Supplementary Information

for

Low-Power Artificial Neural Network Perceptron Based on Monolayer MoS_2

Guilherme Migliato Marega[1,2], Zhenyu Wang[1,2], Maksym Paliy[4], Gino Giusi[5], Sebastiano Strangio[4], Francesco Castiglione[4,6], Christian Callegari[6], Mukesh Tripathi[1,2], Aleksandra Radenovic[3], Giuseppe Iannaccone[4,6*], Andras Kis[1,2*]

[1]*Institute of Electrical and Microengineering, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland*
[2]*Institute of Materials Science and Engineering, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland*
[3]*Institute of Bioengineering, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland*
[4]*Department of Information Engineering, University of Pisa, I-56122 Pisa, Italy*
[5] *Engineering Department, University of Messina, I-98166 Messina, Italy*
[6]*Quantavis s.r.l., Largo Padre Renzo Spadoni snc, I-56123 Pisa, Italy*

**Correspondence should be addressed to Andras Kis, andras.kis@epfl.ch and Giuseppe Iannaccone, giuseppe.iannaccone@unipi.it*

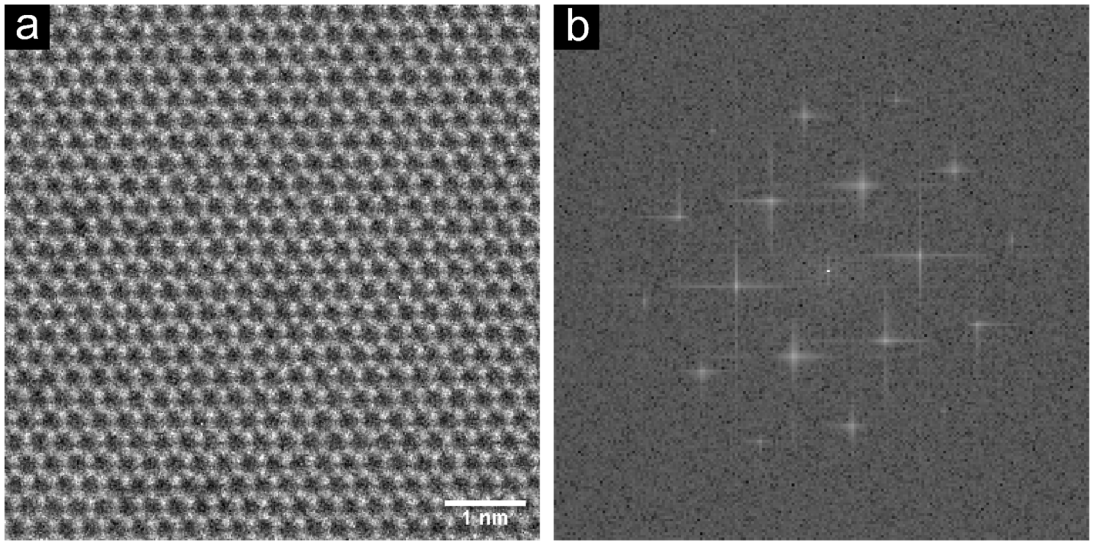

Figure S1. STEM imaging of MoS_2 a, Annular-dark field scanning transmission electron microscopy (ADF-STEM) image shows the hexagonal lattice of single-layer MoS_2. The ADF-STEM image intensity is directly proportional to the atomic number thus the Mo (Z = 42) atoms give brighter contrast than the S atoms (Z=16). **b,** The fast Fourier transform (FFT) pattern from the corresponding image in a demonstrates the highly crystalline nature of the film.

