nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-23719-3 OPEN

Check for updates

An in-memory computing architecture based on two-dimensional semiconductors for multiply-accumulate operations

Yin Wang[1,3], Hongwei Tang[1,3], Yufeng Xie[1,3], Xinyu Chen[1], Shunli Ma[1], Zhengzong Sun[1], Qingqing Sun[1], Lin Chen[1], Hao Zhu[1], Jing Wan[1], Zihan Xu[2], David Wei Zhang[1], Peng Zhou[1]✉ & Wenzhong Bao[1]✉

In-memory computing may enable multiply-accumulate (MAC) operations, which are the primary calculations used in artificial intelligence (AI). Performing MAC operations with high capacity in a small area with high energy efficiency remains a challenge. In this work, we propose a circuit architecture that integrates monolayer MoS_2 transistors in a two-transistor–one-capacitor (2T-1C) configuration. In this structure, the memory portion is similar to a 1T-1C Dynamic Random Access Memory (DRAM) so that theoretically the cycling endurance and erase/write speed inherit the merits of DRAM. Besides, the ultralow leakage current of the MoS_2 transistor enables the storage of multi-level voltages on the capacitor with a long retention time. The electrical characteristics of a single MoS_2 transistor also allow analog computation by multiplying the drain voltage by the stored voltage on the capacitor. The sum-of-product is then obtained by converging the currents from multiple 2T-1C units. Based on our experiment results, a neural network is ex-situ trained for image recognition with 90.3% accuracy. In the future, such 2T-1C units can potentially be integrated into three-dimensional (3D) circuits with dense logic and memory layers for low power in-situ training of neural networks in hardware.

[1] State Key Laboratory of ASIC and System, School of Microelectronics, Fudan University, Shanghai, China. [2] Shenzhen Sixcarbon Technology, Shenzhen, China. [3] These authors contributed equally: Yin Wang, Hongwei Tang, Yufeng Xie. ✉email: pengzhou@fudan.edu.cn; baowz@fudan.edu.cn

Artificial intelligence (AI) algorithms require significant computing power for running successive matrix calculations. Multiply accumulate (MAC) is the most critical operation in AI computation at the chip level. In-memory computing is a technology that uses memory devices assembled in an array to execute MAC operations[1]. As such, it has triggered extensive research interests because data transfer in a conventional von Neumann architecture has a bottleneck between memory and logic circuits[2,3], and a memory device capable of in-memory computing can be used to carry out high-throughput MAC operations directly[4,5]. For an ideal in-memory computing, various features are preferred for its memory portion, including a nonvolatile characteristic, multi-bit storage capability, long cycling endurance, simple erase/write operation, etc.[1,4,6].

Various types of memory devices have been investigated for performing MAC operations. Among them, nonvolatile memory devices, include resistive random-access memory (RRAM)[7–9], phase change RAM (PCRAM)[10–13], spin-transfer torque magnetoresistive RAM (STT-MRAM)[14,15], and conventional FLASH[16–18]. Most nonvolatile memories can realize multi-bit storage, but they usually exhibit a stochastic nature, resulting in a learning accuracy loss in the neural network applications[1,5]. Their limited cycling endurance (FLASH ~10^5, RRAM and PCRAM 10^6–10^9) and relatively complex memory operation[19] are also unsuitable for frequent weight update processes required for in-memory computing[20]. For example, FLASH usually requires high voltages for the write operation. RRAM/PCRAM requires continuous voltage pulses to tune the conductive filaments to control the electrical conductance, which complicates the multiplication operation[5]. STT-MRAM requires a relatively large current to program information in the storage element, which carries greater dynamic power dissipation and overall write energy cost[4,15]. On the other hand, volatile memory devices can also execute in-memory computing, such as static random accesses memory (SRAM)[21–23] and dynamic random-access memory (DRAM)[24–26]. Theoretically, they have much higher programming speed and superior endurance(>10^{16})[1,4], but in volatile memories, the stored information dissipates quickly, and a periodic refresh operation is required[24]. Furthermore, SRAM and DRAM belong to binary memory, and their main applications are limited in the binary-weighted network[1,25,26]. An overall comparison among different types of in-memory computing technologies is also concluded in Supplementary Table S1.

Other than exploring different memory technologies for in-memory computation, suitable channel material is also critical. Two-dimensional layered materials (2DLMs), well-known for their intrinsic nature of atomic thickness, allow aggressive channel length scaling owing to its superior electrostatic control that can substantially suppress short-channel effects[27]. In addition, unlike rigid silicon CMOS, 2DLMs can enable flexible electronic circuitry with multiple sensing functionalities, adding value towards a multifunctional hardware platform[28]. Among various 2DLMs, semiconductive transition metal dichalcogenides (TMDs) are promising due to their rich band structures and tunable bandgaps[29], and molybdenum disulfide (MoS_2) is one representative that has been extensively investigated in the past few years[30,31]. Compared to silicon and other TMDs with a narrower bandgap, monolayer MoS_2 has a relatively wide bandgap (~1.8 eV) to enable a large current on/off ratio in its field-effect transistors (FETs)[32]. Now wafer-scale continuous MoS_2 films can already be synthesized by chemical vapor deposition (CVD) methods[33] and transferred to arbitrary substrates[34]. The device processing techniques have also been intensively investigated to address early criticism of 2D-FETs, such as the realization of Ohmic contact and integration of high-k dielectrics[35–37]. Therefore, recent exploration of 2DLMs has been expanded from fundamental investigations to the demonstration of circuit-level device applications, such as memories, logic gates, and sensors[35,38,39]. A 1T-1R structured in-memory computation unit has also been demonstrated lately, in which a MoS_2 FET is used as a selector, and a HfOx-based RRAM is used to perform analog calculation[40].

In this work, we explored and designed a MAC circuit architecture in a 2T–1C configuration, which includes two MoS_2 FETs and one metal-insulator-metal capacitor. In such a structure, the 1T–1C portion acts as a DRAM cell. Owing to the ultralow leakage current of the MoS_2 FETs, a voltage with 8-level (3 bits) quantization can be stored on a capacitor with longer than 10 s retention time, enough for additional complex operations. The stored voltage is connected to the gate of the second MoS_2 transistor, in which the input drain bias V_d and gate bias V_g can determine the drain current I_d to realize an analog multiplication operation. Moreover, the current in multiple 2T–1C rows can be converged together, giving an addition operation. Based on two identical 2T–1C cells, we demonstrate a simple MAC operation circuit, which is the core module for the convolution operation in an artificial neural network. A more complicated MAC array was trained against the MNIST handwritten digit database and used for image recognition. The successful recognition rate was found to reach 90.3%. Our 2T–1C MoS_2 cells highlight the promising potential of in-memory computing and in situ training of neural networks based on emerging 2D semiconductors to overcome the bottleneck of von Neumann computing.

