Implementation of Discrete Fourier Transform using RRAM Arrays with Quasi-Analog Mapping for High-Fidelity Medical Image Reconstruction

Han Zhao†, Zhengwu Liu†, Jianshi Tang*, Bin Gao, Ying Zhou, Peng Yao, Yue Xi, He Qian, Huaqiang Wu*
School of Integrated Circuits, Beijing Innovation Center for Future Chips, BNRist, Tsinghua University, Beijing, China;
†These authors contributed equally: Han Zhao, Zhengwu Liu; *Email: jtang@tsinghua.edu.cn; wuhq@tsinghua.edu.cn

***Abstract*—**In this paper, we report the first experimental implementation of discrete Fourier transform (DFT) on analog resistive switching random-access memory (RRAM) arrays with computing-in-memory (CIM). Considering the features of transform matrix, we developed a novel conductance mapping strategy, namely quasi-analog mapping (QAM), to realize high-precision mapping by taking advantage of the analog switching characteristics of our RRAM. Based on the RRAM-based DFT models, high-fidelity medical image reconstruction was further demonstrated, achieving a software-comparable peak signal-to-noise ratio (PSNR) of 26.1 dB. Compared to the commonly used quantized mapping (QM), QAM enhanced the image reconstruction quality and showed strong robustness to RRAM read noise. RRAM-implemented DFT also achieved ~128× higher energy efficiency than CPU. This work provides a general strategy for using RRAM array with CIM feature to accelerate signal processing algorithms.

I. Introduction

DFT is one of the most widely used signal processing algorithms with broad applications in multimedia processing and communication (**Fig. 1**). Generally, DFT is implemented on digital hardware based on Si CMOS technology and von Neumann architecture. However, as the Moore's law scaling is approaching its physical limit, the performance improvement of such digital implementation is facing serious bottlenecks such as memory wall in the era of big data [1]. Fortunately, analog RRAM array-based CIM might provide an alternative and promising paradigm to implement DFT efficiently (**Fig. 2**), considering its outstanding performance in accelerating deep learning algorithms [2, 3].

To implement an algorithm in RRAM array-based hardware, there are often three steps involved. First, high-precision model parameters are calculated in software. Second, those parameters are transferred onto the device conductance of RRAM array, known as the mapping process. Third, computations are carried out on the array for different inputs. In practice, due to the limited number of conductance states and the presence of read noise, the parameters to be mapped have to be quantized to only a few conductance levels with pre-defined write margins. Such mapping strategy, known as QM (**Fig. 4**), has been commonly used for implementing deep neural networks, where it is the final classification accuracy rather than the precision of individual weight that matters. In this case, QM is effective by adopting low-bit quantization in training to accommodate the limited number of RRAM conductance levels available at the current stage [4]. However, for the implementation of DFT, the model parameters (i.e., elements in the transform matrix) are pre-calculated and fixed rather than obtained by training, and also the mapping precision of transform matrix largely affects the final DFT results (**Fig. 2b**). Therefore, as the required number of conductance levels dramatically increases with the number of DFT points, the QM strategy becomes inapplicable.

In this work, we proposed a novel mapping strategy QAM to implement DFT on RRAM arrays with CIM for acceleration of DFT calculations. Different from QM, QAM directly maps the target values without quantization, and thus it can eliminate the quantization error during mapping to achieve a much higher precision (**Fig. 3**). Here, to be rigorous, QAM is used rather than "full analog mapping" as the mapping precision is still limited by the bit resolution of peripheral analog-to-digital converters (ADCs). To validate the feasibility of QAM, we demonstrated high-fidelity medical image reconstruction based on RRAM-implemented DFTs and compared the above two mapping strategies. QAM achieved a higher PSNR of 26.1 dB, and also showed ~128× higher energy efficiency than CPU.

