
Supplementary information

Energy-efficient memcapacitor devices for neuromorphic computing

In the format provided by the
authors and unedited

Energy-Efficient Memcapacitor Devices for Neuromorphic Computing

Kai-Uwe Demasius¹, Aron Kirschen², Stuart Parkin¹

¹Max Planck Institute of Microstructure Physics, 06120 Halle (Saale), Germany

²SEMIRON GmbH, 01309, Dresden, Germany

[§]Correspondence to be addressed to: kudemasius@semron-ai.com; akirschen@semron-ai.com; Stuart.Parkin@mpi-halle.mpg.de

S1. Broadening of CV curves with applied V_{AK}

In Supplementary Fig. 1 the half peak broadness of the CV curves is plotted against the applied Anode-Cathode Voltage V_{AK} . The broadness increases with ~ 1.5 V for a 1 V increase of V_{AK} . One can deduce by integration of equation (2) in the main paper the electric field in the semiconductor and by multiplication with the dielectric constant the Charge Q_s with respect to the surface potential ψ_s :

$$Q_s = \pm \frac{2\epsilon_0\epsilon_r}{U_T} \cdot \sqrt{\frac{q}{2\epsilon_0\epsilon_r U_T}} \cdot \sqrt{p \cdot \left[\exp\left(-\frac{\psi_s}{U_T}\right) + \frac{\psi_s}{U_T} - 1 \right] + n \cdot \left[\exp\left(\frac{\psi_s}{U_T}\right) - \frac{\psi_s}{U_T} - 1 \right]} \quad (Supp 1)$$

Since the concentration of holes p and electrons n is brought out of equilibrium during application of V_{AK} one can conclude for the concentrations:

$$p = n_i \cdot \exp\left(\frac{\varphi_i - \varphi_p}{U_T}\right) \quad (Supp 2)$$

$$n = n_i \cdot \exp\left(\frac{\varphi_n - \varphi_i}{U_T}\right) \quad (Supp 3)$$

With φ_i being the intrinsic Fermi potential and φ_p the quasi-Fermi potentials of holes and φ_n of electrons. For a semiconductor in equilibrium the quasi-Fermi potentials are equal: $\varphi_p = \varphi_n$. In case of a pn-junction with applied voltage V_{AK} the difference amounts to:

$$V_{AK} = \varphi_n - \varphi_p \quad (\text{Supp 4})$$

Since the voltages on cathode and anode V_K and V_A are antisymmetric one can conclude:

$$\varphi_n = -\frac{V_{AK}}{2} \quad \varphi_p = \frac{V_{AK}}{2} \quad (\text{Supp 5})$$

So, equation (Supp 1) becomes:

$$Q_s = \pm \frac{2\varepsilon_0\varepsilon_r}{U_T} \cdot \sqrt{\frac{q}{2\varepsilon_0\varepsilon_r U_T}} \cdot \sqrt{n_i \cdot \exp\left(\frac{\varphi_i - \frac{V_{AK}}{2}}{U_T}\right) \cdot \left[\exp\left(-\frac{\psi_s}{U_T}\right) + \frac{\psi_s}{U_T} - 1\right] + n_i \cdot \exp\left(\frac{-\frac{V_{AK}}{2} - \varphi_i}{U_T}\right) \cdot \left[\exp\left(\frac{\psi_s}{U_T}\right) - \frac{\psi_s}{U_T} - 1\right]} \quad (\text{Supp 6})$$

The charge will rapidly increase during the strong inversion/accumulation regime, which takes place when the exponential terms in the bracket start to dominate:

$$\psi_{s,-} < \varphi_i - \frac{V_{AK}}{2} \quad (\text{Supp 7})$$

$$\psi_{s,+} > \varphi_i + \frac{V_{AK}}{2} \quad (\text{Supp 8})$$

So, the onset of strong inversion/accumulation is shifted linearly with respect to the applied V_{AK} according to equation Supp 7 and 8, as in the experimental (and simulation-wise) observation.

The curves in Fig. 2b of the main paper are derivatives of a sigmoid curves, which play an important role in modeling neurons in artificial neural networks, and furthermore, adjusting their broadness by application of reverse or forward voltages provides additional functionality: A common practice during training is the dropout of certain neurons to avoid overfitting¹, which can be implemented

though the forward direction of the diode. Furthermore, the application of the voltage V_{AK} can also enable selection and deselection of memory cells, if necessary.

The actual sigmoid function can be obtained by applying a DC bias of 3 V and increasing the AC amplitude (0.1 V to 6 V) to the gate electrode (n+ and p+ are connected to ground in this case) and the observed AC current on the read-out electrode saturates. The obtained sigmoid curve is shown in Supplementary Fig. 14.

S2. Scaled Device with 45 nm Gate length/width

In order to preserve a high on/off (dynamic) ratio and a high read-out capacitance, high- k dielectrics are necessary for the top and bottom insulator. Introducing high- k dielectrics into the memory dielectric will reduce the EOT and thus the short channel effect. Introduction of high- k dielectric for the bottom insulator will increase the read-out capacitance. Supplementary Fig. 2a shows the simulated device structure and the thicknesses of the layer. Hafniumoxide was used as a high- k dielectric. The obtained CV curve is revealed in Supplementary Fig. 2b and a capacitive on/off ratio of 1:60 is achieved (3V voltage shift). Furthermore, the read-out capacitance has approximately the same value as the 90 nm Device from the main paper, which ensures the same detectability. Regarding matrix level simulation some data is provided in Supplementary Section S5 for 45 nm.

S3. Sensitivity calculation of the sense amplifier

A typical charge integrating amplifier can achieve up to 50 mV output for 1 μ A input current within 3 ns or a sensitivity S of 16.67 mV/fC:²

$$S = 16.67 \frac{mV}{fC} \quad (Supp\ 9)$$

Assuming 100 distinct values of the digitized output value (7 Bit) a voltage of ± 1.5 V must be achieved on the sense amplifier output (30 mV steps by Schmitt-Trigger)³:

$$V_{max/min} = \pm \frac{1}{2} \cdot 30 \text{ mV} \cdot 100 = \pm 1.5 \text{ V} \quad (\text{Supp 10})$$

This output amplifier voltage corresponds according to the sensitivity (equation Supp 9) to a charge of:

$$Q_{max/min} = \frac{V_{max/min}}{S} = \pm 90 \text{ fC} \quad (\text{Supp 11})$$

In neuromorphic systems many memory cells (several 100 to 1000) are read out at once and the accumulated charge is relevant for further processing. According to Fig. 5f in the main paper the accumulated charge over one half-period is $Q_{acc,max} = 6.32 \text{ aC}$. Assuming the read-out of $N_{dev} = 100$ devices at once and accumulation over $N_{per} = \pm 142$ periods (180° phase shift for the negative sign), one obtains:

$$Q_{max/min} = Q_{acc,max} \cdot N_{dev} \cdot N_{per} = 89.7 \text{ fC} \quad (\text{Supp 12})$$

This value fits well with the one obtained in equation Supp 11. Hence for 100 memory cells one needs to apply 142 periods to achieve sufficient accumulated charge. The value of $N_{per} = \pm 142$ periods fits well into a 8 bit input precision.

For less than 100 memory cells (very sparse neural networks/convolutional layers) one might need to consider other detection schemes, like direct integration on a gate capacitance of a MOSFET and subsequent changed channel resistance, or direct integration on a capacitor which is sufficiently large to enable low impedance read-out.

S4. Noise consideration for maximum possible energy efficiency

The effective thermal noise voltage for resistive devices equals to:

$$V_{n,res} = \sqrt{4k_B T R \Delta f} \quad (\text{Supp 13})$$

With the resistance R and the bandwidth Δf . For the noise current one obtains similarly:

$$I_{n,res} = \sqrt{\frac{4k_B T \Delta f}{R}} \quad (Supp 14)$$

The measured read-out signal is:

$$I_{meas,res} = I_{s,eff} + I_{n,res} \quad (Supp 15)$$

A measurement procedure is very similar to Nyquist–Shannon sampling⁴ and thus concepts from that theory can be applied here to obtain the bandwidth Δf . A measurement (integration) over the measurement time T_{meas} would lead to the following charge:

$$Q_o = \int_{t_0}^{t_0+T_{meas}} I_{meas,res} dt \quad (Supp 16)$$

This integration can be expressed as a convolution with a rect-pulse with the duration T_{meas} :

$$Q_o = I_{meas,res} * \text{rect}\left(\frac{t}{T_{meas}}\right) = \int_{t_0}^{t_0+T_{meas}} I_{meas,res} * \text{rect}\left(\frac{t}{T_{meas}}\right) dt \quad (Supp 17)$$

The frequency spectrum of the output charge Q_o can be obtained by Fourier-transform:

$$\mathcal{F}(Q_o) = \mathcal{F}\left(I_{meas,res} * \text{rect}\left(\frac{t}{T_{meas}}\right)\right) = \mathcal{F}(I_{meas,res}) \cdot \mathcal{F}\left(\text{rect}\left(\frac{t}{T_{meas}}\right)\right) \quad (Supp 18)$$

The Fourier transform turns the convolution into a simple multiplication and the Fourier transform of a rect-pulse gives the following function:

$$\mathcal{F}\left(\text{rect}\left(\frac{t}{T_{meas}}\right)\right) = T_{meas} \cdot \text{sinc}(f \cdot T_{meas}) \quad (Supp 19)$$

The sinc function is plotted in Supplementary Fig. 3b, with the applied rect pulse for the measurement in Supplementary Fig. 3a. Obviously the frequency spectrum of the $I_{meas,res}$ signal will get a band limitation, which is approximately the same as the first zero point of the sinc function in Supplementary Fig. 3b:

$$\Delta f \approx \frac{1}{T_{meas}} \quad (Supp\ 20)$$

So in the end the white thermal noise of equation Supp (14) becomes:

$$I_{n,res} = \sqrt{\frac{4k_B T}{RT_{meas}}} \quad (Supp\ 21)$$

In order to achieve a certain number of levels (B bits) the maximum measurement signal needs to have at least the following effective value:

$$I_{s,eff} = I_{n,res} \cdot 2^B \quad (Supp\ 22)$$

The noise level dictates the minimum step height that is possible. From equation Supp (21) and (22) one can conclude for the Energy of one MAC operation (over one integration cycle T_{meas}):

$$E = R \cdot I_{s,eff}^2 \cdot T_{meas} = 4k_B T \cdot 2^{2 \cdot B} \quad (Supp\ 23)$$

The maximum energy efficiency can be calculated by the inverse and considering, that one MAC operation consists out of two operations: one multiplication and one addition. It is 1842 TOPS/W for $B = 8\ bit$ and 29472 TOPS/W for $B = 6\ bit$ (at room temperature).