Results and discussion

Figure 1a shows a wafer-scale MoS_2 film grown using the CVD method (see SI). Raman spectra (Fig. 1b) gathered from different positions in the MoS_2 film show acceptable spatial uniformity, which is vital for performing accurate analog calculations in our circuit. The transfer characteristics (Fig. 1c) of 24 MoS_2 FETs on a $1 \times 1\ \mathrm{cm}^2$ wafer exhibit large on/off current ratios (~10^7) and an acceptable homogeneity level. We fabricated a 2T–1C cell (optical microscopic image shown in Fig. 1d) to provide charge storage and analog computation. Figure 1e shows a circuit schematic of such a 2T–1C cell; the left 1T–1C structure forms a dynamic memory in which the MoS_2 FET is labeled T_1, and the MoS_2 FET T_2 on the right side is used to accomplish the multiplication calculation. Figure 1f schematically illustrates its 3D structure, and the fabrication process is described in the “Methods” section.

The refresh voltage V_{re} on the refresh line (RL) controls the ON/OFF state of transistor T_1. During a write operation, T_1 is turned on and the signal V_w applied by the weight line (WL) then charges the capacitor, which indicates the weight has been written into this 2T–1C cell. During the hold operation, T_1 is turned off by applying a negative V_{re}. Due to the ultralow leakage current in the MoS_2 channel in the OFF state (see Fig. S1), the charge stored in the capacitor can be held for a long time to maintain the voltage that acts as a gate voltage for T_2. Since the input V_x is applied as a drain voltage to T_2, the drain current (I_d) in T_2 is controlled with a combination of V_w and V_x. If the applied V_w and V_x locate in a relatively linear range of the output and transfer characteristics for the MoS_2 FET, an analog multiplication operation between I_d, V_x, and V_w can be realized, which will be discussed in detail later in this paper.

We now propose an array circuit based on such a MoS_2 2T–1C unit cell to implement a MAC operation in an electrical circuit. The circuit diagram is displayed in Fig. 1g, which corresponds to a MAC operation $Y_m = \sum_{k=1}^{n} V_{xk} \times W_{km}$ (Fig. 1h). In each unit, the weight W_{nm} is stored in the capacitor and updated using the RL and WL. The input voltage V_{xn} is then applied to the entire

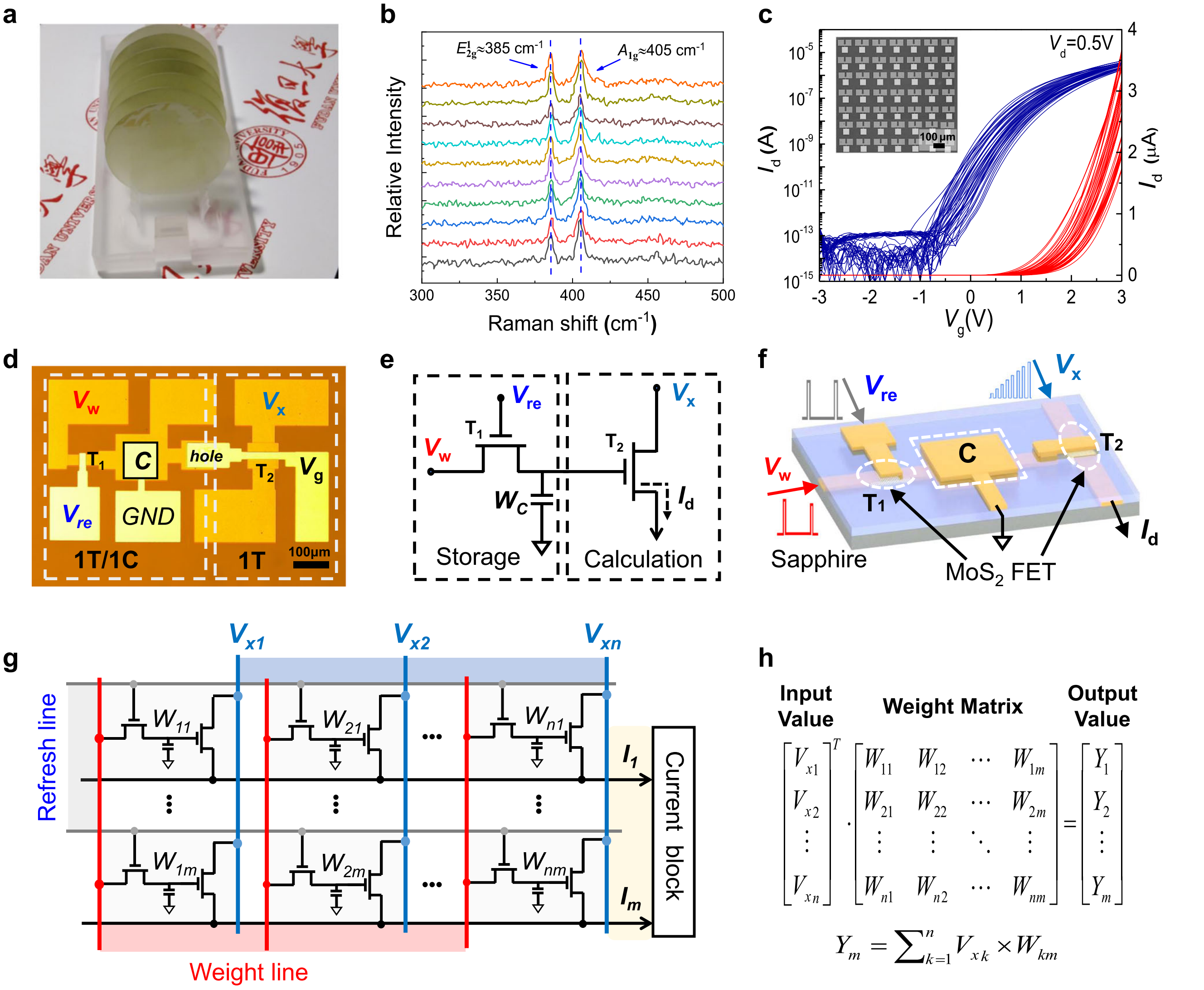

Fig. 1 2T-1C unit cell and circuits fabricated on a wafer-scale MoS$_2$ film. **a** Wafer-scale MoS$_2$ continuous films are batch-synthesized by a CVD method. **b** Raman spectra from different positions on the MoS$_2$ film. **c** Transfer characteristics for 24 MoS$_2$ transistors spread on a 2 in. wafer. **d** Microscope image of the fabricated 2T-1C cell. Scale bar: 100 μm. **e** Circuit schematic of a 2T-1C cell containing storage and calculation modules. **f** 3D schematic illustration of a 2T-1C unit cell, including two MoS$_2$ FETs and one capacitor. **g** Circuit diagram of the proposed 2T-1C cell array. **h** A typical diagram of a matrix convolution operation.