II. Quasi-Analog Mapping with Analog RRAM

To study different mapping strategies, we fabricated 1K-bit RRAM array with a material stack of TiN/TaO_x/HfO_2/TiN [5]. Here TaO_x served as the thermal enhanced layer to improve the analog switching characteristics of HfO_2. Each RRAM cell has a one-transistor-one-resistor (1T1R) structure with 3 terminals connected to bit line (BL), word line (WL) and source line (SL) (**Fig. 4**). Set and reset operations were performed to program the RRAM device conductance to the target value. To perform computation, read voltages were applied in the BLs, and the current results were obtained from the SLs. **Fig. 5** shows the typical DC I-V switching curve of a RRAM device. The analog switching characteristics under a series of set and reset pulses is shown in **Fig. 6**. **Fig. 7** illustrates the linear I-V relationship at different conductance states within a read voltage range of 0~0.3V. All the above results confirm that the RRAM devices have the desired analog switching characteristics to perform CIM, providing an excellent platform for comparing different mapping strategies and further implementing DFT.

Fig. 8 shows the photograph of our customized test system for mapping [2]. To fairly compare the performance of QM and

2021 IEEE International Electron Devices Meeting (IEDM) | 978-1-6654-2572-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IEDM19574.2021.9720547

QAM, we used the same write margin (±100 nA at V_{read} = 0.2 V) when mapping the RRAM conductance in our experiments. As an example in **Fig. 9**, QM had only 3 conductance levels that matched the target values in the range of 400~1000 nA, while QAM directly mapped the 5 target values without quantization. Ideally, the QAM-mapped conductance values should have the same precision as software-calculated parameters. In practice, the precision was limited by the ADC (14 bits) used for sensing the conductance state in the post-write verify process. Also, there could be overlaps between adjacent conductance intervals, depending on the write margin and specific target levels.

III. RRAM Array-based DFT Implementation

Standard DFT algorithm could be decomposed into vector-matrix multiplication operations [6, 7], and the transform matrix remains unchanged during the processing, so it can be naturally implemented on RRAM array for acceleration. **Fig. 10** shows the RRAM-based DFT implementation schematic for processing a complex number input vector $\boldsymbol{x}$. The elements in the transform matrix are also complex numbers, so four RRAM cells are needed to represent each element: one differential pair consisted of two RRAM cells for the real part, and another pair for the imaginary part. After mapping, DFT results are directly obtained from the RRAM array after one read process. As an example, **Fig. 11** compares the typical results of QM-based, QAM-based and software-calculated one-dimensional (1-D) DFT. It is seen that QAM DFT shows more consistent results with software calculations, where an average correlation of 0.9990, higher than QM DFT (0.9980), was achieved (**Fig. 12**).

Besides the above 1-D case, 2-D DFT is also widely used in image processing. **Fig. 13** illustrates the implementation of 2-D DFT on RRAM arrays, where the input signal is processed successively with two 1-D DFTs, along with one operation of matrix transpose. The correlations between RRAM-based and software-calculated 2-D DFT results are 0.9985, and 0.9979 for QAM and QM, respectively (**Fig. 14**). These results demonstrate the feasibility of RRAM-array based DFTs, where QAM-DFT shows a higher processing accuracy than QM-DFT.

IV. Demostration of CT Image Reconstruction

To explore the application of RRAM-based DFT, we then demonstrated the high-fidelity image reconstruction in compute tomography (CT). In this task, projections of organs, acquired from X-ray scanner in CT machine, were used to reconstruct CT images. **Fig. 15** shows the RRAM-based CT reconstruction processing flow, where both 1-D DFT and 2-D inverse DFT (IDFT) were realized by RRAM arrays as described above. The lung section of CT images from an open-access database were used [8]. As an example, **Fig. 16** shows the original and reconstructed images along with intermediate results.