In the case of shot noise which many resistive devices have (MOSFET in subthreshold regime, Memristors and tunnel junctions) one has to consider the following noise level:

$$I_{n,shot} = \sqrt{2qI_{s,eff}\Delta f} \quad (Supp\ 24)$$

This current depends on the actual current flowing through the device. With equation Supp (22), (24) and (20) one obtains:

$$\frac{I_{s,eff}}{2^B} = \sqrt{2qI_{s,eff}\Delta f} \quad (Supp\ 25)$$

$$I_{s,eff} = 2q\Delta f 2^{2\cdot B} = \frac{2q2^{2\cdot B}}{T_{meas}} \quad (Supp\ 26)$$

This leads to the following energy efficiency per MAC:

$$E = I_{s,eff} \cdot U_{s,eff} \cdot T_{meas} = 2q \cdot U_{s,eff} \cdot 2^{2\cdot B} \quad (Supp\ 27)$$

with the operating voltage $U_{s,eff}$ of the device. Assuming $U_{s,eff} = 0.35V$ one obtains 269.4 TOPS/W for 8 bit or 4310 TOPS/W for 6 bit. So, the energy efficiency is significantly lower with shot noise compared to thermal noise.

For a capacitive load the kTC noise level is described by equation (3) in the main paper and thus the necessary voltage level is:

$$V_{s,eff} = V_{n,res} \cdot 2^B = 2^B \cdot \sqrt{\frac{k_B T}{C}} \quad (Supp\ 28)$$

For the stored energy per MAC on the capacitor we obtain:

$$E = \frac{1}{2} C V_{s,eff}^2 = \frac{1}{2} \cdot k_B T \cdot 2^{2B} \quad (Supp\ 29)$$

which is 8 times less than the thermal noise of a resistor and much lower compared to the shot noise limit. It results in 14736 TOPS/W for 8 bit and 235774 TOPS/W for 6 bit (at room

temperature). Most of the energy stored on the capacitor can be recovered as explained in the main text, which gives in principle even higher energy efficiencies.

S5. Spice Simulations

The Spice model of the device can be found in Supplementary Fig. 4a. For the gate leakage and buried oxide leakage resistances a surface current of 10^8 A/cm² at a voltage of 2V was assumed. The non-linear capacitances C_{n+} and C_{p+} contain models for the inversion and accumulation capacitances. As can be seen from Supplementary Fig. 4b the gate capacitance and read-out electrode capacitance fits well with the Synopsys simulation (90 nm device). Supplementary Fig. 4c only shows the coupling capacitance to the read-out electrode in comparison to Synopsys (90 nm device). Furthermore, it was confirmed with AC simulation in Synopsys that resistive losses within the device are very small in the investigated frequency range and can be neglected (the resistive component is 346 times lower than the capacitive component at 1 GHz).

The Spice model was arranged in the crossbar vector along the shielding line (Supplementary Fig. 4d). The parasitic elements of the wordline are negligible compared to the shielding line and bitline, due to the much lower resistivity of copper interconnects and the dominating gate oxide capacitance in the time delay and energy consumption.

For this purpose each single device model was connected at its n+ and p+ connector by the Ti₂Si, which is used in standard CMOS. The thickness was set to 28nm and the resistivity to 14μΩcm. The bitline resistivity was assumed to be 100μΩcm, corresponding to highly n-doped silicon, and the thickness to be 90 nm. The model does also include parasitic capacitances for the bitline, namely the capacitance within the buried oxide to the n+ and p+ region ($C_{\text{box,par}}=21$ aF) and the depletion capacitance to the p region surrounding the BL ($C_{\text{depl,par}}=97$ aF).

Supplementary Fig. 5a shows the accumulated charge during read-out, which is the integration of the BL current over one half period ($T_{\text{per}} = 30\text{ns}$) for a written memory cell (shielding line is

grounded during read-out). For very long bitlines the charge cannot be supplied to the bitline sufficiently, which leads to a lower totally accumulated charge (2500 cells in a row in Supplementary Fig. 5a) and a lot of the charge gets lost on the parasitic capacitors. Also, for an erased memory cell the shielding charge cannot be supplied from the Ti_2Si for high frequency, thus there will be more capacitive coupling for long bitlines (Supplementary Fig. 5b, $T_{per} = 30\text{ns}$). Supplementary Fig. 5c shows the read-out charge for erased (dashed line) and written memory cells for different period length T_{per} . For 1000 cells in a row a period time of $T_{per} = 30\text{ns}$ is necessary, for 2500 cells the necessary time increases to $T_{per} = 200\text{ns}$ and for smaller arrays the time can be smaller. From the time delay the areal efficiency A_η can be calculated with an area of $8F^2$ for each memory cell (One weight consists out of two memory cells and one MAC out of two operations – multiplication and addition):

$$A_\eta = \frac{2}{2 \cdot 8F^2 \cdot T_{per} \cdot N_{per}} \quad (30)$$

Besides the time delay the simulation gives also results on the energy consumption of the crossbar. Since the circuit consists out of resistors and capacitors one needs to distinguish between the complex, reactive and active energy. The active energy (Supplementary Fig. 6a) increases for longer arrays and shorter T_{per} , indicating larger resistive losses. The decrease in on/off ratio is in direct correlation to the active energy. The active energy was calculated from the total current $i_{tot}(t)$ and applied voltage $v_{WL}(t)$ to the wordline:

$$W_p = \int_{t+T_{per}}^{T_{per}} i_{tot}(t) \cdot v_{WL}(t) dt \quad (\text{Supp 31})$$

Supplementary Fig. 6b shows the reactive energy over one period versus the T_{per} for different number of cells. The reactive component per memory cell approximately stays the same no matter how many cells there are in the array or how long the period length T_{per} is. The state (written or

erased) of the memory cell has an influence on the reactive power, due to the larger Gate capacitance of an erased cell (see Supplementary Fig. 4b). The total amount of complex energy was calculated from the effective values of AC input currents and voltages:

$$W_s = \sqrt{\int_{t+T_{per}}^{T_{per}} i_{tot}(t)^2 dt \cdot \int_{t+T_{per}}^{T_{per}} v_{WL}(t)^2 dt} \quad (Supp 32)$$

The reactive energy can be calculated from the complex and active energy:

$$W_r = \sqrt{W_s^2 - W_p^2} \quad (Supp 33)$$

Assuming a maximum gate capacitance of 44.8aF (Supplementary Fig. 4b) one can estimate the AC current per memory cell for an applied $U_{ac} = \frac{0.5V}{\sqrt{2}} = 0.35V$:

$$I_{ac} = U_{ac} \cdot 2\pi f \cdot C \quad (Supp 34)$$

This leads to the following reactive energy, which matches very well with the simulated values in Supplementary Tab. 1:

$$W_r = I_{ac} U_{ac} \cdot T_{per} = U_{ac}^2 \cdot 2\pi \cdot C = 35.19 \text{ aJ} \quad (Supp 35)$$

For calculating the energy/MAC one needs to consider that a total number of $N_{per} = 142$ periods is applied and that each weight includes two memory cells (differential topology). In this calculation the energy recovery of the harmonic power source is not included:

$$E_{MAC} = \frac{W_s \cdot N_{per}}{N_{cells}} \cdot 2 \quad (Supp 36)$$

The energy efficiency in TOPS/W can be calculated from equation Supp 36 by considering that each MAC operation consists out of two operation (one multiplication and one addition):

$$\eta = \frac{2}{E_{MAC}} \cdot \left(10^{-12} \frac{OP}{TOP} \right) \quad (Supp 37)$$

The energy recovery of the AC power source depends on its quality factor q and dictates how much of the reactive power needs to be supplied, in order to maintain the oscillation:

$$E_{MAC,rec} = 2 \cdot \frac{W_r \cdot N_{per}}{N_{cells} \cdot q} + 2 \cdot \frac{W_p \cdot N_{per}}{N_{cells}} \quad (Supp\ 38)$$

Supplementary Fig. 6c summarizes the η without energy recovery and Fig. 6d η_{rec} with energy recovery (assuming $q = 20$)⁵⁻⁷ for different number of memory cells and period length. For shorter period length and longer bitlines the η_{rec} decreases, due to larger resistive losses, which cannot be recovered. For 1000 cells and $T_{per} = 30\ ns$ a total energy efficiency of 3452.6 TOPS/W for the erased state and 7468.1 TOPS/W for the written state can be achieved with energy recovery. Without energy recovery the values are 198.5 TOPS/W for the erased state and 400.3 TOPS/W for the written state.