column n. Both W_{nm} and V_x determine the drain current I_d in each MoS$_2$ FET. Finally, the output currents in all rows are added to give a total current I_m. The collected current then flows into the current block for further calculation. The relationship between I_m, W_{nm}, and V_x is $I_m = \sum_{k=1}^{n} g(V_{xk}, W_{km})$, where $g(x)$ is a current–voltage transform function that depends on the transfer and output characteristics of transistor T$_2$. Below we will try to build a correlation between Y_m and I_m.

We first characterize the properties of the 1T–1C storage module. Figure 2a shows a schematic diagram of the measurement circuit, in which one end of the capacitor is connected to an external oscilloscope (see Fig. S4 for more details). The internal resistance of the oscilloscope R_{in} is used to estimate the current flow (I_Q) during read/write operations by measuring the voltage of R_{in}. To measure I_Q, voltage signals V_w and V_{re} are applied to T$_1$ (Fig. 2b) with pulse widths of 12 and 10 ms, respectively. V_w rises 1 ms earlier than V_{re} and falls 1 ms later than V_{re} to ensure the charge is entirely written onto the capacitor and prevent leakage current through T$_1$. V_{re} and V_w were both set to 3 V during the write operation. The high V_{re} value turns on T$_1$, allowing V_w to charge the capacitor to the same potential. A positive current pulse ($I_Q{}^+$) during the write operation indicates a charge flows into the capacitor. After the write operation completes, V_{re} is switched to −3 V to turn off T$_1$. Due to the ultra-low leakage current (Fig. S1), the charged voltage on the capacitor can be stably maintained during the write operation. After waiting for 10 s, a read operation is triggered, where $V_{\text{re}} = 3$ V and $V_{w\text{–read}} = 2$ V. The polarity of the measured I_Q pulse is now negative, indicating the capacitor potential is higher than 2 V and charge flows out of the capacitor. In contrast, if the capacitor potential is less than 2 V, the capacitor will be recharged again, giving a positive current pulse. To further characterize the dependence of $V_{w\text{–write}}$ for reading I_Q, the above measurements were repeated. Figure 2c shows the I_Q pulses for reading under various values of writing $V_{w\text{-wirte}}$. To estimate the charge in the capacitor, after waiting for 10 s, $V_{w\text{–read}} = 2$ V is applied to compare with the retained capacitor voltage to read the remaining charge. The amplitude of the I_Q pulse becomes larger as V_w increases. It is also noted that all I_Q pulses are under 2 ms

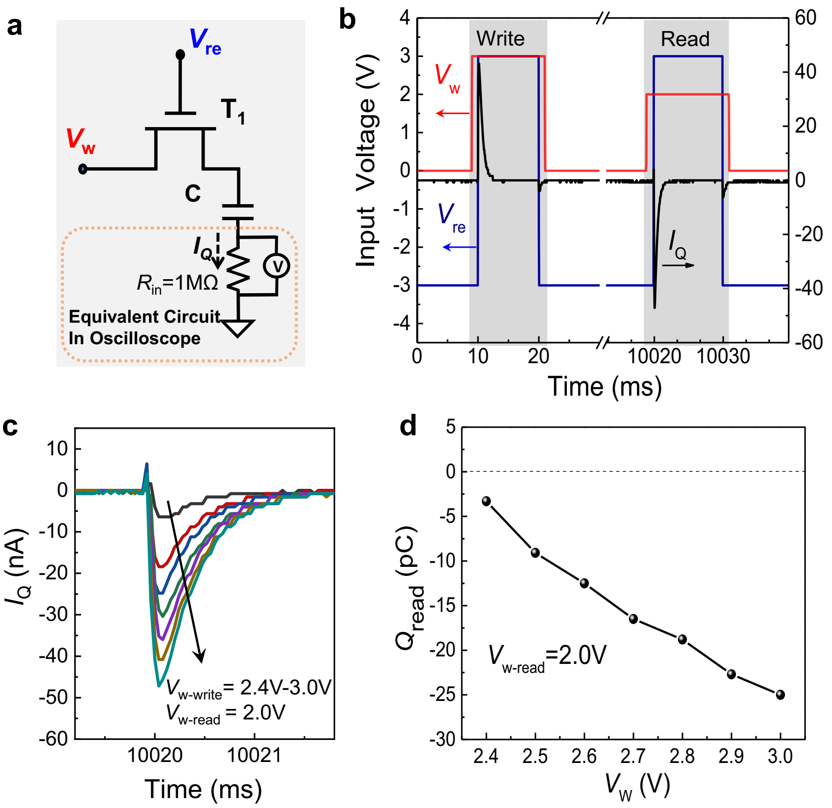

Fig. 2 Characterization of the 1T–1C storage module. a Schematic diagram of the electrical circuit used to measure the 1T–1C storage module (shadow area). The equivalent circuit in the dashed box equals an external oscilloscope connected to the capacitor. **b** Input voltage waveform (V_w, V_{re}) and readout current (I_Q) vs. measurement time. **c** I_Q spikes at $V_{re} = 3$ V while $V_{w-write}$ ranges from 2.4 to 3 V in 0.1 V steps. **d** Calculated retained charge (Q_{read}) in the capacitor as a function of $V_{w-write}$ (compared with $V_{w-read} = 2.0$ V when $I_Q = 0$ A).

(Fig. 2c), which approximately equals the write time. The integral of the current overtime during a read cycle equals the charge Q_{read} remaining after the waiting interval (10 s). In Fig. 2d, the calculated Q_{read} vs. V_w curve is linear, indicating that the charge saved on the capacitor can still be differentiated after 10 s.