To compare the CT image reconstruction quality with different mapping strategies, we used a 26-point DFT, which required 14 conductance values for the transform matrix. The RRAM device was programmed to be within a current range of 400~4000 nA (corresponding to conductance range of 2~20 μS). With a write margin of ±100 nA, the QM strategy yielded 18 quantized levels from 500 to 3900 nA. In contrast, the QAM strategy directly mapped the target values with the same write margin. **Fig. 17** compares the reconstructed images with QAM and QM. The PSNR of QAM-based image was 26.1 dB, which was close to the value of software-reconstructed image (26.3 dB) and higher than that of QM-based image (25.4 dB). The slight degradation in PSNR for QAM was likely attributed to the non-ideal device characteristics such as noise and relaxation [9]. We further simulated the effects of read noise and mapping deviation on the PSNR of reconstructed images (**Fig. 18**). It is found that QAM showed a strong robustness to read noise (up to 600 nA), and the PSNR was largely affected by the mapping deviation, which highlighted the importance of high-precision mapping for the implementation of DFT on RRAM arrays. **Fig. 19** compares the mapping results of QM and QAM strategies. **Fig. 20** plots the histograms of mapping deviations, which shows that QM has larger mapping errors (0.09±1.02 μS) than QAM (0.03±0.55 μS) due to quantization. **Fig. 21** further examines read disturbances of two different mapping strategies.

Certainly, the higher-accuracy DFT results and higher-fidelity image reconstruction for QAM came with the cost of higher-bit-resolution ADCs, which would lead to higher energy consumption than QM. To evaluate the overall performance, we performed benchmarks on both mapping and calculation processes using the circuitry shown in **Fig. 22**. **Fig. 23** shows that QAM consumes 3.03× more energy in average than QM. Such overhead, although affordable in most applications since mapping is usually one-time operation, can be further reduced when using larger RRAM arrays to implement DFT with more points and also adopting more advanced technology nodes for the array peripheral circuits including ADCs. During DFT calculations after deployment, RRAM-implemented DFT could achieve an energy efficiency of 1.15 TOPS/W, showing >11× advantage over GPU and ~128× advantage over CPU (**Fig. 25**).

V. Conclusion

In sum, for the first time, we have implemented both 1-D and 2-D DFTs on analog RRAM arrays for acceleration by taking the advantage of its CIM capability. We have developed a novel mapping strategy of QAM to achieve the required high-precision mapping for DFT. The feasibility and effectiveness of the QAM strategy has been verified by comparison with the commonly used QM strategy. Furthermore, high-fidelity CT image reconstruction has been demonstrated based on the RRAM-implemented DFTs, achieving software-comparable PSNR results and ~128× higher energy efficiency than CPU.

Acknowledgment

This work was supported in part by China Key Research and Development Program (2019YFB2205403).

References

[1]X. Xu *et al.*, *Nat. Electron.*, 2018. [2]P. Yao *et al.*, *Nature*, 2020. [3]Z. Wang *et al.*, *Nat. Electron.*, 2019. [4]Y. Cai *et al.*, *IEEE TCAD*, 2020. [5]W. Wu *et al.*, *IEEE VLSI*, 2018. [6]M. Hu *et al.*, *IEEE ICRC*, 2016. [7]S. Gao *et al.*, *IEEE IEDM*, 2019. [8]S. G. Armato III *et al.*, *Med. Phys.*, 2011. [9]M. Zhao *et al.*, *Appl. Phys. Rev.*, 2020.

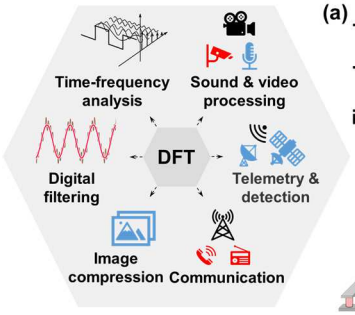


Fig. 1. Various applications of discrete Fourier transform (DFT) from multimedia information processing to communication.

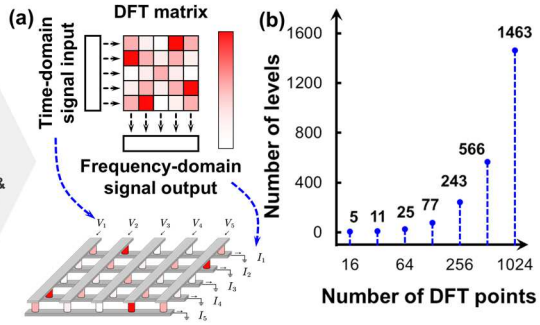


Fig. 2. (a) Schematic of RRAM array-based DFT implementation. **(b)** The mapping of DFT matrix demands for a large number of RRAM conductance levels with the increasing number of DFT points.

