RRAM-Based High-Precision Analogue Eigenvalue Decomposition Solver

Pushen Zuo[1], Ruiqing Xie[1], Congcong Hong[1], Yuchen Qiu[1], Yubiao Luo[1], Shiqing Wang[1], Yuhang Yang[1], Qishen Wang[1], Zongwei Wang[1,2,3*], Yimao Cai[1,2*] and Zhong Sun[1,2*]
[1]Peking University, Beijing, China, email: wangzongwei@pku.edu.cn; caiyimao@pku.edu.cn; zhong.sun@pku.edu.cn
[2]Beijing Advanced Innovation Center for Integrated Circuits, Beijing 100871, China
[3]YanXin Microelectronics Co., Ltd. (YXME), Shanghai, China

***Abstract*—Eigenvalue decomposition (EVD) serves as a foundational computational kernel for multivariate analysis, enabling a wide range of applications. Its implementation, however, confronts dual constraints: digital von Neumann architectures suffer from O(N³) complexity scaling, while analog matrix computing (AMC) approaches exhibit limited precision. This work presents a fully-AMC high-precision EVD solver using foundry-fabricated 40 nm one-transistor-one-RRAM (1T1R) arrays. Our architecture integrates coarse eigenvalue sweeping with iterative residual refinement, leveraging low-precision generalized inverse (LP-GINV) and high-precision matrix-vector multiplication (HP-MVM). Experimental results demonstrate 10^{-6} eigenvector precision within tens of iterations, enabling lossless principal component analysis (PCA) dimensionality reduction. The solver further enables high-precision solutions for ill-conditioned matrix systems, thus supporting reliable massive MIMO detection under strong channel correlation conditions. Benchmark analyses reveal tens-fold improvements in throughput and energy efficiency over a single NVIDIA H100 GPU core at equivalent precision.**

***Keywords*—eigenvalue decomposition, analog matrix computing, RRAM, high-precision**

I. Introduction

EVD, a fundamental linear algebra problem, underpins critical applications in PCA [1], wireless communications [2], and large language model (LLM) fine-tuning [3] (Fig. 1), demanding efficient high-precision solutions. However, digital implementations face cubic complexity, power-hungry operation, and the von Neumann bottleneck. Recent advances in RRAM-based AMC enable accelerated and energy-efficient matrix inversion and eigenpair solving with O(1) computational latency [4]. Although AMC faces an intrinsic precision bottleneck of device programming errors and analog circuit noise, the recently-developed high-precision matrix inversion (HP-INV) scheme achieves high accuracy through iterative refinement combining low-precision INV and bit-slicing-enhanced HP-MVM, converging to precise results within only a few cycles [5]. However, this success in inversion cannot be straightforwardly extended to eigenpair solving. Prior analog eigenpair (EGV) circuits remain insufficient due to error propagation inherent in sequential deflation and precision bottleneck [1, 6-7] (Fig. 1). This motivates a high-precision EGV solver analogous to HP-INV, yet HP-INV proves inapplicable. AMC-friendly approaches such as inverse iteration with shift for interior eigenvalues introduce ill-conditioned INV that causes HP-INV to diverge.

In this work, we propose HP-EVD, a high-precision eigenvalue decomposition solver using fully analogue matrix operations. Our approach leverages generalized matrix inversion (GINV) as the core computational primitive. EVD

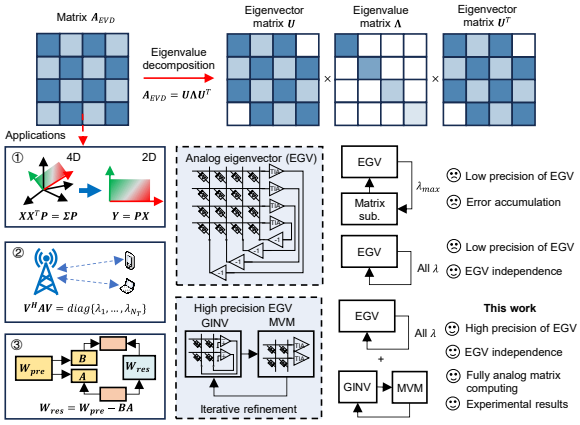

Fig. 1. AMC circuits for solving EVD.

begins with independent coarse eigenvalue sweeping via modified LP-GINV circuits, circumventing error propagation inherent in sequential deflation. Through following iterative refinement combining LP-GINV and bit-slicing-enhanced HP-MVM, the eigenpairs receive reliable incremental updates from LP-GINV and undergo precision refinement by HP-MVM, with both operations retaining low complexity merit. This architecture resolves the dual challenges of limited precision and error accumulation in prior analog EVD implementations. Our system employs high-performance 1T1R chips fabricated in a foundry. The RRAM devices exhibit 8 distinct conductance levels (3-bit) with sufficient margin. We validate the entire workflow through hardware implementation, demonstrating applications in PCA and MIMO detection, highlighting the reliability of the approach.