To test whether the voltage stored in the capacitor can effectively drive T_2, we examined the time evolution of the drain current I_d in T_2 after completing a storage operation. Figure 3a shows a complete diagram of the measurement circuit used to measure a 2T–1C cell's electrical behavior, and a storage cycle is shown in Fig. 3b. The magnified area in Fig. 3b shows the storage operation in detail. $V_w = 2.4$ V with a pulse width of 140 ms, and $V_{re} = 3$ V with a pulse width of 100 ms, i.e., V_w rises 20 ms earlier and falls 20 ms later than V_{re}. One should note that I_d has a steep pulse during a storage operation. Since it synchronizes with V_{re}, this is mainly due to the parasitic capacitance between the gate electrode and the capacitor. After the storage operation completes and the capacitor is charged to 2.4 V, T_1 is then turned off by applying a negative V_{re} (−3 V), and V_w is set to 0 V. Thus, the voltage potential on the capacitor entirely controls I_d of T_2, without the influence of V_w. During the 10 s holding time, the output current I_d decreases from 302 to 292 nA, approximately a 3% loss. It indicates that most of the charge stored in the capacitor can be maintained over a 10 s period, which keeps its voltage potential nearly constant and provides persistent control of the channel current in T_2. Such charge storage persists even the holding time is extended to 100 s with a loss of I_d less than 10% (Fig. S5). Reproducibility tests show that I_d in T_2 remains nearly constant after more than 100 cycles (Fig. S6). Such desirable storage characteristics indicate that, upon tuning V_w and V_x, different values of I_d in T_2 could be obtained and maintained with an acceptable loss in 10 s, which provides various differentiable states.

To demonstrate this, we first explored the electrical characteristics of T_2. Figure 3c shows the output characteristics with V_g ranging from 2.4 to 3.0 V in 0.1 V increments, where one electrical probe is added separately to apply V_g directly to T_2 as V_w (Fig. S7a). A relatively small V_x is applied to obtain linear I_d–V_d output characteristics. Then V_w is fixed at 2.4 V, and V_x varies from 0.05 to 0.35 V in 0.05 V increments. Figure 3d shows I_d–t curves (similar to that in Fig. 3b) under different applied V_x values. For each I_d–t curve, V_x is fixed to monitor the decrease of I_d during one cycle (~10 s) to tell if the I_d at each level can be distinguished without overlap with neighboring states. The right graph shows the variation in I_d during one cycle. We then investigated the corresponding transfer characteristic, as plotted in Fig. 3e. V_x is fixed from 0.05 to 0.35 V in 0.05 V increments while V_g varies from 2.4 to 3 V, in which range the I_d–V_g curves are all nearly linear. Figure 3f again shows the measured I_d–t curves in which V_x is fixed at 0.1 V, and V_w pulse varies from 2.4 to 3 V in 0.1 V increments. Like the results in Fig. 3d, the I_d at each level can be distinguished in one cycle. In Fig. 3d, f, it is noteworthy that there remain charges on the capacitor at the beginning time due to the previous cycle's operation, so that each I_d–t curve has an initial value equals to that after 10 s retention time.

As illustrated in Fig. 4a, we used two nearly identical 2T–1C cells to demonstrate a simple MAC operation. The sources of the two T_2 cells are connected to sum up I_{d1} and I_{d2}. Figure 4b shows that when a test step-waveform is applied to V_x, and V_w is set as various values, I_d from T_2 can be accurately controlled. V_x ranges from 0.05 to 0.35 V in 0.05 V increments during every test cycle,

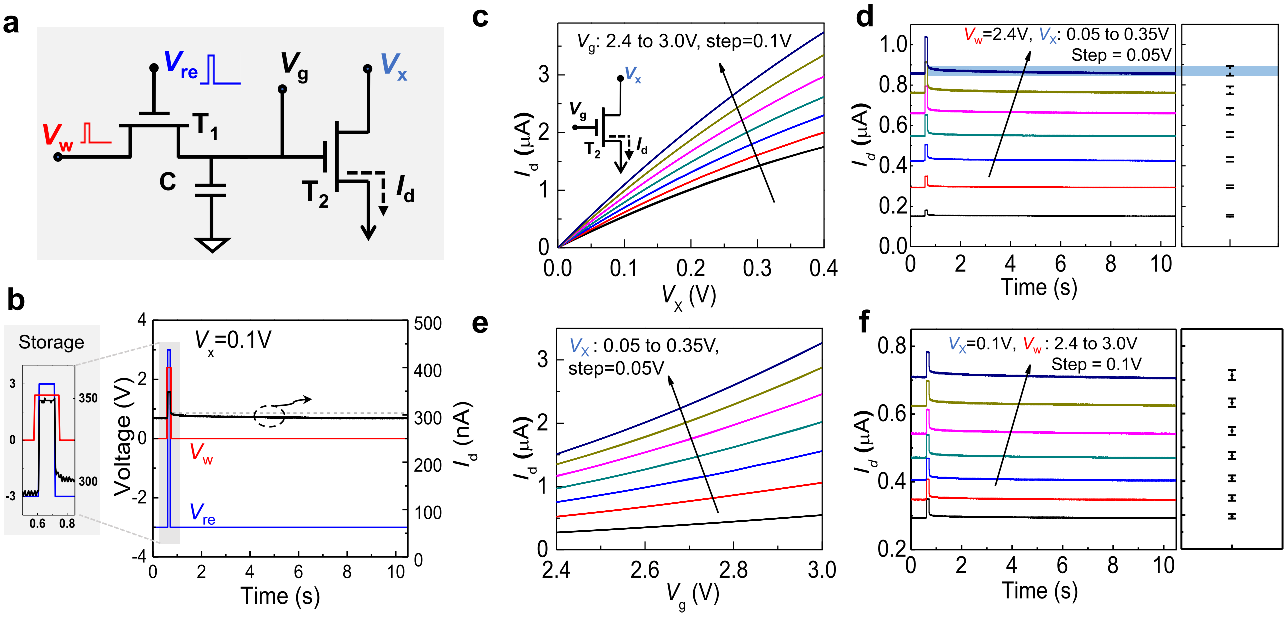

Fig. 3 Characterization of the 2T-1C unit cell. **a** Schematic diagram of the circuit used to gather measurements from a 2T-1C unit cell. **b** A complete storage and calculation operation for a 2T-1C unit. The input voltage (V_w, V_{re}, V_x) and drain current (I_d) are shown vs. measurement time for a 10 s cycle. The magnified inset shows details of the refresh operation. **c** The output characteristics for T_2 with V_g ranging from 2.4 to 3 V in 0.1 V increments. **d** Drain current I_d in T_2 with V_x ranging from 0.05 to 0.35 V, where $V_w = 2.4$ V. The bars in the right panel indicate the variation of I_d for each curve after T_1 is turned off. **e** Transfer characteristics of T_2 with V_x ranging from 0.05 to 0.35 V in 0.05 V increments. **f** Drain current I_d in T_2 with V_w ranging from 2.4 to 3 V, where $V_x = 0.1$ V.

