Check for updates

OPEN
Energy-efficient memcapacitor devices for neuromorphic computing

Kai-Uwe Demasius [1]✉, Aron Kirschen [2]✉ and Stuart Parkin [1]✉

Data-intensive computing operations, such as training neural networks, are essential for applications in artificial intelligence but are energy intensive. One solution is to develop specialized hardware onto which neural networks can be directly mapped, and arrays of memristive devices can, for example, be trained to enable parallel multiply–accumulate operations. Here we show that memcapacitive devices that exploit the principle of charge shielding can offer a highly energy-efficient approach for implementing parallel multiply–accumulate operations. We fabricate a crossbar array of 156 microscale memcapacitor devices and use it to train a neural network that could distinguish the letters 'M', 'P' and 'I'. Modelling these arrays suggests that this approach could offer an energy efficiency of 29,600 tera-operations per second per watt, while ensuring high precision (6–8 bits). Simulations also show that the devices could potentially be scaled down to a lateral size of around 45 nm.

B rain-inspired computing—often termed neuromorphic computing—based on artificial neural networks and their hardware implementations could be used to solve a broad range of computationally intensive tasks. Neuromorphic computing can be traced back to the 1980s (refs. [1,2]), but the field gained considerable momentum after the development of memristive devices[3] and the proposal of convolutional layers in deep neural networks at the algorithmic level[4,5]. Since then, several resistive neuromorphic systems and devices have been implemented using oxide materials[6–8], phase-change memory[9], spintronic devices[10,11] and ferroelectric devices (tunnel junctions[12,13] and ferroelectric field-effect transistors (FeFETs)[14,15]), and such systems—namely, ferroelectric tunnel junctions[13] and SONOS (that is, silicon–oxide–nitride–oxide–silicon) transistors[16]—have exhibited energy efficiencies of up to 100 tera-operations per second per watt (TOPS W⁻¹). All these approaches rely on the analogue storage of synaptic weights, which can be used in multiplication operations, and use Kirchhoff's current law for the summation of currents implemented via crossbar arrays[17].

Memcapacitive devices[18] are similar to memristive devices but are based on a capacitive principle, and could potentially offer a lower static power consumption than memristive devices. There have been theoretical proposals for memcapacitor devices[18–22], but few practical implementations[23–26]. Memcapacitor devices can be realized through the implementation of a variable plate distance concept, as demonstrated in micro-electromechanical systems[27], a metal-to-insulator transition material in series with a dielectric layer[22], changing the oxygen vacancy front in a classical memristor[20], and a simple metal–oxide–semiconductor capacitor with a memory effect[24,25]. To obtain a high dynamic range, these devices either have a large parasitic resistive component[20] at small plate distances or limited lateral scalability due to large plate distances. Similar problems occur with memcapacitors having varying surface areas[23] or varying dielectric constants[26].

In this Article, we report memcapacitor devices based on charge shielding that can offer high dynamic range and low power operation. We fabricate devices on the scale of tens of micrometres and use them to create a crossbar array architecture that we use to run an

image recognition algorithm. We also assess the potential scalability of our devices for use in large-scale energy-efficient neuromorphic systems using simulations.

Memcapacitive device based on charge shielding
Our memcapacitive device consists of a top gate electrode, a shielding layer with contacts and a back-side readout electrode (Fig. 1a). These layers are separated by dielectric layers. The top dielectric layer can have a memory effect, for example, charge trapping or ferroelectric, which may influence the shielding layer, or the shielding layer itself can exhibit a memory effect (in this paper, only the first principle is investigated). A very high on/off ratio of electric field coupling and therefore the capacitance between the gate electrode and readout electrode can be obtained with either total shielding or transmission. The lateral scalability is substantially better compared with the previously mentioned concepts, since the thickness of each layer can be readily optimized, while the dynamic ratio is mainly dependent on the shielding efficiency of the shielding layer.

Generally, charge screening depends on the Debye screening length L_D:

$$L_D = \sqrt{\frac{\varepsilon_0 \varepsilon_r U_T}{n^2 e}}, \tag{1}$$

where U_T is the thermal voltage, n is the charge carrier concentration, ε_0 is the electric field constant, ε_r is the relative electric field constant and e is the elementary charge. The electric field drops exponentially within the shielding layer and drops to 37% within the screening length L_D under the condition $\Psi \ll U_T$. In practice, in semiconductors, the relationship is highly nonlinear depending on potential ψ at depth x, as follows:

$$\frac{d^2\psi}{dx^2} = \frac{-e}{\varepsilon_0 \varepsilon_r} \left(p_0 \left[\exp\left(\frac{-\psi}{U_T}\right) - 1 \right] - n_0 \left[\exp\left(\frac{\psi}{U_T}\right) - 1 \right] \right), \tag{2}$$

where p_0 and n_0 are the charge carrier concentrations of holes and electrons in thermal equilibrium, respectively. Therefore, the Debye

[1]Max Planck Institute of Microstructure Physics, Halle (Saale), Germany. [2]SEMRON GmbH, Dresden, Germany. ✉e-mail: kudemasius@semron-ai.com; akirschen@semron-ai.com; Stuart.Parkin@mpi-halle.mpg.de
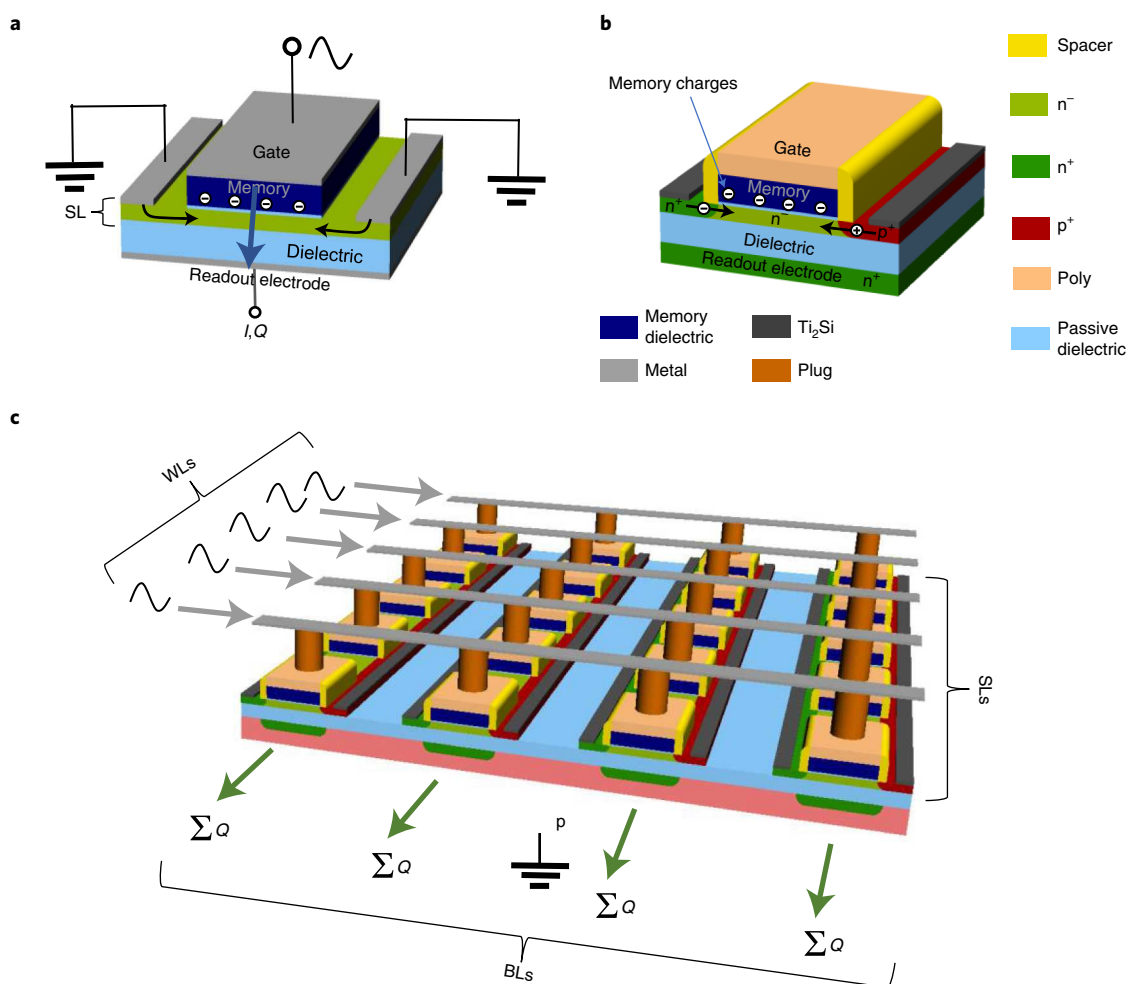
Fig. 1 | Structure of the memcapacitor device. a, General device structure with a gate electrode, shielding layer (SL) and readout electrode (*I*, current; *Q*, charge). The electric field coupling is indicated by the blue arrow. **b**, Device structure with a lateral pin junction as well as electron and hole injection. **c**, Crossbar arrangement of the device in **b**, where a.c. input signals are applied to the word lines (WLs) and the accumulated charge is read out at the bit lines (BLs). During readout, the SL is mostly connected to GND.