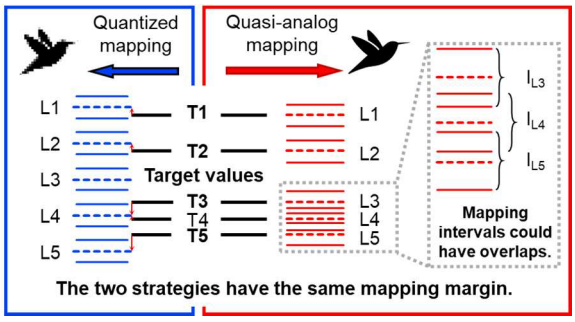


Fig. 3. Illustration of two different mapping strategies on RRAM array (with the same write margin for mapping). **(a)** Typical quantized mapping (QM) strategy, which could suffer from considerable precision loss. **(b)** Quasi-analog mapping (QAM) strategy, where the target values are directly used for mapping.

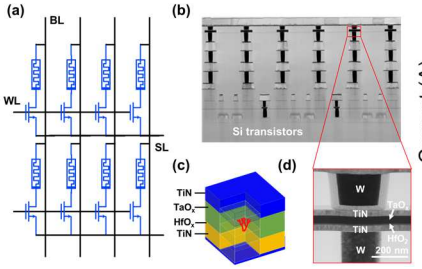


Fig. 4. Schematics **(a, c)** and transmission electron microscope (TEM) images **(b, d)** of RRAM 1T1R array and device structure, respectively.

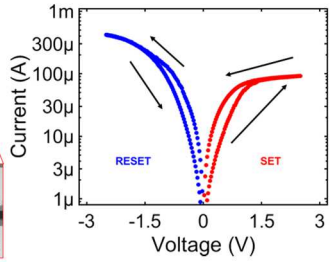


Fig. 5. A typical DC I-V characteristics of RRAM with gradual SET and RESET

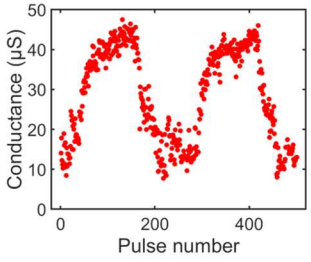


Fig. 6. Analog switching characteristics of RRAM. V_{SET}−1.4V, V_{RESET}−1.5V.

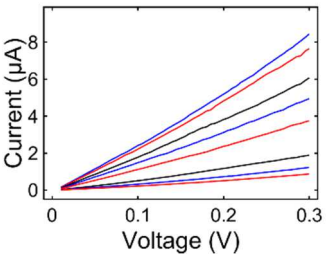


Fig. 7. I-V linearity of RRAM at different conductance levels.

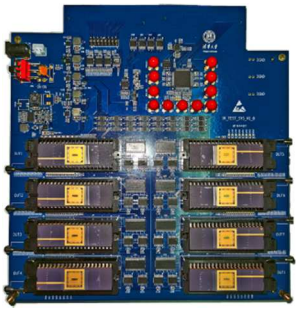


Fig. 8. Photograph of the integrated system with eight chips of 2K RRAM array to implement DFT.

