts: imbalancer, equalizer and latch. Through the imbalancer, a noise voltage is injected at the right-hand-side foot of the latch when both the EQL and LOCKL are disabled, triggering the latch to flip, as shown in Figure 7. When the storage node of SA reaches a sufficiently stabilized threshold state, the result bit is considered to be stably stored in the latch of SA. At this time, the corresponding lock-line (LOCKL) is enabled, to decouple the latch from the imbalancer. Then the SA conformation is changed, and N2 and N3 transistors are now parallel connected via equilibrium. Thus, the stored content in the SA latch will not be disturbed by the further decaying of bitline voltage, while the bitline voltage can be utilized by lower-significance SA for sensing.

The decoupling functionality is three-fold. *First*, the floated bitline will not be biased by the storage node of the flipping SA when the SA is turned on, thus the developed bitline voltage will not be restored and destroyed by the SA. The followed lower-significant SAs are able to compare the bitline voltage with their reference voltage for sensing in the same row cycle, and the tRCD latencies

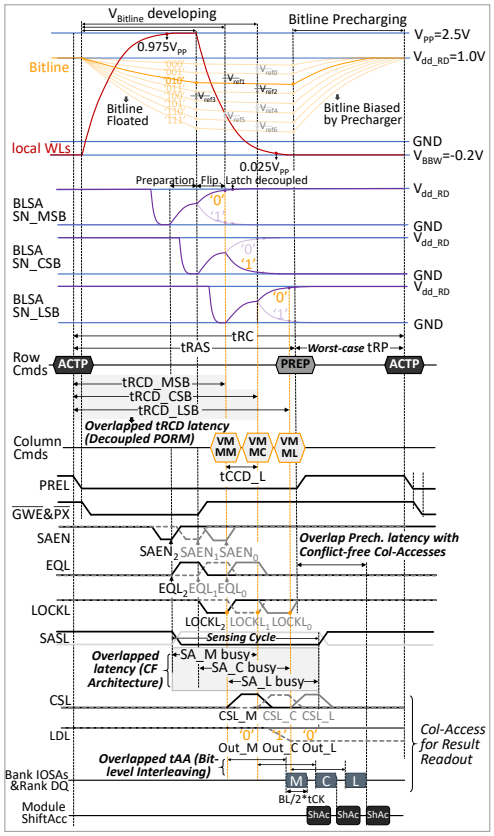

Figure 8: PORM-based Intra-Subarray VMM Access Timing

of different significant bits sensed out on the same bitline can be overlapped, which enables the PORM functionality. *Second*, as the bitline voltage continues to decay with time during bitline voltage development, the bitline voltage may under cross the corresponding reference voltage in the time period if the to-be-sensed result bit is zero. Without the built-in equalizer, the altering of the imbalancer gate input to the SA may reversely pull down the foot of the opposite CMOS inverter arm in the latch and may cause unwanted flip. By the introduction of the Wheatstone bridge as the built-in equalizer, the continous voltage decaying of the floated bitline does not affect the storage nodes of the SA when the bridge is enabled. *Third*, the SA is fully decoupled from the precharger by the imbalancer gate input, and the bitline precharging process will not affect the SA storage node readout process, thus the row buffer storage nodes can be enabled during bitline precharging, and now bitline precharging is able to parallelize with intra-subarray VMM column accesses in a conflict-free manner, as shown in Figure 8.

3.2 SA-Centric Cross-level Interleaving Scheme

3.2.1 Dual-Buffering-based Significant-Bit Interleaving. The SA plus the SA pass-gate (SAPG) transistor together is actually a SRAM cell. We observe that the CSL is actually the SAs' wordline, and the LDL is actually the SAs' bitline. The LDL is first precharged to V_{dd_RD}. To prevent read disturbance and keep the read static noise margin (RSNM) of the SA storage node when the CSL is enabled, the SA storage node should be more decoupled from the LDL, thus the resistance of the SAPG transistor should be much larger than that

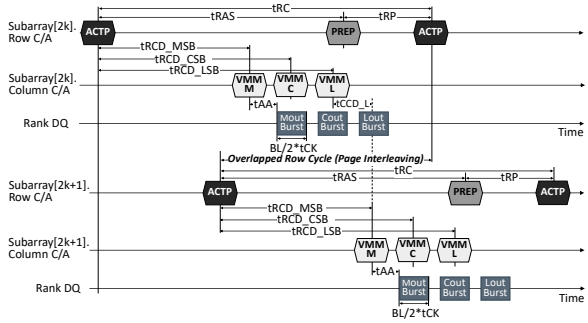

Figure 9: Inter-Subarray Page Interleaving in the Same Bank

of the SA pull-down (PD) transistor, as shown in Figure 6(b). To ensure this, the Width/Length ratio of the SAPG transistor is set to be four times smaller than that of the PD transistor in the SA for our design, i.e., the β ratio of the "SRAM cell" is four. The large channel resistance of the SAPG transistor induces large *RC* latency. The turn-on/turn-off latency of small-size SAPG during "SRAM read access" driven by CSL signal is 3.75 ns, which restrains both the column cycle time tCCD_L and the tAA of a VMM column-access.