screening length (equation (1))—given the exponential spatial dependence of the field in the material—is only a linear approximation of nonlinear differential equation (2). Especially for strong inversion and accumulation within the shielding layer, the length scales of screening become much smaller than the Debye length. This nonlinearity with respect to the applied gate voltage or charge stored in the memory dielectric leads to either strong shielding or fairly good transmission.

A more detailed device structure is shown in Fig. 1b with lateral $p^+n^-n^+$ junctions in the shielding layer. The p^+- and n^+-doped regions act as reservoirs for electrons and holes, respectively, and can inject each carrier type for the purposes of shielding. This enables additional device functionality; however, more importantly, it also allows a symmetric device response for positive and negative gate voltages. This is a crucial feature for neuromorphic devices, because the weight update is then undistorted and the training accuracy is thus higher[17]. During readout, the shielding layer is connected to the ground (GND). During writing and training, the voltages applied to the p^+ and n^+ contacts can differ and can also act as a selector, as explained in Supplementary Section 1. As shown in Fig. 1c, the single device can be arranged into a crossbar for highly parallel multiply–accumulate (MAC) operations. In this case, the gate electrode becomes the word line (WL), where input signals are

applied, and the shielding layer becomes a shielding line (SL) in a direction vertical to the WL. The readout electrode functions as the bit line (BL), which is parallel to the SL, and the accumulated charge out of one BL is the calculated result of accumulated multiplications at each crossing point. The multiplication is conducted between the input signal of the WL and the state of the shielding layer, which, in turn, is adjusted by the memory material. The weights are encoded in the capacitance of each crossing point. In contrast to resistive devices, capacitive devices only react on dynamic voltage or current signals; therefore, an alternating current (a.c.) voltage is applied to the WL during readout. Writing of the memory material is achieved by a voltage difference between the SL and WL.

CV curves and gradual programming of single devices
Single devices on the micrometre scale were fabricated on a silicon-on-insulator wafer, whereas the handle wafer containing a highly n-doped epitaxial layer acts as the readout electrode and the buried oxide acts as the bottom dielectric layer. As a memory principle, ferroelectric-assisted charge trapping (polarization charge attracts carriers and thus promotes trapping) was used to combine the advantages of both principles[28,29], whereas the tunnelling oxide was 2.5 nm thick to avoid charge detrapping. Details of the fabrication can be found in Methods.
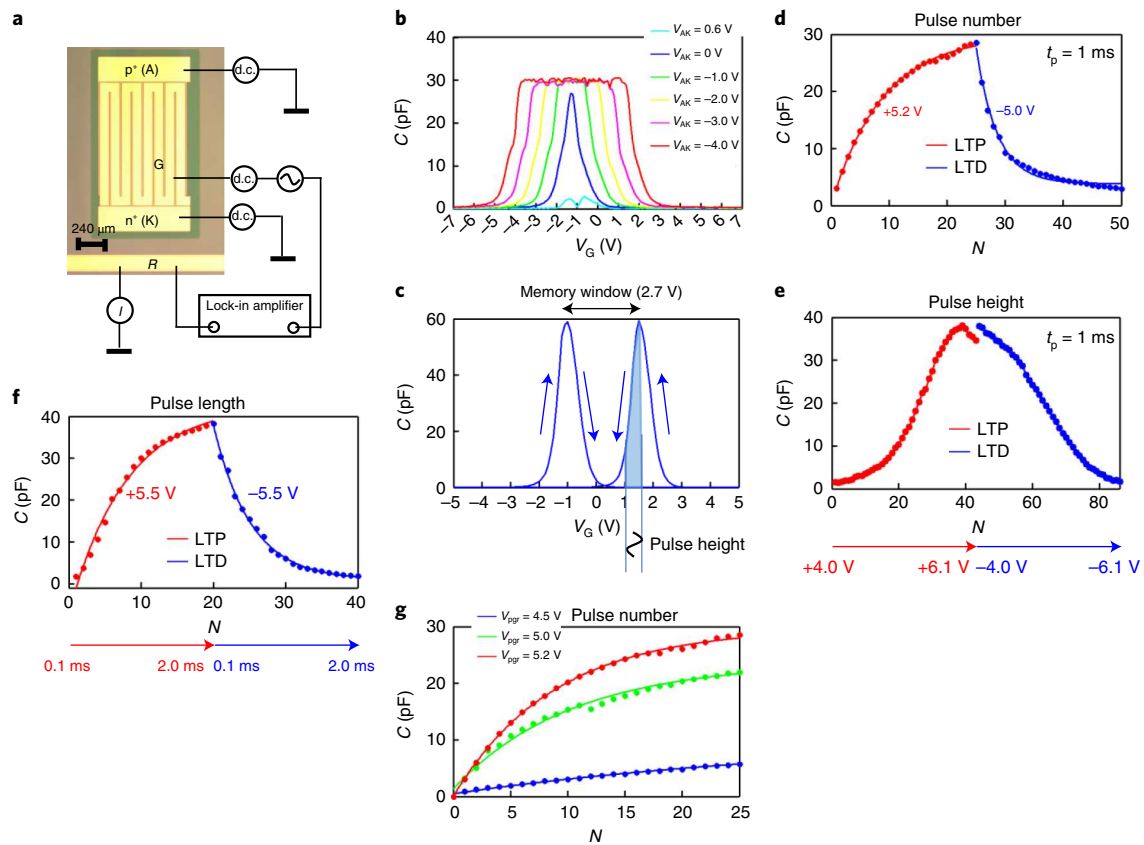
Fig. 2 | Measurement setup and CV curves of single devices. a, Microscopy image of the measured single device and measurement setup. **b**, Measured CV curves for a device without memory at different V_{AK} values; V_{AK} is applied antisymmetrically, the d.c. voltage of the gate was swept between −7 and 7 V, and the small a.c. voltage had an amplitude of 100 mV with a frequency of 1 kHz. **c**, CV curve shifting due to the injection of charges. The device had a memory in this case. **d–f**, Analogue value writing with pulse number modulation (constant write height) (**d**), pulse height modulation (the voltage is increased/decreased from ±4.0 to ±6.1 V) (**e**) and pulse length modulation (**f**). In **d–f**, the shielding layer was grounded, and readout was performed between each pulse with an a.c. signal, as shown in **c**. **g**, Pulse number modulation for different write pulse heights.