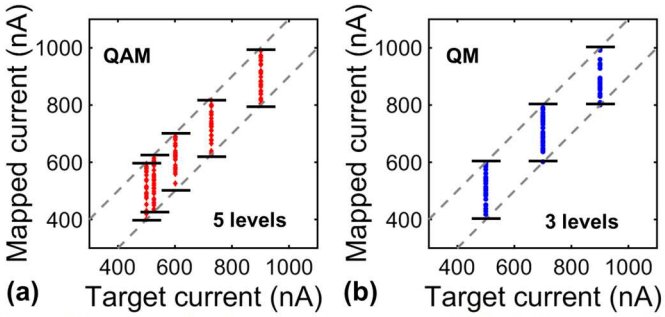
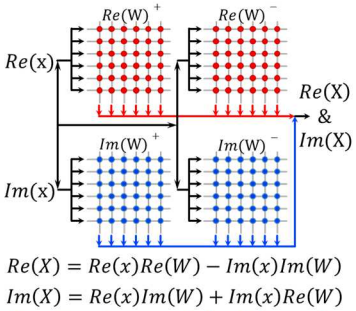


Fig. 9. (a) Mapping results of quasi-analog mapping (QAM). **(b)** Mapping results of quantized mapping (QM). QAM could achieve mapping with 5 coductance levels in the range of 400~1000 nA, which is more than that of QM strategy (3 levels), showing a higher mapping precision than QM.



$$Re(X) = Re(x)Re(W) - Im(x)Im(W)$$
$$Im(X) = Re(x)Im(W) + Im(x)Re(W)$$

Fig. 10. Schematic of RRAM array-based implementation of one-dimensional (1-D) DFT. Here the real (Re) part and imaginary (Im) part are calcualted separately.

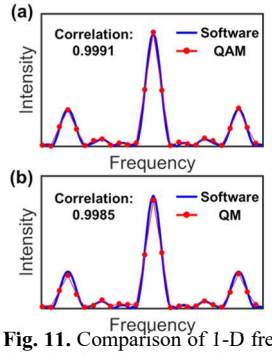


Fig. 11. Comparison of 1-D frequency-domain signals computed by QM-based **(a**, red dots**)**, QAM-based **(b**, red dots**)** and software-based DFTs (blue line). The corrlations between QAM, QM and software results are 0.9991 and 0.9985, respectively. Here a 26-point DFT is used.

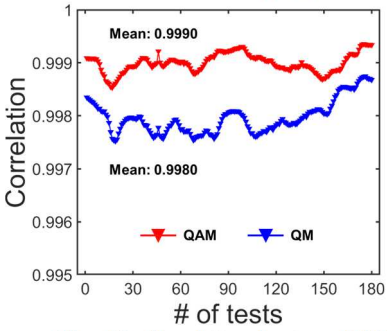


Fig. 12. Correlation bewteen QAM-based, QM-based and software-based DFT results under 180 different input test signals. The mean correlation values of QAM and QM results are 0.9990 and 0.9980, respectively. Here a 26-point DFT is used.

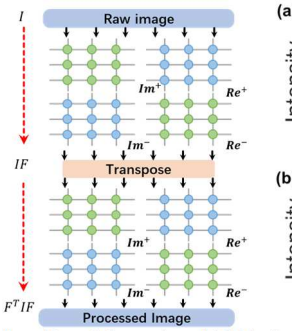


Fig. 13. Schematic of RRAM-implemented 2-D DFT. *I*, the input raw image. *F*, the transform matrix. F^T, the transpose matrix of *F*. The raw image is first processed with a 1-D DFT. After being trsposed, the image is further processed with aother 1-D DFT.

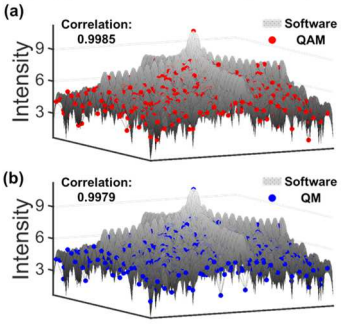


Fig. 14. Comparison between 2-D frequency-domain signals computed by RRAM-based DFT (**a**, QAM; **b**, QM) and software-based FFT (surface). The correlations between QAM, QM and software results, are 0.9985 and 0.9979, respectively.

