Extended Abstracts of the 2023 International Conference on Solid State Devices and Materials, Nagoya, 2023, pp47-48

Heterogeneous Integration of 32 × 32 1S1R Crossbar Array using 2D Hafnium Diselenide on Si Platform and its Compute-in-Memory Hardware Featuring Low Latency and High Energy Efficiency

Samarth Jain[1]*, Sifan Li[1]*, Jianze Wang[1], Xuanyao Fong[1], Kah-Wee Ang[1,2,#]
[1]Department of Electrical and Computer Engineering, 4 Engineering Drive 3, National University of Singapore, Singapore 117583
[2]Institute of Materials Research and Engineering, A*STAR, 2 Fusionopolis Way, Singapore 138634
*Equal Contribution Phone: +65 6516 2575, Fax: +65 6779 1103, [#]E-mail: eleakw@nus.edu.sg

Abstract: For the first time, we report a heterogeneous 1 kbit (32 × 32) crossbar array (CBA) combining silicon (Si) based selectors and memristors made of two-dimensional (2D) hafnium diselenide ($HfSe_2$). The integration of one-selector-one-memristor (1S1R) *via* three-dimensional (3D) stacking of 2D $HfSe_2$ achieves low energy (70 pJ) write/read operations and fast speed performance (< 20 ns). Additionally, a low-power sensing circuit is realized to address the power consumption limitation of ADC for the scaling of CBA architecture. The ADC-free circuit comprises a high-accuracy differential subtractor to minimize quantization loss and a time-to-digital converter (TDC)-based converter using FPGA. The hardware successful achieves scalable, multiplexer (MUX)-free parallel compute-in-memory (CIM) and image processing.

Introduction: The rise of 2D materials-based memristors is expected to play an important role in CIM due to their potential to enhance energy efficiency [1]. Despite these benefits, several challenges still hinder their widespread use in large-scale CBA computing systems. These include the low scalability (limited to 10×10 due to the lack of selectors) and high power consumption and latency caused by the analog-to-digital converter (ADC)-based peripheral sensing circuits during computing [2]. Given the maturity of the Si process, it is essential to explore the heterogeneous integration of Si and 2D materials-based memristors for large-scale CBA [3]. This work presents a scalable integration approach that combines wafer-scale polycrystalline $HfSe_2$ and 3D stacking process on a Si platform. With the integration of Si-based selectors, 1 kbit 1S1R CBA is demonstrated with successful write/read capability and fast response time. Additionally, customized analog current subtractors and low-power TDC-based sensing circuits are utilized to reduce the energy consumption in periphery circuits. To showcase the potential of 2D materials in low-power CIM, parallel image processing with four kernels is demonstrated using the proposed hardware. This work highlights the potential of heterogeneous integration of 2D material and Si to spur the energy efficiency of CIM.

Device Fabrication and Heterogeneous Integration

Fig. 1a presents a 3D schematic of the integrated Si-$HfSe_2$ cell, where the top $HfSe_2$ serves as the memristor switching layer. A Schottky contact was formed between the Si and the bottom electrode (BE), acting as the selector, which is defined by the boron implantation into the Si. To establish an ohmic contact, a nickel silicide (NiSi) was formed below the middle electrode (ME). **Fig. 1b** displays the equivalent circuit diagram of the device. **Fig. 1c** shows an as-fabricated array of Si selectors, and **Fig. 1d** shows a 2-inch polycrystalline $HfSe_2$ before the transfer process. **Fig. 1e** presents a schematic view of the metal-assisted 3D stacking process. The Au/$HfSe_2$ film was delaminated from the SiO_2 substrate and transferred to the target Si substrate [4]. The $HfSe_2$ was grown using molecular beam epitaxy at 750 °C, but the transfer process was conducted at a low temperature (<150 °C) to prevent degradation to the bottom silicon selectors. Finally, the top electrode (TE) was patterned to define each memristor cell, and the interconnection of the TEs in each row was completed by interlayer oxide (SiO_2) and word line (WL) deposition. **Fig. 1f** shows the world's first demonstration of a 1 kbit (32 × 32) Si-$HfSe_2$ heterogeneously intergrated CBA. The inset provides a closer view of the 1S1R unit. **Fig. 2** displays the cross-sectional transmission electron microscopy (TEM) images of the stacked structure, with the WL, interlayer oxide, TE, ME, and the bottom NiSi from top to bottom. The inset of Fig. 2 offers a zoomed-in image of the 2D memristor, displaying a clear Ti/$HfSe_2$/Au structure.

