nature electronics

Article

https://doi.org/10.1038/s41928-023-01064-1

A large-scale integrated vector–matrix multiplication processor based on monolayer molybdenum disulfide memories

Received: 23 March 2023

Accepted: 10 October 2023

Published online: 13 November 2023

Check for updates

Guilherme Migliato Marega[1,2], Hyun Goo Ji[1,2], Zhenyu Wang[1,2], Gabriele Pasquale[1,2], Mukesh Tripathi[1,2], Aleksandra Radenovic[3] & Andras Kis[1,2]

Data-driven algorithms—such as signal processing and artificial neural networks—are required to process and extract meaningful information from the massive amounts of data currently being produced in the world. This processing is, however, limited by the traditional von Neumann architecture with its physical separation of processing and memory, which motivates the development of in-memory computing. Here we report an integrated 32 × 32 vector–matrix multiplier with 1,024 floating-gate field-effect transistors that use monolayer molybdenum disulfide as the channel material. In our wafer-scale fabrication process, we achieve a high yield and low device-to-device variability, which are prerequisites for practical applications. A statistical analysis highlights the potential for multilevel and analogue storage with a single programming pulse, allowing our accelerator to be programmed using an efficient open-loop programming scheme. We also demonstrate reliable, discrete signal processing in a parallel manner.

Over the past decade, billions of sensors from connected devices have been used to translate physical signals and information to the digital world. Due to their limited computing power, sensors integrated into embedded remote devices often transmit raw and unprocessed data to their hosts. However, the high energy cost of wireless data transmission[1] affects device autonomy and data transmission bandwidth. Improving their energy efficiency could open a new range of applications and reduce their environmental footprint. Furthermore, data processing will move from remote hosts to local sensor nodes; therefore, data transmission would be limited to structured and valuable data, which is desirable for such purposes.

The von Neumann architecture—in which memory and logic units are separate—is seen as the critical factor limiting the efficiency of computing systems in general devices and particularly in edge-based devices. The separation between processing and memory imposed by the von Neumann architecture requires the data to be sent back and forth between the two during data and signal processing or inference in neural networks. This data communication between memory and processing units already accounts for one-third of the energy spent in scientific computing[2].

To overcome the von Neumann communication bottleneck[3,4], in-memory computing architectures—in which memory, logic and processing operations are collocated—are being explored. Processing-in-memory devices are especially suitable for performing vector–matrix multiplication, which is a key operation for data processing and the most intensive calculation in machine-learning algorithms. By taking advantage of the memory's physical layer to perform the multiply–accumulate (MAC) operation, this architecture overcomes the von Neumann communication bottleneck. So far, this processing strategy has been used in applications such as solving linear[5,6] and differential equations[7], signal and image processing[8] and artificial neural network accelerators[9–12]. However, the

[1]Institute of Electrical and Microengineering, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. [2]Institute of Materials Science and Engineering, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. [3]Institute of Bioengineering, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. e-mail: andras.kis@epfl.ch