To hide the latency of frequently turned-on SAPGs and improve data bus efficiency, we propose a significant-bit interleaving scheme for burst readout of the VMM result bits from the BLSA when executing column VMM accesses. We design a bit-interleaved column-access connector (BIC), as shown in Figure 6(a). The MSB is first transferred out to the corresponding MSB IOSA of the BBB through local datalines and starts to trigger the single-ended IOSA to flip, and at this time, the shared local dataline is freed by disconnecting from the IOSA and ready for the CSB to transfer out to the corresponding CSB IOSA of the BBB. Finally, after the LSB is transferred to BBB, the MSB SA result readout can be started again. The result bits stored in MSB SA, CSB SA and LSB SA are stabilized successively, therefore they can be serially burst out by time-multiplexing a single local dataline. Therefore, with the dual buffers, through SA bit interleaving to the BBB, the VMM column accesses of different significant bits are partially parallelized and the peripheral column access latencies are overlapped, as shown in Figure 8.

3.2.2 Parity-Staggering-based Inter-Subarray Page Interleaving. The open-bitline ReRAM has parity-staggered bitlines without bitline isolation-transistors. In the conventional bitline-SN-connected SA design, the $(2j+1)$th bitline in the $(2k+1)$th subarray is occupied to serve as the voltage reference-line for the $2j$th bitline in the $2k$th subarray, when the jth SA is enabled, and both the bitline voltage and the corresponding bitline voltage are restored by the SA, as shown in Figure 5. Then the bitline cannot be sensed again by the same SA in the same row cycle. This dependent relationship makes it impossible to let the SA switch to the bitline MSB result sensing when the corresponding bitline LSB result sensing is completed.

To further improve the data bus efficiency, we design the inter-subarray page interleaving scheme. We decouple a pair of bitline and bitline by introducing global reference voltage generators (GRVG) and global reference-lines (GRL) in between two adjacent subarrays, as shown in Figure 6(a). In this manner, there is no need for two dummy edge subarrays for constructing reference bitlines. The corresponding bitline is freed out and can be parallelized with the true bitline by multiplexing the bitline and bitline to the SA, using the sense amp select line (SASL) to control, as shown in Figure 6(a). After the LSB of the bitline result is sensed and the LSB SA becomes idle, the SASL is toggled to connect the bitline to the MSB SA for sensing its MSB result. In this way, two page activations falled in adjacent subarrays are interleaved, and the row cycles of two page accesses that fall in two adjacent subarrays in the same bank are overlapped, as shown in Figure 9. Page interleaving enables the burst transfer of the results of both bitline and bitline sharing the same BLSA, which further doubles the data rate of LDL.

Table 1: Device Configuration

Global data bus rate: 6400 MT/s/wire, burst length (BL) = 16, 2 ranks/module, 8 chips/rank, 8 bank-groups/chip, 4 banks/bank-group, bank size: 65536×8192, subarray size: 1024×8192, 7-bit $\overline{\text{GWE}}$, 3-bit $\overline{\text{PX}}$, CSL-selected column size: 128 BLSAs, TaO_x ReRAM [7], G_{on}=7×10^{-4} S, G_{off}=7×10^{-5} S, one-bit-per-cell, binary input, cell-array feature size 14 nm, Tungsten R_l=14.3 Ω, C_l=0.4883 fF, V_{PP}=2.5 V, V_{dd_RD}=1.0 V, V_{BBW}= -0.2 V, V_{ref}[0]-[6]=(0.83 V, 0.725 V, 0.61 V, 0.5 V, 0.38 V, 0.275 V, 0.16 V)
Timing Params: tRCD_MSB = 19.375 ns, tRCD_CSB = 23.75 ns, tRCD_LSB = 28.125 ns, tAA = 11.25 ns, tRP = 14.375 ns, tRAS = 29.0625 ns, tRC = 43.4375 ns, tRRD_S = 1.25 ns, tRRD_L = 1.875 ns