Results and Discussion

Si-$HfSe_2$ Heterogenous Device Performance: Fig. 3a and **3b** present the current-voltage (*I-V*) characteristics of the Si-based selector and $HfSe_2$-based memristor, respectively. The inset schematics illustrate the direction of voltage supply. Figure 3c displays the *I-V* curves of the integrated device, where the program current during set is limited by reverse-biased current of the selector, thereby improving the device selectivity in array. As demonstrated in **Fig. 4**, the device retains fast write and read time after Si-based selector integration (<20 ns), making it suitable for high-speed multiply-and-accumulate (MAC) operations, which can result in reduced energy consumption through further device downscaling and parasitics optimization. The endurance of the device was tested through a current-visible pulsed voltage scheme (PVS), proving its ability to sustain more than 2.5×10^4 cycles of effective switching (**Fig. 5**). **Fig. 6** shows the retention of the device (more than 10^4 seconds) measured under 85 °C, demonstrating its resistance stability during read and compute operations. **Fig. 7** displays more than 3 bits of multiple level cell (MLC) with overlap probability less than 0.651%, showing that the device is well-suited for synaptic weight storage. Additionally, **Fig. 8** highlights the low cycle-to-cycle variation (1.9%) and low non-linearity (<1) of the device, demonstrating its potential as the synaptic device for neuromorphic computing applications [6].

1 kbit 1S1R Crossbar Array and the Drive Circuit: To drive the CBA and implement CIM, a system prototype was demonstrated with a digital-to-analog converter (DAC), TDC-based sensing circuit and a FPGA on a customized PCB (as shown in **Fig. 9**). Test pattern of '2D-Si NUS' was written using V/2 programming scheme. The readout current map (**Fig. 10)** shows that the pattern is stored in array with a tolerable error, indicating effective device selectivity and a successful write/read operation in the 1S1R array. The proposed ADC-free compute scheme is illustrated in **Fig. 11**, consisting of a 1S1R CBA, a current-mirror-based subtractor, and a TDC-based sense circuit. The transient response simulated using SkyWater foundry's SKY130 PDK is shown in the right panel of the figure, and the compute time is determined by the delay between V_{start} and V_{stop1} , V_{stop2} etc, to enable parallel asynchronous computing.. To compute four different kernels in parallel, single start signal triggers current measurements and various stop signals are received proportional to MAC current (as shown in **Fig. 12**). Given the state-of-the-art response time of our device, a low-latency TDC-based sensing circuit was implemented using FPGA and a current controlled buffer. By scaling to system level using the TDC-based sensing circuit and FPGA, the propotype enables simultaneous MAC computing operation in all the 32 BLs without the use of multiplexers, demonstrating the true parallel computing. As shown in **Fig. 13**, the TDC-based sense circuit has 2.56 times lower power consumption compared to ADC-based sensing. Hence, a low power-centric peripheral circuit design along with low latency device will play an important role in neuromorphic computing [1].

Fig. 14 shows a parallel computing hardware demonstration of convolutional image processing. The same input image was applied to four different kernels simultaneously. The kernel values were defined by the conductance of memristors in the CBA. The image processing functions are demonstrated by adjusting the kernel values, including soft, horizontal edge detection, vertical edge detection and all edge detection. Fig 14a shows the kernel design for different image filtering kernels and Fig 14b displays the physical readout current map of filters in array. The parallel image processing results are shown in Fig. 14c with clear features.

Conclusion: A heterogenous integration of 2D $HfSe_2$ memristors with silicon-based selectors has been experimentally demonstrated using a 32 × 32 (1 kbit) CBA. The 1S1R achieves state-of-the-art response time as benchmarked in Table 1. Additionally, a low-power sensing circuit and scalable parallel computing have also been demonstrated with FPGA. The realization of such crossbar array with an integrated drive/sense platform enables read/write operations with high energy efficieny. The combination of the low-latency device and low-power peripheral circuits manifests its potential for neuromorphic computing.

References: [1] *IRDS* 2022. [2] S. Chen *et al. Nat. Electron.*, 2020. [3] S. Wang *et al. Nat. Mater.*, 2022. [4] S. Li *et al. SSDM*, 2022. [5] P. Chen *et al. IEDM*, 2017. [6] L. Sun *et al. Nat. Commun.*, 2019. [7] B. Tang *et al. Nat Commun.*, 2022.