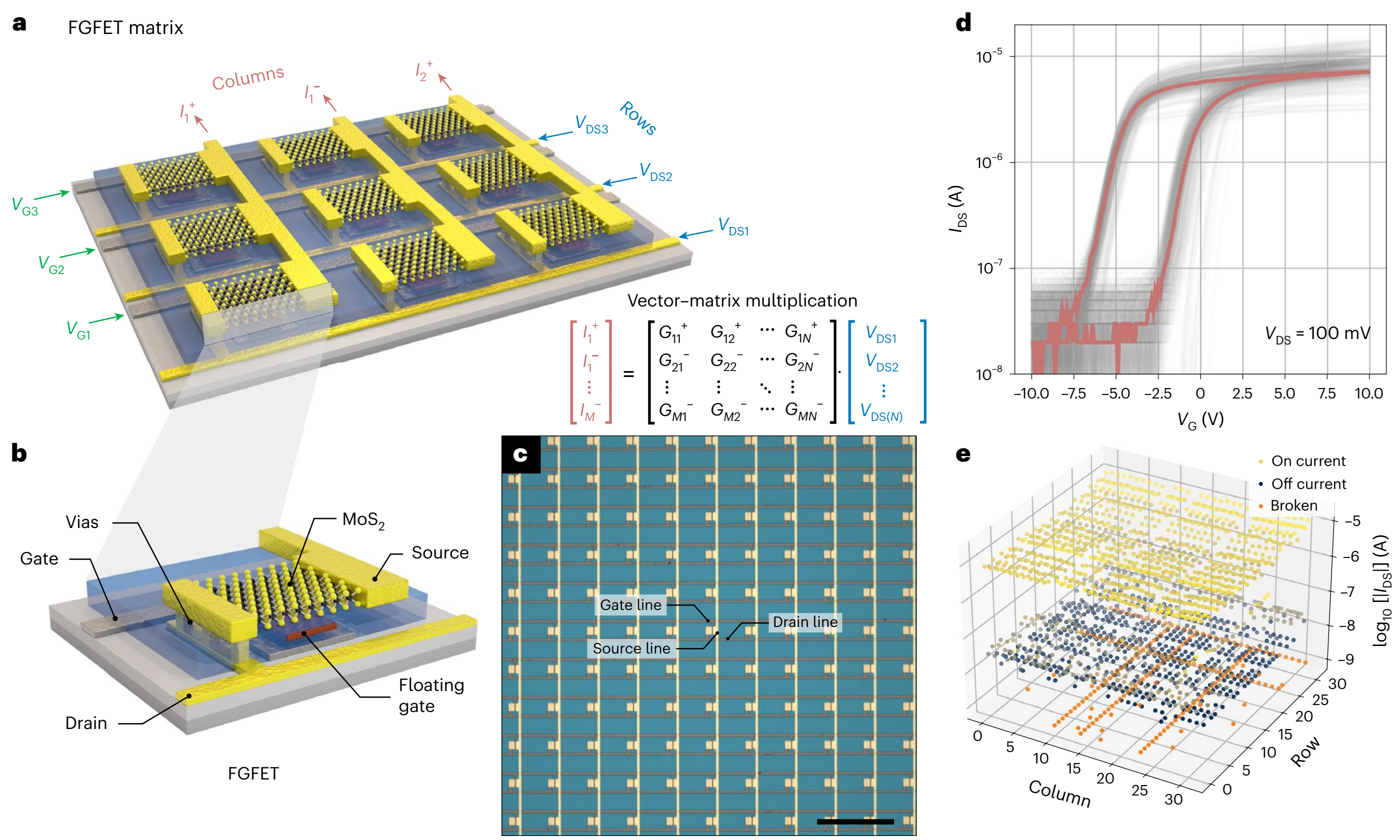

Fig. 1 | Device and matrix description and characterization. a, Three-dimensional rendering of the FGFETs connected into a matrix array. Both gate and drain contacts are organized in rows and the source signal is applied to the columns. The gate signals are applied on the left side and drain signals, on the right. The drain–source current is read from the column. The inset shows the correspondence between signals and vector–matrix multiplication. **b**, Three-dimensional rendering of the FGFET cross section. It shows the different device parts. **c**, Optical image of the memory matrix configuration. Scale bar, 500 µm **d**, I_{DS} versus V_G hysteresis curves of the 851 working devices; the red curve describes the highlighted behaviour of one of the 851 memory devices. The curves coloured in grey correspond to the remaining devices. **e**, Three-dimensional plot shows the mapping of the ON and OFF currents on the 32 × 32 chip. Devices in orange are disconnected.

search for the best materials and devices for this type of processor is still ongoing.

Several devices have been studied for in-memory computing, including standard flash memories, emerging resistive random-access memories and ferroelectric memories[3,13–18]. More recently, two-dimensional (2D) materials have shown promise in the field of beyond-complementary metal–oxide–semiconductor (CMOS) devices[19–24], as well as in-memory and in-sensor computing[25–28]. Due to their atomic-scale thickness, floating-gate field-effect transistors (FGFETs) based on monolayer molybdenum disulfide (MoS_2) offer high sensitivity to charge variations in the floating gate and reduced cell-to-cell interference. Such devices could be scaled down to sub-100 nm lengths without loss of performance[27,29,30]. Moreover, the van der Waals nature of MoS_2 allows devices based on these materials to be integrated into the back-end-of-line[31]. This would allow processors to be fabricated with multiple levels of memory cores directly integrated with the required interfaces, creating dense in-memory networks.

FGFETs based on MoS_2 have been used in logic-in-memory[32] and in-memory computing as well as as the main building blocks of perceptron layers[27,33] where they are projected to offer more than an order of magnitude improvement in power efficiency compared with CMOS-based circuits[30]. These demonstrations have highlighted the promise of 2D materials for in-memory computing, but further progress and practical applications require wafer-scale fabrication and large-scale or very-large-scale system integration. Currently, demonstrations of the wafer-scale and large-scale integration of 2D-semiconducting-materials-based circuits have been limited to photodetectors[34–37] or traditional analogue and digital integrated circuits[38–42]; hardware implementations[43] with full-wafer and large-scale system integration involving 2D-materials-based non-volatile memories that can perform computation are missing.

In this Article, we report a chip containing a 32 × 32 FGFET matrix with 1,024 memory devices per chip and an 83.1% yield. The working devices show similar I_{DS} versus V_G characteristics and hysteresis. During fabrication, we use wafer-scale metal–organic chemical-vapour-deposited (MOCVD) monolayer MoS_2 as the channel material, and the entire fabrication process is carried out in a 4-inch line cleanroom. We also demonstrate multibit data storage in each device with a single programming pulse. Finally, we show that our devices can be used in in-memory computing by performing discrete signal processing with different kernels in a highly parallelized manner.

Memory matrix

We approach in-memory computing by exploiting charge-based memories using monolayer MoS_2 as a channel material. Specifically, we fabricated FGFETs to take advantage of the electrostatic sensitivity of 2D semiconductors[19]. To enable the realization of larger arrays, we organized our FGFETs in a matrix in which we can address individual memory elements by carefully choosing the corresponding row and column. Figure 1a,b shows a three-dimensional rendering of the memory matrix and the detailed structure of each FGFET, respectively. The use of a matrix configuration allows a denser topology and