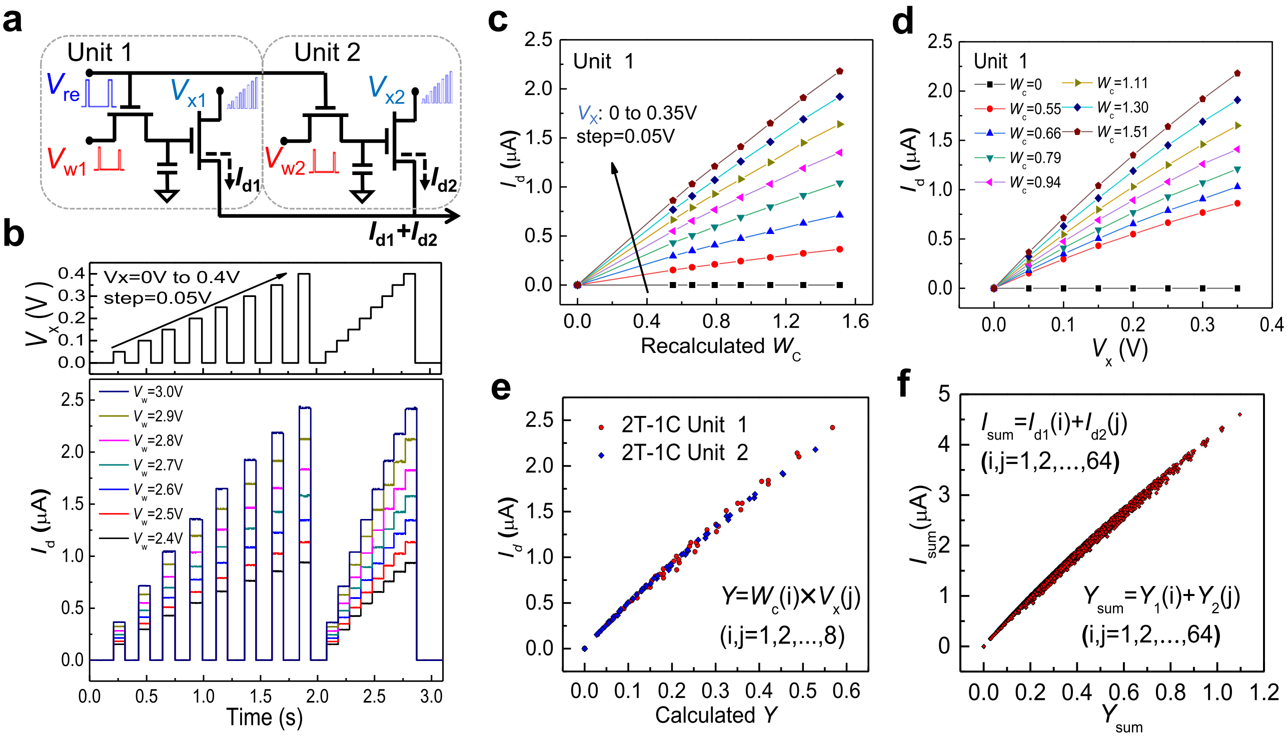

Fig. 4 Demonstration of multiply accumulate operation using two 2T-1C cells. **a** Schematic showing two identical 2T-1C cells. The sources of the two cells are connected to sum the drain current. **b** Top graph: a test multi-step voltage waveform applied to V_x, ranging from 0.005 to 0.35 V with 0.05 V increments. The pulse width is 0.1 s. Bottom graph: The corresponding I_d waveform, while V_w is fixed at a series of values. Both V_x and V_w exhibit eight distinguishable voltage levels (3 bits). **c** The output current I_d as a function of W_c, and the fixed input V_x ranges from 0 to 0.35 V with 0.05 V increments. **d** The output current I_d is plotted as a function of V_x under different W_c values. **e** The measured I_d values as a function of their calculated $Y(W_c \times V_x)$ for two different 2T-1C cells on one MoS_2 wafer. **f** The total output current I_{sum} as a function of Y_{sum} from the two different 2T-1C cells.

and the weighted voltage V_w ranges from 2.4 to 3.0 V in 0.1 V increments. The waveform V_x exhibits eight voltage levels (3 bits) with a pulse width of 0.1 s, while V_w also exhibits eight levels, spanning 7 voltage levels plus a zero level. This measurement imitates when V_w is stored in the 1T–1C unit, a series of operations can be performed to V_x to accomplish multiple calculations in a storage period. The overall speed depends on the response speed of T_2 and the writing speed of T_1. One should note that the output I_d changes almost simultaneously with the input V_x, indicating a fast operation speed. The calculation speed depends on the response speed of the transistor T_2, which is mainly determined by the cut-off frequency $f_T = \frac{g_m}{2\pi C_G}$, where g_m is the transconductance, C_G is the equivalent gate capacitance[41]. Thus the upper limit of f_T approximately equals 127.47 kHz for our current transistor scale (details see Fig. S8), which can act as a reference value for the calculation speed. It is much lower than previously reported MoS_2 RF devices[42,43], mainly because the C_G is significantly influenced by the device size and overlap region of the gate electrode. Thus the speed improvement has a large room through fabrication optimization and further down-scaling.

We have demonstrated storage and calculation capabilities with our 2T–1C cell. We now demonstrate how to implement a MAC operation in detail. Based on the above electrical characterization of a MoS_2 FET, we can obtain linear I_d–V_x curves at small V_x, which approach zero when $V_x = 0$. To realize the multiplication function between I_d and the production of V_w and V_x, a linear correlation between I_d and V_w is also anticipated, i.e., a linear transfer characteristic. However, similar to previous literature results[44–46], I_d has a quadratic dependence on V_w, despite under a relatively low drain voltage regime. To achieve the required linearity, we can propose a recalculated weight

$$W_c = (V_w - 1.9)^2 + 0.3 \tag{1}$$

Now, I_d and the product of W_c and V_x can fulfill the requirements of multiplication operation, i.e., $I_d = \bar{k} W_c V_x$. The conversion between W_c and V_w can be realized by an additional peripheral circuit design (Fig. S9a). Figure 4c shows the output current I_d as a function of W_c, where data was extracted from Fig. 4b, and the fixed input V_x ranges from 0 to 0.35 V in 0.05 V increments. For each V_x value, the output current I_d and the recalculated W_c show satisfying linearity. We then further investigated the relationship between V_x and the output current I_d for different W_c values. As shown in Fig. 4d, I_d is plotted as a function of eight V_x values with different W_c values. For each W_c value, the output current I_d and V_x are also relatively linear. Similar electrical characteristics for the second 2T-1C cell are shown in Fig. S9b, c. In the future, more linear transfer characteristics can be investigated by surface treatment and contact engineering of MoS_2 FETs, or using gapless graphene as an alternative channel material for T_2. So the additional peripheral circuit for linearity conversion can be simplified or removed to realize MAC operation more efficiently.