The fabricated devices had a gate length ranging from 10 to 60 μm, and the gate width was enlarged by winding it around several highly p+- and n+-doped finger-shaped regions, thus forming several parallel pin junctions. The larger area leads to a readily detectable capacitance and the minimum capacitance of turned-off devices could also be precisely measured (capacitive dynamic range). Figure 2a shows a microscopic image of the fabricated device. Capacitance–voltage (CV) measurements were carried out by applying an a.c. signal with a direct current (d.c.) bias (sweep) to the gate: the resulting a.c. current of the readout electrode was measured either by lock-in amplification or by an oscilloscope and current pre-amplifier. Data from the resulting fundamental CV curves for different d.c. voltages (V_{AK}) on the n+ and p+ regions are shown in Fig. 2b (note that a normal silicon dioxide dielectric layer was used here instead of a memory dielectric). The CV curves get broader or are nearly extinguished depending on whether the pin junction is used in the reverse or forward bias direction, respectively; this behaviour is further explained in Supplementary Section 1. Generally, a capacitive coupling window is observed, which is high for depletion (and therefore for transmission through the shielding layer) and low during inversion or accumulation. The curves are derivatives of a sigmoid curve, which play an important role in modelling neurons in artificial neural networks. A direct measurement of the sigmoid curve and further uses are explained in Supplementary Section 1.

Replacing the normal silicon dioxide dielectric with a memory dielectric and with a CV sweep from −5 to 5 V, one can observe a shifting of the capacitive coupling window with a memory window of 2.7 V (Fig. 2d), while the pin junction was grounded. Due to the shifting direction, one can conclude that charge trapping is the memory principle (for purely ferroelectric switching, the curves would shift in the opposite direction). By contrast, capacitive devices can only be read out by a.c. voltages or current signals. For this reason, an alternating voltage (0.5 V) is applied to the gate for readout, together with a bias voltage (1.0 V) to adjust the readout window, as indicated by the shaded area in Fig. 2d (note that the pin junction is grounded during readout). In Supplementary Fig. 11a,b, the readout current of a written and erased cell is shown, and a capacitive dynamic range of ~1:1,478 was experimentally achieved.

To store analogue values, one can apply short pulses with the same amplitude (Fig. 2d,g), apply pulses with increasing height (Fig. 2e) or change the pulse length (Fig. 2f) applied to the gate. The resulting curves exhibit some similarities to those obtained from pure ferroelectric switching[14], indicating the ferroelectric assistance in the memory storage process. The curve in Fig. 2d shows a typical nonlinear long-term potentiation (LTP) curve with an exponential dependence.

$$C_{LTP} = C_{min} + \Delta C \left(1 - \exp\left(\frac{-N_{pgr}}{\beta_{pgr}} \right) \right) \qquad (3)$$

The same applies for the long-term depression (LTD)

$$C_{\text{LTD}} = C_{\text{max}} - \Delta C \left(1 - \exp \left(\frac{-N_{\text{er}}}{\beta_{\text{er}}} \right) \right), \quad (4)$$

where N_{pgr} and N_{er} denote the number of programming or erase pulses, respectively; β_{pgr} and β_{er} are the stretching factors; and C_{min} and C_{max} denote the minimum and maximum capacitance, respectively. Here ΔC describes the maximum change in capacitance. Changing the write pulse height of the pulse number modulation leads to more flattened or steepened curves (Fig. 2g). Write/erase pulse height modulation (Fig. 2e) can lead to relatively symmetric and—in certain regions, linear—behaviour with respect to the pulse height steps. This is highly beneficial for implementing neuromorphic algorithms[17]. Pulse length modulation shows similar behaviour to pulse number modulation (Fig. 2f). In Supplementary Fig. 11c, the measured readout current is illustrated for LTP and LTD for different pulse numbers of pulse height modulation (Fig. 2e) and reveals the pinch-off and increase.

Other memory parameters, like device-to-device variation, endurance and retention can be found in Supplementary Section 9.

Crossbar array and implementation of training algorithm

Crossbar devices—used to execute an image recognition algorithm—were fabricated and wire bonded onto a chip carrier. A printed circuit board (PCB) was designed and controlled by a data acquisition system. An image of the fabricated chip with the bonding pads, a zoomed-in microscopy image of the crossbar and a scanning electron microscopy image are shown in Fig. 3a. Each memory cell had a size of $50 \times 50 \, \mu\text{m}^2$.

A schematic of the device cross section is shown in Fig. 3b. The BLs of the memory array were separated by refilled deep trenches. Details of the fabrication process can be found in Methods.

The matrix comprised 26 WLs and 6 BLs (Fig. 3c). A differential weight topology[17] was used with the positive and negative value of each weight separated in two memory cells. The values of these two BLs were subtracted from each other.

$$W_{ij} = C_{ij}^+ - C_{ij}^- \quad (5)$$

The input values are separated by a sign with a 180° phase shift. For the desired 'four-quadrant multiplication' (input × weight), a global clock signal is used together with the switched capacitor approach (Fig. 3c). Further details are explained in Supplementary Section 11. The integration capacitance of the amplifier is charged up in each period of the input sine signal, and hence, the number of periods (N_{per}) encodes the value of the input signal. This effect also leads to an averaging of the noise level and improvement in the signal-to-noise ratio, as explained later. This theoretical concept of 'four-quadrant multiplication' was confirmed with the following measurement (Fig. 3d): the input number of periods (N_{per}) and the number of programming pulses (N_{pgr}), which adjust the actual weight, were varied in positive and negative values, while the output voltage is read. Positive and negative N_{per} values were encoded by a 180° phase shift and positive/negative programming pulses (N_{pgr}) only changed the positive/negative weights, while the counterpart was in an erased state. Supplementary Fig. 12a,b shows the cross sections of the 3D plot in Fig. 3d. The curves along the input period number behave in a highly linear manner, and this linearity was also confirmed for the accumulation operation (Supplementary Fig. 12c), demonstrating a highly linear MAC operation with the proposed switched capacitor approach.

The first 25 WLs enable a vectorized input feature map for images of 5×5 pixels; thus, one single fully connected layer is carried out. Dark pixels are represented by positive values and bright pixels, by negative values. The bias input is mapped to the 26th WL.