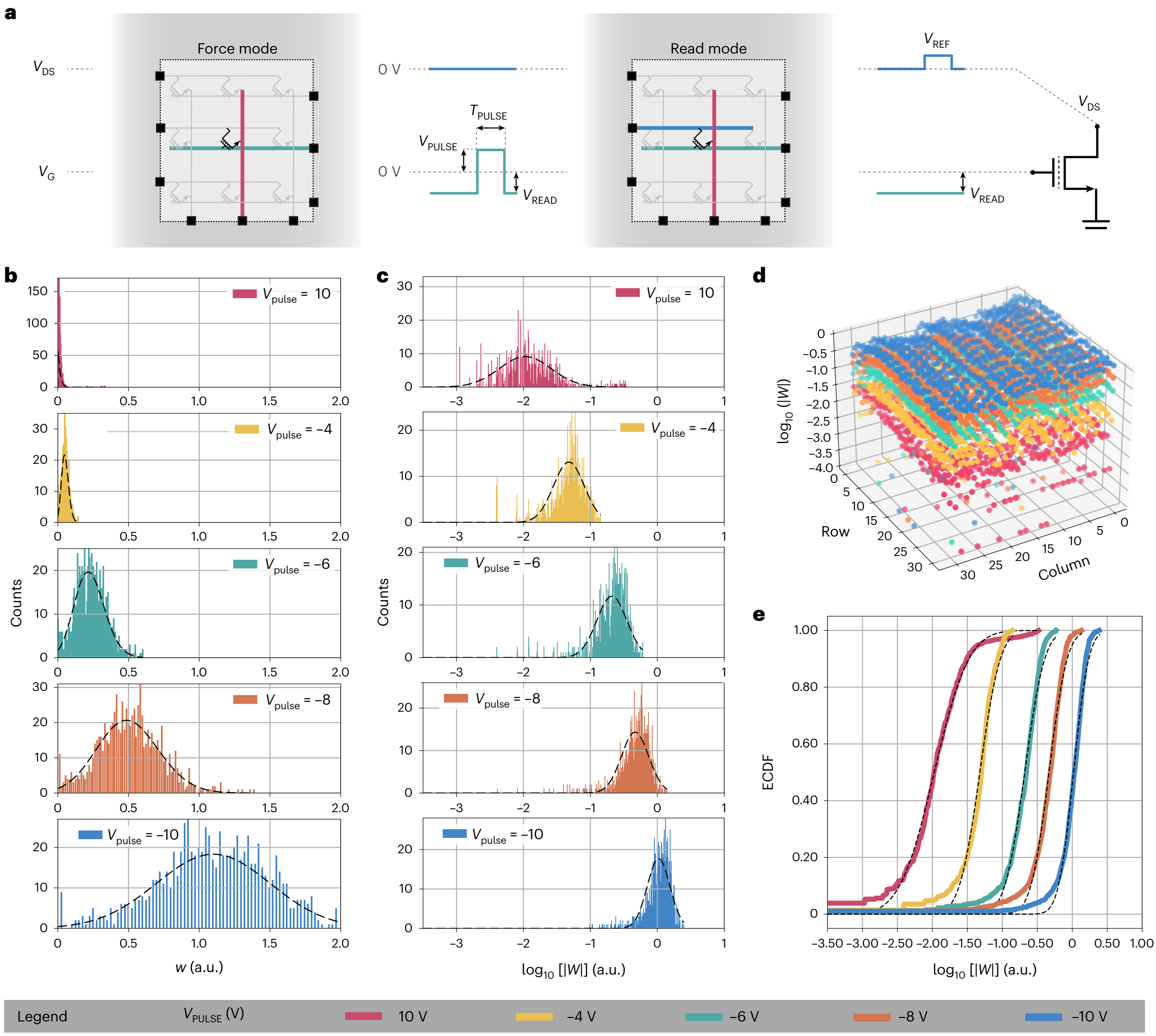

Fig. 2 | Open-loop programming. a, Schematic of the two-state operation of the open-loop programming scheme. In the programming phase, the interface board is used to set the gate and source lines to the low-impedance state and the drain line to the high-impedance state, whereas in the reading phase, all three lines are set to the low-impedance state. **b**, Distribution of output states (w_{OUT}) in the linear scale. The data are fitted with a gamma distribution. **c**, Distribution of output states (w_{OUT}) in the $\log_{10}$ scale. The distributions are fitted with a Gaussian distribution. **d**, Three-dimensional map of $\log_{10}$ value of w_{OUT} as a function of device position and different programming voltages. **e**, Empirical cumulative distribution function (ECDF) as a function of the programmed states in the $\log_{10}$ scale.

directly corresponds to performing vector–matrix multiplications. Our memories are controlled by local 2 nm/40 nm Cr/Pt gates fabricated in a gate-first approach. This allows us to improve the growth of the dielectric by atomic layer deposition[38] and to minimize the number of processing steps that the 2D channel is exposed to, resulting in an improved yield. The floating gate is a 5 nm Pt layer sandwiched between 30 nm HfO_2 (block oxide) and 7 nm HfO_2 (tunnel oxide). Next, we etch vias on the HfO_2 to electrically connect the bottom metal (M1) and top metal (M2) layers. This is required for routing the source and drain signals without an overlap. Wafer-scale MOCVD-grown MoS_2 is transferred on top of the gate stack and etched to form the transistors' channels. Supplementary Figs. 1 and 2 provide details about the material quality and characterization. Finally, 2 nm/60 nm Ti/Au is patterned and evaporated on top, forming the transistors' drain–source contacts as well as the second metal layer. Methods provides further details about the fabrication and Supplementary Figs. 3–8 show the characterization details. Figure 1c shows the optical image of the fabricated chip containing 32 rows and 32 columns for a total of 1,024 memories. In the image, source channels are accessed from the bottom; drain channels, from the right; and gate channels, from the left.