When we multiply each W_c (3-bit) with each V_x (3-bit), we obtain the mathematical product Y with 64 different values

$$Y = W_c(i) \times V_x(j) \cdot (i, j = 1, 2, \ldots, 8) \tag{2}$$

Figure 4e shows the measured I_d values of the two 2T–1C cells as a function of their corresponding Y values separately. I_d is relatively linear with Y for both cells. We then accumulate Y_1 (cell 1) and Y_2 (cell 2), defined as $Y_{\text{sum}} = Y_1(i) + Y_2(j)$ $(i, j = 1, 2, \ldots, 64)$, while the corresponding sum of the output current is defined as $I_{\text{sum}} = I_{d1}(i) + I_{d2}(j)$ $(i, j = 1, 2, \ldots, 64)$. Figure 4f shows a linear relationship between I_{sum} and Y_{sum}.

Thus, we have shown that MAC operations can be successfully performed based on our MoS_2 2T–1C units. Furthermore, during the retention period, it is enough to implement multiple MAC operations upon inputting a sequence of V_x on T_2. Thus our 2T–1C MoS_2 device can be potentially used for in-situ training that can significantly improve the recognition accuracy of neural networks[47]. Therefore, our results suggest a potential path of 2D semiconductors for future post-Moore applications.

Finally, we built a fully connected neural network (FNN) model with a 3-layer network for handwritten digit recognition. As shown in Fig. 5a, the 400 input neurons correspond to the 20 × 20 pixels in one image while 10 output neurons corresponded to the recognition of digits 0–9, respectively. Here, each pixel has a grayscale value from 0 and 255 (8 bits). We used 4000 images to train the simulation model and another 1000 images for testing.

To process the 8-bit grayscale data, we established an 8-bit MAC composed of 32 2T/1 C cells (Fig. S11). The trained W_{in} (weight of the simulation model) corresponded to quantized voltage with 256 levels (8-bit) stored in the cells. The FNN structure is shown in Fig. S12. Each 8-bit MAC works as a neuron to process the input grayscale data for each pixel. The complete FNN diagram consists of 400 × 20 neurons to form forward propagation from the input layer to a hidden layer. We used back-propagation to train our FNN simulation (see Supplementary Notes for more details). A flowchart for the training and test is shown in Fig. S13. After the FNN completed 100 training epochs against 4000 handwritten images, we performed a recognition test using 1000 handwritten images. The average recognition accuracy of our neural network simulation model reached 90.3%. Figure 5b shows the recognition confusion matrix for the 1000 images test. Figure 5c shows the relationship between recognition rate and training epoch, where the recognition rate rises quickly during the initial 10 training epochs primarily due to a large number of training images.

Considering that the size of an 8-bit grayscale input image occupies too many 2T–1C cells, we attempted to reduce the bit depth of the input grayscale images. We find that when an 8-bit input grayscale image is reduced to 1-bit, there is no evident decrease in recognition rate (Fig. 5d). We also simulated the influence of noise in our neural network by randomly choosing pixels and resetting them to random values. As shown in Fig. 5e, the in-set displays images with 0% and 30% noise levels. In the simulation, each well-trained weight is a 32-bit floating type by ex situ training, and it needs to be quantized to meet the finite weight levels. When the trained weights are re-quantized from 8 bits to 1 bit, as shown in Fig. 5f, we find that a 16-level (4-bit) weight is sufficient for our neural network to reach high recognition accuracy. The in-set in Fig. 5f shows the interval distribution of the 20 × 200 quantized 16-level weights (the quantized 256-level weights are shown in Fig. S14). Figure 5g shows a color map of the trained weights after being quantized to 16-levels. The size of the colormap is 20 × 200. These results suggest that two 2T–1C cells are enough for a neuron to store a 4-bit quantized weight.

In conclusion, we experimentally demonstrated an in-memory computing architecture that integrates MoS_2 FETs in a 2T–1C configuration for MAC operations. Owing to the large current on-off ratio of MoS_2 FETs, the charge stored on the capacitor leaks slowly to present a long retention time so that a multi-level voltage can be retained. Based on the electrical characteristics of MoS_2 FETs and an additional peripheral circuit, the analog multiplication operation can be realized with a re-calculated weight parameter. By connecting two or more 2T–1C unit cells in parallel, the output current is summed to provide the accumulation portion of a MAC operation. In addition, a neural network model was built based on the experimental data to provide image