Regarding the implemented training algorithm, the Manhattan update[8,30] rule was chosen, due to its simplified training procedure. In conventional backpropagation training, the weight update is calculated as follows:

$$\Delta W_{ij} = -\alpha \delta_i(n) X_j(n), \quad (6)$$

where α describes the learning rate, $\delta_i(n)$ is the backpropagated error and $X_j(n)$ is the current input for the nth input image, which is randomly chosen from the training set. The weights are updated after each sample (stochastic training). The backpropagated error for a one-layer perceptron can be calculated as follows:

$$\delta_i(n) = \left[f_i(n) - f_i^{\text{d}}(n) \right] \left. \frac{\text{d}f_i}{\text{d}v} \right|_{v=v_i(n)}, \quad (7)$$

where $f_i^{\text{d}}(n)$ is the desired output value and $f_i(n)$ is the current output. Function f_i is related to the voltage output $v_i(n)$ of the ith sense amplifier and the activation function of the neuron (in this case, tanh):

$$f_i(v_i) = \tanh(\kappa v_i(n)), \quad (8)$$

where κ is the steepness factor. With the Manhattan update rule, the weight update from equation (6) is coarse-grained by using the following signing.

$$\Delta W_{ij}^{\text{M}} = \text{sgn} \Delta W_{ij} \quad (9)$$

Therefore, all the weights are updated by the same amount based on their sign. Figure 4a illustrates the pulse scheme for implementing the algorithm. The term $\delta_i(n) X_j(n)$ in equation (6) becomes positive if both error $\delta_i(n)$ and input $X_j(n)$ are positive or it becomes negative for the opposite sign if both $\delta_i(n)$ and $X_j(n)$ are negative . Hence, one can describe this by an XNOR combination. To update the weights, the error signal is applied to the SL, as shown in Fig. 4a. The corresponding input signals are applied to the WL. The differential signal at the crossing points follows the XNOR operation, while the specific signals (shown in Fig. 4a) ensure that the maximum disturbance level is not higher than 1/3 and thus effectively prevents the overwriting of cells in the same column or row (the memory cell acts as the selector itself; see Supplementary Sections 7 and 8). As a 5×5 image recognition task, the letters M, P and I were chosen, and one pixel in each of the samples was flipped, which results in a total set of 78 samples. These pseudo-images were separated into a test and training set; the test images are indicated by a blue frame (Fig. 4b). The resulting misclassified images versus training epochs for the training and test images are shown in Fig. 4c. Evidently, the number rapidly decreases after one training epoch and stays almost zero throughout the training epochs. Figure 4d shows the obtained mean neuron activations for the three classifications over the training epochs. The slightly higher simulated average misclassification rate (Fig. 4c) is the consequence of single steep climbs of the misclassification rate after an arbitrary number of epochs with 100% accuracy in some runs. Misclassifications after epoch 1 are caused by the very similar expected value for individual presynaptic neurons for letters M and P. Measurements also confirm the more stable results for the classification of letter I, as shown in Fig. 4d. The results are in accordance with other studies[7,8].

Thus, experimental results on micrometre-sized devices demonstrate the working principle. For demonstrating scalability to the nanometre regime and superior energy efficiency, detailed and extensive simulations were performed, which are explained in the upcoming sections.
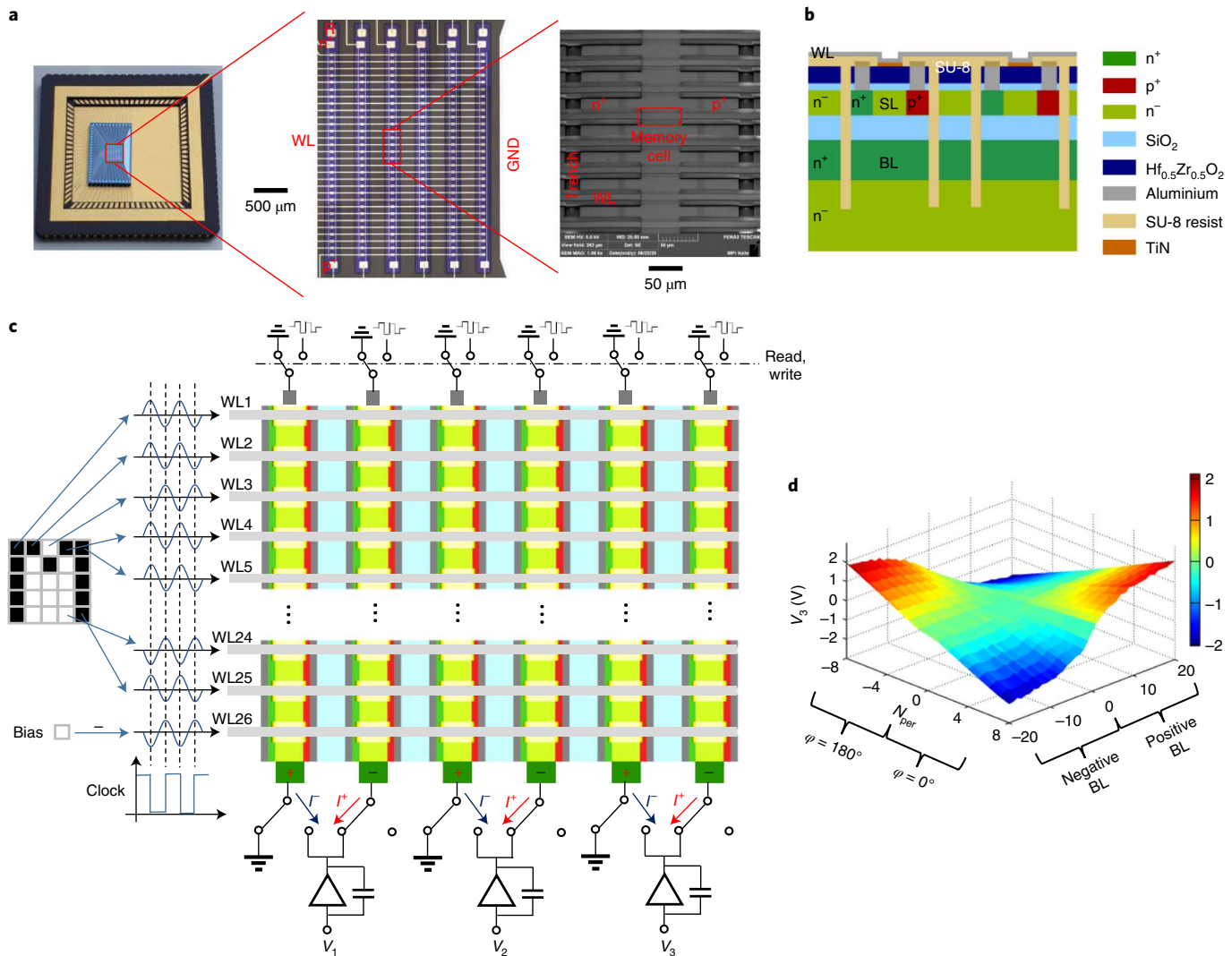
Fig. 3 | Crossbar arrangement and fundamental measurements. a, Wire-bonded chip with microscopy and scanning electron microscopy images.
b, Device cross section. **c**, Neuromorphic system for accomplishing 'four-quadrant multiplication': positive and negative inputs are 180° phase shifted
with each other. The a.c. conditions are the same as in Fig. 2, and the number of periods encodes the amount of input. The clock signal is high for a
rising edge in the positive signal and the switches are in the left position during a high clock signal. The SL is connected to GND during readout.
d, Measured 'four-quadrant multiplication' for different input period numbers N_{per} and programming pulse numbers (pulse number modulation) N_{pgr}.
For negative N_{per}, the input signal is 180° phase shifted, and for positive N_{pgr}, a positive BL is programmed; a negative BL is kept in an erased state (vice
versa for negative N_{pgr}).

