Supplementary Materials

S1 Modeling of FeFET devices

The TEM micrograph of the FeFET co-integrated into the 28 nm technology node is given in Fig. S1**a** and the associated gate stack Fig. S1**b**. The $I_{ds} - V_{gs}$ transistor curves of the 2 states- the low V_{th} (LVT) and the high V_{th} is given in Fig. S1 **c**. The 1FeFET-1R structure and the layout of the segment are shown in Fig. S1**d** and **e** respectively. Fig. S1**f** shows the simulated crossbar structure with the accumulation capacitor and the ADC connected to it. Fig. S1 shows the overall structure of the FeFET and the associated connections for the proposed MAC macro.

The FeFET is modeled by separately modeling the ferroelectric capacitor (Fe-Cap) and the underlying transistor. The Fe-Cap is modeled using the Preisach model to capture the polarization (P) - Voltage (V) characteristics [17]. This macro-level model captures the electrical relationship between the applied voltage and polarization in the ferroelectric layer. An auxiliary voltage (V_{aux}) is calculated to capture the switching characteristics. V_{aux} represents the voltage to which the ferroelectric dipoles respond after relaxation. This is given using Eq. (1).

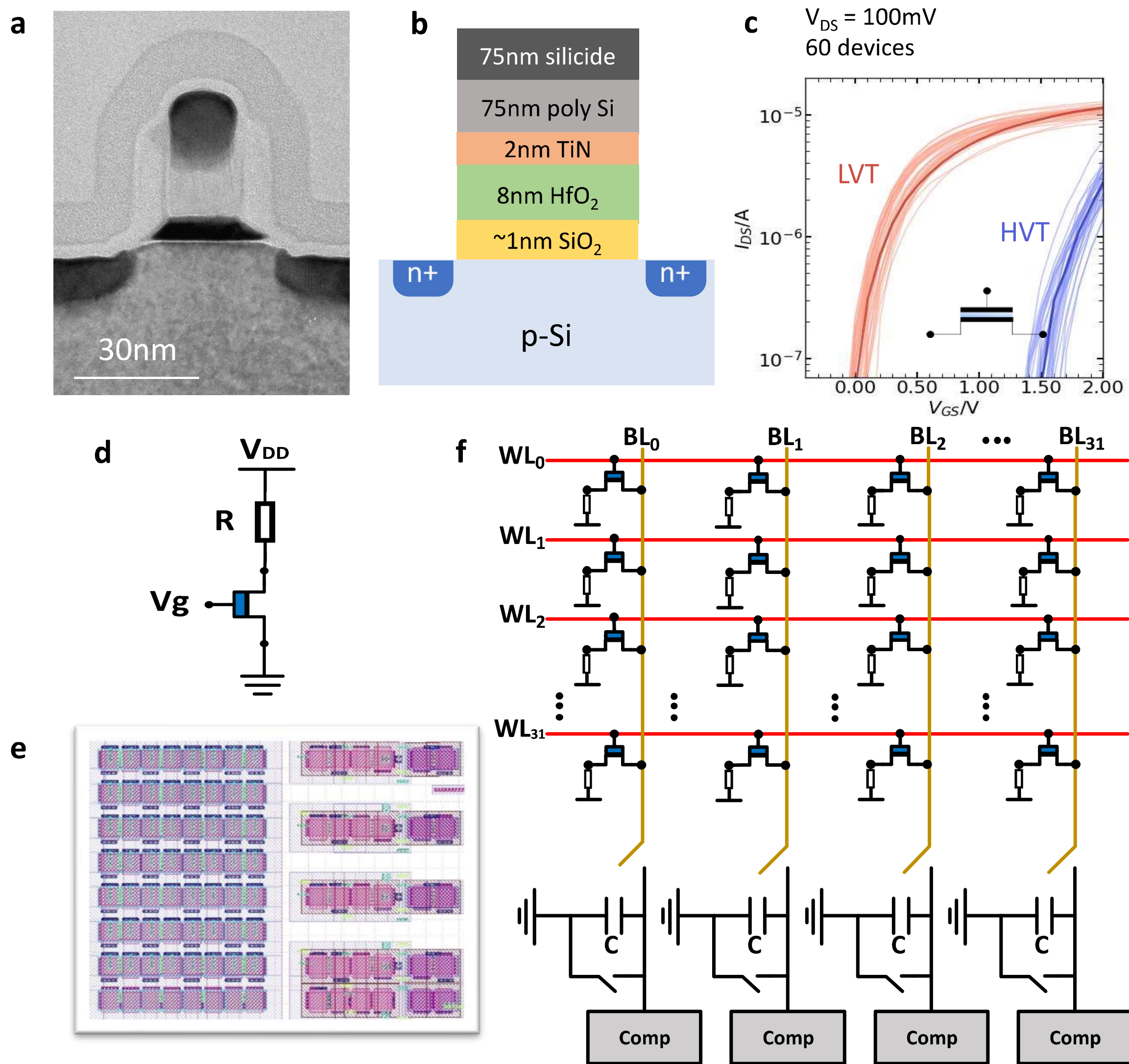

Figure S1: **Structure of the FeFET and proposed design. a.** TEM image of the FeFET integrated in 28 nm HKMG technology, **b.** material stack of theFeFETs, **c.** $I_{ds} - V_{gs}$ curves of 2 states (LVT/HVT) of 60 devices, **d.** 1FeFET-1R structure that forms the basis of our cell. **e.** Layout of the FeFET segment, and **f.** crossbar of the 32x32 cells with the accumulation capacitor.