Our memories are based on standard flash memories. The memory mechanism relies on shifting the neutral threshold voltage (V_{TH0}) by changing the number of charges in the trapping layer (ΔQ), that is, the platinum floating gate in our case. When a high positive/negative bias

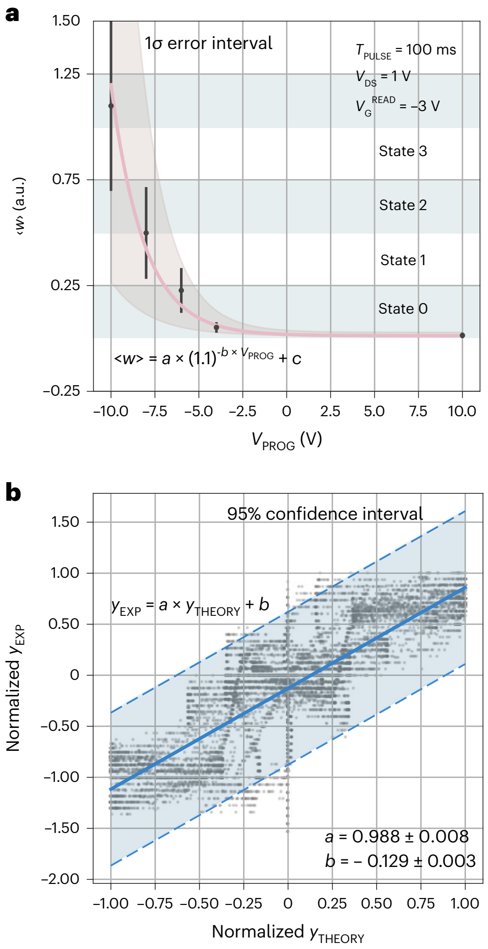

Fig. 3 | MAC operations. a, Output memory states with programming error (<w>) as a function of programming voltage (V_{PROG}). To define the state positions, we perform a fit and select the corresponding state branches for a two-bit open-loop operation **b**, Normalized y_{EXP} versus y_{THEORY} plot, comparing the experimental theoretical results of the MAC operation. The curve is fitted with a linear function with parameters $a = 0.988 \pm 0.008$ and $b = -0.129 \pm 0.003$. The shaded area corresponds to the 95% confidence interval of the linear fitting.

is applied to the gate, the band alignment starts favouring the tunnelling in/out of electrons from the semiconductor to the floating gate, changing the carrier concentration in the trapping layer. We define our memory window (ΔV_{TH}) by taking the difference between the threshold voltage from the forward and reverse paths, which are taken at a constant current level. Our previous work verified the programming mechanism by fitting our experimental curves in a device simulation model[27,29]. Since the memory effect entirely relies on a charge-based process, flash memories tend to have better reliability and reproducibility than emerging memories that are material dependent such as resistive random-access memories and phase-change memories[3]. We designed and manufactured a custom device interface board to facilitate the characterization of the memory array (Supplementary Figs. 9 and 10 provide a detailed description). Figure 1d shows the I_{DS} versus V_G sweeps performed for each device. The fabrication presents a yield of 83.1% and the devices are statistically similar (Supplementary Section 4). The relatively high OFF-state current is due to a lack of resolution of the analogue-to-digital converters used in the setup. High-resolution single-device measurements confirm the typical OFF-state currents on the order of picoamperes. Figure 1e shows the ON and OFF current distribution over the memory matrix. Both ON and OFF currents are taken at $V_{DS} = 100$ mV, forming two distinct planes. The ON and OFF current shows a good distribution over the entire matrix. Supplementary Figs. 13 and 14 show further detailed single-device characterization, confirming the performance of devices as memories with good retention and endurance stabilities. We show that the devices have a statistically similar memory window $\Delta V_{TH} = 4.30 \pm 0.25$ V. This value is smaller compared with the one extracted from single-device measurements due to the higher slew rates (5 V s^{-1}) required for the time-effective characterization of 1,024 devices in the matrix.

Open-loop programming

The similarity of the devices motivates us to pursue a statistical study of the memories' programming behaviour. In the context of in-memory computing, an open-loop programming analysis is fundamental. Standard write–verify approaches may be too time-consuming when programming a large flash memory array. A statistical understanding of memory states in an open loop is essential to improve the performance and speed.

We perform the experiment such that each device is independently excited by selecting the corresponding row (i) and column (j). Analogue switches in the device interface board keep a low-impedance path in the selected row (i)/column (j) and high impedance in the remaining rows and columns. This ensures that a potential difference is applied only to the desired device, avoiding unwanted programming. For the same reason, we divide the device programming and reading into two independent stages. During the programming phase, the corresponding gate line (row) and the corresponding source line (column) are selected and programming pulses with parameters T_{PULSE} and V_{PULSE} are applied in the gate. Due to the tunnelling nature of the device, only two terminals are required to generate the band bending needed for charge injection into the floating gate. After the pulse, the gate voltage is changed to V_{READ}, which is low enough to prevent reprogramming the memory state. In the reading phase, the drain line is also connected, and the conductance value is probed by applying voltage V_{DS} to the drain. This two-stage procedure is required because we are using a three-terminal device; therefore, both gate and drain share the same row, and consequently, the entire row is biased when the gate and drain lines are engaged. If high voltages in the gate were applied when the drain line is connected, the whole row would be reprogrammed, causing a loss of information in the memories. Figure 2a shows the description of this two-stage programming procedure.

For the subsequent measurements, we used $V_{READ} = -3$ V, $V_{DS} = 1$ V and $T_{PULSE} = 100$ ms. Before each measurement, we reset the memories by applying a positive 10 V pulse, which puts the devices into a low-conductance state. Due to parasitic resistances in the matrix, a linear compensation in the digital gains is applied (Supplementary Figs. 17 and 18 provide further details). The compensation method improves the programming reliability of the devices by an order of magnitude. We estimate a programming error of 500 errors per million for programming one bit and having one error per million for programming the erase state. Figure 2b,c shows the distribution of memory states after different pulse intensities, namely, V_{PULSE} = +10 V, −4 V, −6 V, −8 V and −10 V, in both linear and logarithmic representations. We observe that on a linear scale, the increase in the pulse amplitude is accompanied by a higher memory state value and a larger spread. On the other hand, by analysing the logarithm of the state value, we can see that the memory has well-defined storage states. This leads us to conclude that this memory has the potential for multivalued storage without write–verify algorithms, especially when used on a logarithmic scale.