TCAD simulations on single devices

A device with 90 nm gate length (Fig. 5a) was simulated by Synopsys.
Figure 5b (where no memory dielectric was integrated for the first
simulations) shows the CV curves of the coupling capacitances
between the gate and readout electrode with respect to the applied
gate voltage (V_G), which are consistent with the observed experi-
mental behaviour (Fig. 2b).

The ratio between the maximum capacitance and lower-state
capacitance obtained by shifting the gate voltage by 3 V is 1:90 in
this device, and this ratio can be further enlarged by using thin-
ner gate oxides or larger gate lengths, as shown in Fig. 5c. In gen-
eral, the capacitive ratio decreases with a smaller gate length due
to the fact that the influence of the space charge region becomes
more pronounced for smaller gate lengths (short channel effect)
and sufficient shielding is hard to achieve in this region (Fig. 5c,
inset). By using high-κ dielectrics for the top and bottom oxides, a
ratio of 1:60 was obtained for a 45 nm device with the same capac-
itance as the 90 nm device, as shown in Supplementary Section 2.

A dynamic range of 1:60–1:90 is sufficient to achieve a precision
of 6–8 bits[31].

Including a memory window (~3 V for charge-trapping memo-
ries and ~1–2 V for ferroelectric memories depending on the thick-
ness and coercive field) leads to shifted CV curves (Fig. 5d). The a.c.
readout voltage is indicated in Fig. 5d; for the positive shifted curve,
the resulting readout current and therefore the accumulated charge
will be very large. The total readout charge over one-half period
of the applied sinusoidal signal versus memory shift is shown in
Fig. 5e. Most of the negative memory window is used for turning
off the device.

Scalability to 45 nm

With regard to lateral scalability, it is necessary to distinguish three
aspects: (1) the scalability of the memory technology in the top
dielectric itself with regard to how many levels can be stored; (2) the
sensitivity of the sense amplifier at the end of each BL for detecting
the accumulated charge; (3) the noise level of one single device during
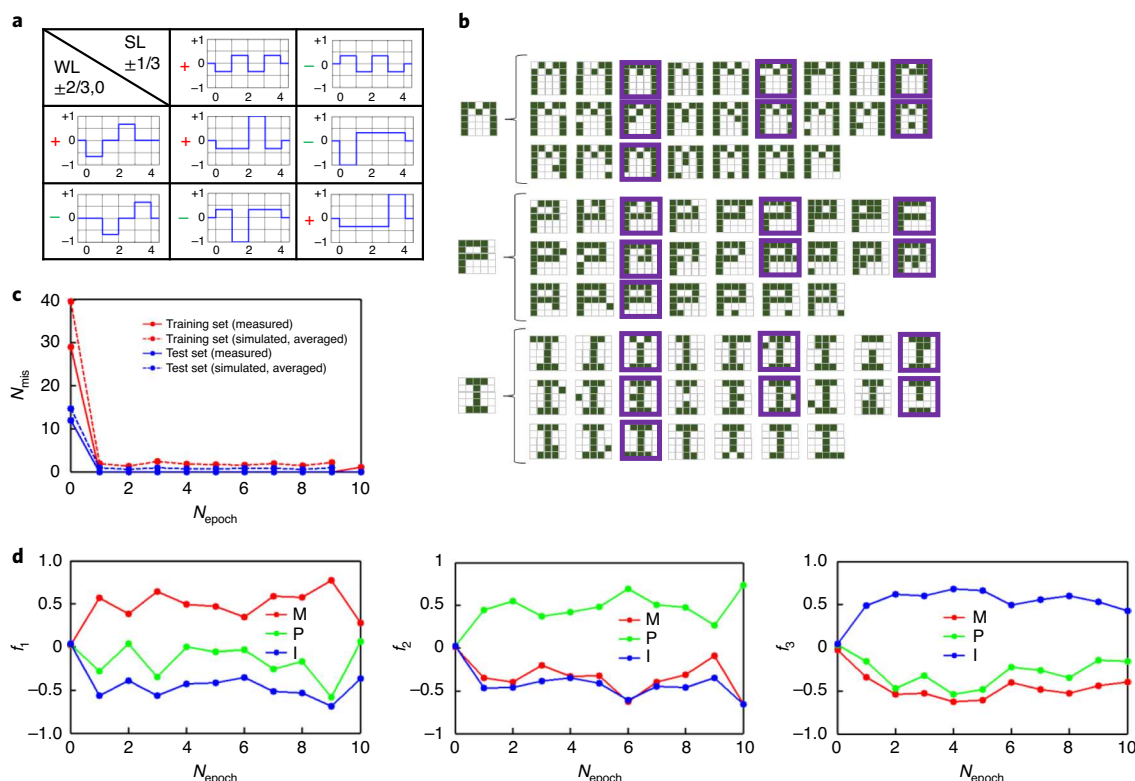
Fig. 4 | Manhattan update training on crossbar. a, Pulse scheme to enable XNOR operation during Manhattan weight update (the write/erase pulse height was ±5.2 V and length was 1 ms). The disturb level is exactly 1/3 of the write/erase voltage. **b**, Training and test set of the letters M, P and I with one flipped pixel. The test images are framed in purple. **c**, Number of misclassified images N_{mis} for the training and test sets over ten training epochs (N_{epoch}). The measured curve is compared with the simulated curves. **d**, Average artificial neuron activation for three classifications (f_1, f_2 and f_3) and three images over ten training epochs.

readout. Fairly common resolutions for input, weight and output signals for neural networks are in the range of 4–8 bits (16–256 levels)[31]. This analogue-like resolution has a significant influence on scalability. Typically, lower precision is needed for inference tasks.

With respect to the memory material, one can generally conclude that charge-trapping memories (for example, SONOS) have shown up to 31 levels down to 40 nm (ref. [16]). The disadvantage of this memory technology is the relatively high write energy and slowness during writing (millisecond regime). However, SONOS might be an alternative for inference-only applications. On the other hand, hafnium oxide (a ferroelectric) has very low write energies and is fast (nanosecond to microsecond regime). Ongoing research is still underway on the scalability of ferroelectric memories with regard to analogue storage. From FeFETs, it is known that they tend to show abrupt switching events below 500 nm, which is attributed to the limited grain size[15].

Regarding capacitive measurement resolutions, some work was done in the context of DNA sensing and chip interconnect measurements with resolutions down to <10 aF (charge-based capacitive measurements, capacitance-to-frequency conversion and lock-in detection)[32–36]. These are similar to a conventional sense amplifier[37,38] and contain an integration capacitor that is charged either by an operational amplifier circuit or a current mirror. Details on the sensitivity calculation can be found in Supplementary Section 3; generally, however, one has to consider that in neuromorphic devices, the accumulated charge from many memory cells (several hundreds to thousands) is read out at once and used for further information processing, which gives rise to much larger charges compared with only one cell. Furthermore, several pulse/period numbers are used for encoding the input value and leads to step-

wise charge integration over many periods. For the device shown in Fig. 5, $N_{per} = 142$ periods is necessary, which fits well into a range of 7–8 bits of the input signal (Supplementary Section 3). Note that 128 periods are sufficient for an 8-bit signed integer due to the use of the 180° phase shift for negative values of the switched capacitor approach.