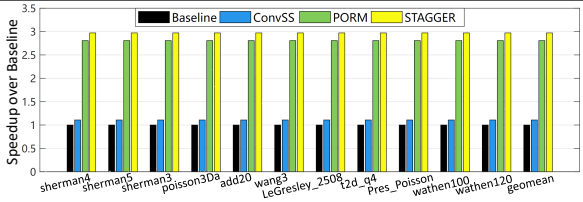

Figure 10: Performance of STAGGER over the Baseline

4 EVALUATION

4.1 Experimental Setup

We generate the HSPICE netlist of TaO_x-based ReRAM cell-array [7] by C-coded script, to determine the latency and power. The zero-biased unselected cells are removed to speed up the simulation, while remaining all the parasitic capacitance and line resistance. The array-level parasitic resistance and capacitance are extracted from [10, 23]. We model the subarray peripheral core including the SWD, precharger, BLSA and connector in HSPICE. We simulate 10 MB 1T-1C eDRAM module burst buffer [5] using CACTI. The global shift-accumulate component, vector logic and control are implemented in Synopsys Design Compiler using TSMC 130 nm cell library and scaled to 14 nm technology size to determine latency, power and area of global periphery. The memory device parameters are listed in Table 1. We evaluate the proposed design by SuiteSparse matrix benchmarks [1] and restructure the Bi-CGSTAB method [16] in MATLAB for processing VMM in cell-arrays with maximum iterations of 15000 for solving linear systems. We implement the double-precision floating-point format similar as [15] and store non-zero matrix blocks in bit-sliced manner. We set ReFloat design [15] under separated S&Hs and ADCs as the Baseline and the bitline-SN-connected conventional subarray sensing (ConvSS) with bit-cascaded SA structure to compare with our proposal.

4.2 Results

4.2.1 Performance Result. Figure 10 shows the performance of STAGGER design compared with the baseline. STAGGER has a uniform

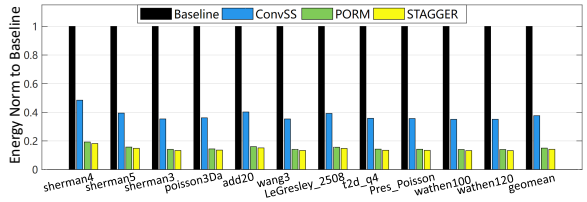

Figure 11: Energy of STAGGER Normalized to the Baseline

performance improvement irrespective to the problem size, since the solution of linear systems is dominated by the $O(n^2)$ VMM kernel operations. The VMM execution process is synchronized among a batch of nonzero block slices, and the latency is determined by the worst-case latency, thus weakening the impact of nonzero pattern on overall performance. Due to the asymmetric inter-group and intra-group bank organization, we used bank-group-first bank interleaving scheme for both the baseline, ConvSS and our design, to overlap the row cycle time for maximizing the parallelism for executing VMM, subject to tRRD_S for inter-group and tRRD_L for intra-group. Overall, PORM improves the performance by 281% and 253% compared with the baseline and ConvSS, and STAGGER improves the performance by 297% and 268% on average compared with the baseline and ConvSS respectively.

4.2.2 Energy Consumption. Figure 11 shows the overall energy results of STAGGER compared with the baseline. The percentage of energy reduction varies across the benchmarks. The energy is the sum of all parts of energy. Although the performance improvement is steady, the energy reduction is not. This is because the power is dynamical, and the transient power consumption varies significantly with block patterns of nonzero elements with a given problem size. PORM reduces the energy by 85% and 60% compared with the baseline and ConvSS, and STAGGER reduces the mean energy by 85.8% and 62.4% compared with the baseline and ConvSS.

4.2.3 Hardware Power and Area Overheads. Compared with ConvSS design, our proposal introduces extra hardware overheads. To support the bitline-SN-decoupled PORM mechanism, four addtional transistors (EQZ, Bridge, N2 and N3) were introduced into each SA, compared with the conventional bitline-SN-connected SA that requires six transistors. To support the cross-level interleaving scheme, we need to set additional two select transistors to pass the to-be-sensed bitline or $\overline{\text{bitline}}$ to the SA. Also, we modified the bank row address decoder to enable two pages in two adjacent subarrays for selecting the corresponding sub-wordline drivers. These totally incur 1.1% extra power and 8.3% extra area overheads.

5 RELATED WORKS

Feinberg et al. [4] studied preconditioning techniques of in-situ VMM operations for linear algebra computation. Feinberg et al. [3] proposed an in-memory scientific computing design with nonzero block optimization. Fan et al. [2] and Wang et al. [18] proposed acceleration-in-memory hardware for algebraic multigrid and conjugate gradient computations. Song et al. [15] proposed an in-memory floating-point iterative linear solver. All these works are based on separated S&Hs and ADCs, which are the fundamental bottleneck of subarray column accesses. Other works such as Xin et al. [19] proposed bitwise logic processing-in-DRAM design leveraging native memory commands, but they do not support VMM memory command as arithmetic kernel for linear algebra.

