ACS NANO

www.acsnano.org

ARTICLE

Low-Power Artificial Neural Network Perceptron Based on Monolayer MoS_2

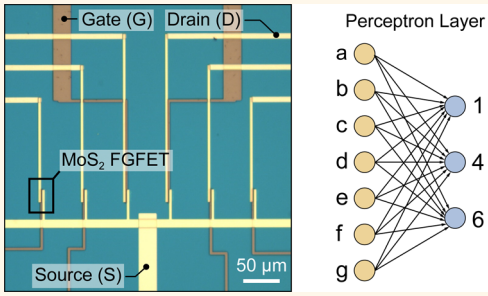
Guilherme Migliato Marega, Zhenyu Wang, Maksym Paliy, Gino Giusi, Sebastiano Strangio, Francesco Castiglione, Christian Callegari, Mukesh Tripathi, Aleksandra Radenovic, Giuseppe Iannaccone,* and Andras Kis*

Cite This: https://doi.org/10.1021/acsnano.1c07065

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: **Machine learning and signal processing on the edge are poised to influence our everyday lives with devices that will learn and infer from data generated by smart sensors and other devices for the Internet of Things. The next leap toward ubiquitous electronics requires increased energy efficiency of processors for specialized data-driven applications. Here, we show how an in-memory processor fabricated using a two-dimensional materials platform can potentially outperform its silicon counterparts in both standard and nontraditional Von Neumann architectures for artificial neural networks. We have fabricated a flash memory array with a two-dimensional channel using wafer-scale MoS_2. Simulations and experiments show that the device can be scaled down to sub-micrometer channel length without any significant impact on its memory performance and that in simulation a reasonable memory window still exists at sub-50 nm channel lengths. Each device conductance in our circuit can be tuned with a 4-bit precision by closed-loop programming. Using our physical circuit, we demonstrate seven-segment digit display classification with a 91.5% accuracy with training performed *ex situ* and transferred from a host. Further simulations project that at a system level, the large memory arrays can perform AlexNet classification with an upper limit of 50 000 TOpS/W, potentially outperforming neural network integrated circuits based on double-poly CMOS technology.**



KEYWORDS: *MoS_2, two-dimensional semiconductors, two-dimensional materials, in-memory computing, nanoelectronics, beyond-Moore*

INTRODUCTION

Modern processors perform many functions needed for the operation of our electronic devices. This flexibility was initially enabled by the separation of processing and memory units in the von Neumann architecture.[1] However, current data-driven applications[2−6] are imposing energy constraints on edge devices due to intensive use of vector matrix-multiplications and access to memory in deep neural networks.[7] The back-and-forth transfer of data between the memory and the processor is now counting for one-third of all energy used in scientific applications.[8] However, the data transfer bottleneck can be avoided by performing computation directly in the memories' physical layer through the combination of Kirchhoff's and Ohm's laws. This type of in-memory processing can benefit calculation-intensive applications such as solving linear system equations,[9] linear and logistic regression,[10] solving partial differential equations,[11] image/signal processing and compression,[12,13] as well as in artificial neural networks (ANN).[14,15]

While many material systems have been explored for in-memory computing,[16] the strong electrostatic sensitivity[17] and intrinsic optoelectronic behavior[18] of two-dimensional (2D) materials present a promising pathway toward reconfigurable and low-power neuromorphic hardware.[19,20] In particular, monolayer transition metal dichalcogenides (TMDCs), such as MoS_2 have been attracting great attention due to their potential to extend Moore's law in advanced technological nodes.[21−24] Moreover, their use in emerging memory devices has also been widely reported. They are being employed from standard flash memories[25−28] to emerging resistive[29] and ferroelectric memories.[30]

Memory devices based on 2D materials have recently been gaining attention in the context of in-memory[20] and neuromorphic computing. However, most of previous reports have focused on a single device and extrapolated their behavior to

Received: August 16, 2021
Accepted: February 7, 2022

Downloaded via ECOLE POLYTECHNIC FED LAUSANNE on February 16, 2022 at 06:17:39 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

© XXXX The Authors. Published by American Chemical Society

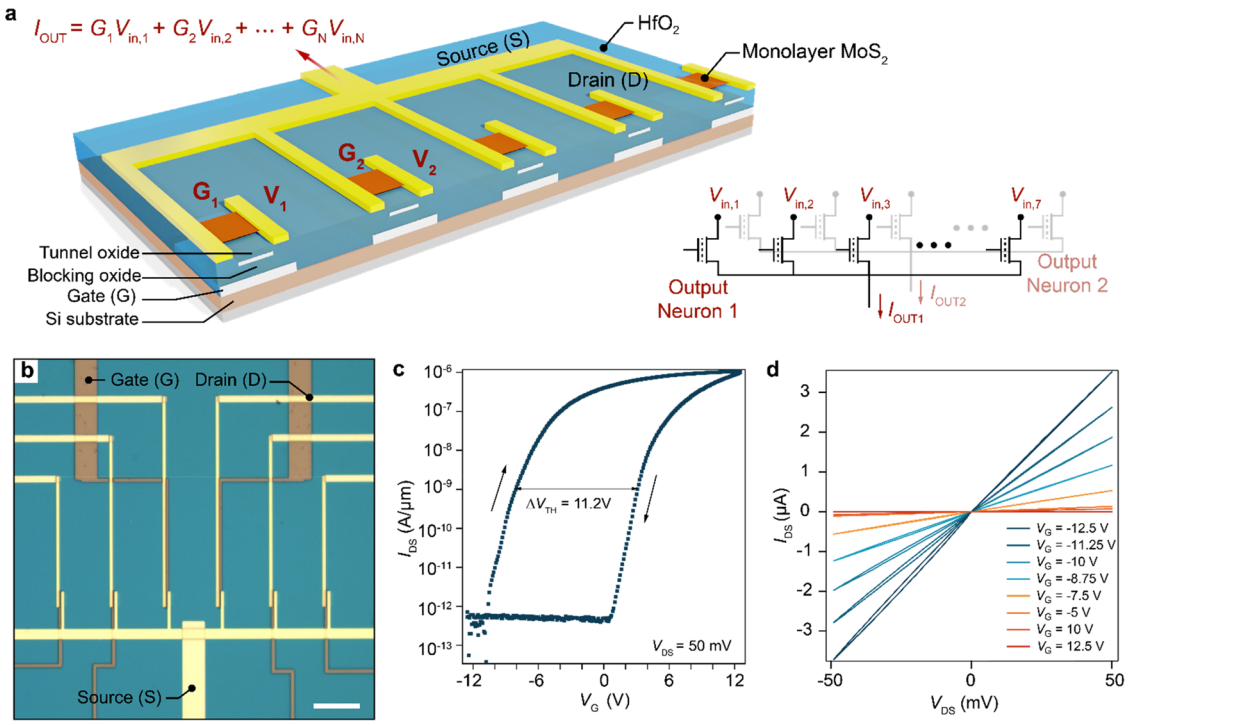


Figure 1. Device structure and characterization. (a) 3D schematic representation of the MoS_2 memory device array and the corresponding circuit schematic for the multiplication-accumulation operation. (b) Optical image of an array of memories connected in parallel (scale bar: 50 μm). (c) I_{DS} as a function of V_G for constant drain-source voltage, V_{DS} = 50 mV. (d) I_{DS} as a function of V_{DS} for different programming voltages, showing the programmable conductance behavior. The device is read using $V_G^{(READ)}$ = 0 V and V_{DS} = 50 mV.

system-level applications using models.[31−33] Exceptions are reports on vision processors based on 2D materials[19,34] in which arrays of photodetectors with programmable conductance were used as artificial neural networks capable of optical pattern recognition. These early examples also used *in situ* training where the training for a neural network was performed directly on the hardware, overcoming any hardware imperfections and device-to-device variability. Although this improves system accuracy for a given chip, training is the most energy-consuming part in the use of artificial neural networks, and it is not desirable to repeat it for every individual chip. In order to conserve energy and time, it would be advantageous to perform training once and transfer it to all the individual processors of the same type. Moreover, a fully electrical processor is preferred for general-purpose applications on the edge since it requires only one excitation source.

Here, we present an in-memory, general purpose processor fabricated on a 2D-material based technology platform. Our processor is based on an array of floating-gate memories with monolayer MoS_2 as an active channel. Simulations predict no significant performance loss as the channel and gate lengths are scaled down to below 100 nm with the scaling trends being experimentally confirmed for devices with gate lengths down to 180 nm, supporting the suitability of 2D materials for scaled in-memory computing circuits. The conductance of the devices can be programmed with a 4-bit precision, allowing them to represent weights for standard dot-product operations needed for in-memory calculations. Finally, we use the memory arrays as artificial networks for seven-segment digit classification with an experimental accuracy of up to 91.5% using transfer of learning from a computer-trained model. Predictions show that large arrays performing the ImageNet classification could potentially outperform silicon counterparts, operating with an upper limit of 50 000 TOpS/W (refs 35 and 36).