II. Design of HP-EVD Scheme with AMC

EVD represents a matrix as $\boldsymbol{A}_{EVD} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, where $\boldsymbol{U}$ contains the orthonormal eigenvectors and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ the corresponding eigenvalues ordered by magnitude. Each (i-th) eigenpair $(\lambda_i, \boldsymbol{x}_i)$ satisfies $\boldsymbol{A}_{EVD}\boldsymbol{x}_i = \lambda_i\boldsymbol{x}_i$, or equivalently $(\boldsymbol{A}_{EVD} - \lambda_i\boldsymbol{I})\boldsymbol{x}_i = \boldsymbol{0}$. The HP-EVD scheme implements this extraction in AMC hardware, beginning with coarse eigenvalues sweeping (Fig. 2). Since the eigenvalues are initially unknown, HP-EVD sweeps a test shift s_i across the spectral range. For each trial value s_i, the circuit solves $(\boldsymbol{A}_{EVD} - s_i\boldsymbol{I})\boldsymbol{x}_o = \boldsymbol{0}$ via GINV. When $s_i \neq \lambda_i$, the matrix $\boldsymbol{A}_{EVD} - s_i\boldsymbol{I}$ is nonsingular and the output $\boldsymbol{x}_o$ vanishes (null response). Conversely, when s_i approaches an eigenvalue, $\boldsymbol{A}_{EVD} - s_i\boldsymbol{I}$ becomes near-singular, producing a pronounced peak in $\max|\boldsymbol{x}_o|$ that exceeds the reference voltage V_{ref} and triggers eigenvalue detection through OPA nonlinearity [6]. For an $N \times N$ full-rank matrix, N eigenpairs

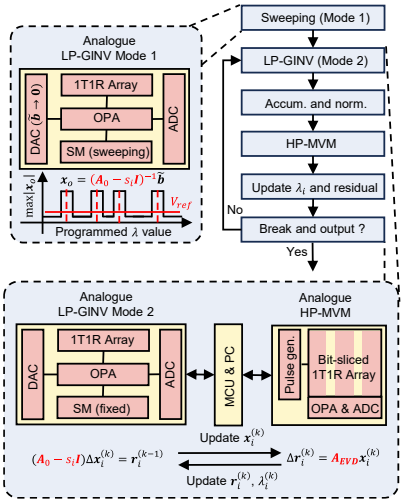

Fig. 2. Schematic of the HP-EVD.

correspond to N high pulses near eigenvalue locations, with pulse midpoints as the confirmed s_i (red dashed line). Due to device quantization and analog circuit non-idealities, the actual matrix implemented is $\boldsymbol{A}_0$ with 3-bit quantized conductance levels, while the input $\tilde{\boldsymbol{b}}$ is close to $\boldsymbol{0}$. In essence, the GINV circuit computes $\boldsymbol{x}_o = (\boldsymbol{A}_0 - s_i\boldsymbol{I})^{-1}\tilde{\boldsymbol{b}}$. The resulting coarse eigenvalue error is strictly upper-bounded by matrix quantization error, ensuring controllable accuracy.