$$V_{aux} = V_{in} - \tau_v \frac{d}{dt} V_{aux} \tag{1}$$

where τ_v represents the relaxation time for V_{aux} and V_{in} is the applied input voltage at the gate of the Fe-Cap. Then, the polarization corresponding to V_{aux} (P_{aux}) is calculated using:

$$P_{aux} = m \cdot P_s \cdot \tanh(w(V_{aux} \pm V_c)) + P_{off} \tag{2}$$

$$w = \frac{T_{fe}}{2V_c} \cdot \ln\left(\frac{P_s + P_r}{P_s - P_r}\right) \tag{3}$$

where V_c is the coercive voltage of the Ferroelectric material, P_s is the saturation polarization, P_r is the remnant polarization, m is the slope of the curve and P_{off} is the offset polarization. The upwards (-) and downwards (+) polarization determine the sign of tanh function, and the values of m and P_{off} are calculated using polarization history, which determines the P-V characteristics of the minor loops. For the main loop, m = 1 and P_{off} = 0. The calculated polarization is equated to the MOSFET gate charge and determines the region of operation. The industry-standard compact model BSIM-IMG [18] is used with calibrated parameters for the underlying MOSFET. Thus, the complete FeFET can be modelled as a Ferroelectric capacitor in series with the gate of MOSFET. Depending on the polarization in the Fe-Cap, the MOSFET is set into different V_{th}.

The shift in V_{th} of the FeFET is observed on applying different gate voltages to the