In addition to the SPICE simulations on the 90 nm device, simulations on the 45 nm device were also performed (Supplementary Fig. 7). Supplementary Fig. 7a shows the comparison between the Spice model and the Synopsys simulation for the 45 nm device. The capacitances are very similar to the 90 nm device, due to the use of high-k dielectrics. Similarly, the time delay for the written state is also very similar to the 90 nm device for different array sizes. Due to the fact that the gate oxide capacitance has a similar value to the 90 nm device, the total reactive energy (equation Supp 35) and, thus, the energy efficiency of the array is expected to be similar to the 90 nm case.

S6. MNIST Simulation

The case from Supplementary section S5 assumes that all wordlines are activated with 142 periods, and thus the worst-case scenario. In a real neural net, not all inputs are activated at the maximum, and one needs to consider the input and weight sparsity. For this purpose, a MNIST simulation of a simple one layer perceptron was conducted.

Supplementary Fig. 8a shows the weights being implemented on the 748x20 crossbar (90 nm device) for each classification. Supplementary Fig. 8b shows the applied numbers to the crossbar. Since many pixels in Fig. 8b are zero, the average sparsity is 86.5%, and improves the energy efficiency. Fig. 8c-d shows the output of each classification for the applied numbers 0, 2 and 4. The SPICE simulated results fit well with the directly calculated ones. An average energy efficiency of $\eta_{rec} = 29600.13 \text{ TOPS/W}$ and $\eta = 1702.25 \text{ TOPS/W}$ was achieved with and without energy recovery.

S7. Pulse scheme consideration

For (mem)-capacitive (as for memristive) devices, one has to consider the potential difference at the crossing points of SL and WL. If *Manhattan update rule* is implemented, the weight update is given by:

$$\Delta w_{ij} = \text{sign}(-\alpha \delta_i X_j)$$

where α is the learning rate, δ_i the (backpropagated) error of the post-synaptical neuron and X_j the applied output from the j^{th} pre-synaptical neuron. Thus, following the multiplication of error and input, a XNOR operation has to be carried out with subtraction⁸. The question is, how to encode positive and negative state of error and input such that XNOR is obtained. Assuming constant signals for $X_j^\uparrow$, $X_j^\downarrow$, $\delta_i^\uparrow$, $\delta_i^\downarrow$ which denote the positive and negative signals for input and error respectively, one can easily see, that a XNOR operation cannot be achieved by adding (I)+(IV) and (II)+(III):

$$(I) \quad X_j^\uparrow - \delta_i^\uparrow \geq a$$

$$(II) \quad X_j^\uparrow - \delta_i^\downarrow < a$$

$$(III) \quad X_j^\downarrow - \delta_i^\uparrow < a$$

$$(IV) \quad X_j^\downarrow - \delta_i^\downarrow \geq a$$

where a is a threshold depending on the implementation (zero in its simplest form).

In the following, it is shown, that each signal $X_j^\uparrow, X_j^\downarrow, \delta_i^\uparrow, \delta_i^\downarrow$ must contain at least 4 phases, denoted by $d_1^\uparrow, d_2^\uparrow, \dots$ for positive error signals $\delta_i^\uparrow$; $t_1^\uparrow, t_2^\uparrow, \dots$ for positive input signals $X_j^\uparrow$ etc. Since we fix the crossing point (i,j), the indices denote signal phases from now. Main conditions results from the consideration, that at least at one phase $1, \dots, n$, the requirement of exceeding the threshold a is satisfied, while side conditions ensure that the threshold for having a negative result b is not exceeded in another phase of the same signal (again, in its simplest form $a=b=0$):

$$\begin{aligned}
 & \text{side conditions} \\
 \text{i.} \quad & \left((t_1^\uparrow - d_1^\uparrow > a) \vee \dots \vee (t_n^\uparrow - d_n^\uparrow > a) \right) \wedge (t_1^\uparrow - d_1^\uparrow \geq b) \wedge \dots \wedge (t_n^\uparrow - d_n^\uparrow \geq b) \\
 \text{ii.} \quad & \left((t_1^\downarrow - d_1^\downarrow > a) \vee \dots \vee (t_n^\downarrow - d_n^\downarrow > a) \right) \wedge (t_1^\downarrow - d_1^\downarrow \geq b) \wedge \dots \wedge (t_n^\downarrow - d_n^\downarrow \geq b) \\
 \text{iii.} \quad & \left((t_1^\uparrow - d_1^\downarrow < b) \vee \dots \vee (t_n^\uparrow - d_n^\downarrow < b) \right) \wedge (t_1^\uparrow - d_1^\downarrow \leq a) \wedge \dots \wedge (t_n^\uparrow - d_n^\downarrow \leq a) \\
 \text{iv.} \quad & \left((t_1^\downarrow - d_1^\uparrow < b) \vee \dots \vee (t_n^\downarrow - d_n^\uparrow < b) \right) \wedge (t_1^\downarrow - d_1^\uparrow \leq a) \wedge \dots \wedge (t_n^\downarrow - d_n^\uparrow \leq a)
 \end{aligned}$$

for a fixed phase k we further define $\varepsilon_k, \gamma_k, \beta_k, \alpha_k$ as follows:

- $(t_k^\uparrow - d_k^\uparrow = \varepsilon_k) \quad (\text{Supp 39})$
- $(t_k^\downarrow - d_k^\downarrow = \gamma_k) \quad (\text{Supp 40})$
- $(t_k^\uparrow - d_k^\downarrow = \beta_k) \quad (\text{Supp 41})$
- $(t_k^\downarrow - d_k^\uparrow = \alpha_k) \quad (\text{Supp 42})$

This leads to the fact, that only one main condition (Supp 39) ... (Supp 42) can be satisfied in one phase i :

$$\exists k: (t_k^\uparrow - d_k^\uparrow = \varepsilon_k > a)$$

assume also:

- $(t_k^\downarrow - d_k^\downarrow = \gamma_k > a)$

$$2a < \varepsilon_k + \gamma_k \equiv \beta_k + \alpha_k \leq 2a \quad \text{✗}$$

- $(t_k^\uparrow - d_k^\downarrow = \beta_k < b)$

$$\begin{aligned} \beta_k - \varepsilon_k &= \overbrace{(t_k^\uparrow - d_k^\downarrow)}^{< b} - \overbrace{(t_k^\uparrow - d_k^\uparrow)}^{> a} = (d_k^\uparrow - d_k^\downarrow) < b - a \\ \gamma_k - \alpha_k &= \overbrace{(t_k^\downarrow - d_k^\downarrow)}^{\geq b} - \overbrace{(t_k^\downarrow - d_k^\uparrow)}^{\leq a} = (d_k^\uparrow - d_k^\downarrow) \geq b - a \quad \text{✗} \end{aligned}$$

(side conditions)

- $(t_k^\downarrow - d_k^\uparrow = \alpha_k < b)$

$$\alpha_k - \varepsilon_k = (t_k^\downarrow - d_k^\uparrow) - (t_k^\uparrow - d_k^\uparrow) = (t_k^\downarrow - t_k^\uparrow) < b - a$$

$$\gamma_k - \beta_k = (t_k^\downarrow - d_k^\downarrow) - (t_k^\uparrow - d_k^\downarrow) = (t_k^\downarrow - t_k^\uparrow) \geq b - a \quad \text{✗}$$

all other cases follow based upon mathematical symmetry deliberations (or easily as above). As a result, $n \geq 4$. Since implementations with 4 phases are easily obtained, 4 phases are sufficient.

Also, it is easy to derive that

$$\forall k: \varepsilon_k + \gamma_k \equiv \beta_k + \alpha_k \quad (\text{Supp 43})$$

S8. Disturbance level consideration

In order to determine the minimal disturbance potential / voltage λ , the relationship (Supp 43) from S7 is recalled. For a phase k it is valid:

$$\forall k: \varepsilon_k + \gamma_k \equiv \beta_k + \alpha_k \quad (\text{Supp 44})$$

The value for ε_k is normalized to the value 1, so $\alpha \leq 0.999\dots$. With that assumption together with the main and side conditions in S7 we estimate λ , an arbitrary value for γ_k , which we define w.l.o.g. as positive value ($0 \leq \lambda \leq 1$):

$$1. \quad \gamma_k = \lambda$$

$$\rightarrow 0 \leq \beta_k, \alpha_k \leq \lambda \quad (\text{otherwise either } \beta_k \text{ or } \alpha_k > 1 \text{ in accordance with (Supp 44)})$$

$$\rightarrow 1 + \lambda = \beta_k + \alpha_k$$

$$\rightarrow 1 + \lambda \leq 2 \cdot \lambda \quad \rightarrow \lambda \geq 1 \quad \text{✗}$$

$$2. \quad \gamma_k = -\lambda, 0 \leq \alpha_k, \beta_k \leq \lambda$$

$$\rightarrow 1 - \lambda = \beta_k + \alpha_k$$

$$\rightarrow 1 - \lambda \leq 2 \cdot \lambda$$

$$\rightarrow \mathbf{1/3} \leq \lambda$$

$$3. \quad \gamma_k = -\lambda, \text{ w.l.o.g.: } -\lambda \leq \alpha_k \leq 0 \leq \beta_k \leq \lambda$$

$$\rightarrow 1 - \lambda = \beta_k + \alpha_k$$

$$\text{if } -\lambda \leq \beta_k + \alpha_k \leq 0: \quad -\lambda \leq 1 - \lambda \leq 0 \quad \text{✗}$$

$$\text{thus } 0 \leq \beta_k + \alpha_k \leq \lambda: \quad 0 \leq 1 - \lambda \leq \lambda$$

$$\rightarrow \lambda \geq 0.5 \quad (\text{not optimal})$$

Since all other cases lead to the same result (put $-\gamma_i$ for α_i etc.), the minimal disturbance voltage in Manhattan update rule is $1/3 \cdot V_{\text{set}}$ (write voltage). A specific implementation with 4 phases and maximum disturbance voltage of $1/3$ is given in Fig. 3a in the main paper.