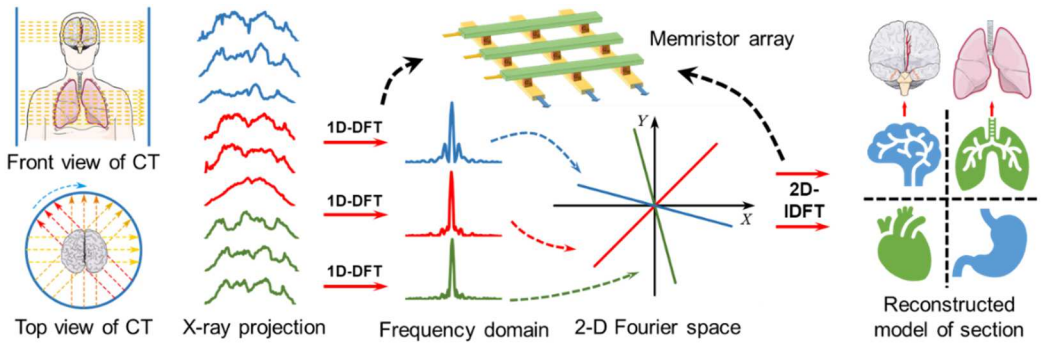


Fig. 15. Schamatic of RRAM-based computed tomography (CT). Projection acquired from X-ray scanner is input into the RRAM array for image (e.g. lung or brain sections) reconstruction. Based on the Fourier Central Slice Theorem, the projection is first transformed to frequency-domain and then mapped onto the 2-D Fourier space. 2-D inverse discrete Fourier transform (IDFT) is then performed to obtain the reconstructed section images.

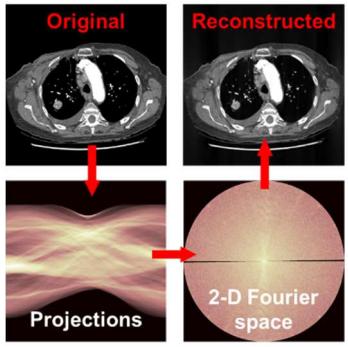


Fig. 16. Illustration of the original and reconstructed images and intermediate results for the CT reconstruction task .

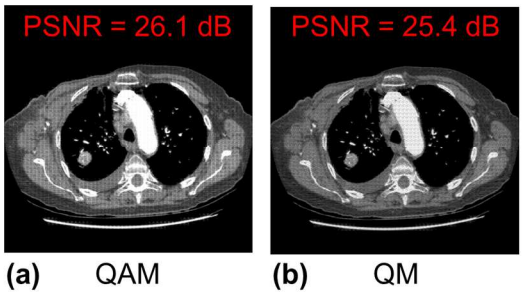


Fig. 17. Reconstructed lung sections by RRAM array-based DFT models using (**a**) QAM and (**b**) QM mapping strategies. The QAM strategy yields a higher PSNR, and such advantage is expected to become more prominent as the number of DFT points further increases.

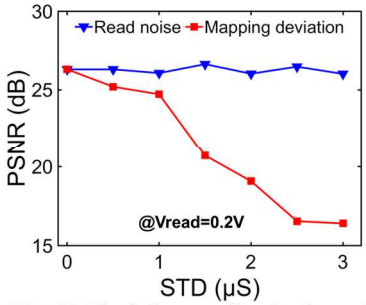


Fig. 18. The influence of read noise and write deviation of RRAM array on the CT image reconstruction. The QAM-enabled DFT shows a strong robustness to RRAM read noise.

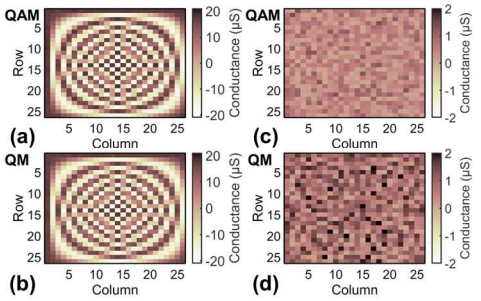


Fig. 19. The mapping results of (**a**) QAM and (**b**) QM strategies. The difference between the target conductance and experimentally mapped results of (**c**) QAM and (**d**) QM. QAM clearly shows a higher precision mapping than QM.