RESULTS AND DISCUSSION

Device Description and Characterization. Figure 1a presents the three-dimensional schematic and the cross-sectional view of our floating-gate memory array,[20] based on a gate stack composed of a 40 nm thick platinum (Pt) gate (G), a 30 nm thick hafnium oxide (HfO_2) blocking oxide layer, 5 nm Pt floating gate, and 7 nm HfO_2 tunnel oxide, chosen to give a good compromise between writing speed and retention. Wafer-scale, continuous and large-grain monolayer MoS_2 grown using metal−organic chemical vapor deposition (MOCVD)[37,38] is transferred on top of the gate stack and contacted using titanium−gold (Ti/Au) drain (D) − source (S) electrodes. The devices have a channel length and width of 1 and 12.5 μm, respectively. Individually addressable devices are connected in parallel for performing in memory the multiply−accumulate (MAC) operations using Kirchoff laws for summation and Ohm's law for multiplication (Figure 1a inset). Raman spectroscopy and high-resolution transmission electron microscopy (HRTEM) is used to ascertain the material thickness and quality of the MoS_2 film (Figures S1 and S3). Gate-stack and electrode fabrication were carried out in a class 100 clean room using standard wafer-scale fabrication tools (more details in the Methods). This combination of both wafer-scale material growth and device fabrication allows scaling toward smaller devices and more complex two-dimensional nanocircuits. Figures S1 and S2 show the cross-sectional TEM image of the fabricated memory gate stack. The image shows a conformal deposition of all layers, including the two-dimensional material. No visible defects and cracks were observed in the material nor in the device, also confirmed by

https://doi.org/10.1021/acsnano.1c07065
ACS Nano XXXX, XXX, XXX−XXX

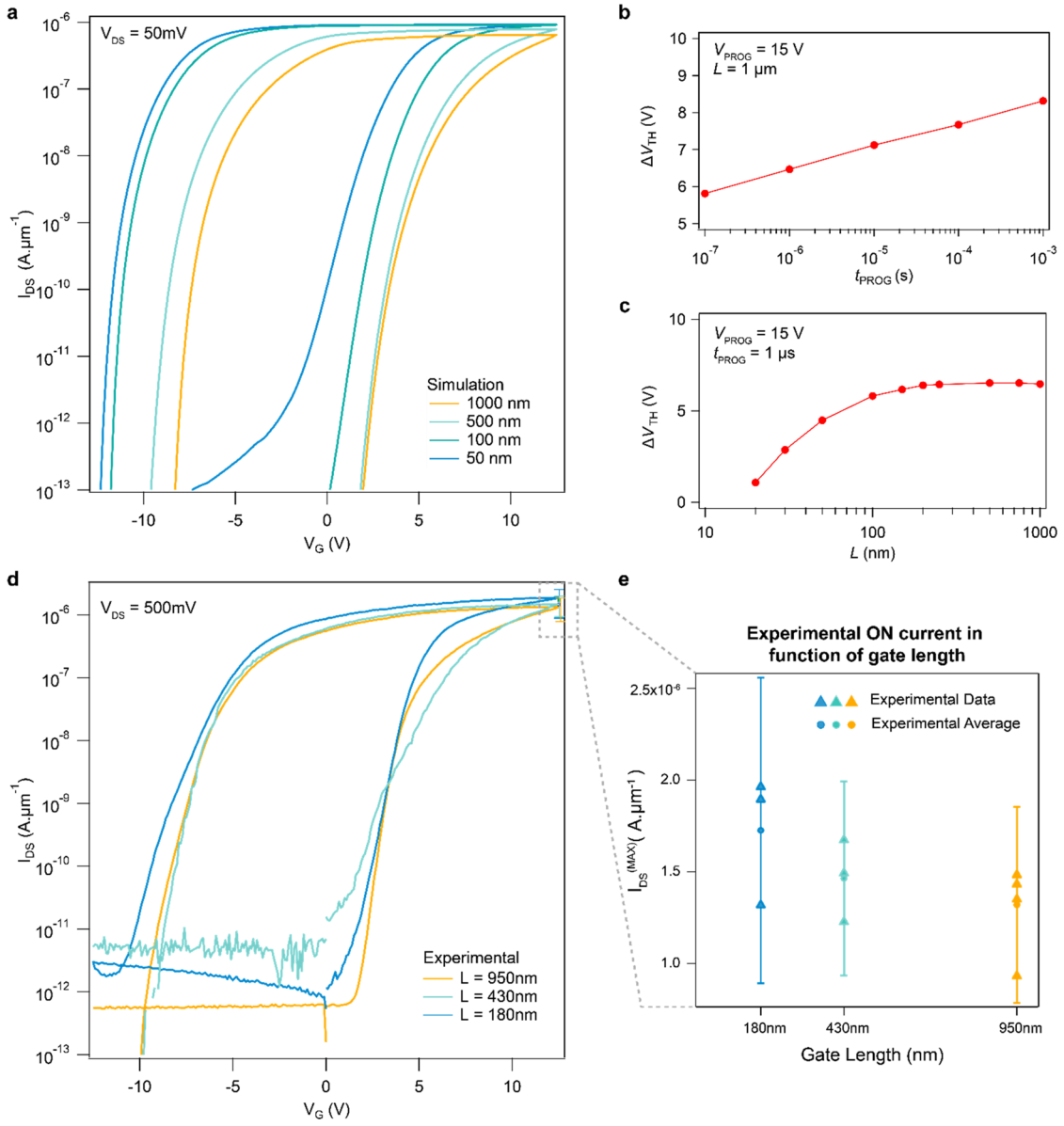


Figure 2. Device scaling. (a) Simulated hysteresis cycle as the device gate length is scaled from L = 1 μm to L = 50 nm. (b) Calculated threshold voltage shift (for $I_{DS} = 10^{-10}$A·μm^{-1}) as a function of programming time t_{PROG}. (c) Calculated threshold voltage shift for different channel lengths with a program time of 1 μs. (d) Experimental hysteresis cycle (I_{DS} versus V_G with V_{DS} = 500 mV) of devices with 950, 430, and 180 nm gate length. The curves shown were select as the median behavior from the experimental data set. (e) Experimental variation of the ON current for different devices with gate lengths demonstrated in (d). Triangle: experimental data. Dot: average value. Error bar: confidence interval with 95% certainty.

electrical measurements. The optical micrograph of a fabricated memory array is shown on Figure 1b.

The operation of the previously described memory device is based on charge transfer between the semiconductor channel and the embedded metallic floating gate. The memory is programmed by applying a control gate voltage such that it bends the bands of the dielectric stack so that direct electron tunnelling can occur through the oxide barrier, from the MoS_2 channel to the platinum floating gate. The charge Q stored in the floating gate causes a shift in the threshold voltage of the MoS_2 transistor $\Delta V_{TH} = -Q/C_{CG-FG}$, where C_{CG-FG} is the capacitance between the control gate and the floating gate.[39] For large gate voltage sweeps, the memory programming operation results in a shift of the threshold voltage between the forward and the reverse paths, creating a hysteresis cycle. The experimental confirmation of the threshold voltage (V_{TH}) shift between the forward and reverse paths are seen in Figure 1d. This creates a 11.2 V memory window that can be tuned depending on the programming voltage that is applied to the device gate. At a constant gate voltage used for reading the memory state ($V_G^{(READ)}$ = 0 V), different values of V_{TH} result in different conductance (G) levels, allowing the memory to be used as a programmable resistor.