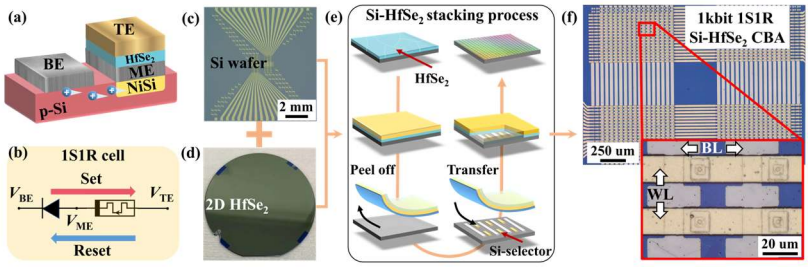

Fig. 1 Schematic of 2D $HfSe_2$ stacking integration with Si-based selectors. (a) Schematic of the Si-$HfSe_2$ 1S1R cell. (b) Equivalent circuit diagram. (c,d) Optical images of the Si-based selector and 2-inch $HfSe_2$ wafer. (e) Si-$HfSe_2$ stacking process for integration. (f) Optical image of the integrated 1kbit 1S1R Si-$HfSe_2$ CBA.

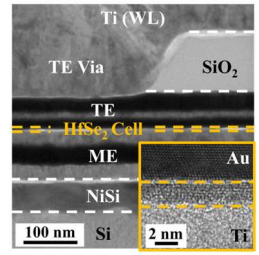

Fig. 2 Cross-section TEM image of 1S1R cell. Inset shows the Ti/$HfSe_2$/Au memristor.

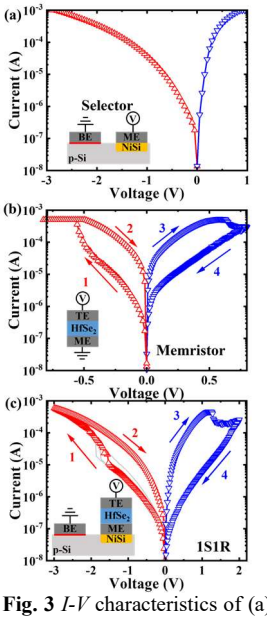

Fig. 3 *I-V* characteristics of (a) the Si-based selector, (b) memristor, and (c) 1S1R cell. The program leakage is reduced to improve the cell selectivity.

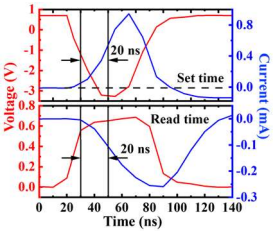

Fig. 4 Fast program/read response (<20 ns) enable high-speed MAC operation.

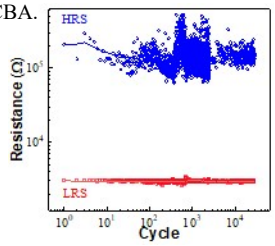

Fig. 5 High endurance (more than 25000 cycles) show potential in CIM. The current was read at 0.7 V.

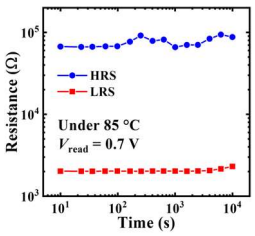

Fig. 6 Stable retention measured under 85 °C.

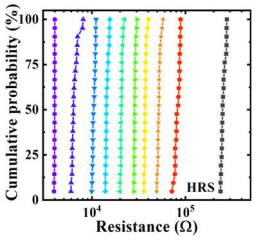

Fig. 7 3-bit MLC with low overlap probability (<0.651%) shows the device can be accurately programmed to multiple synaptic weights.

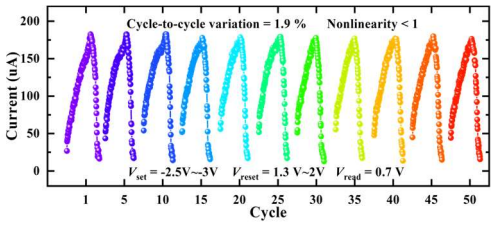

Fig. 8 Gradual set and reset of devices for 50 cycles with low temporal variation of 1.9% and nonlinearity <1, suitable for high-accuracy neural network training [5].