Regarding the noise level of capacitive devices, one has to consider kTC noise.

$$v_n = \sqrt{\frac{k_B T}{C}} \tag{10}$$

where k_B defines the Boltzmann constant, T the temperature and C the capacitance. For a 6.65 aF device (Fig. 5d), one obtains a noise voltage of 25.00 mV (at room temperature), which is 14 times lower than the effective readout value of 0.35 V. However, one has to consider that the noise level decreases with the number of repetitive measurements, namely, $1/\sqrt{N_{per}}$, which results in a noise level of 2.20 mV (at room temperature) or 169 times lower than the effective readout value; this defines a precision of ~7 bits. Based on this minimum amplitude necessary to distinguish between different levels, it also becomes possible to assess the theoretical energy efficiency of resistive and capacitive devices in general (Supplementary Section 4): capacitive devices are at least eight times more energy efficient than resistive devices.

Simulation of ultrahigh energy efficiency
Much of the energy sourced to 'memcapacitors' can be recovered since it is stored in the capacitor; this is an important difference from resistors in which the readout operation is inherently
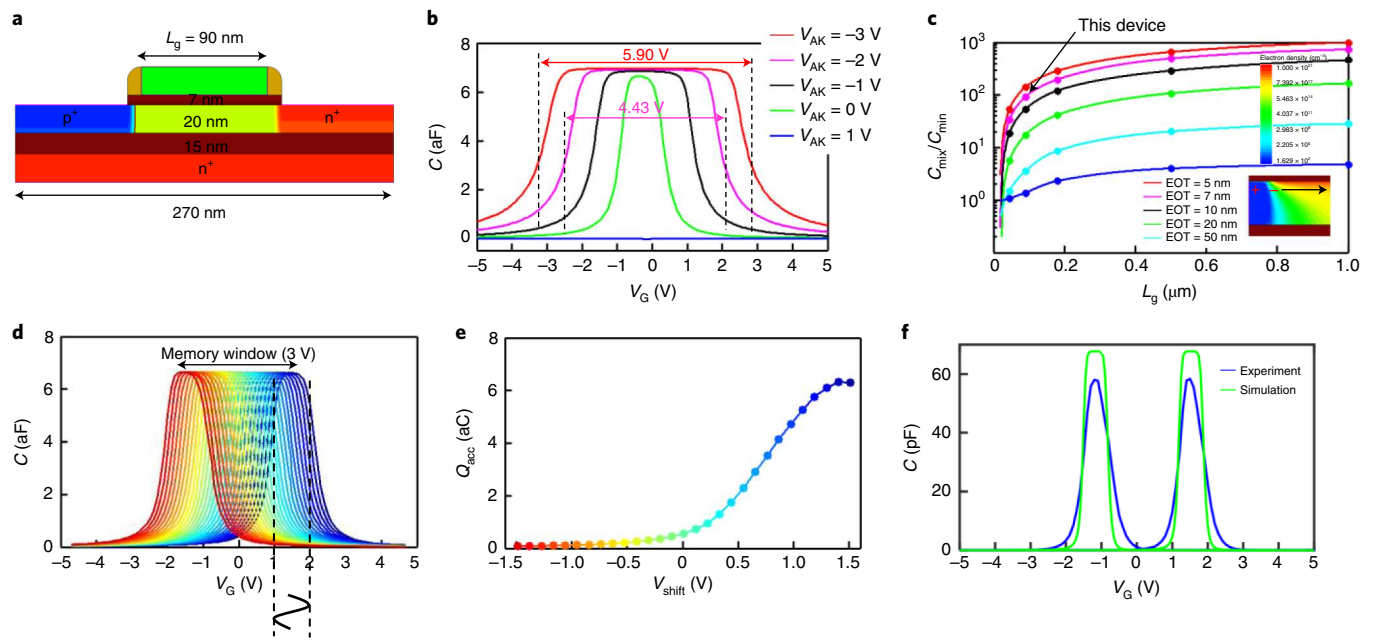
Fig. 5 | TCAD simulation results. a, Simulated structure with gate length $L_g = 90$ nm. **b**, Obtained CV curves with respect to the gate voltage for different voltages V_{AK} along the p$^+$n$^-$n$^+$ diode (quasi-static simulation). The voltage V_{AK} was applied antisymmetrically, as that in Fig. 2. **c**, Capacitive dynamic ratio (maximum capacitance/minimum capacitance of the CV curves with p$^+$n$^-$n$^+$ connected to GND) for different gate lengths and gate oxide thicknesses. The inset shows the electron density, and the short channel effect becomes obvious. EOT, equivalent oxide thickness. **d**, Shifting of the CV curves for $V_{AK} = 0$ V for different memory charges in the gate oxide. Note the applied readout a.c. signal with bias. **e**, Accumulated charge (Q_{acc}) for different voltage shifts (V_{shift}; caused by memory charges) over one-half period of the a.c. signal in **d**. **f**, Comparison of the simulated and experimental capacitive coupling curves for the micrometre-scaled device shown in Fig. 2.

dissipative due to Joule heating. The energy fed in during charging can be, in principle, recovered during discharging. This concept of energy recovery is also present in adiabatic circuit designs[39,40], which are at the core of the reversible computing paradigm[41,42]. The limiting factor of energy recovery in adiabatic circuits are resistive losses in the circuit, as well as in the inductances used for the power clock generators. The inductances have limited quality factors (q factor) in the order of dozens to hundreds. In common adiabatic realizations, energy recovery of the supply clock generators is of the order of 95% for harmonic signals[43–45], which means the supplied active power is $q = 20$ times lower than the reactive power.

To estimate the time delay, areal efficiency and energy efficiency (Table 1) of a realistic crossbar arrangement (including parasitic elements), a SPICE model (Supplementary Fig. 4a) for the 90 nm device was developed (Supplementary Section 5). One can conclude that extremely fast readout transitions can suppress shielding in the SL, since charge cannot be supplied any longer (silicide lines are a critical resistive path). In the table, the energetically worst-case scenario was assumed: all the WLs are activated at once and all the weights are zero with a resulting shielding effect, which, in turn, would lead to charging in the top gate oxide. Table 1 summarizes the minimum period of time for different matrix sizes, which is proportional to the RC delay, with R being the resistance and C the capacitance. The areal efficiency A_η in TOPS mm^{-2} can be derived from the memory footprint (2×8 F^2), assuming differential weights and the earlier mentioned time delay. The active (W_p) and reactive (W_r) energy per cell for 142 periods is also summarized in Table 1. With this estimate in mind, we can conclude a minimum energy efficiency η_{rec} of 3,452.6 TOPS W^{-1} in the worst-case scenario for 0% input signal sparsity and 100% weight sparsity and an energy recovery of 95% (Supplementary Section 5). Without any charge recovery, the energy efficiency η would amount to 198.5 TOPS W^{-1}. In a realistic neural network scenario, for example, a one-layer perceptron trained

on the Modified National Institute of Standards and Technology (MNIST) database, the energy efficiency is 29,600 TOPS W^{-1} including charge recovery (Supplementary Section 6). Without recovery, the efficiency amounts to 1,702 TOPS W^{-1} for MNIST.

Comparison of simulation and experimental results

To verify the functionality of the simulator, we performed simulations of the device with 60 μm gate length (Fig. 2). As shown in Fig. 5f, experimental data from Fig. 2d match well with the simulated data.