Once coarse eigenvalues are located, precision enhancement proceeds through iterative refinement of individual eigenpairs. During refinement, each eigenpair is processed independently with fixed shift s_i. The LP-GINV circuit yields a low-precision estimate $\boldsymbol{x}_i$ satisfying $(\boldsymbol{A}_{EVD} - \lambda_i\boldsymbol{I})\boldsymbol{x}_i \approx \boldsymbol{0}$, leaving residual $\boldsymbol{r}_i = -(\boldsymbol{A}_{EVD} - \lambda_i\boldsymbol{I})\boldsymbol{x}_i$. This residual drives the next iteration, with LP-GINV generating incremental corrections. Precise residual computation therefore necessitates bit-slicing-enhanced HP-MVM [8]. Here, $\boldsymbol{A}_{EVD}$ is partitioned into multiple 3-bit sliced matrices. The most significant slice $\boldsymbol{A}_0$ is mapped to LP-GINV arrays, while all slices are mapped to HP-MVM arrays. In each iteration, LP-GINV computes the increment $\Delta\boldsymbol{x}_i^{(k)} = (\boldsymbol{A}_0 - s_i\boldsymbol{I})^{-1}\boldsymbol{r}_i^{(k-1)}$, where k is the cycle index, $\boldsymbol{r}_i^{(k-1)}$ and $\Delta\boldsymbol{x}_i^{(k)}$ are the input and output of LP-GINV. Starting from zero, the eigenvector is updated by accumulation $\boldsymbol{x}_i^{(k)} = \boldsymbol{x}_i^{(k-1)} + \Delta\boldsymbol{x}_i^{(k)}$ and normalization. HP-MVM performs the core task $\Delta\boldsymbol{r}_i^{(k)} = \boldsymbol{A}_{EVD}\boldsymbol{x}_i^{(k)}$, where digitized $\boldsymbol{x}_i^{(k)}$ is 1-bit-sliced. Partial results of MVM are combined by shift & add. Following by computing residual $\boldsymbol{r}_i^{(k)} = -\Delta\boldsymbol{r}_i^{(k)} + \lambda_i^{(k)}\boldsymbol{x}_i^{(k)}$ and eigenvalue $\lambda_i^{(k)} = \boldsymbol{x}_i^{(k)T}\Delta\boldsymbol{r}_i^{(k)}$, these two values are computed by the result of core tasks. The residual is fed back to LP-GINV until convergence. As $\boldsymbol{A}_0 - s_i\boldsymbol{I}$ becomes increasingly ill-conditioned and slows convergence, the shift is adaptively perturbed to nearby values $s_i + \delta$ during refinement iterations, balancing convergence speed with eigenvalue accuracy.

Notably, this iterative structure mirrors the coarse sweeping phase, where identical LP-GINV operations enable unified dual-mode hardware: sweeping mode (Mode 1) for coarse eigenvalue sweeping and GINV mode (Mode 2) for iterative refinement, integrating both core functions in a single circuit module. In Mode 1, to realize precise shifting, sweeping module (SM) is designed and integrated with LP-GINV circuit (Fig. 3). the SM is designed with 8-bit precision,

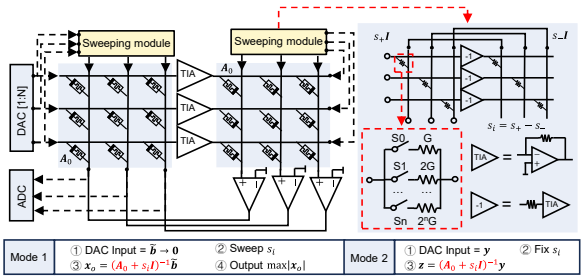

Fig. 3. The RRAM-based closed-loop circuit for solving LP-GINV in one step for sweeping coarse eigenvalues (Mode 1) and computing low-precision eigenvectors (Mode 2).

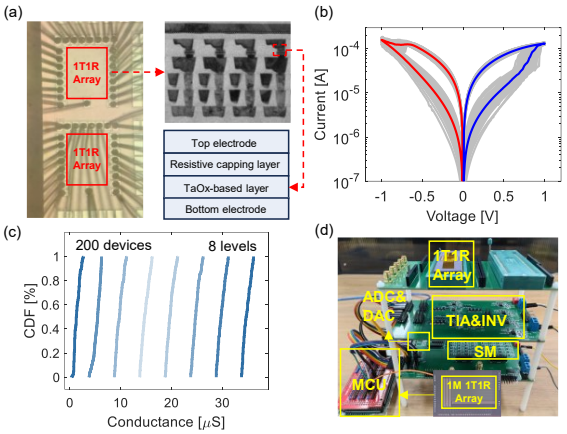

Fig. 4. (a) Microstructure of the proposed 1T1R integrated chip. (b) I-V characteristics of RRAM device. 50 cycles of results are tested. (c) CDF plot of 8 conductance levels of the 200 tested devices. (d) Hardware implementation of HP-EGVD solver system. 1Mb 1T1R array for HP-MVM is shown in inset.