To ensure minimum write disturb for partly selected cells, the change of the written and erased state of the memory for small (1.5 and 2 V; $1/3$ voltage is approx. 1.7 V) pulses in the opposite direction was tested (Supplementary Fig. 9). A lot of the drop in the memory window of the written state is also associated to the limited retention of our memory stack (see next section). For the Manhattan update rule, the disturbance is low enough. Using active depletion of the PIN Diode is another possibility to further deselect the memory cells, but is not considered further here.

S9. Other memory parameters: Device-to-Device variation, Retention and Endurance

In Supplementary Fig. 10 the memory parameters Device-to-Device variation (a), endurance (b) and retention (c) are shown. The Device-to-Device variation was measured for 10 devices in the pulse number programming scheme and was low enough for the Manhattan update training.

The CV curves for various numbers of writing/erasing cycles are shown in Supplementary Fig. 10b and the memory window first seems to decrease for 10^5 cycles and for 10^8 cycles the curves broaden significantly. We expect the endurance to be in the range of 10^5 cycles.

The retention for the different levels can still be improved, but was good enough for the proposed Manhattan update training (also due to the fact that the cells are written over and over again with each training cycle). We expect the retention to be significantly improved with deeper traps in the hafnium oxide or another nitride layer, as shown by other groups⁹.

S10. Influence of Quantum confinement and Band-to-Band tunnelling

Quantum confinement leads to lower inversion charge carrier concentration in the channel in MOSFETs. In Supplementary Fig. 13 we show the influence of Quantum confinement and Band-to-Band tunneling on the capacitive coupling curves. The influence of the Band-to-Band tunneling is nearly not visible. Quantum confinement does not change the read-out capacitance (Fig. 13a) much, but the Gate oxide capacitance (Fig. 13b) is lower, especially during electron/hole inversion, since at the interface to the Gate oxide, there are less states for the electrons, thus the channel is shifted slightly below the interface. The Quantum capacitance is added in series to the Gate oxide capacitance; thus, the total Gate capacitance is lower. However, the effect leads to only a $\sim 2\%$ change and is negligible (every experimental variation would be larger). Also, the energy efficiency would increase by $\sim 2\%$ because there is a lower Gate capacitance to be charged.

Band-to-Band tunneling has some influence on the reverse current of the pin diode, but no influence on the read-out energy efficiency, where the shielding layer is grounded.

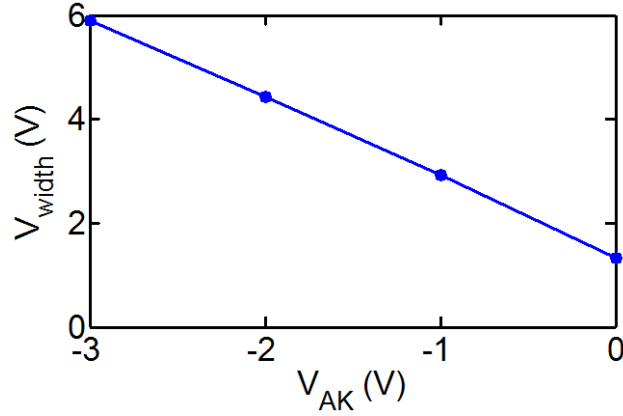
S11. Further explanation of switched capacitor approach

For the desired ‘four-quadrant multiplication’ (input $\times$ weight), a global clock signal is used together with the switched capacitor approach (Fig. 3c). The clock signal is defined as high for a rising edge in positive signals and as low for negative signals. For high clock signals the switches in Fig. 3c are in the left position, which means grounded positive BL and a connection between negative BL and the amplifier with an integration capacitance.

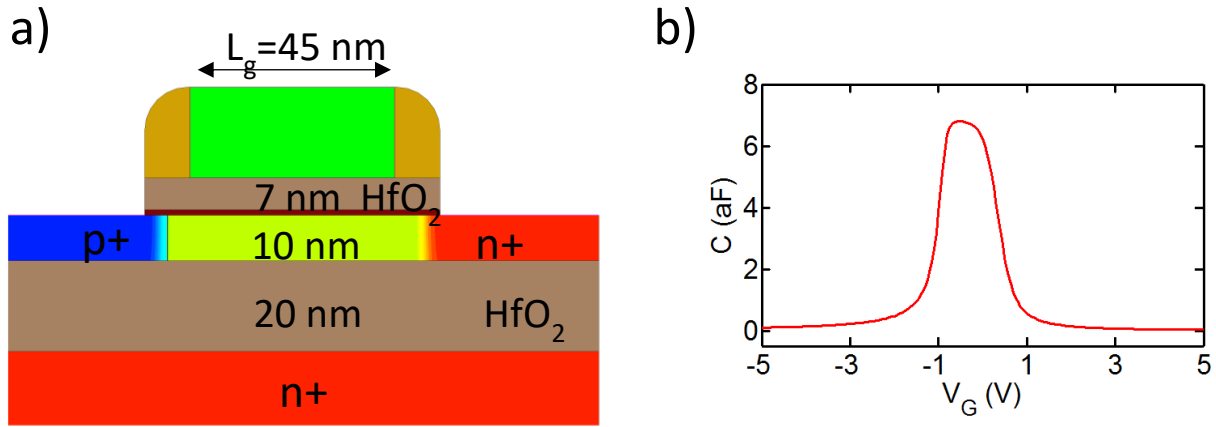
For the case of positive input signals and negative weights, the corresponding negative BL is positively charged and connected to the integration capacitance during the high clock signal. Due

to an inverting amplifier the resulting output voltage is negative. The positive BL is connected to the amplifier during the low clock signal (falling edge for positive input signal) and hence negatively charged, which turns into a positive voltage. The signs of the output voltages are inverted for a negative input signal (180 ° phase shift). With this switched capacitor approach one can subtract capacitances (C_{ij}^+ and C_{ij}^-) and achieve a ‘four-quadrant multiplication’ without a differential amplifier (with i indexing the post-synaptic neurons / BL and j the presynaptic neurons / WL).

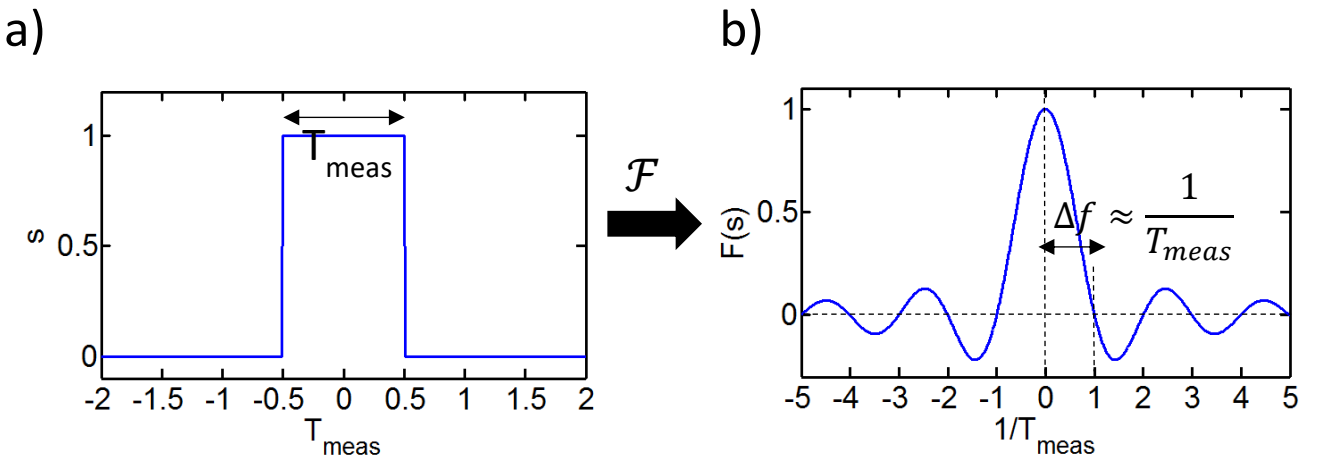
The integration capacitance of the amplifier is charged up in each period and hence the number of periods (N_{per}) encodes the value of the input signal.



Supplementary Fig. 1: Half width of capacitive coupling versus applied V_{AK} .