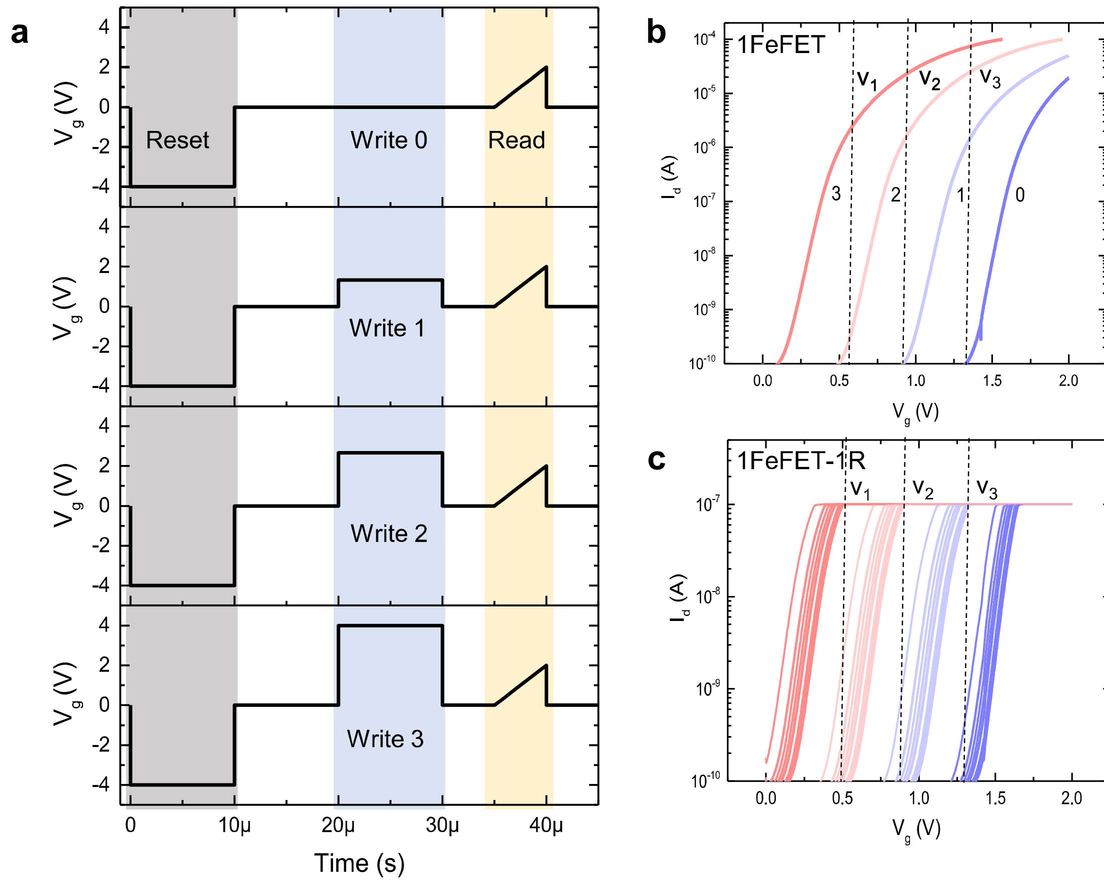

Figure S2: **Simulation results utilizing the proposed FeFET model. a.** The gate voltage applied at the FeFET to write it into 4 distinct states. Initially, a reset pulse is applied, followed by a write pulse of suitable magnitude and voltage sweep. **b.** I-V characteristics on sweeping the gate voltage (V_g) for the four different states of single FeFET during read. **c.** I-V characteristics of 1FeFET-1R for the four different states of single FeFET during read. The current is limited to 100 nA.

gate of the FeFET. Fig. S2**a** and **b** demonstrate the write and read functionality of the FeFET. The magnitude of the Write Voltage sets the different levels of polarization in the Ferroelectric layer according to Eq. (3), and the corresponding shift in V_{th} is observed in the $I_{ds} - V_{gs}$ characteristics.

To account for the variability due to the extrinsic sources of variation and the variation due to ferroelectric polarization, the V_{th} of the underlying MOSFET is varied. This can be varied using the V_{th} deviation parameter "DELVTRAND". The measured variation of V_{th} of mean standard deviation of 40 mV is used for this work. An external resistor connected to the FeFET cell (1FeFET-1R) can drastically limit ON current variability. The corresponding $I_{ds} - V_{gs}$ curve for 100 Monte Carlo runs is shown in Fig. S2**c** As expected, due to the external Resistor ON current is constant at 100 nA despite having variability in the V_{th}.

S2 Measurement setup

For the electrical characterization a measurement setup consisting of a PXIe System from NI is used. The test-structures, consisting of 25 pads, contain up to 8 individual FeFETs. A separate NI PXIe-4143 Source Measure Unit (SMU) can access each contact of the test-structures. Source selection for each contact is handled by a custom switch-matrix that is controlled by NI PXIe-6570 Pin Parametric Measurement Units (PPMU). The external resistor is connected at the source-terminal contact on the switch-matrix. The switch-matrix connects to the FeFET-structures via a probe-card (Fig. S3 and Fig. S4).