adopting a highly-efficient 8T8R structure, *i.e.,* eight parallel switch-resistor subunits. Each resistor is set to a conductance value of G, $2G$, …, 2^7G, enabling configurable and synchronized diagonal matrix mapping. Since eigenvalues may reside on either side of the origin, the sweep must cover both positive and negative shifts. To represent a signed diagonal shift matrix $s_i\boldsymbol{I}$, the positive $s_+\boldsymbol{I}$ module is directly paralleled with RRAM array, while the negative $s_-\boldsymbol{I}$ is connected to the positive module via analog inverters formed by operation amplifiers (OPAs). In Mode 2, the circuit executes standard inverse computation $\Delta\boldsymbol{x}_i^{(k)} = (\boldsymbol{A}_0 - s_i\boldsymbol{I})^{-1}\boldsymbol{r}_i^{(k-1)}$ under fixed s_i to provide the approximate initial results and successive incremental updates of eigenvectors.

III. RRAM Fabrication, Characteristics and Experimental Setups

The proposed 1T1R integrated chips are fabricated on a commercial 40 nm CMOS manufacturing platform, with the detailed structure shown in Fig. 4a. TaO_x-based RRAM arrays were integrated between the M4 and M5 metal layers. To reduce switching voltages and improve device stability, we implement novel structural enhancements, including a tri-heterojunction stack (THS) and a self-formed passivation sidewall (SPS) [9]. The fabrication process remains fully compatible with standard CMOS BEOL processing. Fig. 4b shows the current-voltage (I-V) switching characteristics of the RRAM devices. Programming is performed using ASAP write-verify algorithm [10], achieving 8 discrete conductance levels via sequential coarse/fine tuning (Fig. 4c). The devices are programmed to 0.5-35 μS with sufficient margins. Fig. 4d

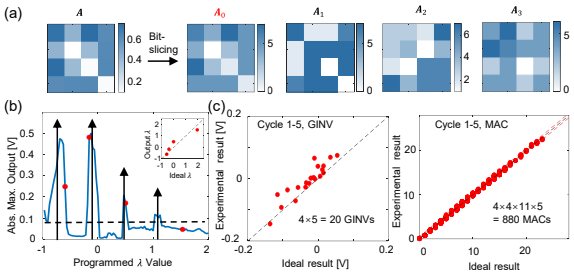

Fig. 5. (a) Example of HP-EVD for a 4×4 12-bit matrix A, along with its bit-slicing results (A0–A3). (b) Results of coarse sweeping of eigenvalues. (c) The experimental results of five cycles of LP-GINV and HP-MVM in dominant eigenpair solving.

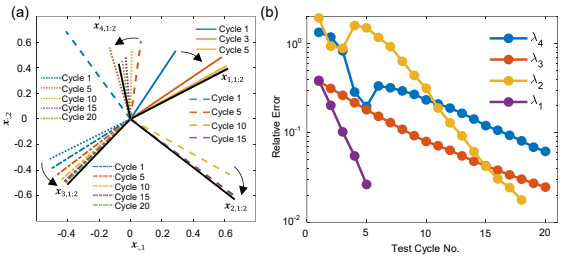

Fig. 6. (a) Convergence of the 4 eigenpairs solution. (b) Convergence behavior of HP-EVD.

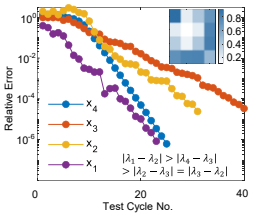

Fig. 7. Convergence behavior of HP-EVD, indicating that the convergence rate decreases as eigenvalue gaps between neighboring eigenvalues decreases.

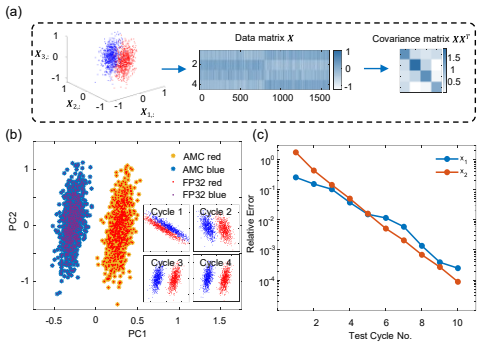

Fig. 8. (a) An example of PCA. (b) The results of dimensionality reduction to 2D. (c) Convergence behavior of HP-EVD in PCA.

presents the HP-EVD solver system. The LP-GINV circuit is implemented across 3 layers of PCBs, including 2 8×8 RRAM arrays, OPAs for TIAs and analog inverters, analog switches and resistors for SMs, DACs & ADCs for input/output interfacing. The HP-MVM is performed by measuring output current of analog multiply-accumulate (MAC) operations of bit-sliced RRAM array. The HP-MVM is implemented using a 1Mb 1T1R array, employing restricted row-wise parallelism to achieve zero arithmetic error [11]. The MCU connects LP-GINV and HP-MVM, governs control flow, and perform scalar computations.