Figure 2d shows the spatial distribution of the states on the entire chip. We observe that the memory states create a constant plane value for the different programming voltages, V_{PULSE}. Finally, Fig. 2e shows the empirical cumulative distribution function (ECDF) of the logarithmic representation. These results support the possibility of multivalued programming, as discussed previously, and indicate that the memory elements can be used for storing analogue weights for in-memory computing.

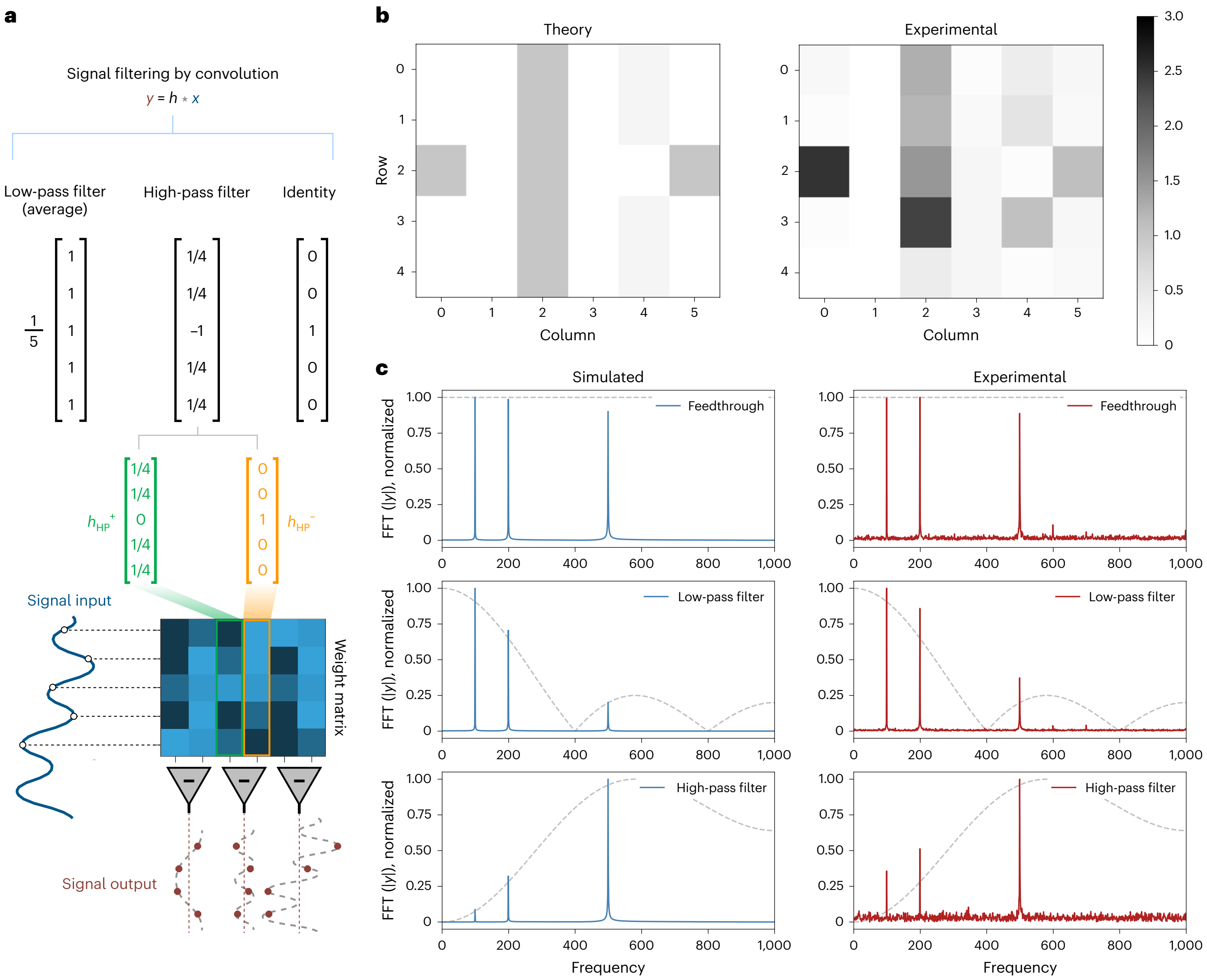

Fig. 4 | Signal processing based on in-memory processing. a, Description of convolution-based signal processing for different filters (low-/high-pass filters and identity). *y*, processed signal; *x*, input signal; *h*, filter kernel. The kernel is split between its positive and negative components; these values are proportionally transferred to the memory weights. The input signal is simultaneously applied to all the memories and the difference between the output of two columns is the result of the processed signal for a given kernel. **b**, Comparison of the theoretical kernel weight mapping and the experimental weight transfer into the conductance of the memories. **c**, Comparison of the fast Fourier transform (FFT) of the simulated and experimental output signals after each kernel.