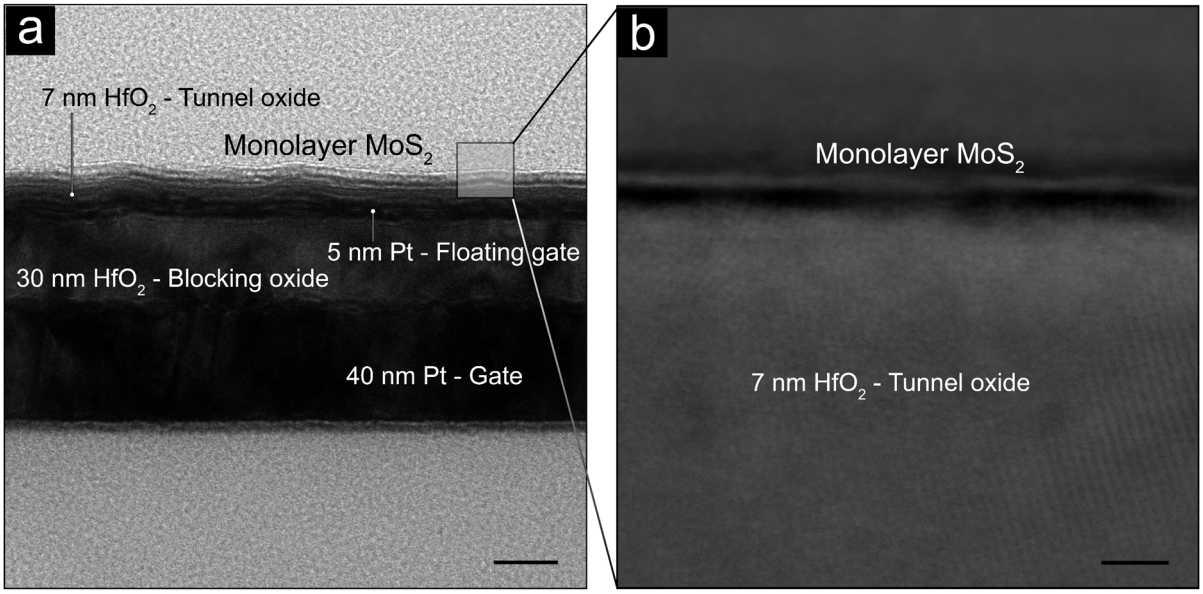

Figure S2. Cross-sectional TEM image of the fabricated device a, Overview of all the constituent layers (scale bar: 100 nm). b, Zoomed-in view of the interface between the gate stack and the monolayer MoS_2 (scale bar: 2 nm).

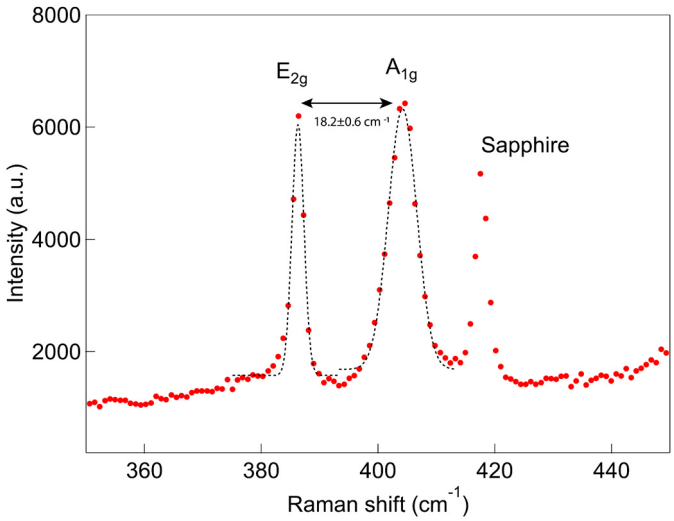

Figure S3. Raman characterization of monolayer MoS_2. Raman spectra of MoS_2 transferred onto a SiO_2 substrate. We used a 523 nm laser excitation and a 3000 line mm^{-1} grating with 10 s acquisition time and averaged from 10 acquisitions. The observed difference between E_{2g} and A_{1g} Raman modes (19.088 cm^{-1}) of MoS_2 is consistent with a monolayer[1].