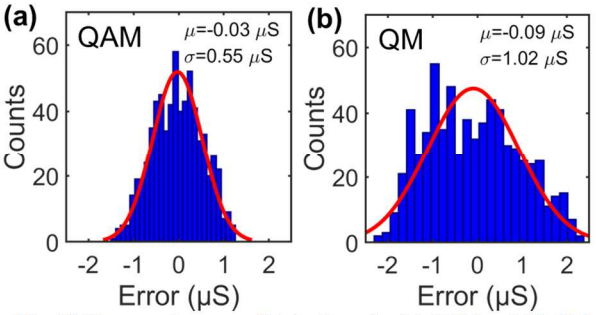


Fig. 20 The mapping error distributions for (**e**) QAM and (**f**) QM, which approximately follows a Gaussian distribution.

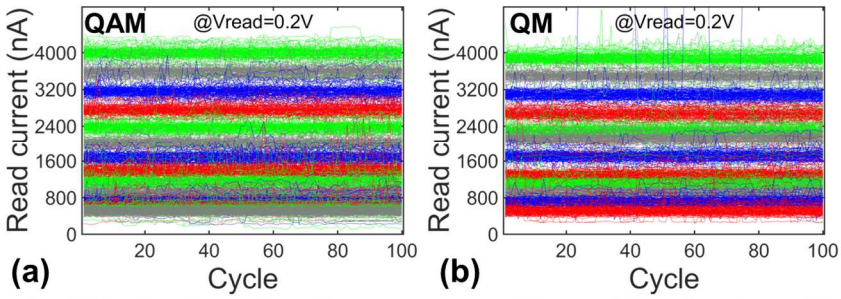


Fig. 21 RRAM read noise of two different mapping strategies: (**a**) *QAM* and (**b**) *QM*. Up to *14* levels can be identified in the QAM results, while only 12 levels can be seen in the QM results.

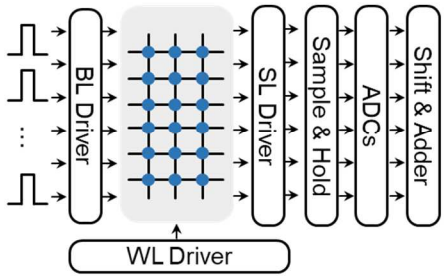


Fig. 22. Hardware framework of the RRAM array-based DFT model and peripheral circuit modules for benchmark. The pulse number encoding method with the bit position-weighted scheme was used for inputs.

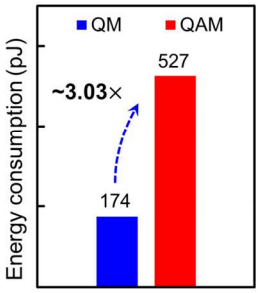


Fig. 23. The comparison of energy consumption between the two mapping strategies. QAM has ~3.03× higher energy consumption due to the use of higher bit resolution ADCs

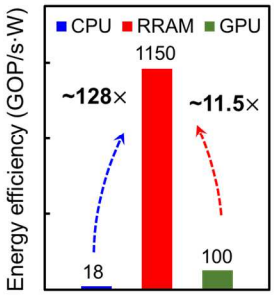


Fig. 24. Comparison of DFT energy efficiency between GPU, CPU and RRAM. RRAM-implemented DFT has large advantage in energy efficiency than CPU/GPU.

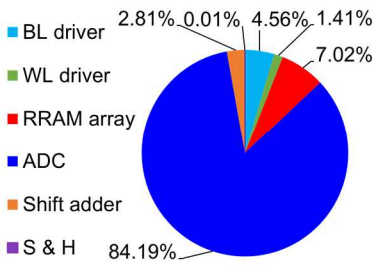


Fig. 25. Energy consumption breakdown of different circuit modules. The ADC portion can be reduced by adopting more advanced technology nodes and using larger RRAM arrays to implement DFT with more points.