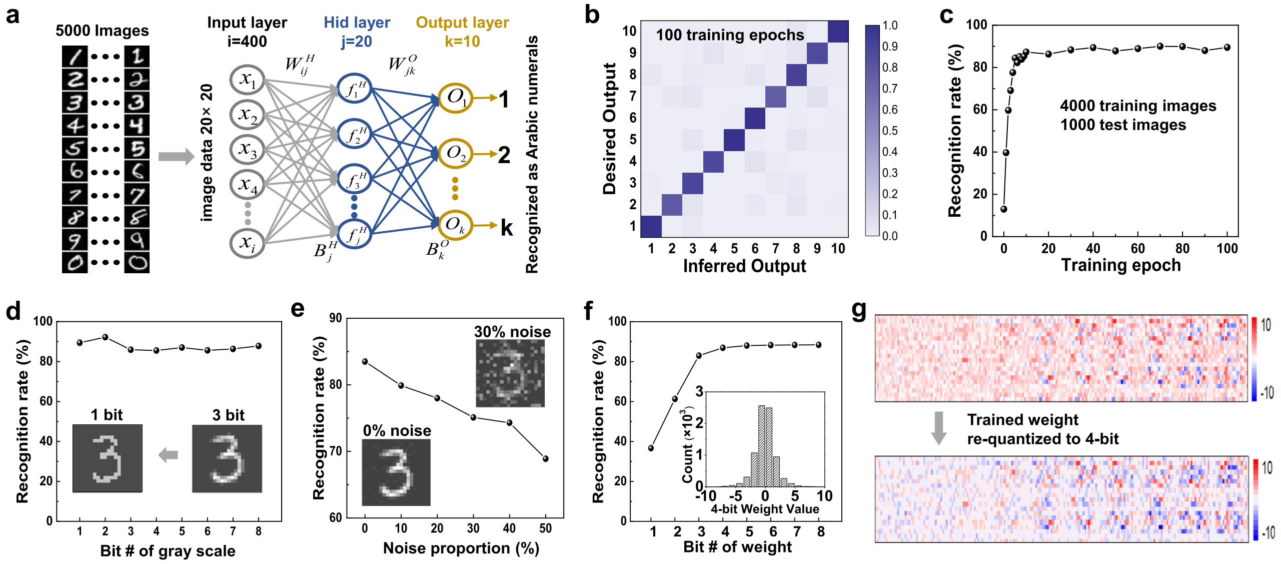

Fig. 5 Fully connected neural network simulation for handwritten digit recognition based on experimental data. a Neuromorphic network with three layers, each containing 400 input neurons, 20 hidden neurons, and 10 output neurons. Where $W_{ij}{}^{H}$ denotes the weight between the input neuron i and the hidden neuron j, f_j^H denotes the convolution value of hidden neurons j, O_k denotes the convolution value of the output layer, and $B_j{}^{H}$ is the bias of hidden neuron j. **b** Confusion matrix of the test results under 100 training epochs, with 4000 images used for training and 1000 images used for testing. The output 0–9 denotes the desired output handwritten digits. **c** Recognition rate as a function of training epoch (0–100), using 4000 images for training and 1000 images for testing. **d** The relationship between the recognition rate and the grayscale bit depth; the inset shows 1-bit and 3-bit grayscale input images. **e** The relationship between recognition rate and noise pixel proportion; the inset shows images with 0% noise and 30% noise. **f** The relationship between the recognition rate and quantized weight levels. The inset shows the interval distribution of 20 × 200 re-quantized 16-level weights. **g** Color map showing trained weights are re-quantized to 4-bits. The size of the colormap is 20 × 200.

recognition with an average 90.3% accuracy. Our MoS2 2T–1C circuit is still a prototype device at the current research stage, and its performance requires further improvement by optimizing material quality and fabrication. Nevertheless, our demonstrated results offer a promising research platform for in-memory computation and in situ training of neural networks.

Methods

Fabrication of MoS_2 2T–1C cell arrays. Device fabrication begins by using photolithography (Microwriter ML3) to pattern the source/drain region and bottom capacitor plate on a monolayer MoS_2 film grown on a sapphire substrate. The channel width/length of T_1 and T_2 are defined as 30/20 and 90/20 μm using ICP etching, respectively. Next, a seed layer (3 nm SiO_2) was evaporated on the MoS_2 film using electron beam evaporation, followed by annealing (200 °C, 10 min) in a high vacuum furnace to remove any resist residue and ensure low contact resistance. A 20-nm-thick HfO_2 layer was then deposited using atomic layer deposition at 180 °C. The oxide stack containing 3 nm SiO_2 and 20 nm HfO_2 serves as a high-*k* gate dielectric of MoS_2 FETs and the capacitor's insulating layer as well. CF_4/Ar plasma etching was used to form an interconnect opening in the dielectric layer to connect the source in T_1 to the gate in T_2. Finally, 30 nm Au was deposited as gate electrodes of the MoS_2 FETs and the top plate of the capacitor.

Characterization and electrical measurements. All measurements were gathered in an ambient environment at room temperature. For capacitor characterization, capacitance–voltage curves were measured with a Keysight E4990A Impedance Analyzer. The MoS_2 FETs were characterized using a semiconductor parameter analyzer (Agilent B1500A). For dynamic memory and 2T–1C cell measurements, the Agilent B1500A was used for supplying voltage signal and detecting the channel current, and a waveform generator (Aligent 33260A) was also used to supply waveforms to the test circuit, while an oscilloscope (DS 1054Z) was used for capturing output signal voltage.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding authors upon reasonable request.

Code availability

The codes used for simulation and data plotting are available from the corresponding authors upon reasonable request.

Received: 22 September 2020; Accepted: 7 May 2021;
Published online: 07 June 2021

References

1. Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R. & Eleftheriou, E. Memory devices and applications for in-memory computing. *Nat. Nanotechnol.* **15**, 529–544 (2020).
2. Wulf, W. A. & McKee, S. A. Hitting the memory wall: implications of the obvious. *SIGARCH Comput. Arch. News* **23**, 20–24 (1995).
3. Mutlu, O., Ghose, S., Gómez-Luna, J. & Ausavarungnirun, R. Processing data where it makes sense: enabling in-memory computation. *Microprocessors Microsyst.* **67**, 28–41 (2019).
4. Wong, H. S. & Salahuddin, S. Memory leads the way to better computing. *Nat. Nanotechnol.* **10**, 191–194 (2015).
5. Ielmini, D. & Wong, H. S. P. In-memory computing with resistive switching devices. *Nat. Electron.* **1**, 333–343 (2018).
6. Berdan, R. et al. Low-power linear computation using nonlinear ferroelectric tunnel junction memristors. *Nat. Electron.* **3**, 259–266 (2020).
7. Yao, P. et al. Fully hardware-implemented memristor convolutional neural network. *Nature* **577**, 641–646 (2020).
8. Prezioso, M. et al. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* **521**, 61–64 (2015).
9. Sheridan, P. M. et al. Sparse coding with memristor networks. *Nat. Nanotechnol.* **12**, 784–789 (2017).
10. Wang, C.-H., Chuang, C.-C. & Tsai, C.-C. A fuzzy DEA–neural approach to measuring design service performance in PCM projects. *Autom. Constr.* **18**, 702–713 (2009).
11. Bichler, O. et al. Visual pattern extraction using energy-efficient "2-PCM synapse" neuromorphic architecture. *IEEE Trans. Electron. Devices* **59**, 2206–2214 (2012).