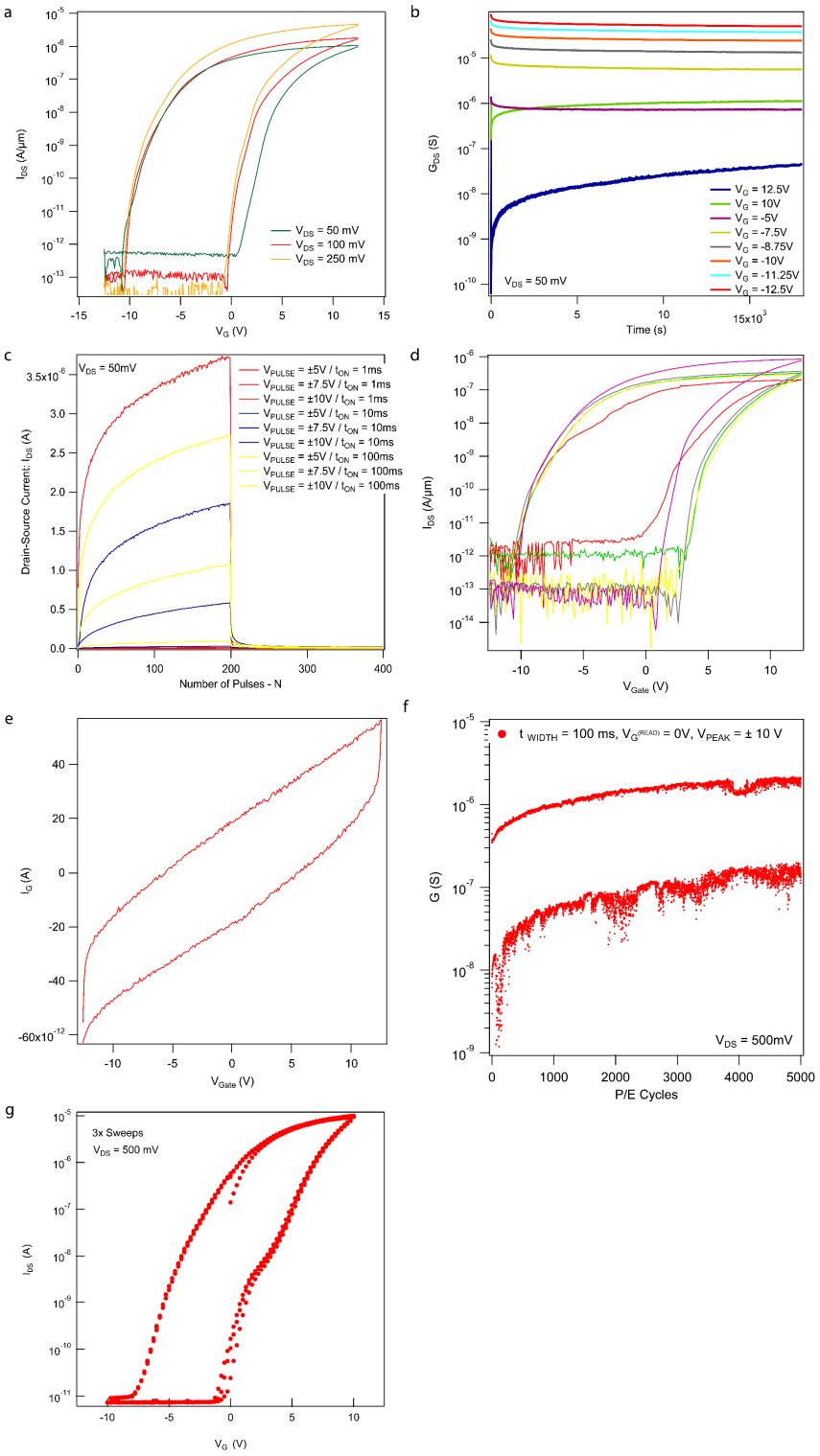

Figure S4. Extended Characterization Curves. a, I_{DS} *versus* V_G for varying V_{DS}**. b,** Retention curves for different programming voltages**. c,** Variation of the device current in function of the number of potentiating and depressive pulses with different peak voltages (V_{PULSE}) and duration (t_{ON})**. d,** Device variability**. e,** I_{DS} *versus* V_G – gate leakage current. **f,** Endurance measurement with pulses with Programming/Erasing V_{PEAK} **=** ± 10 V and t_{WIDTH} = 100 ms. The graph shows the evolution of the channel's conductance G_{DS} *versus* number of Programming and Erasing cycles consisting of two opposite pulses. **g,** Repeatability measurement for the I_{DS} *versus* V_G characteristic curve. Endurance and Repeatability were done in a different device from the previous one but followed the same fabrication steps**.**