IV. Experimental Results and APPPlications

A. Experimental results of AMC based HP-EVD

HP-EVD experimental validation uses a randomly generated 4×4 positive matrix. Matrix $\boldsymbol{A}$ is truncated to 12-bit and partitioned into $\boldsymbol{A}_0$~$\boldsymbol{A}_3$, with the most significant slice $\boldsymbol{A}_0$ mapped into LP-GINV and fixed during computation (Fig. 5a). In coarse eigenvalue sweeping, the LP-GINV performs a sweep across 96 discrete values spanning -1 to 2 (Fig. 5b). V_{ref} is configured to 0.09V. The estimated eigenvalues manifest as voltage pulse trains, whose midpoints are eigenvalue approximations. The fixed s_i (red points) are selected near the pulses as nearby values (Fig. 5b).

After coarse eigenvalue sweeping, successive analog LP-GINV and HP-MVM iterations are performed for each eigenpair. Notably, the results of LP-GINV and the bit-sliced HP-MVM during 5-cycle computation of the dominant eigenvector are presented in Fig. 5c. 20 LP-GINV outputs reflect remarkable errors. Thanks to the reliable HP-MVM, the iterative refinement improves the solution accuracy. For the unsigned 12-bit matrix $\boldsymbol{A}$ and signed 12-bit $\boldsymbol{x}_1$, as the elements of $\boldsymbol{x}_1$ is positive digits, a total of (12/3) ×(11/1)×5=220 LP-MVMs are performed, yielding 880 MACs, each precisely correct. Four eigenvectors convergence trajectories are shown in Fig. 6a, highlighting the evolution of the first two vector components as representative examples, where $\boldsymbol{x}_{i,b}$ denotes the b^{th} element of the i^{th} eigenvector. After 20 cycles, the residual errors of all eigenvectors are reduced to the order of 10^{-2}, quantified by $\|\boldsymbol{x}_i - \boldsymbol{x}_i^*\|/\|\boldsymbol{x}_i^*\|$, where $\boldsymbol{x}_i^*$ and $\boldsymbol{x}_i$ denote the ideal and experimental results (Fig. 6b).

It is worth noting that while experimental results validate the feasibility of the HP-EVD, its convergence behavior is inherently governed by distribution of eigenvalues. For the matrix in Fig. 7, the eigenvalues are λ_1=2, λ_2=0.4, λ_3=0.1, and $\lambda_4 = -0.7$, respectively. The observed convergence deceleration is inversely correlated with the spectral gap between adjacent eigenvalues. Here the 2nd and 3rd eigenpairs has the worse convergence performance. This is consistent with the theoretical behavior of power iteration dynamics.

Building upon the preceding experiments, we demonstrate HP-EVD experimentally for PCA using a dataset $\boldsymbol{X}^{4\times1600}$ with 4 features (first three shown in $\boldsymbol{X}_{1\sim3,:}$) and two classes (red/blue) (Fig. 8a). 2D reduction relies on the Top-2 eigenpairs $(\boldsymbol{\Sigma}, \boldsymbol{P})$ of the covariance matrix $\boldsymbol{X}\boldsymbol{X}^T$ and the projection of $\boldsymbol{X}$ onto the eigenvectors $\boldsymbol{P}$. With 10 HP-EVD iterations, the reconstructed top-2 principal components (PC1/PC2) achieve lossless reconstruction versus FP32 results, while only 4 iterations suffice for accurate classification (Fig. 8b). After 10 cycles, the relative error of eigenvectors is converged to the order of 10^{-4}, showing the prospect of HP-EVD in applications (Fig. 8c).