6 CONCLUSION

Efficient and low-cost subarray sensing is the major challenge for achieving high-performance VMM performed on ReRAM cell-arrays. Based on the observation that the coupling effect between bitlines and sense amplifier storage nodes is the key limiter of the subarray VMM sensing latency, we proposed an all-in-one bitline-SN-decoupled subarray sensing architecture that fully enables the precharge once, readout multiple functionality within a single row cycle. We further proposed cross-level interleaving scheme to improve overall hardware utilization by exploiting the open-bitline layout. Experimental results show that the proposed memory-native subarray VMM sensing mechanism and the corresponding memory timing optimization techniques can significantly improve the overall device performance and efficiency for scientific computing.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China No. 2023YFB4502100; the NSFC under Grant 62202195, Grant 61821003, Grant 61832007; the Pre-Research Project No. 31511090201.

REFERENCES

[1] T. A Davis et al. 2011. The University of Florida sparse matrix collection. *ACM TOMS* 38, 1 (2011), 1–25.
[2] M. Fan et al. 2023. AmgR: Algebraic Multigrid Accelerated on ReRAM. In *Proc. DAC*. 1–6.
[3] B. Feinberg et al. 2018. Enabling scientific computing on memristive accelerators. In *Proc. ISCA*. 367–382.
[4] B. Feinberg et al. 2021. An analog preconditioner for solving linear systems. In *Proc. HPCA*. 761–774.
[5] G. Fredeman et al. 2015. A 14 nm 1.1 Mb embedded DRAM macro with 1 ns access. *IEEE JSSC* 51, 1 (2015), 230–239.
[6] H. Ha et al. 2016. Improving energy efficiency of DRAM by exploiting half page row access. In *Proc. MICRO*. 1–12.
[7] M. Hu et al. 2018. Memristor-based analog computation and neural network classification with a dot product engine. *Adv. Mater.* 30, 9 (2018), 1705914.
[8] JEDEC Solid State Technology Association. Oct. 2021. JEDEC Standard: DDR5 SDRAM. *JESD79-5A* (Oct. 2021), 1–490.
[9] D. Kim et al. 2022. Read disturbance in cross-point phase-change memory arrays—Part II: array simulations considering external currents. *IEEE T-ED* 70, 2 (2022), 521–526.
[10] N. S. Kim et al. 2019. LL-PCM: Low-latency phase change memory architecture. In *Proc. DAC*. 1–6.
[11] D. Lee et al. 2013. Tiered-latency DRAM: A low latency and low cost DRAM architecture. In *Proc. HPCA*. 615–626.
[12] W. Li et al. 2022. A 40-nm mlc-RRAM compute-in-memory macro with sparsity control, on-chip write-verify, and temperature-independent ADC references. *IEEE JSSC* 57, 9 (2022), 2868–2877.
[13] A. Redaelli et al. 2022. *Semiconductor Memories and Systems.* Woodhead Publish.
[14] A. Shafiee et al. 2016. ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars. *ACM SIGARCH Comput. Archit. News* 44, 3 (2016), 14–26.
[15] L. Song et al. 2023. ReFloat: Low-cost floating-point processing in ReRAM for accelerating iterative linear solvers. In *Proc. SC*. 1–15.
[16] H. A Van der Vorst. 1992. Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. *SIAM J. Scientific & Statistical Comput.* 13, 2 (1992), 631–644.
[17] T. Vogelsang. 2010. Understanding the energy consumption of dynamic random access memories. In *Proc. MICRO*. 363–374.
[18] C. Wang et al. 2023. CorcPUM: Efficient processing using cross-point memory via cooperative row-column access pipelining and adaptive timing optimization in subarrays. In *Proc. DAC*. 1–6.
[19] X. Xin et al. 2020. ELP2IM: Efficient and low power bitwise operation processing in DRAM. In *Proc. HPCA*. 303–314.
[20] S. Yu. 2022. *Semiconductor Memory Devices and Circuits.* CRC Press.
[21] S. Yu et al. 2021. Compute-in-memory chips for deep learning: Recent trends and prospects. *IEEE Circuits & Syst. Mag.* 21, 3 (2021), 31–56.
[22] J. Zahurak et al. 2014. Process integration of a 27nm, 16Gb Cu ReRAM. In *Proc. IEDM*. 140–143.
[23] P. Zheng et al. 2017. The anisotropic size effect of the electrical resistivity of metal thin films: Tungsten. *J. Appl. Phys.* 122, 13 (2017).