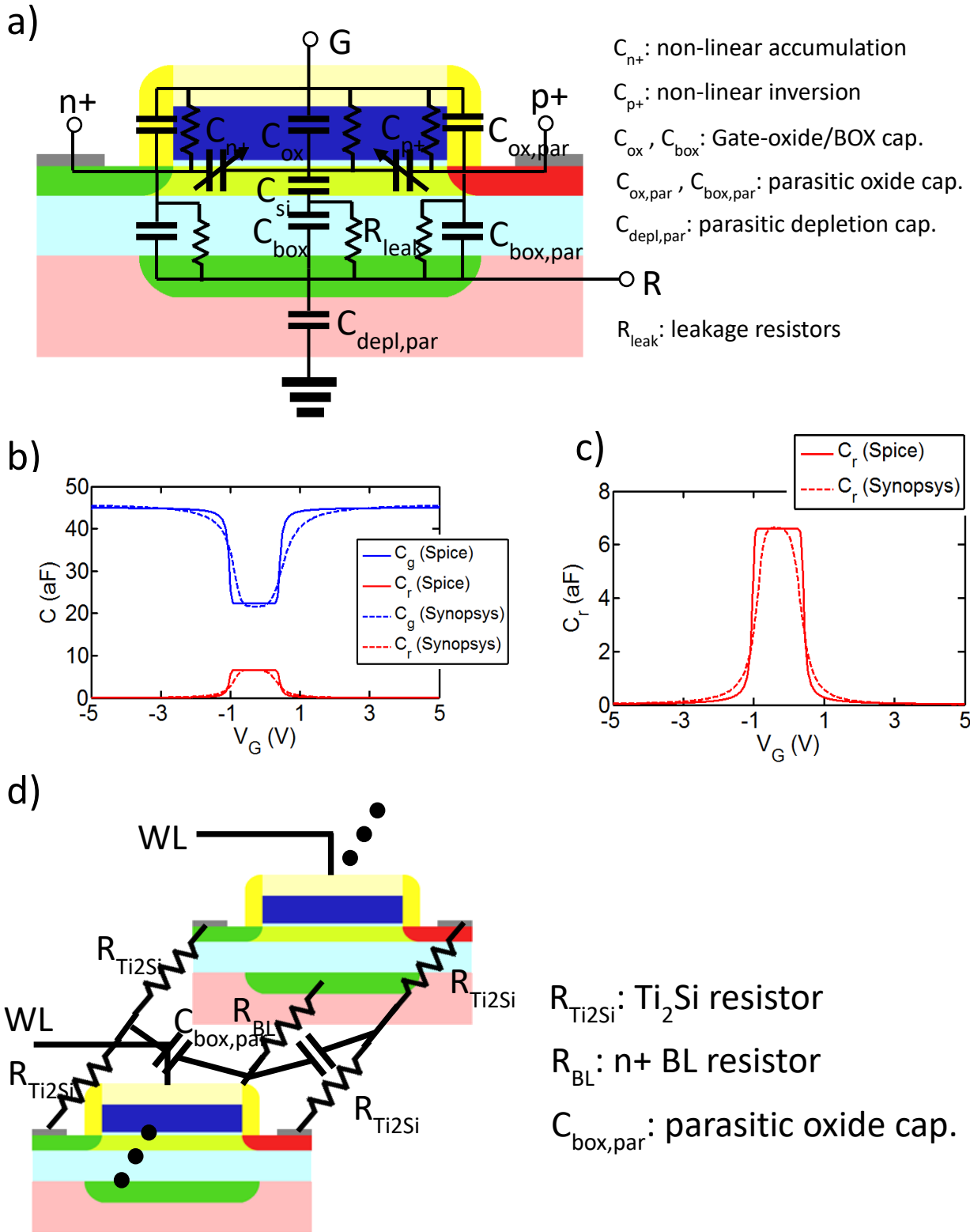


Supplementary Fig. 2: a) Simulated 45 nm Device b) Obtained CV curve.

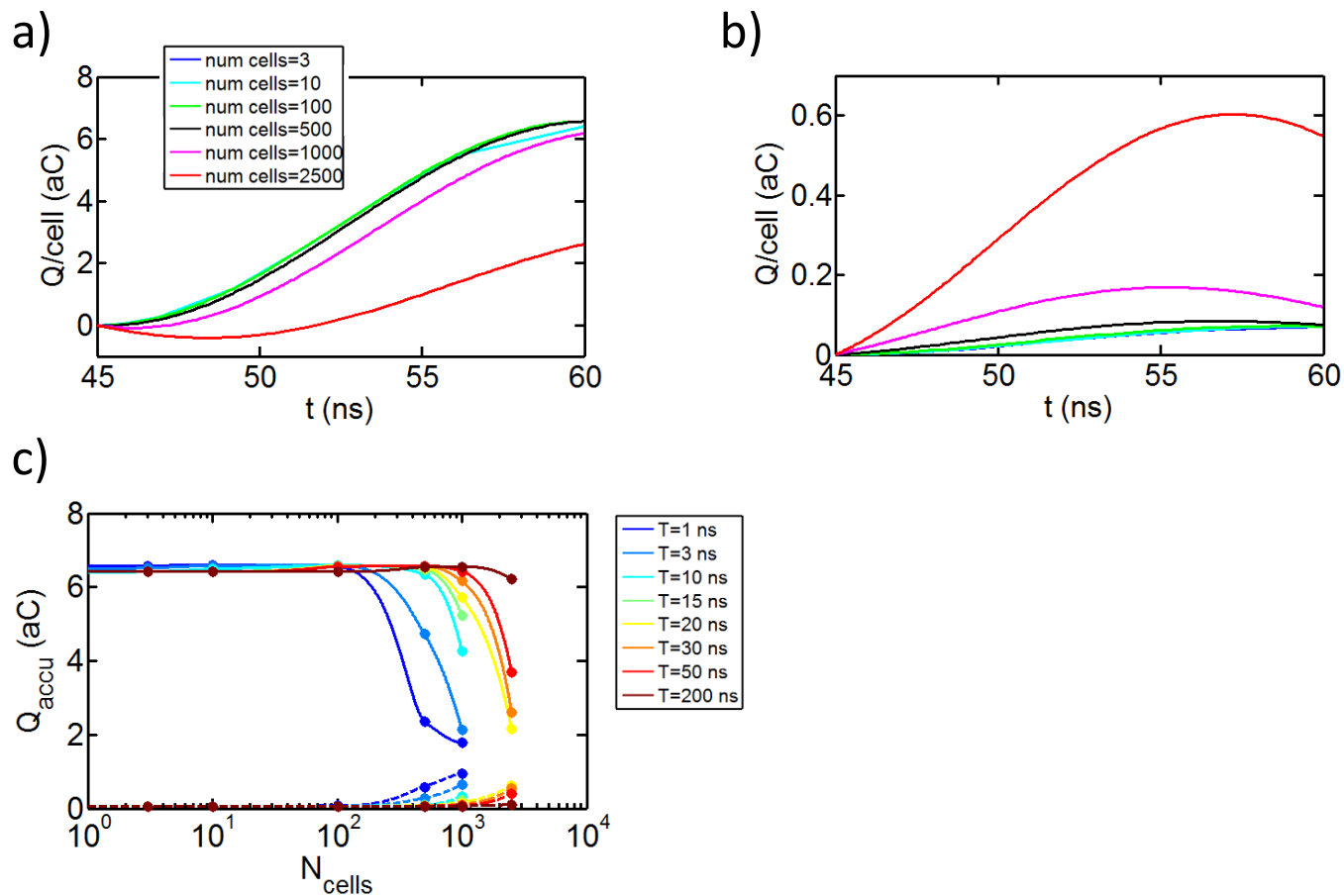


Supplementary Fig. 3: a) Read-pulse rect over which the input signal is integrated. b)

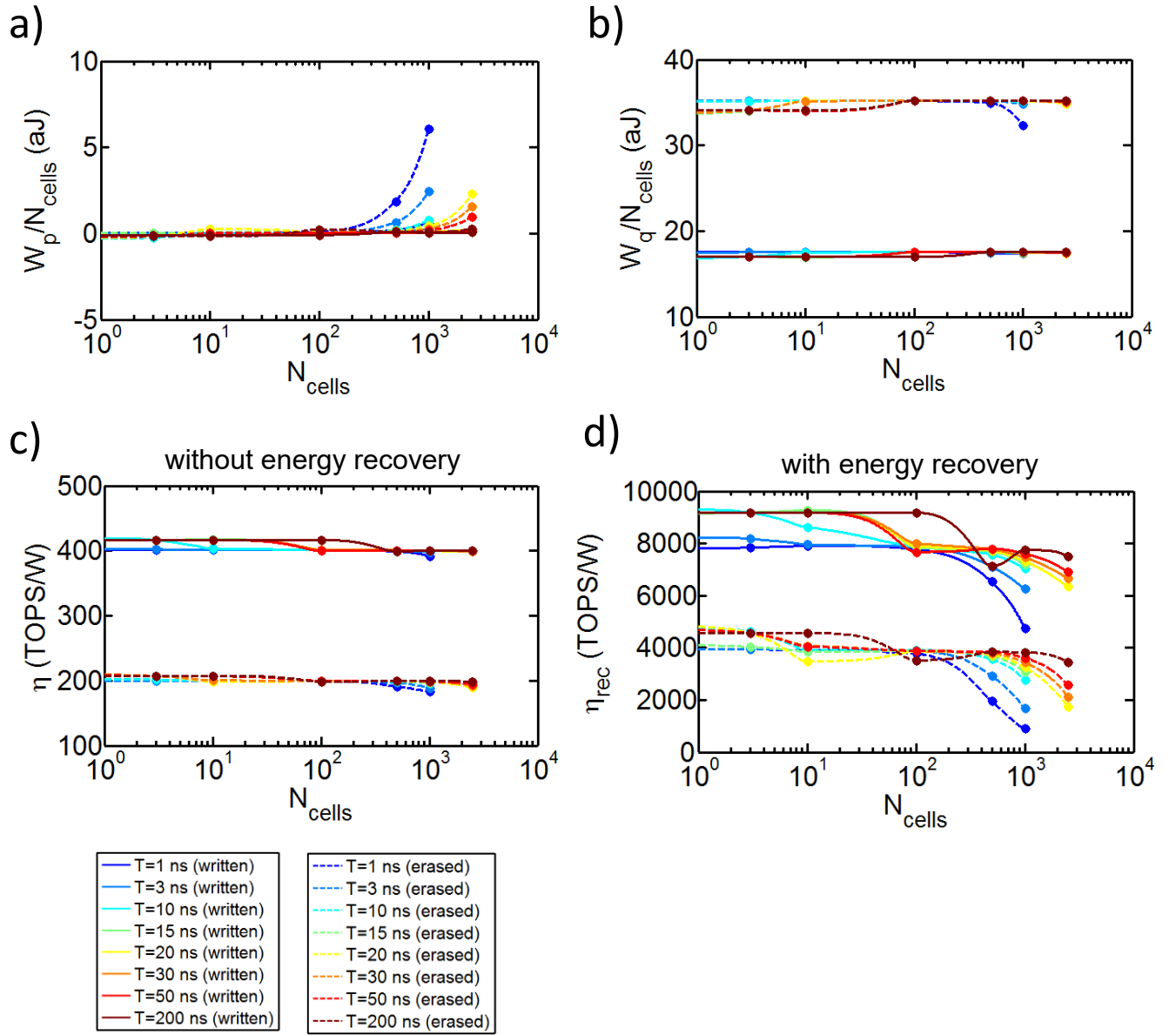
Corresponding Fourier transform of read-rect.