B. Application of ill-conditioned matrix inversion to MIMO detection

Massive MIMO represents an ideal application scenario for AMC. Although AMC enables efficient massive MIMO signal processing for 6G (Fig. 9a) [12], practical channels $\boldsymbol{H}$ produce ill-conditioned Gram matrices $\boldsymbol{H}^H\boldsymbol{H}$ (e.g., high antenna correlation and non-orthogonal channel) that existing AMC approaches cannot satisfy precision requirements. Here we demonstrate an example of AMC-based MIMO detection with ill-conditioned channel matrix, higher precision and higher-order modulation than previous study. For $K \times K$ ill-conditioned systems $\boldsymbol{A}_{IC}\boldsymbol{m} = \boldsymbol{b}$, the HP-EVD mitigates

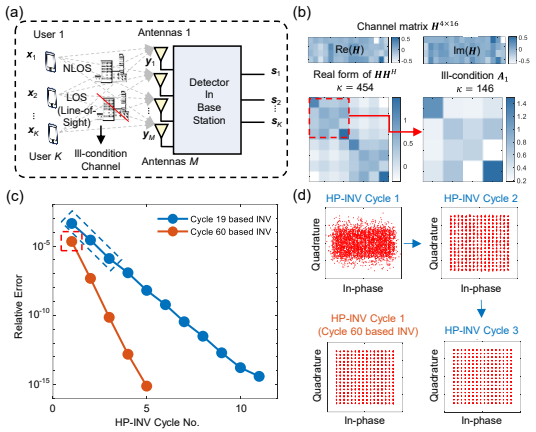

Fig. 9. (a) Illustration of a massive MIMO system. (b) An example of the uplink detection of a 16 × 4 MIMO. (c) Convergence of INV of $\boldsymbol{A}_1$. (d) Constellation diagram for symbols recovered by ZF detection.

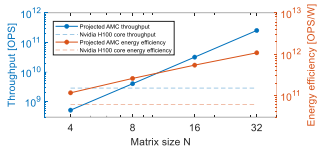

Fig. 10. Benchmarking details of throughput and energy efficiency of AMC and state-of-the-art GPU (FP32).

inaccuracies through the following approach. First, it constructs preconditioned LP-INV solutions as $\Delta\boldsymbol{y}^{(k)} = \sum_{i=1}^{K} \boldsymbol{x}_i\left(\boldsymbol{\lambda}_i^{-1}\boldsymbol{x}_i^T\boldsymbol{r}^{(k)}\right)$, where eigenpairs are computed by AMC, and $\boldsymbol{r}^{(k)}$ denotes the residual. Then, iterative residual refinement $\boldsymbol{r}^{(k)} = \boldsymbol{b} - \boldsymbol{A}_{IC}\boldsymbol{m}^{(k)}$ enables accurate inversion of the ill-conditioned matrix with an MVM. While its core method remains HP-INV iterative refinement, the construction of high-precision preconditioned results via EVD method fulfills convergence condition, thereby enabling the implementation of HP-INV for such matrices.

Experimental validation of HP-INV is demonstrated for a 4×16 MIMO system (Fig. 9b). In zero-forcing (ZF) detection under strong inter-antenna correlation, the signal recovery is expressed as $\boldsymbol{s} = (\boldsymbol{H}^H\boldsymbol{H})^{-1}\boldsymbol{H}^H\boldsymbol{y}$, where $\boldsymbol{y}$ and $\boldsymbol{s}$ are received and recovered signal vector. The heaviest computational burden resides in inverting the Gram matrix, corresponding to double size real-valued matrix. To solve this real-valued matrix by BlockAMC [13], the solution of 4×4 ill-conditioned systems (κ=146) is required. Applying 19 HP-EVD preconditioning cycles requires 11 cycles of additional ill-conditioned HP-INV to 10^{-15} relative error of INV (Fig. 9c), whereas applying 60 HP-EVD cycles with robust preconditioning require only 5 cycles. For MIMO detection, 1000 random 256-QAM symbols are transmitted over a channel with SNR = 60 dB. With 19 cycles of eigenpairs, all symbols are correctly detected with no errors by only 3 cycles of HP-INV (Fig. 9d). When 60 cycles of eigenpairs solving are executed, error-free detection is achieved with only 1 cycle of HP-INV, indicating that the EVD result can be directly utilized without iterative refinement of HP-INV. These findings break through the ill-conditioning bottleneck inherent to AMC and facilitating robust signal processing for massive MIMO in 6G era.