States and vector–matrix multiplications

With the open-loop analysis completed (Fig. 3a), we plot the memory states ($<w>$) as a function of the programming voltage (V_{PROG}). We define four equally distributed states (two-bit resolution) to be programmed as discrete weights in the matrix for the vector–matrix multiplication (Supplementary Fig. 20). To analyse the effectiveness of the processor for performing vector–matrix operations, we compare (Fig. 3b) the normalized theoretical (y_{THEORY}) value with the normalized experimental (y_{EXP}) value obtained on several dot-product operations. The linear regression of the experimental points shows a line with parameters $\mathbf{a} = 0.988 \pm 0.008$ and $\mathbf{b} = -0.129 \pm 0.003$ for $y_{\mathrm{EXP}} = \mathbf{a} \times y_{\mathrm{THEORY}} + \mathbf{b}$, whereas the shaded area corresponds to a 95% confidence interval. The ideal processor should converge to $\mathbf{a} = 1$ and $\mathbf{b} = 0$ with a confidence interval that converges to linear fitting. In our case, the processor has a linear behaviour converging to the ideal case, with a large spread and slight nonlinearity of the experimental values. We explain this behaviour by the non-ideality of the memories and the quantization error due to the limited resolution of the states. This shift in parameter **b** can be explained by the intrinsic transimpedance amplifier offset with memory leakage seen at $y_{\mathrm{THEORY}} = 0$, but it does not affect the observed linear trend. We conclude that we can perform MAC operations with reasonable accuracy. This operation is needed for performing diverse types of algorithms, such as signal processing and inference in artificial neural networks.

Signal processing

Next, we configure this accelerator to perform signal processing to demonstrate a real-world scenario and application. For signal processing, the input signal (*x*) is convoluted with a kernel (*h*), resulting in the processed signal (*y*). Depending on the nature of the kernel elements, different types of processing can be achieved. Here we limit ourselves to three different kernels that perform low-pass filtering, high-pass filtering and feedthrough. All the kernels run in parallel within a single processing cycle, demonstrating the efficiency of this processor targeting data-centric problems by parallelized processing. More kernels could be added in parallel, limited only by the size of the matrix. Figure 4a shows the convolution operation and the different kernels used for processing the input signal. The strategy to encode negative kernel values into the conductance values of the memories is to split the kernel (h) into a kernel with only the positive values (h^+) and one with the absolute values of the negative numbers (h^-) and encode only the positive numbers with a direct relation with the conductance values (G). After the processing is realized, the outputs of the positive (y^+) and negative (y^-) kernels are subtracted ($y^+ - y^-$), resulting in the final signal (y).

Figure 4b shows the comparison between the original weights and the weights transferred into the memory matrix using the previously described open-loop programming scheme. To simplify the transfer, we normalize the weight values at each kernel by its maximum value. As a result, we observe a good agreement between the original and experimental values. Next, to verify the effectiveness of processing, we first construct our input signal (x) as a sum of sinusoidal waves with different frequencies. In this way, we can easily probe the behaviour of the filters at different frequencies without creating an overly complex signal. Since the signal has positive and negative values, the signal amplitude must fall within the linear region of device operation. Thus, we restrict the signal range from −100 to 100 mV at $V_{\mathrm{READ}} = 0$. Figure 4c shows the fast Fourier transform of the simulated processed signals (left) and experimental signals (right). The grey line in both simulated and measured signals is the fast Fourier transform of each kernel, giving a guideline for the predicted behaviour of each operation. We highlight that the experimental processing of all three filters matches fairly well with the theoretical values as well as the prototype filter. Altogether, large-scale arrays of FGFETs based on 2D materials could be used for other applications such as image processing and inference with artificial neural networks.

Conclusions

We have reported the large-scale integration of 2D materials as the semiconducting channel in an in-memory processor. We demonstrated the reliability and reproducibility of our devices both in terms of characterization and statistical similarity of the programming states in open-loop programming. The processor carries out vector–matrix multiplications and illustrates its functionality by performing discrete signal processing. Our approach could allow in-memory processors to reap the benefits of 2D materials and bring new functionality to edge devices for the Internet of Things.

Methods

Wafer-scale memory fabrication

The fabrication starts with a p-doped silicon substrate with a 270-nm-thick SiO_2 insulating layer. The first metal layer and FGFET gates were fabricated by photolithography using an MLA150 advanced maskless aligner with a bilayer 0.4-µm-thick LOR 5A/1.1-µm-thick AZ 1512 resist. The 2 nm/40 nm Cr/Pt gate metals were evaporated using an electron-beam evaporator under a high vacuum. After resist removal by dimethyl sulfoxide, deionized water and O_2 plasma are used to further clean and activate the surface for HfO_2 deposition. The 30-nm-thick HfO_2 blocking oxide is deposited by thermal atomic layer deposition using TEMAH and water as precursors with the deposition chamber set at 200 °C. The 5 nm Pt floating gates were patterned by photolithography and deposited using the same process as described previously. With the same atomic layer deposition system, we deposit the 7-nm-thick HfO_2 tunnel oxide layer with the same process mentioned before. Next, vias are exposed using a single-layer 1.5-µm-thick ECI 3007 photoresist and etched by Cl_2/BCl_3 chemistry reactive ion etching. After the transfer of MoS_2 onto the substrate, patterning it with photolithography using a 2-µm-thick nLOF resist and etching by O_2 plasma. Drain–source electrodes are patterned by photolithography and 2 nm/60 nm Ti/Au is deposited by electron-beam evaporation. To increase the adhesion of contacts and MoS_2 onto the substrate, a 200 °C annealing step is performed in a high vacuum. The devices have a width/length ratio of 49.5 µm/3.1 µm.

Device passivation

The fabricated device is first wire-bonded onto a 145-pin pin-grid-array chip carrier. The device is heated inside an Ar glovebox at 135 °C for 12 h, which removes the adsorbed water from the device surface. After in situ annealing in the glovebox, a lid is glued onto the chip carrier using a high-vacuum epoxy and cured in an Ar atmosphere. This protects the device from oxygen and water.