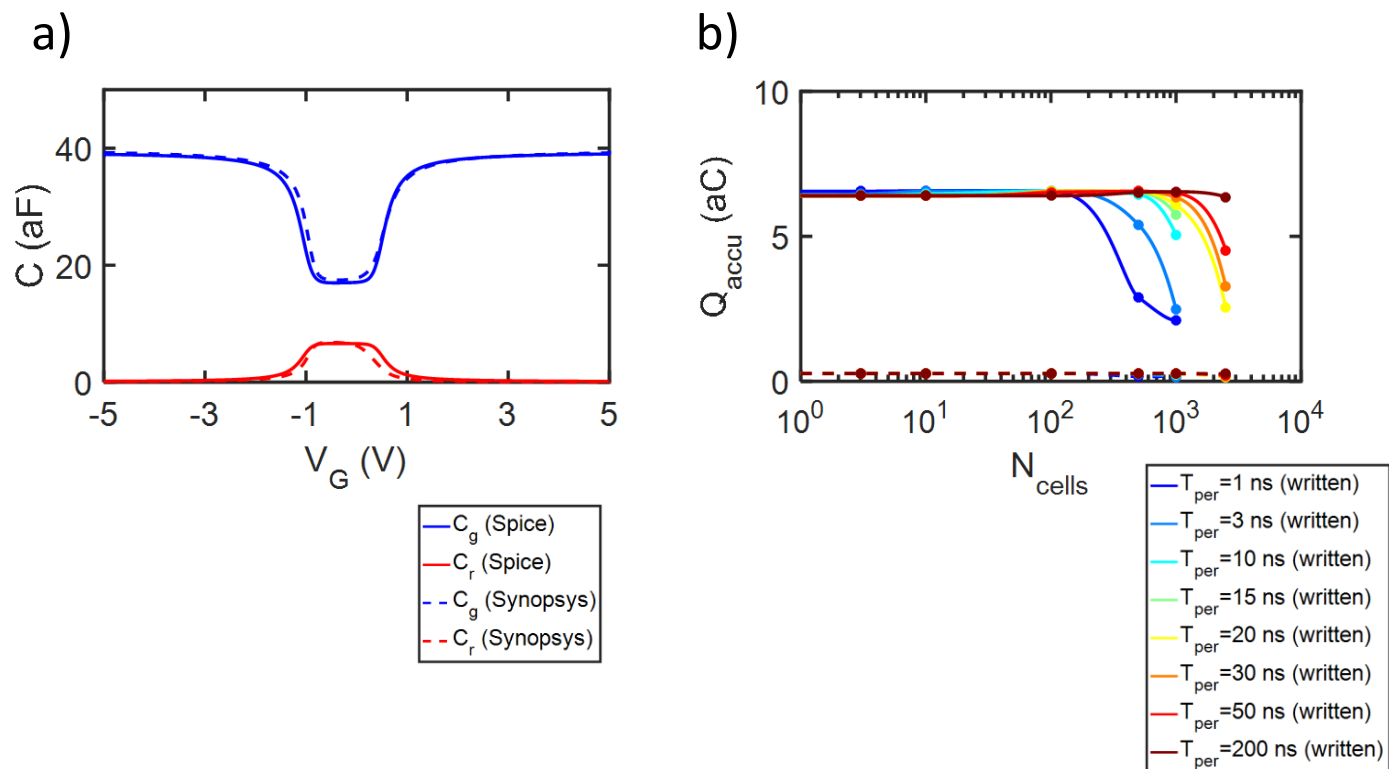
Supplementary Fig. 4: a) Spice model of the device. b) and c) Comparison of the SPICE model with the TCAD simulation (90 nm) d) Arrangement of single device model into an array.



Supplementary Fig. 5 (90 nm): a) Accumulated charge for different array length and written state of all cells ($T=30$ ns). b) Accumulated charge for different array length and erased state of all cells ($T=30$ ns). c) Comparison of accumulated charge for different array length for the written and erased state. The degradation of the on/off ratio becomes obvious.

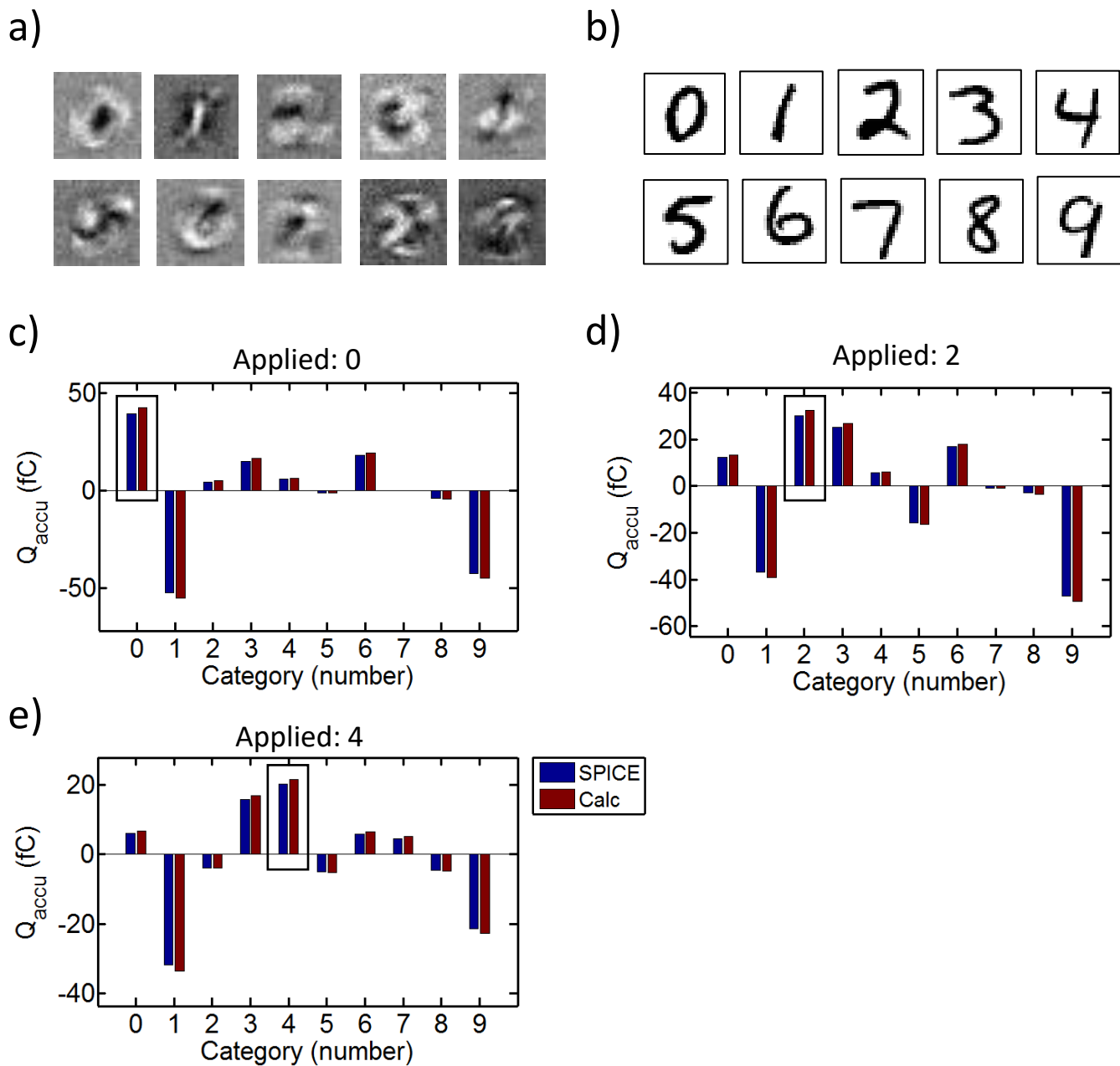


Supplementary Fig. 6 (90 nm): a) Active energy for different array sizes. b) Reactive energy for different array sizes. c) Energy efficiency without charge recovery. d) Energy efficiency with charge recovery.

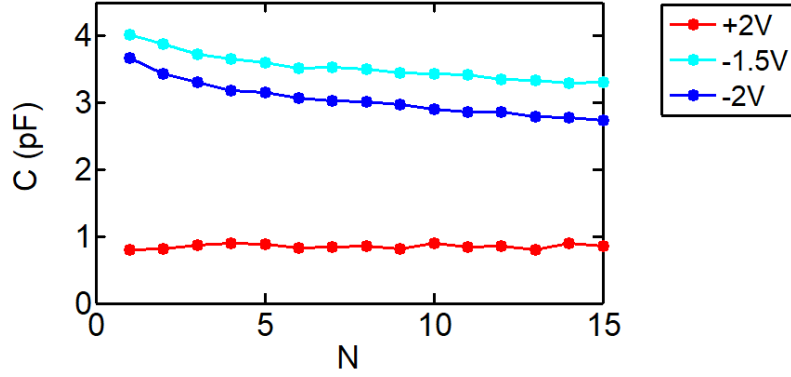


Supplementary Fig. 7 (45 nm): a) Comparison between Spice and TCAD simulated CV curves.