12. Oh, S., Shi, Y., Liu, X., Song, J. & Kuzum, D. Drift-Enhanced Unsupervised Learning of Handwritten Digits in Spiking Neural Network With PCM Synapses. *IEEE Electron Device Lett.* **39**, 1768–1771 (2018).
13. Wang L., Gao W., Yu L., Wu J.-Z. & Xiong B.-S. Multiple-matrix vector multiplication with crossbar phase-change memory. *Appl. Phys. Express* **12**, 105002 (2019).
14. Pan, Y. et al. A multi-level cell STT-MRAM-based computing in-memory accelerator for binary convolutional neural network. *IEEE Trans. Magn.* **54**, 1–5 (2018).
15. Khvalkovskiy, A. V. et al. Basic principles of STT-MRAM cell operation in memory arrays. *J. Phys. D* **46**, 074001 (2013).
16. Guo, X. et al. Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology. in *2017 IEEE International Electron Devices Meeting (IEDM)*) (2017).
17. Lin, Y.-Y. et al. A novel voltage-accumulation vector-matrix multiplication architecture using resistor-shunted floating gate flash memory device for low-power and high-density neural network applications. in *IEEE International Electron Devices Meeting (IEDM)* 2.4.1–2.4.4 (2018).
18. Wang, P. et al. Three-dimensional nand flash for vector–matrix multiplication. in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 27, 988–991 (2019).
19. Bez, R., Camerlenghi, E., Modelli, A. & Visconti, A. Introduction to flash memory. *Proc. IEEE* **91**, 489–502 (2003).
20. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
21. Yin S., Jiang Z., Seo J.-S., Seok M. XNOR-SRAM: in-memory computing SRAM macro for binary/ternary deep neural networks. *IEEE J. Solid-State Circuits* **55**, 1–11 (2020).
22. Biswas, A. & Chandrakasan, A. P. CONV-SRAM: an energy-efficient sram with in-memory dot-product computation for low-power convolutional neural networks. *IEEE J. Solid-State Circuits* **54**, 217–230 (2019).
23. Zhang, J., Wang, Z. & Verma, N. In-memory computation of a machine-learning classifier in a standard 6T SRAM array. *IEEE J. Solid-State Circuits* **52**, 915–924 (2017).
24. Liu, J., Jaiyen, B., Veras, R. & Mutlu, O. *RAIDR: Retention-Aware Intelligent DRAM Refresh* 40, 1–12 (2012).
25. Li, S. et al. DRISA: a DRAM-based reconfigurable in-situ accelerator. in *2017 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*) (2017).
26. Seshadri, V. et al. Ambit: in-memory accelerator for bulk bitwise operations using commodity DRAM technology. in *2017 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*) (2017).
27. Lin, Y. C., Dumcenco, D. O., Huang, Y. S. & Suenaga, K. Atomic mechanism of the semiconducting-to-metallic phase transition in single-layered MoS2. *Nat. Nanotechnol.* **9**, 391–396 (2014).
28. Li, N. et al. Large-scale flexible and transparent electronics based on monolayer molybdenum disulfide field-effect transistors. *Nat. Electron.* **3**, 711–717 (2020).
29. Kumar A., Ahluwalia P. K. Electronic structure of transition metal dichalcogenides monolayers 1H-MX2 (M = Mo, W; X = S, Se, Te) from ab-initio theory: new direct band gap semiconductors. *Eur. Phys. J. B* **85**, 186 (2012).
30. Liu, C. et al. A semi-floating gate memory based on van der Waals heterostructures for quasi-non-volatile applications. *Nat. Nanotechnol.* **13**, 404–410 (2018).
31. Liu, C. et al. Two-dimensional materials for next-generation computing technologies. *Nat. Nanotechnol.* **15**, 545–557 (2020).
32. Radisavljevic, B., Radenovic, A., Brivio, J., Giacometti, V. & Kis, A. Single-layer MoS2 transistors. *Nat. Nanotechnol.* **6**, 147–150 (2011).
33. Wang, L. et al. Electronic devices and circuits based on wafer-scale polycrystalline monolayer MoS$_2$ by chemical vapor deposition. *Adv. Electron. Mater.* **5** (2019).
34. Zhang, S. et al. Wafer-scale transferred multilayer MoS2 for high performance field effect transistors. *Nanotechnology* **30**, 174002 (2019).
35. Wachter, S., Polyushkin, D. K., Bethge, O. & Mueller, T. A microprocessor based on a two-dimensional semiconductor. *Nat. Commun.* **8**, 14948 (2017).
36. Xu, H. et al. High-performance wafer-scale MoS2 transistors toward practical application. *Small* **14**, e1803465 (2018).
37. Tang, H. et al. Realizing wafer-scale and low-voltage operation MoS2 transistors via electrolyte gating. *Adv. Electron. Mater.* **6**, 1900838 (2019).
38. Mennel, L. et al. Ultrafast machine vision with 2D material neural network image sensors. *Nature* **579**, 62–66 (2020).
39. Xiang, D. et al. Two-dimensional multi-bit optoelectronic memory with broadband spectrum distinction. *Nat. Commun.* **9**, 2966 (2018).
40. Smithe, K. K. H., Suryavanshi, S. V., Munoz Rojo, M., Tedjarati, A. D. & Pop, E. Low variability in synthetic monolayer MoS2 devices. *ACS Nano* **11**, 8456–8463 (2017).
41. Neamen, Donald A. *Semiconductor Physics and Devices: Basic Principles*. (Publishing House of Electronics Industry, 2011).
42. Chang, H. Y. et al. Large-area monolayer MoS2 for flexible low-power RF nanoelectronics in the GHz regime. *Adv. Mater.* **28**, 1818–1823 (2016).
43. Zhang, X. et al. Two-dimensional MoS2-enabled flexible rectenna for Wi-Fi-band wireless energy harvesting. *Nature* **566**, 368–372 (2019).
44. Di Bartolomeo, A. et al. Hysteresis in the transfer characteristics of MoS$_2$ transistors. *2D Materials* **5**, 015014 (2017).
45. Roh, J., Lee, J.-H., Jin, S. H. & Lee, C. Negligible hysteresis of molybdenum disulfide field-effect transistors through thermal annealing. *J. Inf. Disp.* **17**, 103–108 (2016).
46. Liu, L. et al. Electrical characterization of MoS2 field-effect transistors with different dielectric polymer gate. *AIP Adv.* **7**, 065121 (2017).
47. Li, C. et al. Efficient and self-adaptive in-situ learning in multilayer memristor neural networks. *Nat. Commun.* **9**, 2385 (2018).

Acknowledgements

This work was supported by the National Key Research and Development Program (2016YFA0203900), Shanghai Municipal Science and Technology Commission (18JC1410300), Innovation Program of Shanghai Municipal Education Commission (2021-01-07-00-07-E00077), and National Natural Science Foundation of China (61925402, 61851402, 62090032, and 61874031).

Author contributions

All authors discussed the results and commented on the paper.

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/s41467-021-23719-3.

Correspondence and requests for materials should be addressed to P.Z. or W.B.

Peer review information *Nature Communications* thanks Jianhua Yang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

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