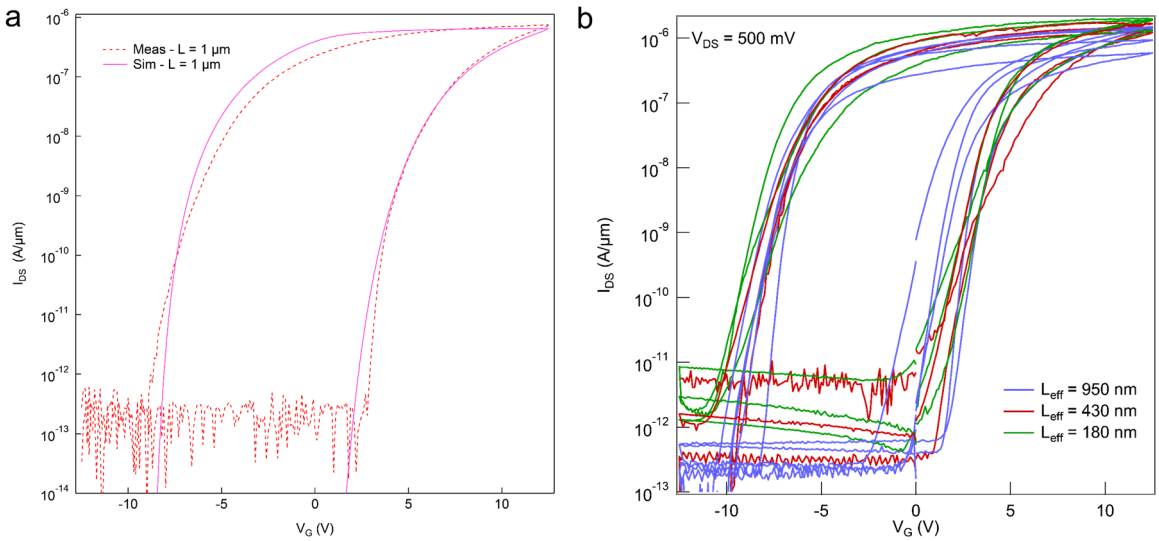

Figure S5. a, Simulated Fitting and Measured values for 1 µm memory. b, I_{DS} *versus* V_G for different gate length (950, 430, 180 nm) and different devices.

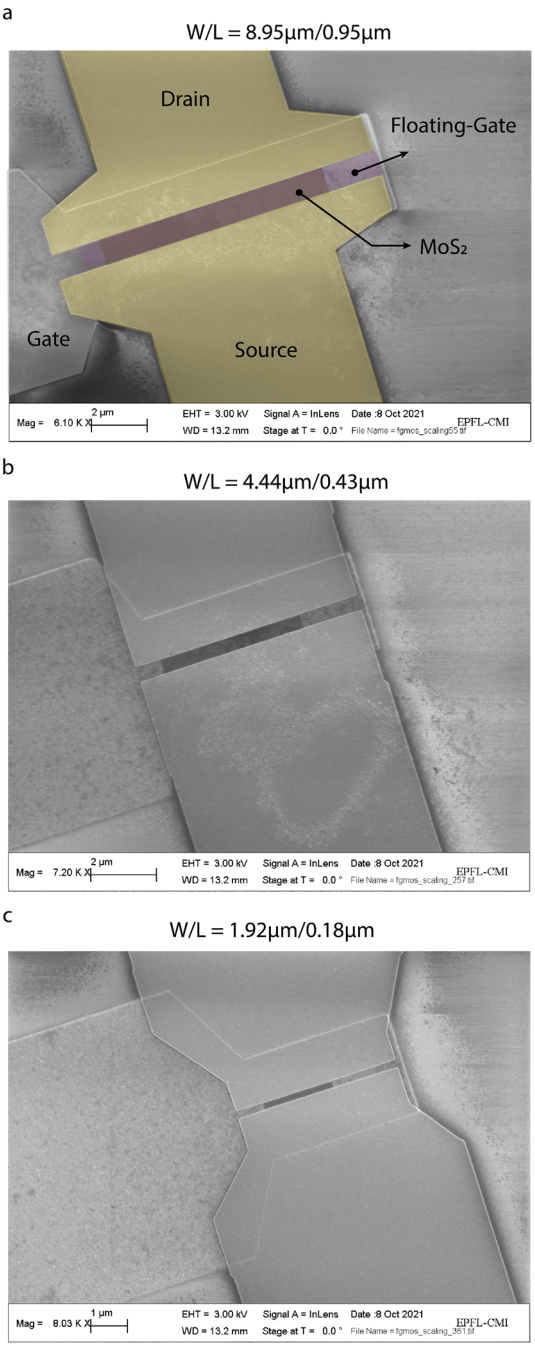

Figure S6. Scanning electron microscopy images of scaled devices. a, Device dimensions L = 950 nm, W = 8.95 µm, fake colouring to indicated contact positions. **b,** L = 430 nm, W = 4.44 µm **c,** L = 180 nm, W = 1.92 µm.

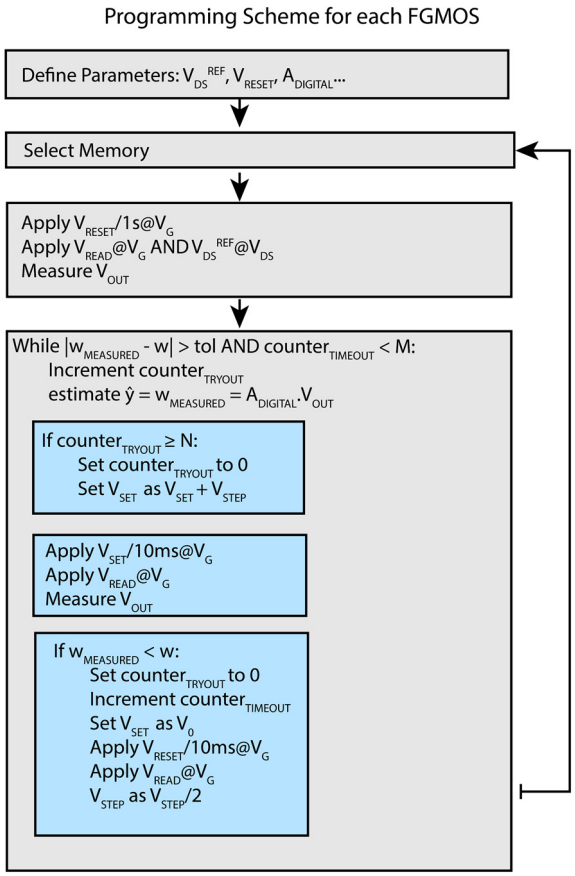

Figure S7. Extended Closed-Loop Programming Block Diagram.