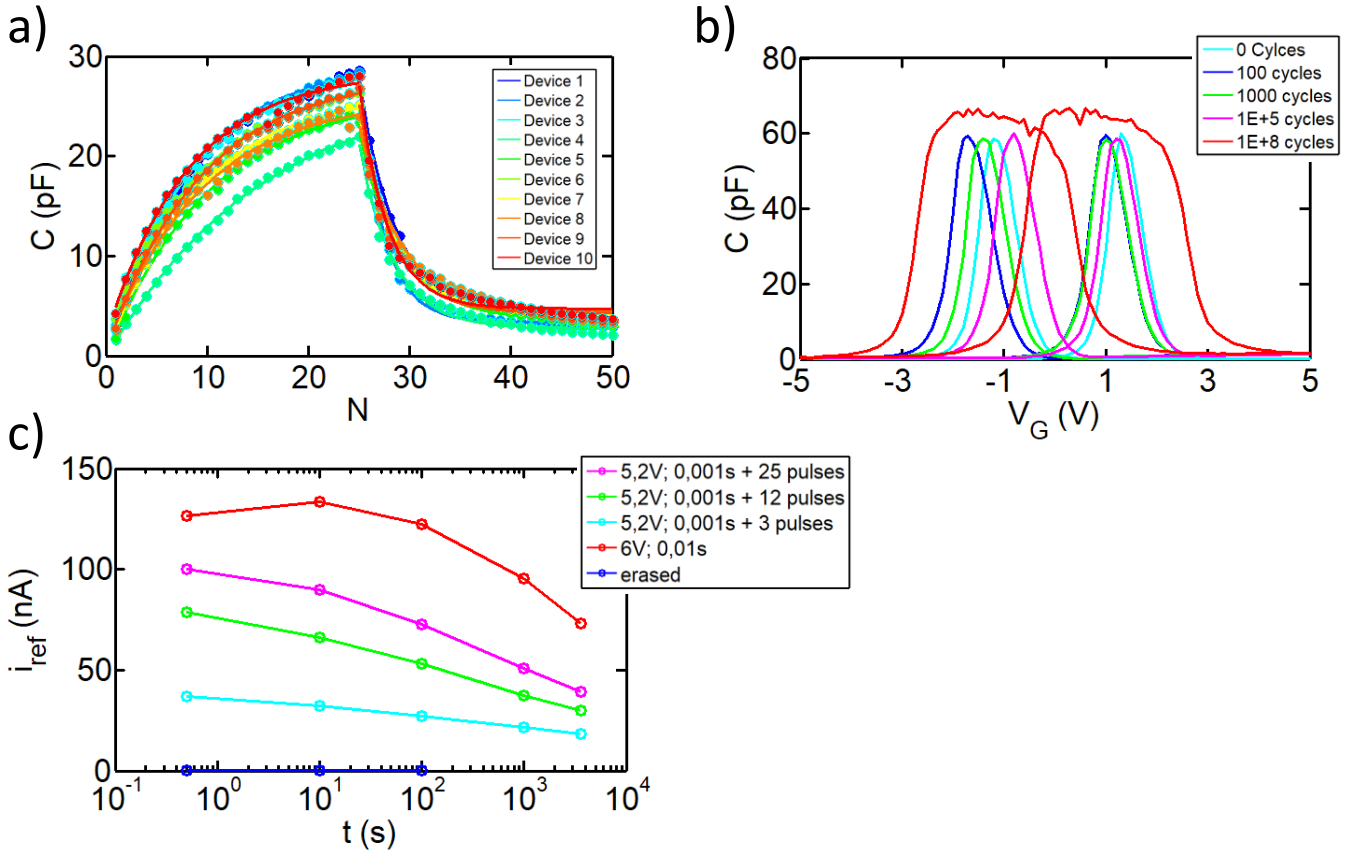
b) Comparison of accumulated charge for different array length for the written (solid line) and erased (dashed line) state. The curves for the written state are very similar to the 90 nm case.



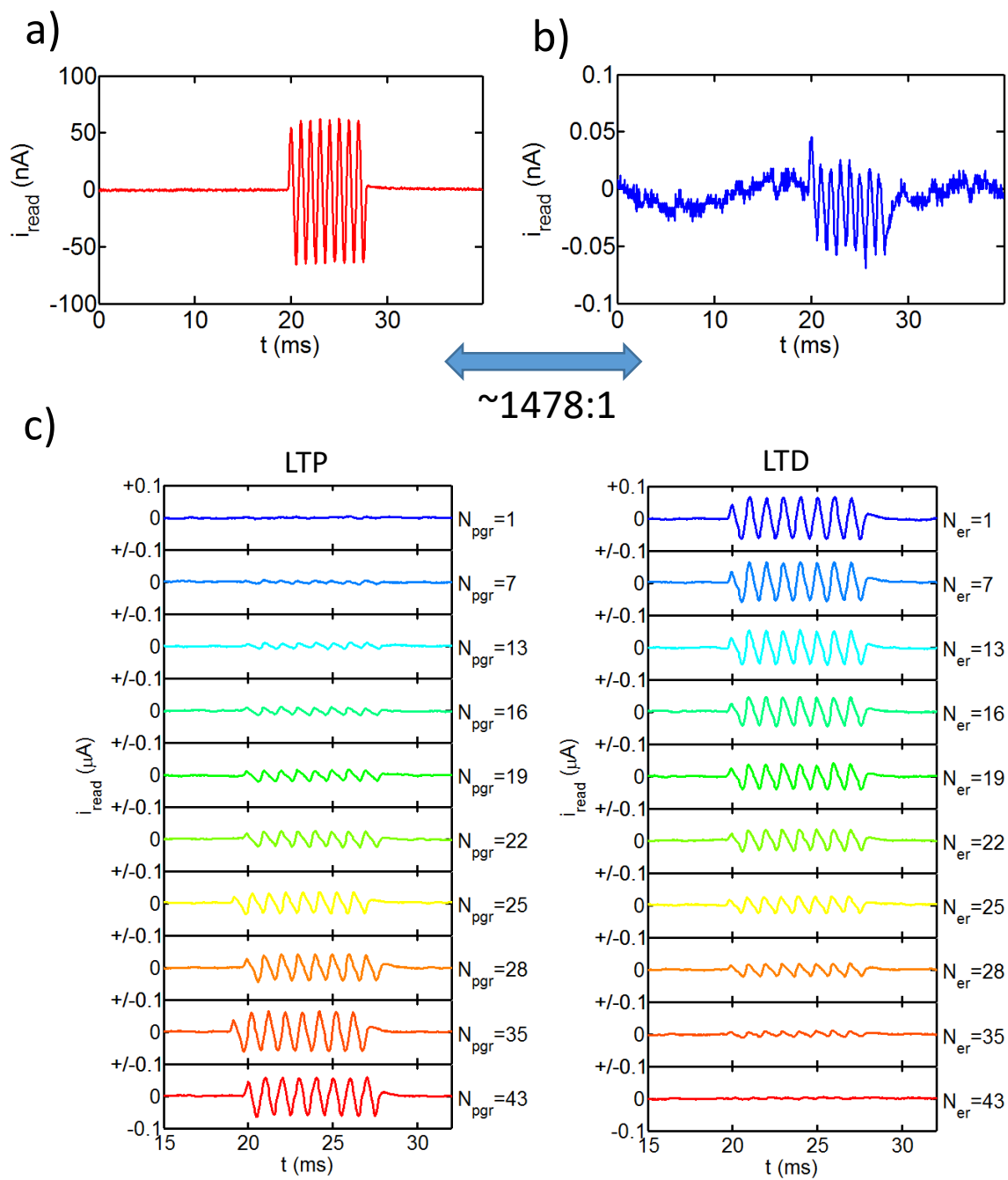
Supplementary Fig. 8: a) Implemented weight map for each classification of MNIST. b) Applied images to SPICE simulated array. c)-e) Output classification result for the numbers 0,2 and 4.



Supplementary Fig. 9: Disturb of written state (blue curves) and erased state (red curve) with a smaller partial select voltage in the 1/3 Matrix write scheme.