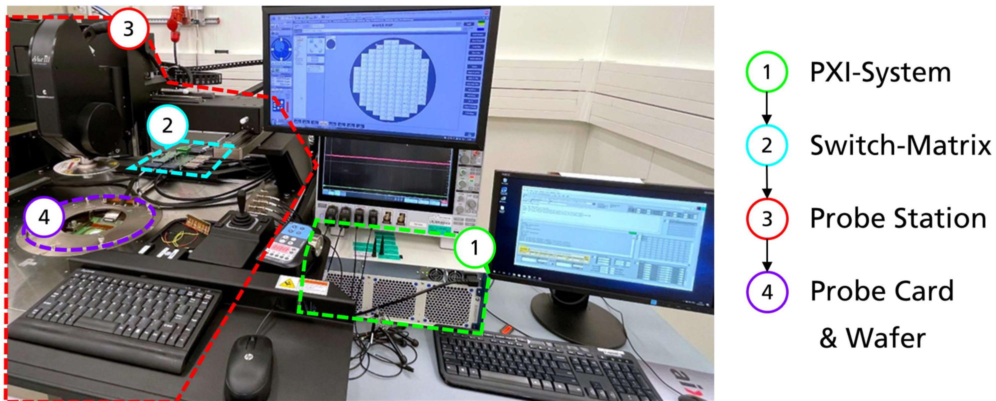

Figure S3: **Measurement setup for FeFET characterization.** A PXI System provides Source Measurement Units (SMU) and Pin Parametric Measurement Units (PPMU). PPMUs are used to configure the Switch Matrix for routing the source signals to the respective contact needles. Test structures are available on 300 mm wafers and connected to the measurement setup on a semi-automatic probe station using a probe card.

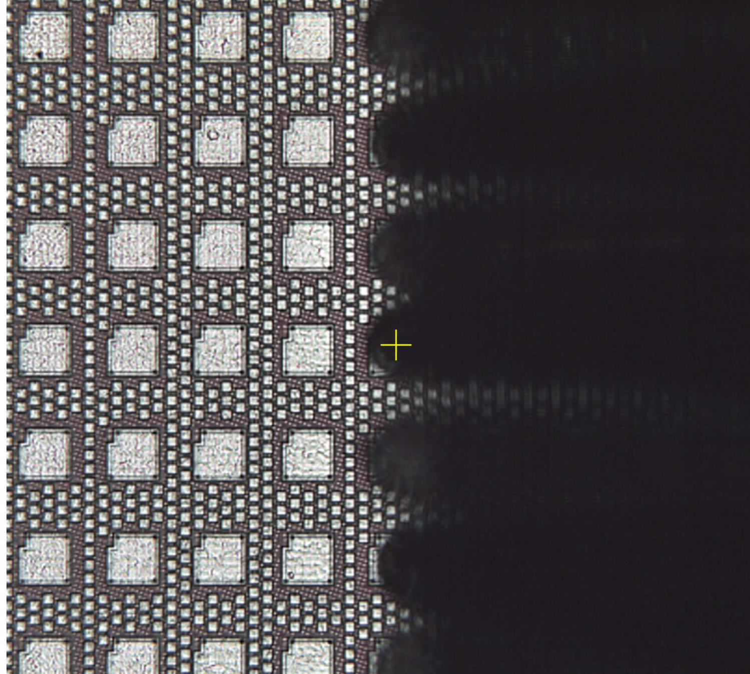

Figure S4: **Top view of the characterised test structure**. Up to 8 FeFETs with the individual source, drain and gate terminals are located in industry-standard test structures with 25 pads.

52 ## S3 Device variations measurements and simulation

53 The achievable device variation is obtained by screening the switching transition of 60
54 FeFETs with an area size of 900x900 nm^2 each. The states are set by erase operation, which
55 is typically more gradual over a wider switching range [31] due to the intrinsic current
56 percolation path effects in MFIS-based FeFETs [32]. The erase voltages range from -2 V to
57 -5.3 V with a pulse length of 200 ns in decrements of -50 mV. This gives us a measured
58 variability on average of about 40 mV.

59 From the measured variability in V_{th}, variability in the sampling voltage is com-
60 puted. 1000 Monte-Carlo runs for each stored and input state is performed. The cor-
61 responding mean, fifth percentile, and ninety-fifth percentile of the sampled voltage are
62 shown in Fig. S5. To reduce simulation complexity, up to 4 cells in the array are simu-
63 lated. For a higher number of cells, the variability is calculated algebraically. This saves
64 the simulation time, which otherwise would have taken considerable time and comput-
65 ing resources.