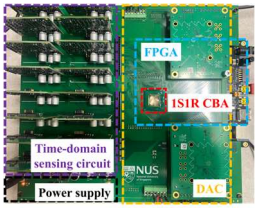

Fig. 9 The drive system on the customized PCB, integrating FPGA, DAC, time-domain sensing circuit, CBA, and power supply.

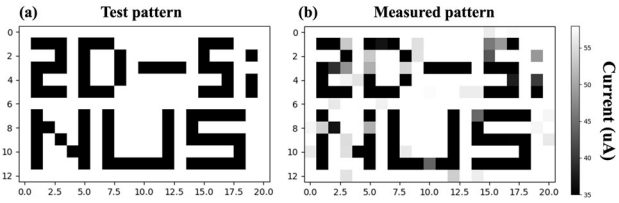

Fig. 10 Test pattern of '2D-Si NUS' is physically written into the CBA, showing negligible cross-talk issue due to the proposed selector.

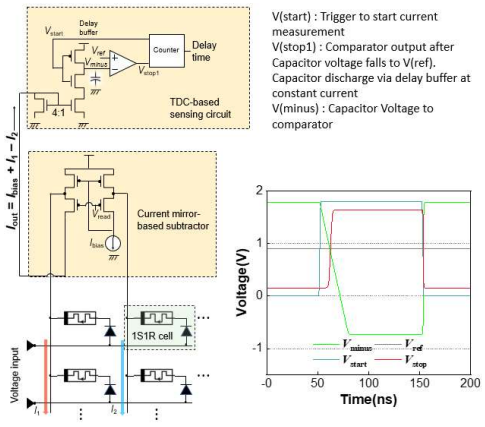

Fig. 11 Structure and operation of the proposed TDC-based compute scheme using the 1S1R array.

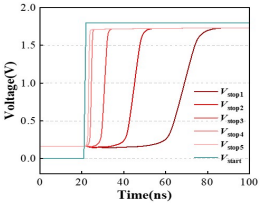

Fig. 12 Parallel Computing of 4 different kernel outputs.

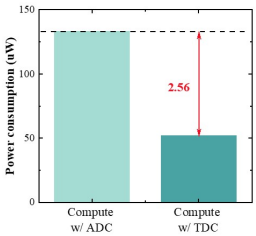

Fig. 13 Power comparison between CIM w/ADC and this work (52 uW).

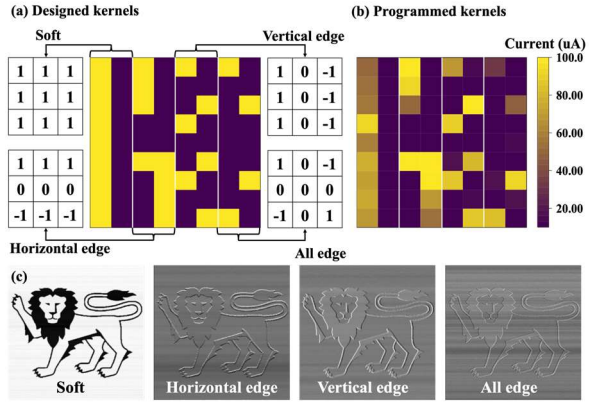

Fig. 14 Parallel computing results. (a) Test kernels written to the CBA. The kernel values are encoded using column differential. (b) Measured current map in CBA. (c) Parallel computing using programmed kernels for image processing. All four images were obtained parallelly using hardware with tolerable error.

Table 1 Comparison with other 2D materials-based CBAs, highlighting the outstanding device performance in our work, including array size, response time, circuit-level demonstration for parallel computing, and sensing power.

	This work	L. Sun [6]	S. Li [4]	B. Tang [7]	S. Chen [2]
Array structure	**1S1R Si/$HfSe_2$**	1S1R hBN/Graphene/hBN	1R $HfSe_2$	1R MoS_2	1R hBN
Array size	**32×32**	12×12	28×4	10×10	10×10
Program/Read Time	**< 20 ns**	~100 ns	< 20 ns	40 ns	200 ns
Program energy	**70 pJ**	1500 pJ	-	-	-
Integration with PCB	**Yes**	No	Yes	No	No
Sensing power	**52 uW**	-	135 uW	-	-

Acknowledgement. This research is supported by A*STAR Science and Engineering Research Council (A2083c0061) and National Research Foundation, Singapore (NRF-CRP24-2020-0002).