Figure 1e shows this programmable conductance feature of the floating-gate memory. Different slopes of linear I_{DS} *versus* V_{DS} can be programmed, using different program and read voltages. Linearity is an important characteristic since the multiplication operation in our in-memory processor is based on the physical relationship between current and voltage. The different conductance states are also stable in a 5 h window

https://doi.org/10.1021/acsnano.1c07065
ACS Nano XXXX, XXX, XXX−XXX

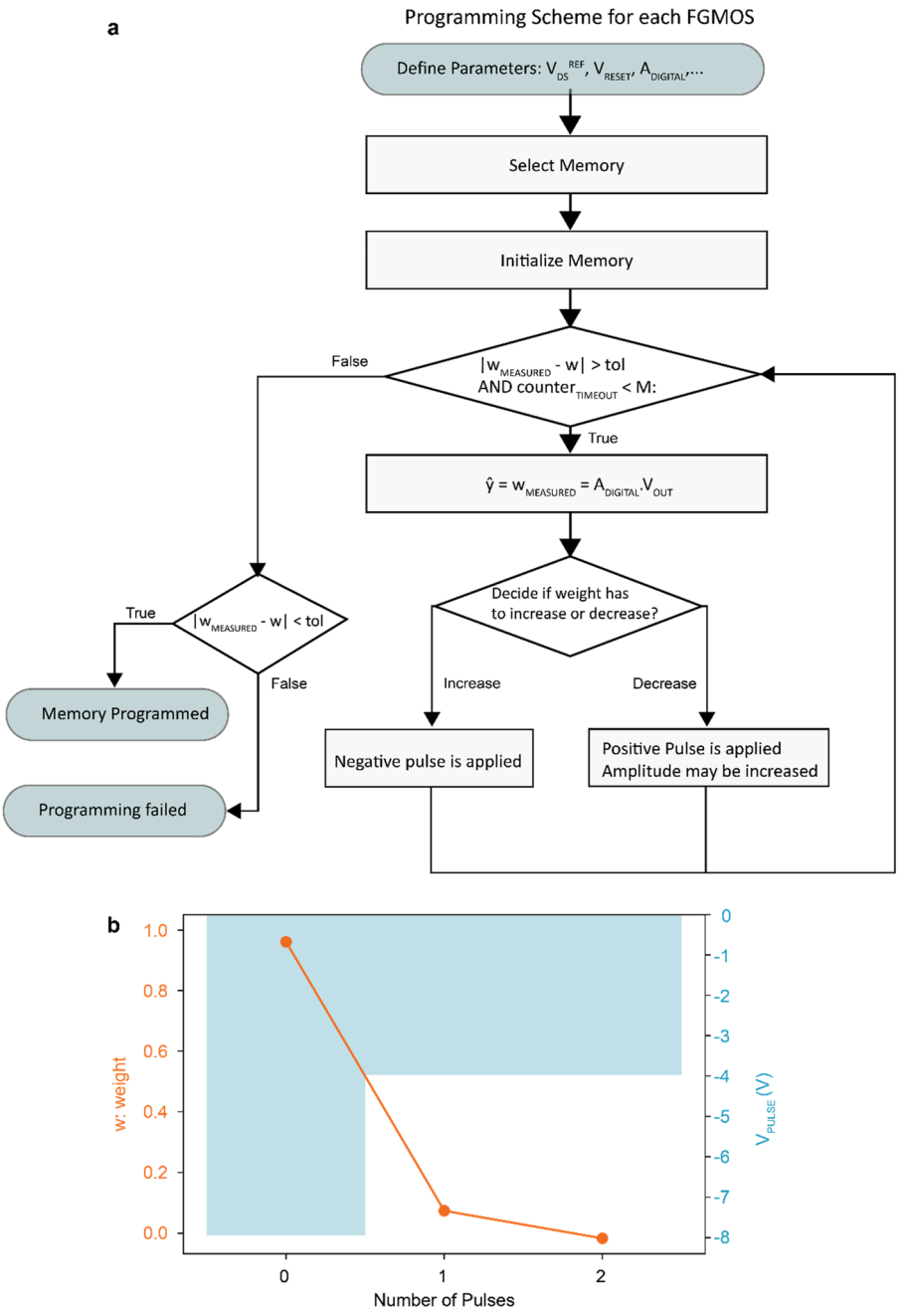


Figure 3. Closed-loop programming. (a) Block diagram explaining the closed-loop programming procedure. (b) Convergence map for overshoot of the weight and progressively decreasing the weight until the correct value has been reached.

without significant degradation. Additional device characteristics are presented in Figure S4.

Device Simulation and Scaling. To advance our understanding of the device behavior and to analyze its performance in advanced technological nodes, we have performed device simulations using a commercial CAD software (Sentaurus by Synopsys, Inc.) by fitting the experimental results for the long-channel floating-gate memories. Figure 2a shows the hysteresis cycle of the transfer characteristics for the simulated long channel device with a channel and gate lengths L = 1 μm. The sweep rate is 3.6 V/min. We obtain a good agreement between the simulated and measured curves for this gate length; see Figure S5. The longitudinal transport is simulated using a drift-diffusion model with Fermi–Dirac statistics, Shockley–Read–Hall recombination, and thermionic Schottky contacts. Interface and intrinsic traps are required to reproduce the gradual subthreshold slope of the transfer characteristics. The charge injection into and from the floating gate is responsible for the observed memory window and is modeled using the Wentzel–Kramers–Brillouin approximation for the electron tunnelling.

After having calibrated the model using the experimental data, we have investigated the scalability of the memory device. Figure 2a shows the simulated hysteresis cycles for gate lengths L down to 50 nm. As the gate length is scaled down, the hysteresis cycle is shifted toward lower gate voltages due to electrostatic degradation, while the peak current increases due to the higher longitudinal electric field in the channel. It is evident from Figure 2a that the large programming window of the long channel is almost maintained down to L = 50 nm. In

https://doi.org/10.1021/acsnano.1c07065
ACS Nano XXXX, XXX, XXX–XXX

order to investigate the programming speed, we have performed transient simulations of $I_{DS}-V_G$ characteristics after the application of a programming pulse with an amplitude V_{PROG} = 15 V and variable width t_{PROG}.

Figure 2b shows the shift of the threshold voltage, extracted at a constant current of 10^{-10} A $\cdot\mu m^{-1}$, for different values of t_{PROG}. The results show that a reasonable programming window can be obtained with a program time of 1 μs but also that the programming window is reduced as the gate length is scaled down. The threshold voltage roll-off is due to the increased semiconductor potential and reduced transverse electric field across the tunnel oxide, which in turn induces a lower tunnel injection into the floating gate. Simulations show that a gate length of about 100 nm still maintains most of the long channel memory window for pulse widths of at least 1 μs. In addition, it is important to highlight that the memory window measured from pulse programming is lower than the one extracted from the hysteresis as discussed in detail in T. Sasaki *et al.*[40]

In other to verify the simulated scaling of our floating-gate memories, we fabricated scaled devices down to 180 nm; see Figure S6 for the microscopy images of our devices.

Figure 2d shows the hysteresis cycle of devices with 950, 430, and 180 nm gate lengths. We show here experimental curves corresponding with the median behavior of the devices. In Figure S5, we show the full data set, indicating the device-to-device variability of the scaled devices. From the I_{DS} *versus* V_G curves, we can observe the threshold roll-off of the scaled devices as a function of the gate length as predicted in the simulations. The electrostatic degradation is more pronounced at a gate length of 180 nm. To analyze the ON current increase, we show the average behavior of a set of devices in Figure 2e. As the gate length decreases, we observe an increase in the ON current due to the increased horizontal fields, as expected.

Closed-Loop Programming. Our individual devices show promising behavior for advanced scaling. However, inevitable process and device-to-device variations will affect the relationship between the device conductance and the programming voltage. In order to reliably perform in-memory the MAC operations, we need to be able to accurately tune the conductance of each device in the network to a predefined conductance value while overcoming device−device variations. The corresponding conductance is then used to map a precise multiplication coefficient used inside filter kernels or as synapse weights in artificial neural networks. In our work, we base our programming technique on previously reported pulsed tuning algorithms using depression and potentiation pulses with a closed-loop convergence procedure.[41] These consist of providing stimuli on the input and probing the device output until it reaches the desired value within a certain tolerance.

First, we map the abstract values (input value: x, output value: y, multiplication factor: w) to physical quantities (input voltage: V, output current: I, memory conductance: G) using a reference voltage, V_{DS}^{REF} and trans-impedance and digital gains, A_{TI} and $A_{DIGITAL}$. The reference voltage is used to convert the input value x to the input voltage as $V = V_{DS}^{REF}\cdot x$. For the reminder of the paper, we use $V_{DS}^{REF} = -1$ V. We have chosen a negative voltage to prevent reprogramming the memory elements during their normal use. For scaling the output current I back to the abstract value y, we transform the current into voltage using a trans-impedance amplifier with a gain A_{TI} = 2.5 MΩ and rescale the obtained voltage with a digital gain $A_{DIGITAL}$ = 10 as $y = A_{DIGITAL}\cdot A_{TI}\cdot I$. With this mapping, the abstract multiplication coefficient w naturally emerges when we set x = 1, $y = w$, allowing the conductance value to be indirectly probed.

We start the algorithm by resetting the conductance value to its highest level by applying a long (1 s) negative pulse (V_{RESET} = −8 V). We successively probe the experimental weight value and compare it to the desired one. If the measured weight is higher than the desired one, the programming pulses are increased to $V_{PULSE} + V_{STEP}$ and applied up to N times. Otherwise, in case that the measured value undershoots the target, a short (10 ms) negative reset pulse (V_{RESET} = −8 V) is applied and V_{STEP} is halved. The next iteration starts until either a maximum of M iterations is reached or the algorithm converges to a desired conductance value, within a tolerance. Our programming tolerance is defined by a discretization of the weight range into 2^{Nbits} values where Nbits is the number of bits of the desired accuracy. Figure 3a shows a simplified block diagram of the previously described algorithm, while the extended block diagram is shown on Figure S4. We present in Figure 3b the evolution of weights and applied voltage pulse values V_{PULSE}. During iterative programming and measurement steps, the gate reading voltage is set to a negative value (≈ −5 V) in order to stabilize the programming values and prevent unintentional reprogramming by operating the device in the subthreshold regime.