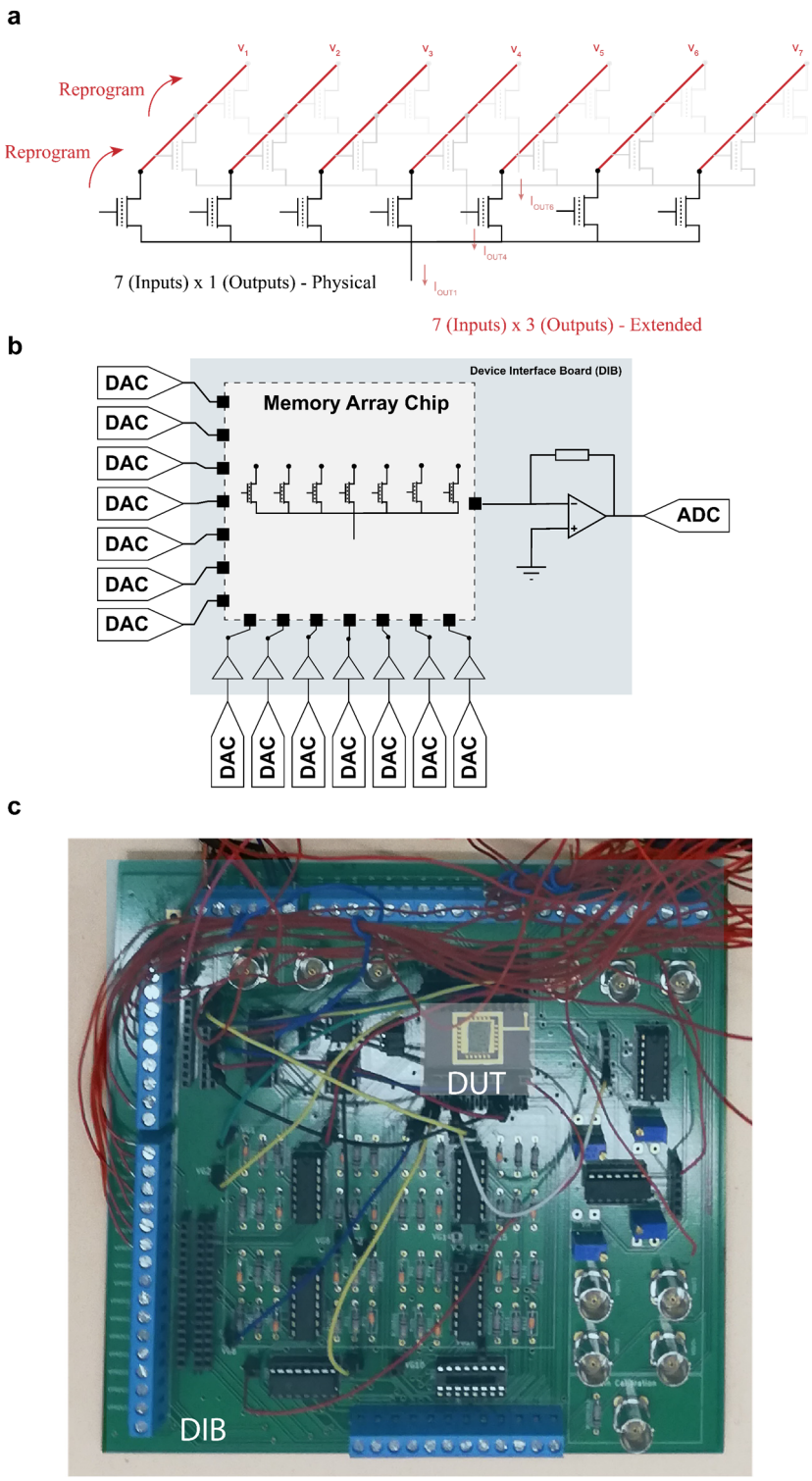

Figure S8. Neural network device interface. a, Example of the used memory array, and expansion of functionality by reprogramming. **b**, Simplified circuit schematics of the chip connected to its peripherals. **c**, Experimental setup with the device-interface-board and the device under test. During measurements, a lid is used on the device to inhibit light sources influencing the measurements.