Transfer procedure

The MOCVD-grown material is first spin coated with PMMA A2 at 1,500 r.p.m. for 60 s and baked at 180 °C for 5 min. Next, we attach a 135 °C thermal release tape onto the MoS_2 sample and detach it from sapphire in deionized water. After this, we dry the film and transfer it onto the patterned substrate. Next, we bake the stack at 55 °C for 1 h. We remove the thermal release tape by heating it on the hot plate at 130 °C. Next, we immerse the sample in an acetone bath for cleaning the tape polymer residues. Finally, we transfer the wafer to an isopropanol bath and dry it in air.

MOCVD growth

Monolayer MoS_2 was grown using the MOCVD method. $Mo(CO)_6$, Na_2MoO_4 and diethyl sulfide were used as precursors. NaCl was spin coated as a catalyst. A pre-annealed three-inch *c*-plane sapphire wafer with a small off-cut angle (<0.2°) was used as a growth substrate (UniversityWafer). The chemical vapour deposition reaction was performed using a home-built furnace system with a four-inch quartz tube reactor and mass flow controllers connected with Ar, H_2, O_2 and metal–organic precursors ($Mo(CO)_6$ and diethyl sulfide). For the MoS_2 crystal growth, a reactor was heated to 870 °C at ambient pressure for 20 min.

Electrical measurements

The electrical measurements were performed using a custom device interface board connected to a CompactRIO (cRIO-9056) running a real-time LabVIEW 2020 server. We installed the NI-9264 (16-channel analogue output), NI-9205 (32-channel analogue inputs) and NI-9403 (digital input/output) modules.

Data availability

The data that support the findings of this study are available via Zenodo at https://doi.org/10.5281/zenodo.8383470.

References

1. Xu, X. et al. Scaling for edge inference of deep neural networks. *Nat. Electron.* **1**, 216–222 (2018).
2. Kestor, G., Gioiosa, R., Kerbyson, D. J. & Hoisie, A. Quantifying the energy cost of data movement in scientific applications. In *2013 IEEE International Symposium on Workload Characterization (IISWC)* 56–65 (IEEE, 2013).
3. Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R. & Eleftheriou, E. Memory devices and applications for in-memory computing. *Nat. Nanotechnol.* **15**, 529–544 (2020).
4. McKee, S. A. Reflections on the memory wall. In *Proc. 1st Conference on Computing Frontiers—CF'04* 162 (ACM Press, 2004).
5. Sun, Z., Pedretti, G., Bricalli, A. & Ielmini, D. One-step regression and classification with cross-point resistive memory arrays. *Sci. Adv.* **6**, eaay2378 (2020).
6. Sun, Z. et al. Solving matrix equations in one step with cross-point resistive arrays. *Proc. Natl Acad. Sci. USA* **116**, 4123–4128 (2019).