Performing the Dot Product Using the In-Memory Circuit. By tuning the conductance of each memory device, we can define the weight vector [w_1, w_2]. Next, we demonstrate the ability of our devices to perform simple multiplication-accumulation operations. In order to do that, we connect two devices in parallel as shown in Figure 4a. We test the calculation for different pairs of x_1 and x_2 with values in the 0−1 range. Parts b−e of Figure 4 show the surface planes representing the results of the dot product operation for different weight matrices. The experimental plots are the raw data showing the linearity of the calculation. The overshoot seen in one of the planes (Figure 4c, for $x_1 = x_2 = 1$) is due to the intrinsic error in the programming of weights and read noise.

Application to a Seven-Segment Display Classification. Next, as a proof of concept, we demonstrate an artificial neural network based on a circuit composed of seven memory devices connected in parallel. We perform digit classification of artificially generated inputs containing noise, corresponding to a seven-segment LCD display, Figure 5. We show additional details related to the physical layout of the memory accelerator in Supporting Section 4. The seven memory devices are reprogrammed to produce up to three different classification outputs in a 7 × 3 perceptron layer. Figure 5a shows the seven-segment display used to define our figure representation. This display configuration was widely used in the past where spurious signal variations cause a noisy representation of numbers that standard classification methods have difficulty of classifying. To perform a robust figure classification, we construct a one-layer perceptron network with a SoftMax activation function in the output layer. The dot-product operation is performed in memory while the nonlinear function is implemented numerically in the acquisition system, for more information see Supporting Section 4. Figure 5b presents the schematics of the one-layer network.

https://doi.org/10.1021/acsnano.1c07065
ACS Nano XXXX, XXX, XXX−XXX

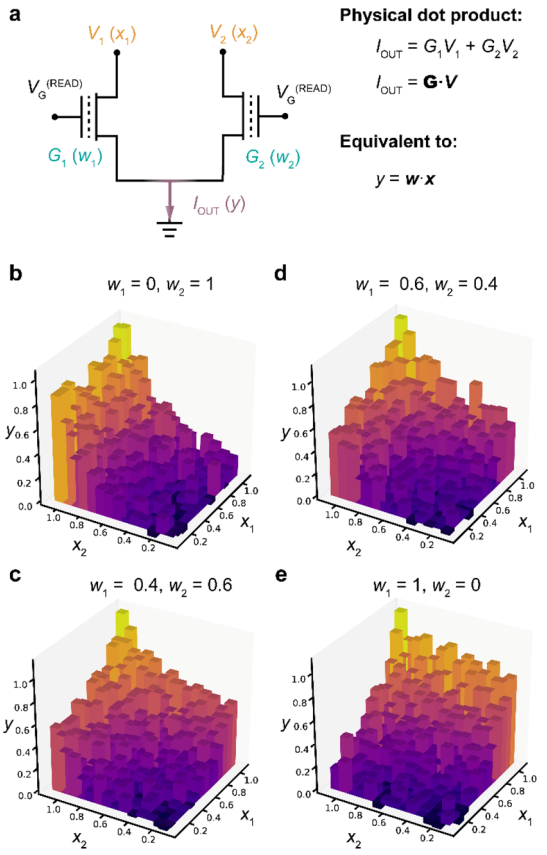


Figure 4. In-memory dot product. (a) Realization of the dot-product operation using two memories connected in parallel. (b–d) Data surface showing the equivalent multiplication-sum planes of a dot-product with the following weights: (b) $w_1 = 1$, $w_2 = 0$; (c) $w_1 = 0.4$, $w_2 = 0.6$; (d) $w_1 = 0$, $w_2 = 1$.

We choose to train the synaptic weight values to each noise-generated data set *ex situ* using the standard TensorFlow and Keras python libraries and transfer the acquired learning to the physical layer. The computer-trained values give an accuracy of 95.5% for an input signal with added white noise having a standard deviation $\sigma = 0.1$, which we use as a baseline for comparison with the measured accuracy of the circuit. This approach performs training only *ex situ*, while the trained weights are then transferred to different neural network processors. This reduces the energy consumption of neuromorphic hardware since training is an extremely power-hungry step in deep neural network algorithms.[42] Figure 5c shows the comparison between the theoretical weight maps, obtained by backpropagation, and the experimental ones after transfer using the previously described programming algorithm with 4-bit precision. A sample of the acquired output signal after the physical multiplication-accumulation operation without the SoftMax function and with the digital gain used for scaling the physical values to the abstract numbers of the neural network is presented in Figure 5d. We achieve a maximum accuracy of 91.5%, compared to the 95.5% accuracy estimated in the software model, classifying up to 10000 numbers/s. This measurement is performed with 4-bit precision programming and an input signal with added white noise having a standard deviation $\sigma = 0.1$. We estimate a resistive power consumption of the memory network of ~74.4 pJ/classification, neglecting the energy expended at the input-output interfaces and on charging the line capacitors (Supporting Sections 5 and 6).

To further analyze the implemented network, we vary both the noise in the input signal and the programming resolution to evaluate their impact on the accuracy of in-memory classification. Figure 5e presents the effect of the input noise on the accuracy of the neural network. We can see that both experimental and computational accuracies follow a linearly decreasing trend as the noise at the input is increased. In addition, the difference between the average experimental values and the theoretically predicted accuracy, as well as the spread of the values, remain similar as the noise standard deviation increases, except for the case of $\sigma = 0.5$ where the smaller spread is due to the saturation of the output analog-to-digital converters. We expect that the spread in measured accuracy is due to variations in each memory weight due to imperfect programming and system noise. Since both experimental and theoretical values are following the same trend, we conclude that the expected behavior has been observed. We show in Figure 5f the effect of the programming resolution (N_{bits}) on the accuracy for a fixed input noise ($\sigma = 0.1$). A more relaxed programming resolution is expected to decrease the precision since the error between the desired and measured conductance value is large. Although this effect is seen from 2-bit to 4-bit data, classification with 1-bit weight programming resolution can be as accurate as for 4 bits. Since the rest of the data follows the predicted behavior, we consider the discrepancy of the 1-bit accuracy data to be due to chance.

Performance of Larger Neural Networks. Encouraged by the promising performance of the demonstrated MoS_2-based artificial neural network accelerator, we evaluate complex neural networks based on the realized FGFET devices. We consider hardware implementations of deep neural networks in which the most frequent large building block is an analogue vector-matrix multiplier (VMM).[43] A network of this type is AlexNet, used for image classification of the large ImageNet benchmark database.[44]

The considered analogue VMM circuit is shown in Figure 6a, where each floating gate memory is used as a programmable resistor. During inference, the control gate voltage is set at V_G, and the input vector is encoded in the voltage values $\{V_{in,1}, \ldots, V_{in,M}\}$. If w_{ij} ($i = 1, \ldots, M$; $j = 1, \ldots, N$) is the conductance of the floating gate memory, then the output current $I_{out,j}$ is given by the matrix multiplication of the voltage vector with the weight matrix as shown in Figure 6a.

To analyze the circuit performance, we have first extracted the SPICE model of the floating gate memory in inversion, Figure 6b, and in the subthreshold region, Figure 6c. We then evaluate the achievable effective number of bits (ENOB) as a function of the gate voltage and of the input voltage full scale. As seen in the previous section, a better linearity is obtained with a lower gate voltage. We find that an ENOB of 5 bits is achieved for a gate voltage of −3 V, which biases the memory in subthreshold, and a maximum input voltage of 50 mV. A system-level simulation of an analogue implementation of AlexNet, performed using TensorFlow and Keras, shows that for a signal-to-noise-and-distortion ratio of 32 dB, corresponding to 5 effective bits, an error rate smaller than 20% can be obtained in ImageNet classification. The difference from the simulated $V_G^{(READ)}$ and the experimentally observed one for an effective 4-bit programming can be understood in terms of the variations of the threshold voltage due to variations in the grown material.