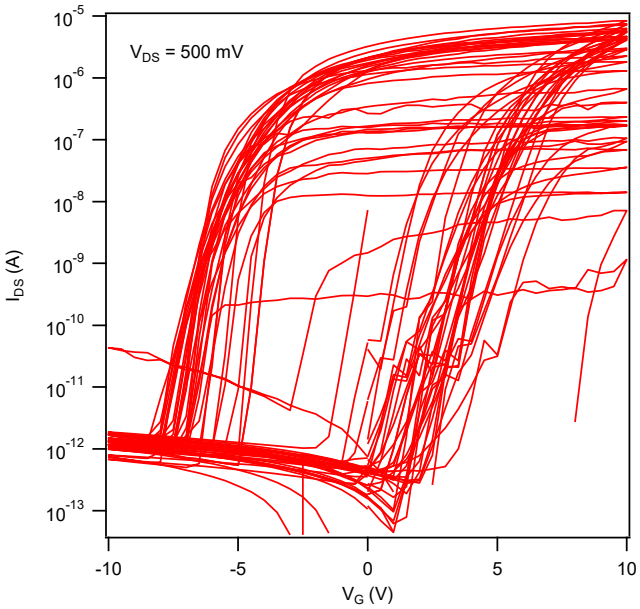

Figure S9. Device variation of 1 μm gate length floating-gate memory spanning over a transferred 2-inch film. Yield is calculated to be 80%.

1. Material Characterization

In order to verify the quality and number of layers of the MOCVD MoS_2, we have performed STEM (Scanning Transmission Electron Microscopy) and Raman spectroscopy in grown materials with the same recipe as the one presented in the main manuscript. In Figure S1, panel a show the crystalline structure of a MoS_2 monolayer and panel b its Fourier transform. These images confirm the atomic quality of the as-grown materials showing only a few sulphur vacancies. Figure S1 shows the TEM image of the cross-section of the gate stack, confirming both thickness and quality of deposited materials. Figure S3 shows the Raman spectrum of the material transferred onto a SiO_2 substrate, confirming its monolayer property.

2. Device Characterization

Figure S4 shows further device characterization of the floating gate memory. Figure S4a presents the I_{DS} *versus* V_G curves for different V_{DS}. Figure S4b shows the retention measurements for programming voltages on the same device. The measurement shows a multistate stability in a 5-hour measurement slot at room temperature. Because of the stability of the states, a much longer retention can be expected. Figure S4c shows tunning of the electrical properties of the memory by symmetrical pulses, allowing an analog modulation of the memory's conductance states. This characteristic is essential for the application of our devices into synapses of the neural network accelerators. Next, in Figure S4d, we show the I_{DS}-V_G for a few devices in the same chip, showing a consistent behaviour. Figure S4e, we show the gate leakage during the sweep shown in Figure S4a. Figure S4f shows the endurance of the memory for cycling. The channel's conductance G_{DS} is probed *versus* the number of programming (V_{PEAK} = +10V) and erasing (V_{PEAK} = -10V) cycles pulses in the gate (T_{PULSE} = 100ms). Finally, the last measurement probes the repeatability of the hysteresis curve (I_{DS} *versus* V_G) after 3 cycles. We note that the endurance and repeatability characterisation was performed with different devices but with the same fabrication process.

3. Device Scaling

Figure S5a shows the device fitting for the simulation, fitting – solid and measurement – dashed. Figure S5b shows the full data set (I_{DS} *versus* V_G) of the scaled devices. Figure S6 shows the SEM images of the scaled devices.

4. Closed Loop Programming

Figure S7 shows the extension of the block diagram described in the main text, used for programming the individual memories to a defined value.

5. Experimental Setup

Due to the limited number of devices, we reuse the same devices 3 times to expand the number of outputs as shown in a Figure S8a. The schematic of the Figure S8b shows the connections and internal circuits to perform the inference/programming of the neural network. We use a custom-made device interface board (DIB) described in Figure S8c to apply gains and route the input and output voltages as shown in the previous schematic. The signals are generated using a CompactDAQ system and a LabVIEW software in the host computer. Although almost all the computation is done directly on hardware, the neuron's activation function is performed numerically after the acquisition of the signals. In the computer, we perform a SoftMax function:

$$f_i(z) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

6. Energy Efficiency Estimation For a Large Neural Network Based on the Fabricated Devices

The energy efficiency *EE* of a neural network is defined as:

$$EE = \frac{N_{EOclass}}{E_{class}},$$

where $N_{EOclass}$ is the number of elementary operations (sums or products) required to complete a single classification, and E_{class} is the energy required by the neural network for a single classification.
If we assume that the main building block of a deep neural network is represented by the VMM, we can approximate the EE of the neural network with the EE of the VMM, and therefore write:

$$EE = \frac{N_{VMM}}{E_{VMM}} = \frac{N_{VMM}}{P_{VMM} t_{lat}}$$

where N_{VMM} and E_{VMM} are the number of elementary operations and the energy required for a vector-matrix multiplication, respectively, P_{VMM} is the power consumption of the VMM, and t_{lat} is the latency time, i.e. the time required to obtain the multiplication result.
With reference to Fig. 6(a) P_{VMM} of a VMM with M rows and N columns is the sum of the power consumed over all the $2 \times M \times N$ memory cells (the factor two takes into account the fact that the weights of a deep neural network can be both positive and negative and therefore we need one column for positive weights and one column for negative weights, leading to a doubling of the number of cells). In this calculation we are not considering the effect of other peripheral blocks (i.e., transimpedance amplifiers converting output currents of one neuron layers into the input voltage of the following neuron layer, or analog-to-digital and digital-to-analog converters), because their presence depends on the particular implementation of the network. This also means that we are estimating an upper value for the EE and that actual implementations will likely have a lower EE, depending on the details of the circuit.
The power $P_{i,j}$ consumed by the (i,j) memory cell is

$$P_{i,j} = V_{in,i}^2 G_{i,j} = V_{in,i}^2 G_{max} w_{i,j}$$

Where $V_{in,i}$ is the input voltage of the i-th row, and $G_{i,j}$is the conductance of the memory cell, that can be written as the maximum conductance G_{max} times the cell weight $w_{i,j}$ normalized from zero to 1. The power consumption therefore depends on the input vector and on the weights. We estimate the average power consumption by maximizing $V_{in,j}$ with the maximum value $V_{in,max}$ (50 mV, in the case considered) and by considering the average weight obtained after training the AlexNet (in our case $\langle w \rangle = 0.2$). Therefore we have:

$$P_{VMM} = \sum_{i,j}^{\text{over } 2\times M\times N \text{ cells}} V_{in,i}^2 G_{max} w_{i,j} \sim 2MNV_{in,max}^2 G_{max} \langle w \rangle$$

The multiplier performs in parallel $M \times N$ elementary products and $(M-1)N$ elementary additions, therefore the number N_{VMM} of elementary operations is:

$$N_{VMM} = MN + (M-1)N = (2M-1)N$$

We can therefore obtain the EE efficiency of the multiplier as:

$$EE = \frac{N_{EOclass}}{E_{class}} \sim \frac{N_{VMM}}{P_{VMM} t_{lat}} \sim \frac{(2M-1)N}{2MNV_{in,max}^2 G_{max} \langle w \rangle t_{lat}} \sim \frac{1}{V_{in,max}^2 G_{max} \langle w \rangle t_{lat}}$$

For $V_{in,max} = 50$ mV, $G_{max} = \frac{1}{R_{min}} = (2.5\ M\Omega)^{-1}$ (e.g. extracted from Fig. 6c considering $V_G = -3\ V$ and $V_{ds} = 500$ mV) , $\langle w \rangle = 0.2$, $t_{lat} = 100$ ns, we obtain $EE = 50$ POps/W.
As we mention in the main text, we are not considering the peripheral circuits, and in particular possible digital-to-analog and analog-to-digital converters, as well as current-to-voltage converters and interface circuits, therefore our estimate is the upper limit of the achievable EE. Let us highlight the fact that for larger VMM the weight of the power consumption of the cells becomes higher with respect to that of peripheral circuits, since the power consumed by the array scales with $N \times M$, whereas the power of the peripheral circuits scale linearly with N and M.

7. Energy Efficiency Calculation For the Fabricated Neural Network – Only Resistive Losses

From the experimental results we obtain the distribution of the output voltages (10000 counts) of the system and we calculate the average output voltage as <V_{OUT}> = 0.62 V.