7. Zidan, M. A. et al. A general memristor-based partial differential equation solver. *Nat. Electron.* **1**, 411–420 (2018).
8. Li, C. et al. Analogue signal and image processing with large memristor crossbars. *Nat. Electron.* **1**, 52–59 (2018).
9. Lin, P. et al. Three-dimensional memristor circuits as complex neural networks. *Nat. Electron.* **3**, 225–232 (2020).
10. Wang, Z. et al. Reinforcement learning with analogue memristor arrays. *Nat. Electron.* **2**, 115–124 (2019).
11. Yao, P. et al. Fully hardware-implemented memristor convolutional neural network. *Nature* **577**, 641–646 (2020).
12. Wang, Z. et al. Fully memristive neural networks for pattern classification with unsupervised learning. *Nat. Electron.* **1**, 137–145 (2018).
13. Khaddam-Aljameh, R. et al. HERMES-Core—a 1.59-TOPS/mm^2 PCM on 14-nm CMOS in-memory compute core using 300-ps/LSB linearized CCO-based ADCs. *IEEE J. Solid-State Circuits* **57**, 1027–1038 (2022).
14. Jung, S. et al. A crossbar array of magnetoresistive memory devices for in-memory computing. *Nature* **601**, 211–216 (2022).
15. Berdan, R. et al. Low-power linear computation using nonlinear ferroelectric tunnel junction memristors. *Nat. Electron.* **3**, 259–266 (2020).
16. Ielmini, D. & Wong, H.-S. P. In-memory computing with resistive switching devices. *Nat. Electron.* **1**, 333–343 (2018).
17. Bavandpour, M., Sahay, S., Mahmoodi, M. R. & Strukov, D. B. 3D-aCortex: an ultra-compact energy-efficient neurocomputing platform based on commercial 3D-NAND flash memories. *Neuromorph. Comput. Eng.* **1**, 014001 (2021).
18. Merrikh-Bayat, F. et al. High-performance mixed-signal neurocomputing with nanoscale floating-gate memory cell arrays. *IEEE Trans. Neural Netw. Learn. Syst.* **29**, 4782–4790 (2018).
19. Radisavljevic, B., Radenovic, A., Brivio, J., Giacometti, V. & Kis, A. Single-layer MoS_2 transistors. *Nat. Nanotechnol.* **6**, 147–150 (2011).
20. Ciarrocchi, A. et al. Polarization switching and electrical control of interlayer excitons in two-dimensional van der Waals heterostructures. *Nat. Photon.* **13**, 131–136 (2019).
21. Bertolazzi, S., Krasnozhon, D. & Kis, A. Nonvolatile memory cells based on MoS_2/graphene heterostructures. *ACS Nano* **7**, 3246–3252 (2013).
22. Sangwan, V. K. et al. Gate-tunable memristive phenomena mediated by grain boundaries in single-layer MoS_2. *Nat. Nanotechnol.* **10**, 403–406 (2015).
23. Shen, P.-C., Lin, C., Wang, H., Teo, K. H. & Kong, J. Ferroelectric memory field-effect transistors using CVD monolayer MoS_2 as resistive switching channel. *Appl. Phys. Lett.* **116**, 033501 (2020).
24. Desai, S. B. et al. MoS_2 transistors with 1-nanometer gate lengths. *Science* **354**, 99–102 (2016).
25. Paliy, M., Strangio, S., Ruiu, P. & Iannaccone, G. Assessment of two-dimensional materials-based technology for analog neural networks. *IEEE J. Explor. Solid-State Computat.* **7**, 141–149 (2021).
26. Feng, X. et al. Self-selective multi-terminal memtransistor crossbar array for in-memory computing. *ACS Nano* **15**, 1764–1774 (2021).
27. Migliato Marega, G. et al. Low-power artificial neural network perceptron based on monolayer MoS_2. *ACS Nano* **16**, 3684–3694 (2022).
28. Mennel, L. et al. Ultrafast machine vision with 2D material neural network image sensors. *Nature* **579**, 62–66 (2020).
29. Giusi, G., Marega, G. M., Kis, A. & Iannaccone, G. Impact of interface traps in floating-gate memory based on monolayer MoS. *IEEE Trans. Electron Devices* **69**, 6121–6126 (2022).
30. Cao, W., Kang, J., Bertolazzi, S., Kis, A. & Banerjee, K. Can 2D-nanocrystals extend the lifetime of floating-gate transistor based nonvolatile memory? *IEEE Trans. Electron Devices* **61**, 3456–3464 (2014).
31. Hu, V. P.-H. et al. Energy-efficient monolithic 3-D SRAM cell with BEOL MoS_2 FETs for SoC scaling. *IEEE Trans. Electron Devices* **67**, 4216–4221 (2020).
32. Migliato Marega, G. et al. Logic-in-memory based on an atomically thin semiconductor. *Nature* **587**, 72–77 (2020).
33. Zhu, K. et al. Hybrid 2D–CMOS microchips for memristive applications. *Nature* **618**, 57–62 (2023).
34. Hinton, H. et al. A 200 ×256 image sensor heterogeneously integrating a 2D nanomaterial-based photo-FET array and CMOS time-to-digital converters. In *2022 IEEE International Solid-State Circuits Conference (ISSCC)* **65**, 1–3 (IEEE, 2022).
35. Dodda, A. et al. Active pixel sensor matrix based on monolayer MoS_2 phototransistor array. *Nat. Mater*. **21**, 1379–1387 (2022).
36. Jang, H. et al. An atomically thin optoelectronic machine vision processor. *Adv. Mater.* **32**, 2002431 (2020).
37. Ma, S. et al. A 619-pixel machine vision enhancement chip based on two-dimensional semiconductors. *Sci. Adv.* **8**, eabn9328 (2022).
38. Yu, L. et al. Design, modeling, and fabrication of chemical vapor deposition grown MoS_2 circuits with E-mode FETs for large-area electronics. *Nano Lett.* **16**, 6349–6356 (2016).
39. Ma, S. et al. An artificial neural network chip based on two-dimensional semiconductor. *Sci. Bull.* **67**, 270–277 (2022).
40. Wang, X. et al. Analog and logic circuits fabricated on a wafer-scale two-dimensional semiconductor. In *2022 International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA)* 1–2 (IEEE, 2022).
41. Polyushkin, D. K. et al. Analogue two-dimensional semiconductor electronics. *Nat. Electron.* **3**, 486–491 (2020).
42. Wachter, S., Polyushkin, D. K., Bethge, O. & Mueller, T. A microprocessor based on a two-dimensional semiconductor. *Nat. Commun.* **8**, 14948 (2017).
43. Chen, S. et al. Wafer-scale integration of two-dimensional materials in high-density memristive crossbar arrays for artificial neural networks. *Nat. Electron.* **3**, 638–645 (2020).

Acknowledgements

We thank Z. Benes (CMI) for help with the electron-beam lithography and R. Chiesa for assistance with the energy-dispersive X-ray measurements. Device preparation was carried out in the EPFL Centre of MicroNanotechnology (CMI). We thank B. Bartova and R. Therisod (CIME) for device cross-sectioning and transmission electron microscopy imaging, which were carried out at the EPFL Interdisciplinary Centre for Electron Microscopy (CIME). We acknowledge support from the European Union's Horizon 2020 research and innovation programme under grant agreement nos. 829035 QUEFORMAL (to G.M.M., Z.W. and A.K.), 785219 and 881603 (Graphene Flagship Core 2 and Core 3) to A.K. and 964735 (EXTREME-IR) to H.J and A.K.; the European Research Council (ERC, grant nos. 682332 and 899775, to H.J., M.T. and A.K.); the CCMX Materials Challenge grant 'Large area growth of 2D materials for device integration' (to A.R. and A.K.); and the Swiss National Science Foundation (grant no. 175822, to G.P. and A.K.).

Author contributions

A.K. initiated and supervised the project. G.M.M. fabricated the devices, designed/prepared the measurement setup and performed the device characterization and remaining measurements. H.J. and Z.W. grew the 2D materials and assisted in materials characterization under the supervision of A.R. M.T. performed the high-resolution transmission electron microscopy for the characterization of devices and materials. G.P. performed the atomic force microscopy imaging and elemental characterization. A.K. and G.M.M. analysed the data. The manuscript was written by G.M.M. and A.K. with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41928-023-01064-1.

Correspondence and requests for materials should be addressed to Andras Kis.

Peer review information *Nature Electronics* thanks Su-Ting Han, Jing-Kai Huang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023, corrected publication 2023