The latency time can be computed with the optimistic assumption that the slow time constants typically associated

https://doi.org/10.1021/acsnano.1c07065
ACS Nano XXXX, XXX, XXX−XXX

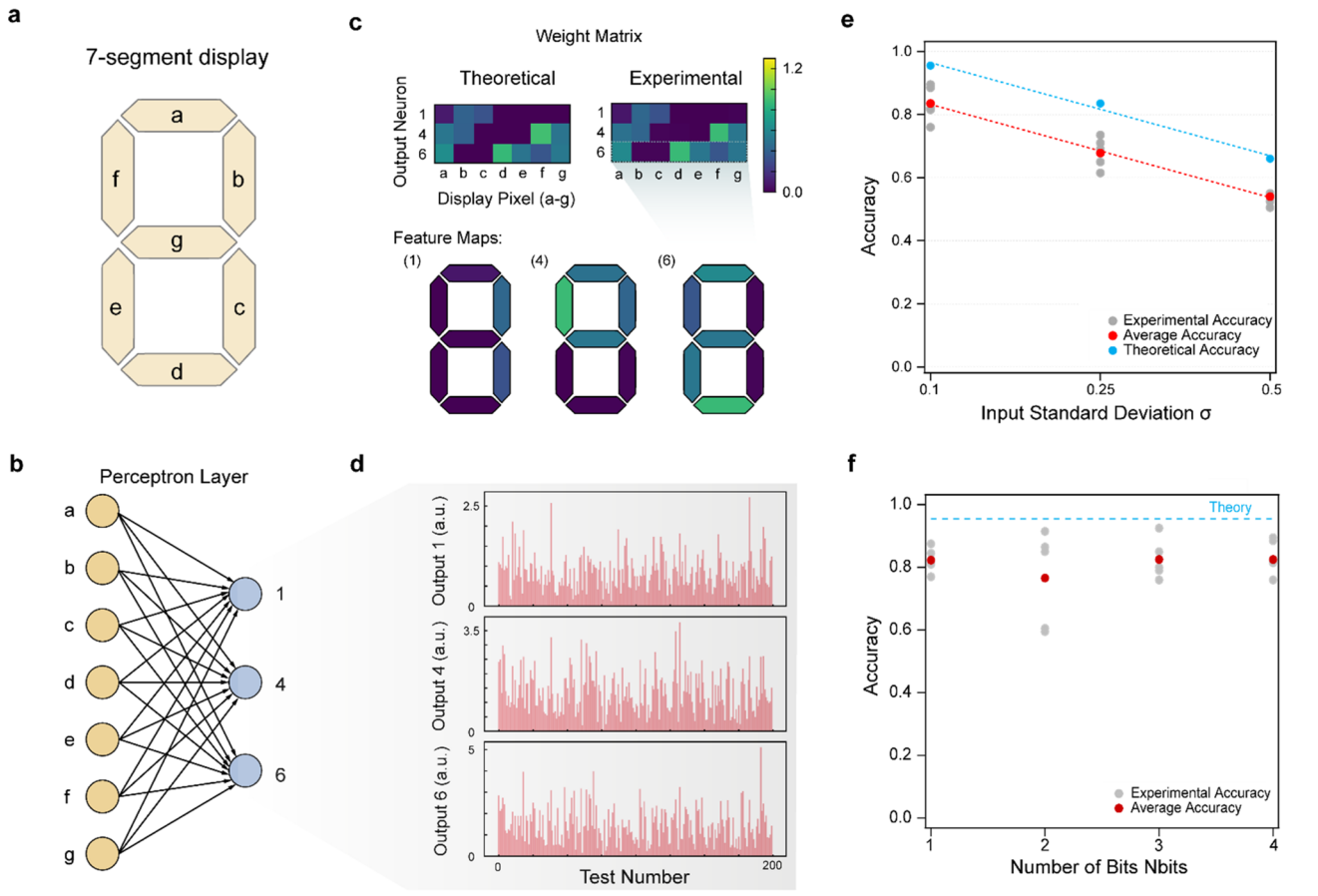


Figure 5. Classification of a seven-segment digit in memory. (a) Representation of a seven-segment display. (b) One-layer perceptron network for seven-segment figure classification. (c) Transfer of learning of the theoretical weight matrix to proportional conductance values of individual memories. (d) Sample of inference operations after different test signals are sent to the input layer and measured in one of the neurons. (e) Effect of the signal noise on the classification accuracy. (f) Effect of the programming resolution on the classification accuracy.

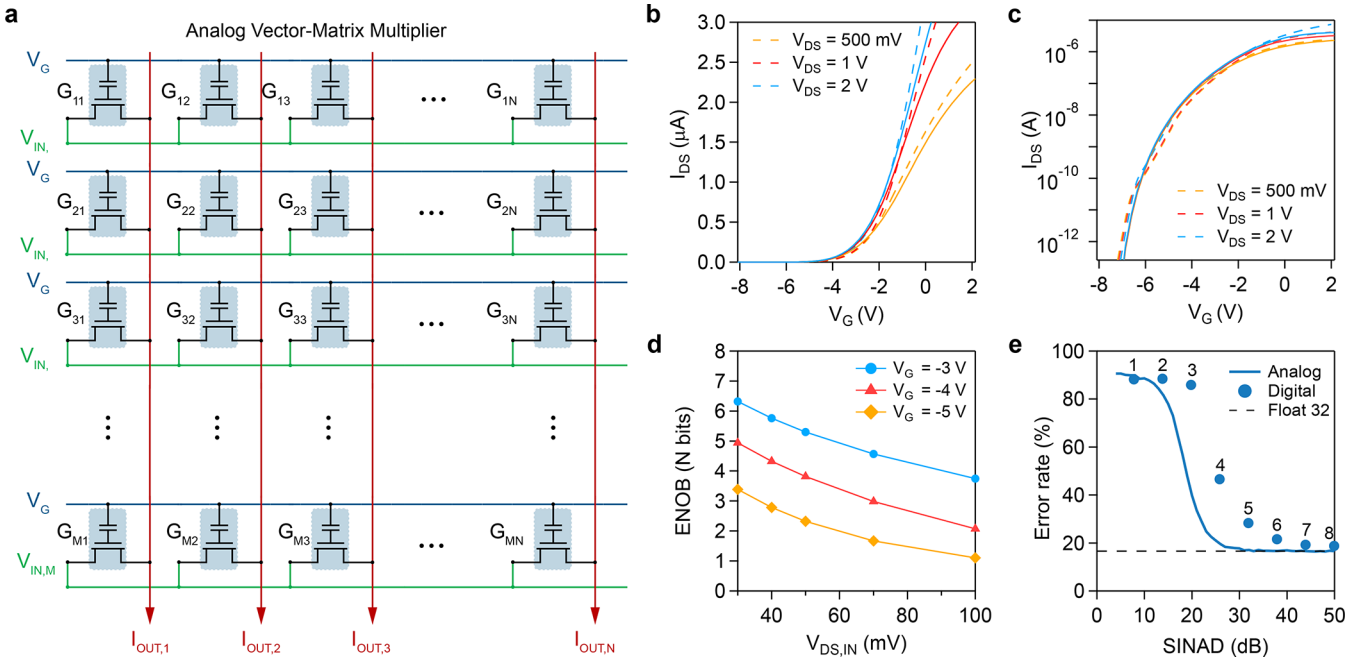


Figure 6. System-level analysis. (a) Analogue vector-matrix multiplier circuit with floating gate memory devices. (b) Transfer characteristics of the memory cells and of the extracted SPICE models in inversion. (c) Transfer characteristics of the memory cells and of the extracted SPICE models in the subthreshold. (d) Achievable ENOB of the multiplier as a function of the cell voltage bias. (e) Error rate in Imagenet classification for an analogue neural network as a function of the signal-to-noise-and-distortion ratio (SINAD) and of the number of bits.

https://doi.org/10.1021/acsnano.1c07065
ACS Nano XXXX, XXX, XXX−XXX

with devices based on 2D materials will be effectively removed as fabrication technology reaches the industrial standards and that therefore, transient behavior can be accurately predicted based on quasi-static device models. With this assumption, transient circuit simulations of the analogue VMM provide a latency time of 100 ns and a record-high energy efficiency of 50 PetaOps/J, where each single operation is either a scalar multiplication or a sum, as is usually assumed. This is a very promising value, considering that the best published estimate is 1.3 PetaOps/J for neural network integrated circuits based on double-poly CMOS technology.[35] We must stress that for estimating the energy consumption we have considered only the analogue VMMs that are the main building blocks, whereas in a full neural network one should also take into account the energy consumption of peripheral circuits, such as the current-to-voltage converters for each column, digital-to-analog and analog-to-digital converters, and interlayer circuitry. While a full implementation of the peripheral circuits is beyond the scope of this work, it would not alter the order of magnitude of the estimated energy efficiency.