Benchmarking details between AMC and state-of-the-art GPU (FP32) for varying N are shown in Fig. 10 [14]. The complexity for single-core GPU-based Gram-Schmidt QR iterations for FP32 HP-EVD is about $2N^3$. 100-point sweeping requires $100t_{LP}$ time (with a parallelism of p_{sw}), while paralleled N HP-EVDs (assuming 20 cycles) demand $20\left(\frac{100t_{LP}}{p_{sw}} + t_{MVM} + 2t_{ADC}\right)$ time. Here $p = 24$ for computing precision, $m_0 = 3$ for RRAM precision and $n_0 = 4$ for input precision. Our scheme enables 88× throughput and 18× energy efficiency at $N = 32$ as projected by scalability analysis.

V. Conclusion

We demonstrate HP-EVD, a high-precision analog EVD solver on foundry-fabricated 1T1R arrays, achieving tens-fold improvements in throughput and energy efficiency over digital processors with validated PCA and massive MIMO applications. These results highlight the potential of AMC for next-generation signal processing and scientific computing.

Acknowledgment

This work has received funding from NSFC (62025401, 62322401, U24B200170, 62572011), Beijing Natural Science Foundation (4252016), and the 111 Project (B18001). The authors thank the support from ICRD and HLMC.

References

[1] C. Hong *et al.*, "Solving All Eigenpairs With Resistive Memory-Based Analog Matrix Computing Circuits," *IEEE Trans. Circuits Syst. I*, vol. 73, no. 4, pp. 2522-2532, Apr. 2026.

[2] L. Lee *et al.*, "Square-root generalized eigenvalue decomposition processor for leakage-based multi-user MIMO precoding with multi-antenna users," *IEEE Trans. Circ. Syst. I: Reg. Papers*, vol. 66, no. 6, pp. 2382-2393, Jun. 2019.

[3] F. Meng *et al.*, "PiSSA: Principal Singular Values and Singular Vectors Adaptation of Large Language Models," in *Advances in Neural Information Processing Systems 37*, Vancouver, BC, Canada, 2024, pp. 121038–121072.

[4] Z. Sun *et al.*, "Solving matrix equations in one step with cross-point resistive arrays," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 116, no. 10, pp. 4123–4128, Mar. 2019.

[5] P. Zuo *et al.*, "Precise and scalable analogue matrix equation solving using resistive random-access memory chips," *Nat. Electron.*, vol. 8, no. 12, pp. 1222–1233, Oct. 2025.

[6] P. Mannocci *et al.*, "In-Memory Principal Component Analysis by Analogue Closed-Loop Eigendecomposition," *IEEE Trans. Circuits Syst. II*, vol. 71, no. 4, pp. 1839–1843, Apr. 2024.

[7] H. Zhao *et al.*, "A novel memristor-based method to compute eigenpairs," *Analog Integr. Circ. Sig. Process*, vol. 118, no. 1, pp. 109–119, Jan. 2024.

[8] M. Le Gallo *et al.*, "Mixed-precision in-memory computing," *Nat. Electron.*, vol. 1, no. 4, pp. 246–253, Apr. 2018.

[9] Q. Wang *et al.*, "A Logic-Process Compatible RRAM with 15.43 Mb/mm2 Density and 10years@150°C retention using STI-less Dynamic-Gate and Self-Passivation Sidewall," *International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2023, pp. 1-4.

[10] J. Sun *et al.*, "ASAP: An Efficient and Reliable Programming Algorithm for Multi-level RRAM Cell," *IEEE International Reliability Physics Symposium (IRPS)*, Grapevine, TX, USA, 2024, pp. 1-4.

[11] Y. Li *et al.*, "Constraints on Bit Precision and Row Parallelism for Reliable Computing-in-Memory," *Electronics*, vol. 14, no. 13, p. 2532, Jun. 2025.

[12] P. Zuo *et al.*, "Extremely-Fast, Energy-Efficient Massive MIMO Precoding with Analog RRAM Matrix Computing," *IEEE Trans. Circ. Syst. II: Exp. Briefs,* vol. 70, no. 7, pp. 2335-2339, July 2023.

[13] D. Bridarolli *et al.*, "High-Density Multilevel 3D Vertical Resistive Switching Memory (VRRAM) for Massively Parallel in-Memory Computing," *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2024, pp. 1-4.

[14] NVIDIA H100 Tensor Core GPU (2024) https://www.nvidia.com/en-us/data-center/h100.