As shown in Supplementary Fig. 14, we measured the gate charging current together with the applied readout a.c. voltage for the single device (Fig. 2), and a perfect 90° phase shift is visible. From the curves, we can calculate the reactive (W_R) power consumption per period (using equations 31–33, Supplementary Section 5) and obtain $W_r = 3.22$ nJ per period. Furthermore, for 142 periods, as in the simulation, we obtain the total reactive energy for one MAC operation, namely, $W_{r,tot} = 457$ nJ per cell. If we scale this value by seven orders of magnitude, we obtain $W_{r,scaled} = 45.7$ fJ per cell (capacitance shown in Fig. 2d is seven orders of magnitude lower compared with the capacitance of the simulated 90 nm device shown in Fig. 5b).

This value is approximately ten times higher than the value shown in Table 1 (5 fJ per cell). One has to consider that the thickness of the buried oxide of the experimental devices is much thicker (190 nm) than in the case of the 90 nm device simulation (15 nm), leading to a 12.7 times lower readout capacitance/area at approximately the same gate oxide capacitance/area. Also considering the different device silicon thicknesses, one can obtain a corrected reactive energy of $W_{r,scaled,corr} = 5.84$ fJ cell, which is very close to the value shown in Table 1. Other influencing phenomena during scaling, like short channel effects (Fig. 5c), quantum confinement and band-to-band tunnelling, are explained in Supplementary Section 10.

Table 1 | Results on areal and energy efficiency obtained from SPICE simulation

Array size	Period T_{per} (total time delay (142 periods))	A_η (TOPS mm^{-2}) (142 periods + 16 F^2)	W_r (fJ) per cell* (W_p (fJ) per cell)*	η_{rec} (TOPS W^{-1})* (η (TOPS W^{-1}))*
100 × 100	1.00 ns (142.00 ns)	108.70 TOPS mm^{-2}	5.000 fJ per cell (0.015 fJ per cell)	3,782.20 TOPS W^{-1} (199.51 TOPS W^{-1})
500 × 500	15.00 ns (2.13 μs)	7.25 TOPS mm^{-2}	5.000 fJ per cell (0.022 fJ per cell)	3,676.80 TOPS W^{-1} (199.19 TOPS W^{-1})
1,000 × 1,000	30.00 ns (4.25 μs)	3.62 TOPS mm^{-2}	5.000 fJ per cell (0.040 fJ per cell)	3,452.60 TOPS W^{-1} (198.54 TOPS W^{-1})
2,500 × 2,500	200.00 ns (28.40 μs)	0.54 TOPS mm^{-2}	5.000 fJ per cell (0.039 fJ per cell)	3,461.70 TOPS W^{-1} (198.59 TOPS W^{-1})

*All cells are erased (worst-case scenario), 95% energy efficiency of power clock source Necessary time period T_{per} and resulting areal efficiency A_η for different matrix sizes. The reactive energy during the readout of arrays, W_r, and active energy, W_p, are obtained from simulations (Supplementary Section 5). The energy is presented per cell and for 142 periods. From this number and assuming a 95% energy recovery of the power source, energy efficiency η_{rec} (in TOPS W^{-1}) can be calculated for the energetically worst-case scenario (erased state). The same applies for energy efficiency η without recovery.

Conclusions

We have reported a memcapacitive device with the potential to deliver high tera-operations per second per watt when scaled. By using a shielding layer between two electrodes, we can achieve high dynamic ratios of ~1,480 for microscale devices and ~90 for simulated 90-nm-sized devices. Furthermore, a 5 × 5 image recognition task was implemented using an experimental crossbar array with 156 memory cells. Circuit-level simulations and noise-level calculations show that our memcapacitive devices can potentially offer superior energy efficiency compared with conventional resistive devices. Using adiabatic charging, most of the charging energy of the capacitors can be recovered. This allows a combination of reversible computing and neuromorphic computing. The energy efficiency of the human brain is estimated to be in the range of ~10 fJ per operation (ref. [46]) (or 100 TOPS W^{-1}), which is similar to current memristive-device-based approaches[13,16]. Our approach could potentially offer an energy efficiency of 1,000–10,000 TOPS W^{-1}. The technology is also compatible with complementary metal–oxide–semiconductor technology and could be fabricated using state-of-the-art processes.

Methods

The technology computer-aided design (TCAD) simulations were performed with Synopsys and SPICE-level simulations were performed with LTspice. In the TCAD simulations, the drift-diffusion equations (electron + hole continuity equation and Poisson equation) were included. Furthermore, Shockley–Read–Hall recombination and electric-field-, temperature- and dopant-dependent mobility models were included. The influence of quantum confinement and band-to-band tunnelling was investigated in Supplementary Section 10.

The devices were fabricated using a silicon-on-insulator wafer with an n⁺-handle, 3.5-μm-thick epitaxial layer; a 190-nm-thick buried oxide layer; and an 88-nm-thick device layer. First, alignment marks were etched into the device layer, followed by boron- and phosphorous-ion implantation and subsequent activation annealing. The interface oxide was chemically grown by Standard Clean 1 solution and O$_2$ oxidation at 750 °C. The Hf$_{0.5}$Zr$_{0.5}$O$_2$ deposition with a TiN capping layer was carried out by atomic layer deposition and annealed at 600 °C. The Hf$_{0.5}$Zr$_{0.5}$O$_2$ was patterned for contact holes and the first aluminium metallization was deposited by sputtering. The SLs were etched by ion beam sputtering and the BLs were separated by the reactive-ion etching of 7-μm-deep trenches. The trenches were refilled by SU-8 resist and the second metallization layer (WLs) were insulated from the first metallization layer by another patterned SU-8 layer.

Measurements were carried out with a function generator (Agilent 33500B), a lock-in amplifier (Stanford Research Systems SR830) and a current pre-amplifier (Stanford Research Systems SR570). A DSO5052A oscilloscope was used for visualizing the measured currents.

The PCB for the neuromorphic chip was designed using EAGLE and manufactured by Eurocircuits GmbH. A data acquisition system (USB-6363, National Instruments) was used for controlling the PCB. The measurement routines were written in LabVIEW. Python was used for simulating the Manhattan algorithm and Keras for MNIST simulation.

Data availability

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

Code availability

The code that supports the findings of this study is available from the corresponding authors upon reasonable request.