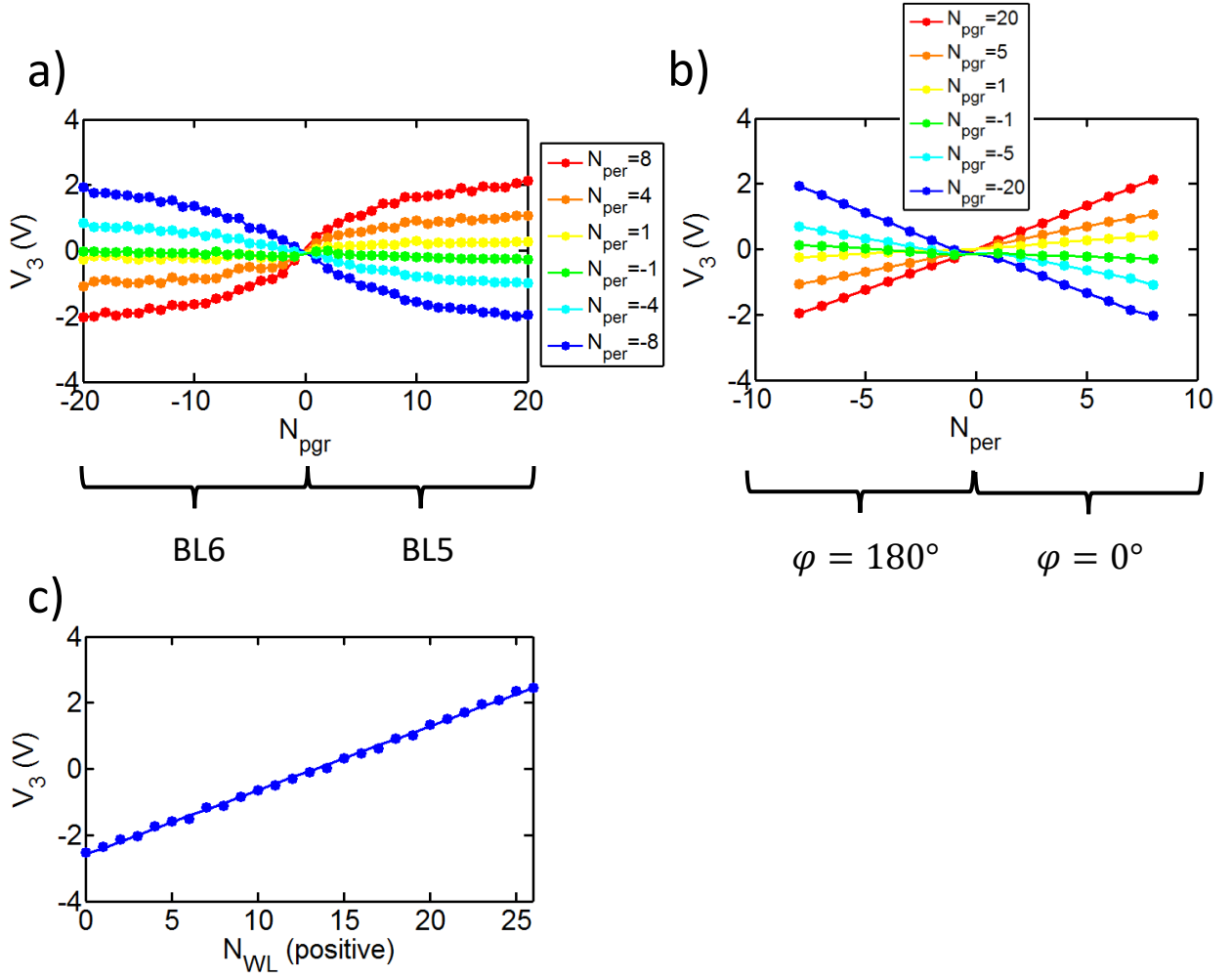


Supplementary Fig. 10: a) Device-to-Device variation for 10 devices in pulse number programming. b) CV curves after a certain number of write-erase cycles (endurance). c) Retention for different capacitive levels. Note the capacitance is proportional to the here plotted read-out ac current (i_{ref}).

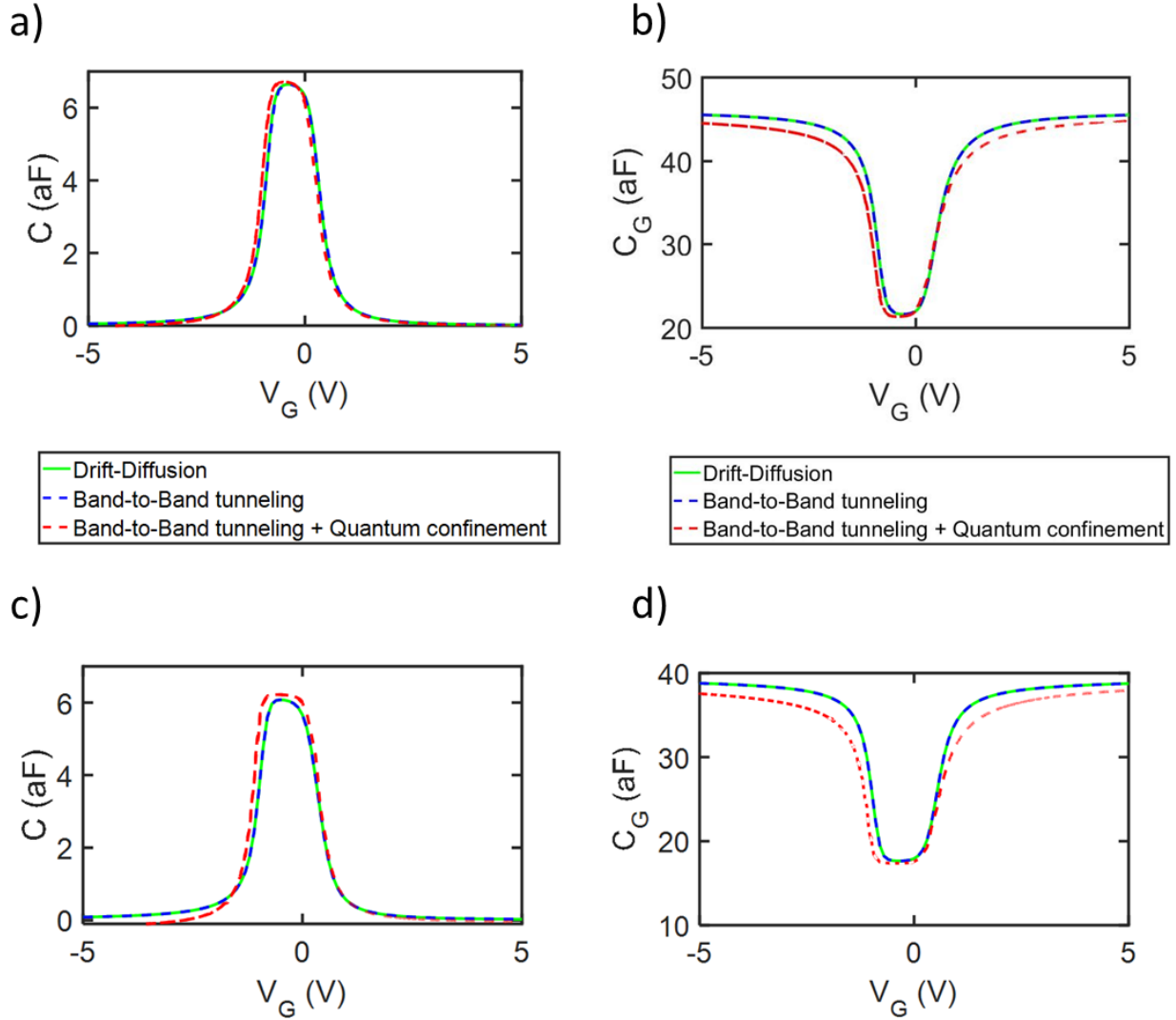


Supplementary Fig. 11: a) Measured read-out current for the written state and the erased state (b).

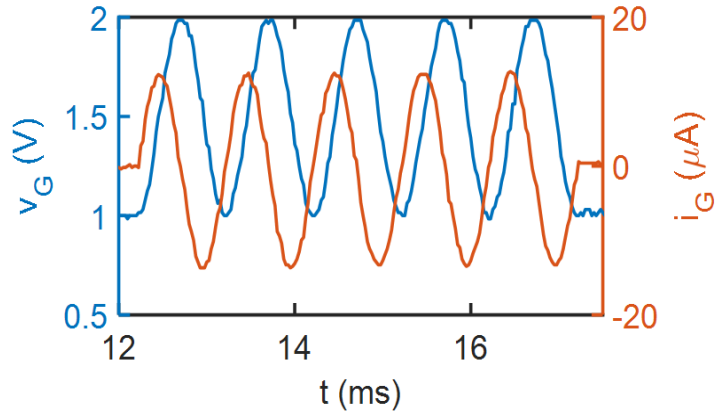
c) Measured read-out current for different LTP/LTD pulses for pulse height modulation.



Supplementary Fig. 12: a) Cross-section of 3D ‘four-quadrant-multiplication’ plot in Fig. 3d (main paper) along the programming pulse number. b) Cross-section along the input period number. c) Linear accumulation operation for different numbers of positive wordlines.



Supplementary Fig. 13: Influence of Band-to-Band tunneling and Quantum confinement for the 90 nm device on the a) read-out capacitance and b) Gate capacitance and for the 45 nm device on the c) read-out capacitance and d) Gate capacitance.



Supplementary Fig. 14: AC current to the Gate electrode and applied voltage for the 60 μm Gate length Device of Fig. 2a) in the main Paper. The energy efficiency is calculated from this plot.

1. Srivastava, N. & Geoffrey Hinton. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
2. Kern, T. Symmetric differential current sense amplifier. vol. 2 1–10 (2010).
3. Kademoti, D. *et al.* Parallel architecture with resistive crosspoint array for dictionary learning acceleration. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **5**, 194–204 (2015).
4. Shannon, C. E. Communication in the presence of noise. *Proc. IEEE* **86**, 447–457 (1998).
5. Ye, Y. & Roy, K. QSERL: Quasi-Static Energy Recovery Logic. *IEEE J. Solid-State Circuits* **36**, 239–248 (2001).
6. Kanchana Bhaaskaran, V. S. Energy recovery performance of quasi-adiabatic circuits using lower technology nodes. in *India International Conference on Power Electronics, IICPE 2010* vol. 1–7 (IEEE, 2011).
7. Maksimović, D., Oklobdžija, V. G., Nikolić, B. & Current, K. W. Clocked CMOS adiabatic logic with integrated single-phase power-clock supply: Experimental results. *High-Performance Syst. Des. Circuits Log.* **8**, 255–259 (1999).
8. Zamanidoost, E., Bayat, F. M., Strukov, D. & Kataeva, I. Manhattan rule training for memristive crossbar circuit pattern classifiers. in *Proc. IEEE* vol. 1–6 (IEEE, 2015).
9. Ji, H., Wei, Y., Zhang, X. & Jiang, R. Improvement of charge injection using ferroelectric Si:HfO₂ as blocking layer in MONOS charge trapping memory. *IEEE J. Electron Devices Soc.* **6**, 121–125 (2018).