CONCLUSION

We have demonstrated floating-gate memory devices based on monolayer MoS_2 with simulations showing no performance degradation down to 100 nm gate length and a useable memory window that persists to sub-50 nm channel lengths. The conductance of each memory can be finely tuned with a 4-bit precision using our closed-loop programming scheme, being limited only by the speed of the experimental setup. Circuits based on the MoS_2 floating-gate devices were used to perform in-memory dot-product calculations and inference. We also realize a simple perceptron layer with weights transferred from a simulated model onto the MoS_2 circuit. Our perceptron layer archives a maximum of 91.5% experimental accuracy, comparing favorably to the modeled 95.5% base accuracy. Finally, we extended our circuit topology to perform ImageNet classification based on the AlexNet architecture. Our network shows an upper limit of computation efficiency, excluding peripheral circuits, of 50 PetaOps/J, almost 2 orders of magnitude higher than for previously reported accelerators. We believe that our findings support the two-dimensional semiconductor material platform for the next generation of in-memory processors where machine learning implementations such as deep neural networks can harness the full potential of this architecture.

METHODS

Material Growth. The continuous monolayer MoS_2 film was synthesized on 2-in. sapphire substrates using the metal−organic chemical vapor deposition (MOCVD) method.[37,38] Before growth, the c-plane sapphire wafers were annealed at 1000 °C in air for 6 h and treated with 3 wt % potassium hydroxide (KOH). A 0.2 mol/L sodium chloride (NaCl) solution was spin-coated onto the wafers to suppress nucleation and enlarge the grain size. During the growth process, molybdenum hexacarbonyl ($Mo(CO)_6$) and diethyl sulfide ($(C_2H_5)_2S$) were used as precursors and carried into the quartz tube by argon with carrier gas flow rates of 10 and 3 sccm, respectively. Both precursors were kept in a water bath at 25 °C to maintain a stable vapor pressure. Hydrogen and oxygen were delivered separately into the growth chamber, with a ratio of 4:1, to balance the growth and etching as well as to achieve the growth of a continuous, wafer-scale monolayer. The reactions proceeded at a temperature of 870 °C and at atmospheric pressure for 30 min.

TEM Imaging. The sapphire substrate with as-grown material was spin-coated with PMMA and baked at 85 °C for 10 min. The MoS_2/PMMA film was detached from the sapphire substrate by submerging it in water. Water surface tension promotes the separation of the grown material from the substrate. Next, the film floating in water is collected using a TEM grid and heated for 15 min at 85 °C. After the transfer is completed, the TEM grid is left in acetone overnight and annealed at 250 °C.

For aberration-corrected scanning transmission electron microscopy (STEM) imaging, an FEI Titan Themis microscope equipped with double Cs corrector, monochromator, and Schottky X-field emission gun was operated at an acceleration voltage of 80 kV. The electron probe current was in the 17−20 pA range. The semi-convergence angle of the probe was 21.2 mrad. High-angle annular dark field detector (HAADF) was used to capture the images using short dwell times (8 μs) with 512 × 512 pixels. The camera length was set to 185 mm which corresponds to the 49−200 mrad collection angle range. Focused ion beam (FIB, Zeiss Nvision40) was used to prepare the cross-section lamella from the device. For the low-resolution cross-sectional TEM imaging, a FEI Talos F200 S G2 microscope was used at 80 kV acceleration voltage.

Transfer Procedure. The MOCVD-grown material is first spin-coated with PMMA A2 at 1500 rpm for 60 s and baked at 180 °C for 5 min. Next, we attach a Gel-pak elastomer film 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 hour. Finally, the sample is immersed in acetone for 2 days and subsequently annealed at 200 °C in a high vacuum to remove the polymer resist and increase adhesion to the surface. For the scaled devices, we used a 130 °C thermal release tape of instead of Gel-pak and removed it by heating on the hot plate.

Processor/Floating-Gate Memory Fabrication. We used a silicon substrate with a 270 nm thick SiO_2 insulating layer. The gate electrodes were fabricated by photolithography using an MLA150 advanced maskless aligner with a bilayer LOR 5A/AZ 1512 resist. The 2 nm/40 nm Cr/Pt gate metals were evaporated using an e-beam evaporator under high vacuum. After resist removal, DI water and O_2 plasma are used to further clean and activate the surface for HfO_2 deposition. The blocking oxide is further deposited by thermal atomic layer deposition using TEMAH and water as precursors. The floating gates were patterned using e-beam lithography in a standard double-layer PMMA/MMA process. The floating-gate metal was deposited in the same evaporator as the gate electrode. With the same atomic layer deposition system, we deposit the 7 nm tunnel oxide layer. For decreasing the e-beam exposure time, the drain-source electrodes are deposited in two steps. First pads and big contacts are exposed using the photolithography procedure described for the gate exposure and 2 nm/60 nm Ti/Au are evaporated in the same machine. After transfer of MoS_2 onto the substrate, patterning it with either e-beam/photolithography and etching by O_2 plasma. Next, the drain-source contacts are patterned using e-beam lithography and 2 nm/100 nm of Ti/Au are further evaporated. To increase adhesion of contact and the MoS_2 onto the substrate, a 200 °C annealing step is performed in high vacuum. The devices have a W/L ratio of 12.5 μm/1 μm.

Fabrication of Scaled Devices Fabrication. We used a silicon substrate with a 270 nm thick SiO_2 insulating layer. The gate electrodes were fabricated using e-beam lithography with standard bilayer polymer PMMA/MMA. The 2 nm/40 nm Cr/Pt gate metals were evaporated using an e-beam evaporator under high vacuum. After resist removal, DI water and O_2 plasma are used to further clean and activate the surface for HfO_2 deposition. The 30 nm blocking oxide is further deposited by thermal atomic layer deposition using TEMAH and water as precursors. The floating gates were patterned using e-beam lithography in a standard double-layer PMMA/MMA process. The floating-gate metal was deposited in the same evaporator as the gate electrode. With the same atomic layer deposition system, we deposit the 7 nm tunnel oxide layer. Next, we transfer MoS_2 onto the substrate, patterning it with negative tone resist (nLOF) using the same MLA150 advanced maskless aligner and etching by O_2 plasma. To achieve sub-1 μm resolution for the drain-source contacts, we

https://doi.org/10.1021/acsnano.1c07065
ACS Nano XXXX, XXX, XXX−XXX

expose them by e-beam lithography with standard bilayer polymer PMMA/MMA mentioned previously. Following the exposure, 2 nm/20 nm Ti/Au are evaporated in the same machine. To increase adhesion of contact and the MoS_2 onto the substrate, a 200 °C annealing step is performed in high vacuum.

Device Characterization. The devices were characterized in a high-vacuum chamber after *in situ* annealing for removing any adsorbents in the surface of the 2D materials which could degraded mobility and induce non controllable memory effects from contaminations. After this, we characterized the devices using an Agilent E5270 Precision Measurement Mainframe.

Circuit Characterization. The electrical characterization of circuits was performed in air with the chip closed with a lid to avoid any light disturbance during the measurements. The device under test (DUT) was connected using a custom device interface board (DIB) described in the Supporting Information. The board serves as a routing medium from both the input and output voltages and has embedded amplifiers to boost voltage and provide current-to-voltage conversion. The analogue voltages were generated and read using a CompactDAQ system with NI-9205 and NI-9264 modules. The CompactDAQ was connected to a host computer running a LabVIEW software to perform the programming and inference of the neural networks using the described closed loop programming algorithm.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsnano.1c07065.

TEM characterization of the material and device cross-section, Raman characterization of MoS_2, extended device characterization, device scaling data, closed loop programming of the devices, description of the experimental setup, and energy efficiency estimation (PDF)

AUTHOR INFORMATION

Corresponding Authors

Andras Kis − *Institute of Electrical and Microengineering, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland; Institute of Materials Science and Engineering, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland;* orcid.org/0000-0002-3426-7702; Email: andras.kis@epfl.ch

Giuseppe Iannaccone − *Department of Information Engineering, University of Pisa, I-56122 Pisa, Italy; Quantavis s.r.l., Largo Padre Renzo Spadoni snc, I-56123 Pisa, Italy*; Email: giuseppe.iannaccone@unipi.it

Authors

Guilherme Migliato Marega − *Institute of Electrical and Microengineering, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland; Institute of Materials Science and Engineering, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland*

Zhenyu Wang − *Institute of Electrical and Microengineering, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland; Institute of Materials Science and Engineering, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland*

Maksym Paliy − *Department of Information Engineering, University of Pisa, I-56122 Pisa, Italy*

Gino Giusi − *Engineering Department, University of Messina, I-98166 Messina, Italy*

Sebastiano Strangio − *Department of Information Engineering, University of Pisa, I-56122 Pisa, Italy*

Francesco Castiglione − *Quantavis s.r.l., Largo Padre Renzo Spadoni snc, I-56123 Pisa, Italy*

Christian Callegari − *Quantavis s.r.l., Largo Padre Renzo Spadoni snc, I-56123 Pisa, Italy*

Mukesh Tripathi − *Institute of Electrical and Microengineering, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland; Institute of Materials Science and Engineering, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland*

Aleksandra Radenovic − *Institute of Bioengineering, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland;* orcid.org/0000-0001-8194-2785

Complete contact information is available at: https://pubs.acs.org/10.1021/acsnano.1c07065

Author Contributions

A.K. initiated and supervised the project. G.M.M. fabricated the devices, designed/prepared the measurement setup, and performed device/neural network characterization. G.G. performed TCAD device modeling based on experimental data and simulated their performance at advanced nodes. M.P., S.S., and F.C. performed system simulations supervised by G.I. Z.W. grew the two-dimensional material and assisted in material characterization under supervision by A.R. M.T. performed cross-sectional TEM and HRTEM for characterization of devices and materials. A.K. and G.M.M. analyzed the data. The manuscript was written by A.K., G.M.M., G.G., and G.I. with the input of all authors.