References

1. Mead, C. Neuromorphic electronic systems. *Proc. IEEE* **78**, 1629–1636 (1990).
2. Mead, C. How we created neuromorphic engineering. *Nat. Electron.* **3**, 434–435 (2020).
3. Strukov, D. B., Snider, G. S., Stewart, D. R. & Williams, R. S. The missing memristor found. *Nature* **453**, 80–83 (2008).
4. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNET classification with deep convolutional neural networks. *Adv. Neural Inf. Process Syst.* **25**, 1097–1105 (2012).
5. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
6. Bayat, F. M. et al. Implementation of multilayer perceptron network with highly uniform passive memristive crossbar circuits. *Nat. Commun.* **9**, 2331 (2018).
7. Cai, F. et al. A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations. *Nat. Electron.* **2**, 290–299 (2019).
8. Prezioso, M. et al. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* **521**, 61–64 (2015).
9. Burr, G. W. et al. Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element. *IEEE Trans. Electron Devices* **62**, 3498–3507 (2015).
10. Borders, W. A. et al. Analogue spin–orbit torque device for artificial-neural-network-based associative memory operation. *Appl. Phys. Express* **10**, 013007 (2017).
11. Grollier, J. et al. Neuromorphic spintronics. *Nat. Electron.* **3**, 360–370 (2020).
12. Garcia, V. & Bibes, M. Ferroelectric tunnel junctions for information storage and processing. *Nat. Commun.* **5**, 4289 (2014).
13. Berdan, R. et al. Low-power linear computation using nonlinear ferroelectric tunnel junction memristors. *Nat. Electron.* **3**, 259–266 (2020).
14. Jerry, M. et al. Ferroelectric FET analog synapse for acceleration of deep neural network training. In *2017 IEEE International Electron Devices Meeting (IEDM)* **6**, 6.2.1–6.2.4 (IEEE, 2018).
15. Mulaosmanovic, H. et al. Novel ferroelectric FET based synapse for neuromorphic systems. In *2017 Symposium on VLSI Technology* T176–T177 (IEEE, 2017).
16. Agrawal, V. et al. In-memory computing array using 40nm multibit SONOS achieving 100 TOPS/W energy efficiency for deep neural network edge inference accelerators. In *2020 IEEE International Memory Workshop (IMW)* 1–4 (IEEE, 2020).
17. Tsai, H., Ambrogio, S., Narayanan, P., Shelby, R. M. & Burr, G. W. Recent progress in analog memory-based accelerators for deep learning. *J. Phys. D* **51**, 283001 (2018).
18. Di Ventra, M., Pershin, Y. V. & Chua, L. O. Circuit elements with memory: memristors, memcapacitors, and meminductors. *Proc. IEEE* **97**, 1717–1724 (2009).
19. Martinez-Rincon, J., Di Ventra, M. & Pershin, Y. V. Solid-state memcapacitive system with negative and diverging capacitance. *Phys. Rev. B* **81**, 195430 (2010).

20. Mohamed, M. G. A., Kim, H. & Cho, T. W. Modeling of memristive and memcapacitive behaviors in metal-oxide junctions. *Sci. World J.* **2015**, 910126 (2015).

21. Pershin, Y. V. & Di Ventra, M. Memcapacitive neural networks. *Electron. Lett.* **50**, 141–143 (2014).

22. Khan, A. K. & Lee, B. H. Monolayer MoS_2 metal insulator transition based memcapacitor modeling with extension to a ternary device. *AIP Adv.* **6**, 095022 (2016).

23. Wang, Z. et al. Capacitive neural network with neuro-transistors. *Nat. Commun.* **9**, 3208 (2018).

24. Kwon, D. & Chung, I. Y. Capacitive neural network using charge-stored memory cells for pattern recognition applications. *IEEE Electron Device Lett.* **41**, 493–496 (2020).

25. You, T. et al. An energy-efficient, $BiFeO_3$-coated capacitive switch with integrated memory and demodulation functions. *Adv. Electron. Mater.* **2**, 1500352 (2016).

26. Zheng, Q. et al. Artificial neural network based on doped HfO_2 ferroelectric capacitors with multilevel characteristics. *IEEE Electron Device Lett.* **40**, 1309–1312 (2019).

27. Emara, A. A. M., Aboudina, M. M. & Fahmy, H. A. H. Non-volatile low-power crossbar memcapacitor-based memory. *Microelectr. J.* **64**, 39–44 (2017).

28. Yurchuk, E. et al. Charge-trapping phenomena in HfO_2-based FeFET-type nonvolatile memories. *IEEE Trans. Electron Devices* **63**, 3501–3507 (2016).

29. Ji, H., Wei, Y., Zhang, X. & Jiang, R. Improvement of charge injection using ferroelectric Si:HfO_2 as blocking layer in MONOS charge trapping memory. *IEEE J. Electron Devices Soc.* **6**, 121–125 (2018).

30. Zamanidoost, E., Bayat, F. M., Strukov, D. & Kataeva, I. Manhattan rule training for memristive crossbar circuit pattern classifiers. In *Proc. 2015 IEEE 9th International Symposium on Intelligent Signal Processing (WISP)* 1–6 (IEEE, 2015).

31. Zhao, M., Gao, B., Tang, J., Qian, H. & Wu, H. Reliability of analog resistive switching memory for neuromorphic computing. *Appl. Phys. Rev.* **7**, 011301 (2020).

32. Chang, Y. W. et al. A novel simple CBCM method free from charge injection-induced errors. *IEEE Electron Device Lett.* **25**, 262–264 (2004).

33. Forouhi, S., Dehghani, R. & Ghafar-Zadeh, E. Toward high throughput core-CBCM CMOS capacitive sensors for life science applications: a novel current-mode for high dynamic range circuitry. *Sensors* **18**, 3370 (2018).

34. Widdershoven, F. et al. A CMOS pixelated nanocapacitor biosensor platform for high-frequency impedance spectroscopy and imaging. *IEEE Trans. Biomed. Circuits Syst.* **12**, 1369–1382 (2018).

35. Nabovati, G., Ghafar-Zadeh, E., Letourneau, A. & Sawan, M. Towards high throughput cell growth screening: a new CMOS 8×8 biosensor array for life science applications. *IEEE Trans. Biomed. Circuits Syst.* **11**, 380–391 (2017).

36. Ciccarella, P., Carminati, M., Sampietro, M. & Ferrari, G. Multichannel 65 zF rms resolution CMOS monolithic capacitive sensor for counting single micrometer-sized airborne particles on chip. *IEEE J. Solid-State Circuits* **51**, 2545–2553 (2016).

37. Kern, T. Symmetric differential current sense amplifier. US patent 7,800,968 (2010).

38. Kadetotad, D. et al. Parallel architecture with resistive crosspoint array for dictionary learning acceleration. *IEEE Trans. Emerg. Sel. Topics Circuits Syst.* **5**, 194–204 (2015).

39. Athas, W. et al. The design and implementation of a low-power clock-powered microprocessor. *IEEE J. Solid-State Circuits* **35**, 1561–1570 (2000).

40. Yadav, R. K., Rana, A. K., Chauhan, S., Ranka, D. & Yadav, K. Adiabatic technique for energy efficient logic circuits design. In *2011 International Conference on Emerging Trends in Electrical and Computer Technology* 776–780 (IEEE, 2011).

41. Bennett, C. H. Logical reversibility of computation. *IBM J. Res. Dev.* **17**, 525–532 (1973).

42. Frank, M. P. The future of computing depends on making it reversible. In *IEEE Spectrum* **25** (IEEE, 31 August 2017).

43. Ye, Y. & Roy, K. QSERL: quasi-static energy recovery logic. *IEEE J. Solid-State Circuits* **36**, 239–248 (2001).

44. Bhaaskaran, V. S. K. Energy recovery performance of quasi-adiabatic circuits using lower technology nodes. In *India International Conference on Power Electronics (IICPE2010)* 1–7 (IEEE, 2011).

45. Maksimović, D., Oklobdžija, V. G., Nikolić, B. & Current, K. W. Clocked CMOS adiabatic logic with integrated single-phase power-clock supply: experimental results. *High.-Perform. Syst. Des. Circuits Log.* **8**, 255–259 (1999).

46. Xu, W., Min, S. Y., Hwang, H. & Lee, T. W. Organic core-sheath nanowire artificial synapses with femtojoule energy consumption. *Sci. Adv.* **2**, e1501326 (2016).

Author contributions

K.-U.D. performed the TCAD and SPICE simulations, device fabrication and measurement. A.K. contributed to the MNIST simulation and pulse scheme of the neuromorphic system. S.P. supervised the work. All the authors wrote 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/s41928-021-00649-y.

Correspondence and requests for materials should be addressed to Kai-Uwe Demasius, Aron Kirschen or Stuart Parkin.

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

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

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