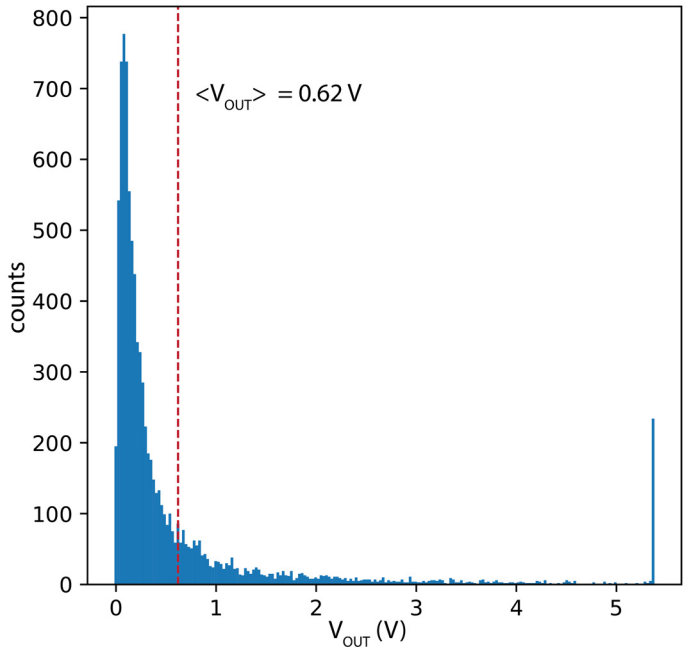

Figure S10 Output voltage distribution of the 7-segment classification perceptron layer.

Total energy consumed by the network

The total energy consumed by the resistor network can be estimated as the sum of the energy dissipated in each individual memory. Current flow to the input $i_{CELL}(t)$:

$$E = \sum_{0}^{NTOTAL} \int_{0}^{T} v_{IN}(t) i_{CELL}(t) dt$$

Since we have $V_{IN}^{MAX} < v_{IN}(t) < 0$ and we assume that $I_{CELL}^{MAX} < i_{CELL}(t) < 0$, we can set the upper limit of the integral as:

$$E = \sum_{0}^{NTOTAL} \int_{0}^{T} v_{IN}(t) i_{CELL}(t) dt \leq \sum_{0}^{NTOTAL} \int_{0}^{T} V_{IN}^{MAX} . i_{CELL}(t) dt$$

Rewriting as the following we can correlate the energy with the mean current in each cell:

$$E = V_{IN}^{MAX} . T . \sum_{0}^{NTOTAL} \frac{1}{T} \int_{0}^{T} i_{CELL}(t) dt$$

$$= V_{IN}^{MAX} \cdot T \cdot \sum_{0}^{NTOTAL} \langle I_{CELL} \rangle = V_{IN}^{MAX} \cdot T \cdot N_{TOTAL} \cdot \langle I_{CELL} \rangle$$

Since we know that the transimpedance amplifier output voltage can be written as

$$\langle V_{OUT} \rangle = -R_{TI} \cdot \langle I_{OUT} \rangle$$

and the output current can be assumed to be constant in each of the cells in one output branch:

$$\langle I_{CELL} \rangle = \langle I_{OUT} \rangle / N_{ARRAY}$$

Therefore,

$$E \leq V_{IN}^{MAX} \cdot T \cdot \sum_{0}^{NTOTAL} \langle I_{CELL} \rangle = V_{IN}^{MAX} \cdot T \cdot N_{TOTAL} \cdot \langle I_{CELL} \rangle = V_{IN}^{MAX} \cdot T \cdot \frac{N_{TOTAL}}{N_{ARRAY}} \cdot \langle I_{OUT} \rangle$$

$$E \leq -V_{IN}^{MAX} \cdot T \cdot \frac{N_{TOTAL}}{N_{ARRAY}} \cdot \frac{\langle V_{OUT} \rangle}{R_{TI}}$$

The energy can be understood as the operation frequency (f_{OP}) times the operation time (T) and the energy per operation (E_{OP}). By analysing only the upper limit:

$$E = E_{OP} \cdot T \cdot f_{OP} = -V_{IN}^{MAX} \cdot T \cdot \frac{N_{TOTAL}}{N_{ARRAY}} \cdot \frac{\langle V_{OUT} \rangle}{R_{TI}}$$

In our system, we have V_{IN}^{MAX}= -1V, f_{OP} = 10000 samples, N_{TOTAL} = 21 memories, N_{ARRAY} = 7 memories (per output), R_{TI} = 2.5MΩ:

$$E_{OP} = \frac{N_{TOTAL}}{N_{ARRAY}} \cdot \frac{V_{IN}^{MAX} \cdot \langle V_{OUT} \rangle}{f_{OP} \cdot R_{TI}} = 74.32 \text{ pJ/op}$$

$$0 \leq E_{OP} \leq 74.32 \text{ pJ/op}$$

References

(1) Li, H.; Zhang, Q.; Yap, C. C. R.; Tay, B. K.; Edwin, T. H. T.; Olivier, A.; Baillargeat, D. From Bulk to Monolayer MoS_2: Evolution of Raman Scattering. *Adv. Funct. Mater.* **2012**, *22* (7), 1385–1390. https://doi.org/10.1002/adfm.201102111.