Notes

The authors declare no competing financial interest.

The data that support the findings of this study are available at Zenodo at https://zenodo.org/record/6033560#.YgqbUt_MKUk and python codes can be found at GitHub at https://zenodo.org/record/6078096#.YgrUxN_MLIU.

ACKNOWLEDGMENTS

We acknowledge support of the microfabrication and electron-beam lithography from EPFL Centre of MicroNanotechnology (CMI) and thank Z. Benes (CMI) for help with the electron-beam lithography. TEM was performed at the EPFL Interdisciplinary Centre for Electron Microscopy (CIME). We thank L. Navrátilová for preparation of device cross sections. We acknowledge support from the European Union's Horizon 2020 research and innovation programme under Grant Agreements 829035 (QUEFORMAL), 785219 and 881603 (Graphene Flagship Core 2 and Core 3), from the H2020 European Research Council (ERC, Grant No. 682332) as well as from the CCMX Materials Challenge grant "Large area growth of 2D materials for device integration".

REFERENCES

(1) Yu, S. Neuro-Inspired Computing with Emerging Nonvolatile Memorys. *Proc. IEEE* **2018**, *106* (2), 260−285.

(2) Kononenko, I. Machine Learning for Medical Diagnosis: History, State of the Art and Perspective. *Artif. Intell. Med.* **2001**, *23* (1), 89−109.

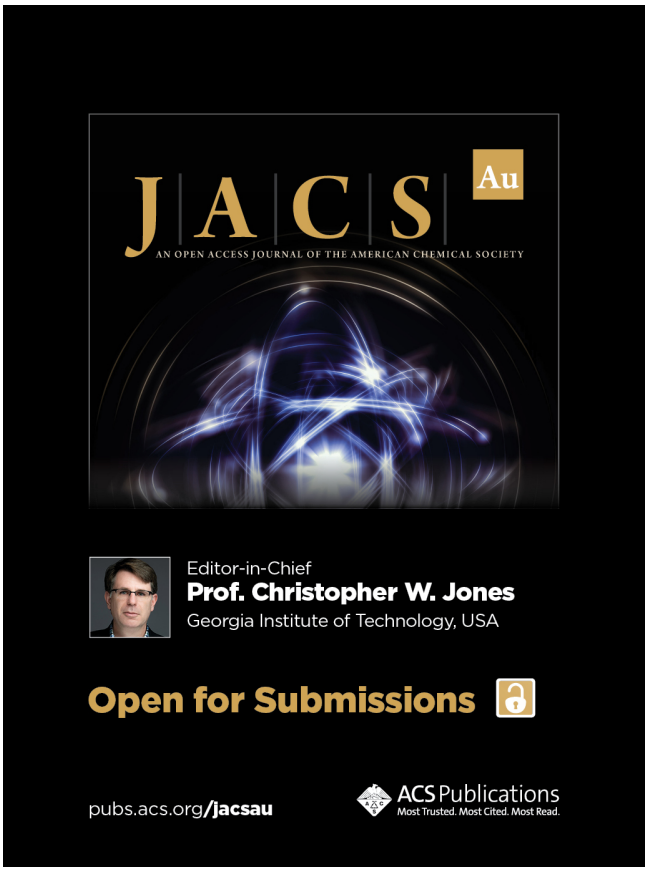
(3) Graves, A.; Mohamed, A.; Hinton, G. Speech Recognition with Deep Recurrent Neural Networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal

https://doi.org/10.1021/acsnano.1c07065
ACS Nano XXXX, XXX, XXX−XXX

Processing; Vancouver, BC, Canada, May 26−31, 2013; IEEE Publishing, New York, 2013; pp 6645−6649. DOI: 10.1109/ICASSP.2013.6638947.
(4) Chen, C.; Seff, A.; Kornhauser, A.; Xiao, J. DeepDriving: Learning Affordance for Direct Perception in Autonomous Driving. In *2015 IEEE International Conference on Computer Vision (ICCV)*, Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, Dec 7−13, 2015; IEEE Publishing, New York, 2015; pp 2722−2730. DOI: 10.1109/ICCV.2015.312.
(5) Zhang, L.; Ai, H.; Chen, W.; Yin, Z.; Hu, H.; Zhu, J.; Zhao, J.; Zhao, Q.; Liu, H. CarcinoPred-EL: Novel Models for Predicting the Carcinogenicity of Chemicals Using Molecular Fingerprints and Ensemble Learning Methods. *Sci. Rep.* **2017**, *7* (1), 2118.
(6) Carleo, G.; Cirac, I.; Cranmer, K.; Daudet, L.; Schuld, M.; Tishby, N.; Vogt-Maranto, L.; Zdeborová, L. Machine Learning and the Physical Sciences. *Rev. Mod. Phys.* **2019**, *91* (4), 045002.
(7) Xu, X.; Ding, Y.; Hu, S. X.; Niemier, M.; Cong, J.; Hu, Y.; Shi, Y. Scaling for Edge Inference of Deep Neural Networks. *Nat. Electron.* **2018**, *1* (4), 216−222.
(8) 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)*, Proceedings of the 2013 IEEE International Symposium on Workload Characterization (IISWC), Portland, OR, Sep 22−24, 2013; IEEE Publishing, New York, 2013; pp 56−65. DOI: 10.1109/IISWC.2013.6704670.
(9) Sun, Z.; Pedretti, G.; Ambrosi, E.; Bricalli, A.; Wang, W.; Ielmini, D. Solving Matrix Equations in One Step with Cross-Point Resistive Arrays. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (10), 4123−4128.
(10) Sun, Z.; Pedretti, G.; Bricalli, A.; Ielmini, D. One-Step Regression and Classification with Cross-Point Resistive Memory Arrays. *Sci. Adv.* **2020**, *6* (5), No. eaay2378.
(11) Zidan, M. A.; Jeong, Y.; Lee, J.; Chen, B.; Huang, S.; Kushner, M. J.; Lu, W. D. A General Memristor-Based Partial Differential Equation Solver. *Nat. Electron.* **2018**, *1* (7), 411−420.
(12) Li, C.; Hu, M.; Li, Y.; Jiang, H.; Ge, N.; Montgomery, E.; Zhang, J.; Song, W.; Dávila, N.; Graves, C. E.; Li, Z.; Strachan, J. P.; Lin, P.; Wang, Z.; Barnell, M.; Wu, Q.; Williams, R. S.; Yang, J. J.; Xia, Q. Analogue Signal and Image Processing with Large Memristor Crossbars. *Nat. Electron.* **2018**, *1* (1), 52−59.
(13) Le Gallo, M.; Sebastian, A.; Cherubini, G.; Giefers, H.; Eleftheriou, E. Compressed Sensing With Approximate Message Passing Using In-Memory Computing. *IEEE Trans. Electron Devices* **2018**, *65* (10), 4304−4312.
(14) Wang, Z.; Li, C.; Song, W.; Rao, M.; Belkin, D.; Li, Y.; Yan, P.; Jiang, H.; Lin, P.; Hu, M.; Strachan, J. P.; Ge, N.; Barnell, M.; Wu, Q.; Barto, A. G.; Qiu, Q.; Williams, R. S.; Xia, Q.; Yang, J. J. Reinforcement Learning with Analogue Memristor Arrays. *Nat. Electron.* **2019**, *2* (3), 115−124.
(15) Yao, P.; Wu, H.; Gao, B.; Tang, J.; Zhang, Q.; Zhang, W.; Yang, J. J.; Qian, H. Fully Hardware-Implemented Memristor Convolutional Neural Network. *Nature* **2020**, *577* (7792), 641−646.
(16) Sebastian, A.; Le Gallo, M.; Khaddam-Aljameh, R.; Eleftheriou, E. Memory Devices and Applications for In-Memory Computing. *Nat. Nanotechnol.* **2020**, *15* (7), 529−544.
(17) Desai, S. B.; Madhvapathy, S. R.; Sachid, A. B.; Llinas, J. P.; Wang, Q.; Ahn, G. H.; Pitner, G.; Kim, M. J.; Bokor, J.; Hu, C.; Wong, H.-S. P.; Javey, A. MoS2 Transistors with 1-Nanometer Gate Lengths. *Science* **2016**, *354* (6308), 99−102.
(18) Lopez-Sanchez, O.; Lembke, D.; Kayci, M.; Radenovic, A.; Kis, A. Ultrasensitive Photodetectors Based on Monolayer MoS2. *Nat. Nanotechnol.* **2013**, *8* (7), 497−501.
(19) Mennel, L.; Symonowicz, J.; Wachter, S.; Polyushkin, D. K.; Molina-Mendoza, A. J.; Mueller, T. Ultrafast Machine Vision with 2D Material Neural Network Image Sensors. *Nature* **2020**, *579* (7797), 62−66.
(20) Migliato Marega, G.; Zhao, Y.; Avsar, A.; Wang, Z.; Tripathi, M.; Radenovic, A.; Kis, A. Logic-in-Memory Based on an Atomically Thin Semiconductor. *Nature* **2020**, *587* (7832), 72−77.
(21) Radisavljevic, B.; Radenovic, A.; Brivio, J.; Giacometti, V.; Kis, A. Single-Layer MoS 2 Transistors. *Nat. Nanotechnol.* **2011**, *6* (3), 147−150.
(22) Agarwal, T. kumar; Soree, B.; Radu, I.; Raghavan, P.; Iannaccone, G.; Fiori, G.; Dehaene, W.; Heyns, M. Material-Device-Circuit Co-Optimization of 2D Material Based FETs for Ultra-Scaled Technology Nodes. *Sci. Rep.* **2017**, *7* (1), 5016.
(23) Smets, Q.; Arutchelvan, G.; Jussot, J.; Verreck, D.; Asselberghs, I.; Mehta, A. N.; Gaur, A.; Lin, D.; Kazzi, S. E.; Groven, B.; Caymax, M.; Radu, I. Ultra-Scaled MOCVD MoS_2 MOSFETs with 42nm Contact Pitch and 250μA/μm Drain Current. In *2019 IEEE International Electron Devices Meeting (IEDM)*, Proceedings of the IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, Dec 7−11, 2019; IEEE Publishing, New York, USA, 2020; p 23.2.1−23.2.4. DOI: 10.1109/IEDM19573.2019.8993650.
(24) Iannaccone, G.; Bonaccorso, F.; Colombo, L.; Fiori, G. Quantum Engineering of Transistors Based on 2D Materials Heterostructures. *Nat. Nanotechnol.* **2018**, *13* (3), 183−191.
(25) Bertolazzi, S.; Krasnozhon, D.; Kis, A. Nonvolatile Memory Cells Based on MoS_2/Graphene Heterostructures. *ACS Nano* **2013**, *7* (4), 3246−3252.
(26) Wang, J.; Zou, X.; Xiao, X.; Xu, L.; Wang, C.; Jiang, C.; Ho, J. C.; Wang, T.; Li, J.; Liao, L. Floating Gate Memory-Based Monolayer MoS_2 Transistor with Metal Nanocrystals Embedded in the Gate Dielectrics. *Small* **2015**, *11* (2), 208−213.
(27) Li, D.; Wang, X.; Zhang, Q.; Zou, L.; Xu, X.; Zhang, Z. Nonvolatile Floating-Gate Memories Based on Stacked Black Phosphorus-Boron Nitride-MoS_2 Heterostructures. *Adv. Funct. Mater.* **2015**, *25* (47), 7360−7365.
(28) Woo, M. H.; Jang, B. C.; Choi, J.; Lee, K. J.; Shin, G. H.; Seong, H.; Im, S. G.; Choi, S.-Y. Low-Power Nonvolatile Charge Storage Memory Based on MoS2 and an Ultrathin Polymer Tunneling Dielectric. *Adv. Funct. Mater.* **2017**, *27* (43), 1703545.
(29) Sangwan, V. K.; Jariwala, D.; Kim, I. S.; Chen, K.-S.; Marks, T. J.; Lauhon, L. J.; Hersam, M. C. Gate-Tunable Memristive Phenomena Mediated by Grain Boundaries in Single-Layer MoS 2. *Nat. Nanotechnol.* **2015**, *10* (5), 403−406.
(30) Shen, P.-C.; Lin, C.; Wang, H.; Teo, K. H.; Kong, J. Ferroelectric Memory Field-Effect Transistors Using CVD Monolayer MoS2 as Resistive Switching Channel. *Appl. Phys. Lett.* **2020**, *116* (3), 033501.
(31) Jayachandran, D.; Oberoi, A.; Sebastian, A.; Choudhury, T. H.; Shankar, B.; Redwing, J. M.; Das, S. A Low-Power Biomimetic Collision Detector Based on an in-Memory Molybdenum Disulfide Photodetector. *Nat. Electron.* **2020**, *3* (10), 646−655.
(32) Choi, C.; Leem, J.; Kim, M. S.; Taqieddin, A.; Cho, C.; Cho, K. W.; Lee, G. J.; Seung, H.; Bae, H. J.; Song, Y. M.; Hyeon, T.; Aluru, N. R.; Nam, S.; Kim, D.-H. Curved Neuromorphic Image Sensor Array Using a MoS 2 -Organic Heterostructure Inspired by the Human Visual Recognition System. *Nat. Commun.* **2020**, *11* (1), 5934.
(33) Feng, X.; Li, S.; Wong, S. L.; Tong, S.; Chen, L.; Zhang, P.; Wang, L.; Fong, X.; Chi, D.; Ang, K.-W. Self-Selective Multi-Terminal Memtransistor Crossbar Array for In-Memory Computing. *ACS Nano* **2021**, *15* (1), 1764−1774.
(34) Jang, H.; Liu, C.; Hinton, H.; Lee, M.-H.; Kim, H.; Seol, M.; Shin, H.-J.; Park, S.; Ham, D. An Atomically Thin Optoelectronic Machine Vision Processor. *Adv. Mater.* **2020**, *32* (36), 2002431.
(35) Bavandpour, M.; Sahay, S.; Mahmoodi, M. R.; Strukov, D. Efficient Mixed-Signal Neurocomputing Via Successive Integration and Rescaling. *IEEE Trans. Very Large Scale Integr. VLSI Syst.* **2020**, *28* (3), 823−827.
(36) Paliy, M.; Strangio, S.; Ruiu, P.; Rizzo, T.; Iannaccone, G. Analog Vector-Matrix Multiplier Based on Programmable Current Mirrors for Neural Network Integrated Circuits. *IEEE Access* **2020**, *8*, 203525−203537.
(37) Kim, H.; Ovchinnikov, D.; Deiana, D.; Unuchek, D.; Kis, A. Suppressing Nucleation in Metal−Organic Chemical Vapor Deposition of MoS_2 Monolayers by Alkali Metal Halides. *Nano Lett.* **2017**, *17*, 5056.

https://doi.org/10.1021/acsnano.1c07065
ACS Nano XXXX, XXX, XXX−XXX

(38) Cun, H.; Macha, M.; Kim, H.; Liu, K.; Zhao, Y.; LaGrange, T.; Kis, A.; Radenovic, A. Wafer-Scale MOCVD Growth of Monolayer MoS_2 on Sapphire and SiO_2. *Nano Res.* **2019**, *12* (10), 2646−2652.
(39) Aritome, S. *NAND Flash Memory Technologies*; Wiley-IEEE Press: Hoboken, NJ, 2015.
(40) Sasaki, T.; Ueno, K.; Taniguchi, T.; Watanabe, K.; Nishimura, T.; Nagashio, K. Understanding the Memory Window Overestimation of 2D Materials Based Floating Gate Type Memory Devices by Measuring Floating Gate Voltage. *Small* **2020**, *16* (47), 2004907.
(41) Merced-Grafals, E. J.; Dávila, N.; Ge, N.; Williams, R. S.; Strachan, J. P. Repeatable, Accurate, and High Speed Multi-Level Programming of Memristor 1T1R Arrays for Power Efficient Analog Computing Applications. *Nanotechnology* **2016**, *27* (36), 365202.
(42) Strubell, E.; Ganesh, A.; McCallum, A. Energy and Policy Considerations for Deep Learning in NLP. *ArXiv* 2019, 1906.02243, http://arxiv.org/abs/1906.02243 (accessed 2021-03-16).
(43) Bavandpour, M.; Mahmoodi, M. R.; Strukov, D. B. ACortex: An Energy-Efficient Multipurpose Mixed-Signal Inference Accelerator. *IEEE J. Explor. Solid-State Comput. Devices Circuits* **2020**, *6* (1), 98−106.
(44) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60* (6), 84−90.



https://doi.org/10.1021/acsnano.1c07065
ACS Nano XXXX, XXX, XXX−XXX