66 In Eq. (2) of the main text shows that the sampling voltage is the summation of the
67 current from all FeFETs. Therefore, the maximum variability in the sampling voltage is
68 the summation of the variability for all the FeFETs. The sampling voltage doubles for
69 doubling the number of cells. Therefore, for a given number of cells (n), the maximum

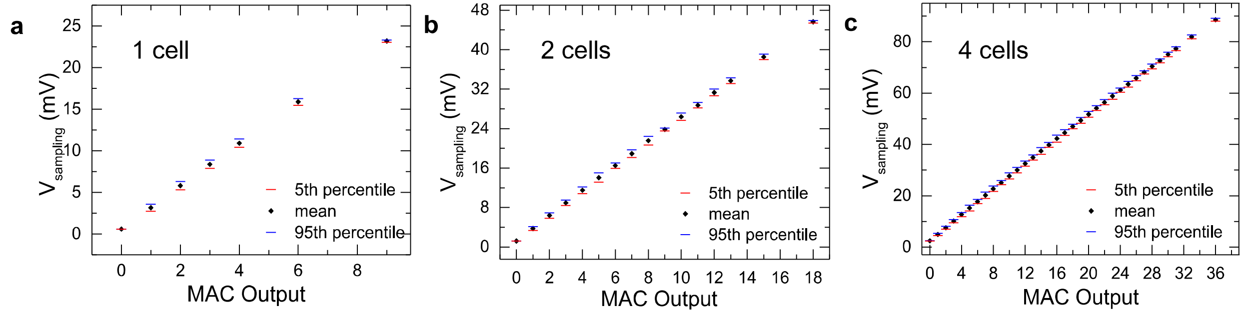

Figure S5: **Variation in MAC output due to the measured variation in V_{th}. a.** Mean, 5th percentile and 95th percentile of Sampling Voltage against MAC output for 1 cell, **b.** 2 cells, and **c.** 4 cells. Monte Carlo simulations of 1000 runs for each input and stored case is performed with a standard deviation of 40 mV in V_{th}.

70 standard deviation is given by :

$$\sigma V_{sampling,ncells} = 2 * \sigma V_{sampling,n/2cells} \tag{4}$$

71 The predicted sampling voltage matches very well with simulations for up to 4 cells.
72 Thus, for predicting the variability in sampling voltage for 32 cells, results from 4 FeFET
73 are extrapolated. The maximum σ $V_{sampling}$ is 0.16 mV for a single cell. For the case of 4
74 cells, the maximum σ $V_{sampling}$ from simulations is 0.44 mV. From our predictions using
75 the equation above, the maximum sigma is 0.48 mV. This validates the equation above to
76 calculate the variability for a higher number of cells. σ $V_{sampling}$ for 32 cells is calculated
77 to be 3.52 mV. The final variability for 32 cells in the array is used to analyze the loss in
78 inference accuracy for the neural network.

79 ## S4 Neural network implementation

80 Weights and activations are quantized to 2 bits with unsigned format. The initial trans-
81 formation from FP to quantized values is computed by (5), where a_{FP}, a_{quant} are FP and
82 quantized tensors respectively, b is the target bit-width and Δ_a is the scaling factor.

$$a_{quant} = \text{clip}\left(\text{round}\left[\left(\frac{a_{FP}}{\Delta_a}\right), \{-2^{b-1}, 2^{b-1}-1\}\right]\right) \tag{5}$$

83 The scaling factor Δ_a is obtained by:

$$\Delta_a = \frac{max(|a_{FP}|)}{2^{b-1}-1} \tag{6}$$

84 To efficiently used the proposed IMC architecture, we need to enable unsigned quanti-
85 zation of weights w and activations x. Therefore, zero-points z_w, z_x are used as an offset
86 to have all quantized values > 0. The resulting unsigned quantization is computed as
87 follows (e.g. for activations x):

$$x_{quant,unsigned} = x_{quant} + z_x = x_{quant} + (-\min(x_{quant})) \tag{7}$$

88 Assuming that weights and activations tensors have been transformed to matrices, The
89 matrix multiplication between activation matrix X_{quant} of size $m \times n$ and weights matrix
90 W_{quant} of size $n \times p$ outputs matrix Y_{quant} $(m \times p)$. The element $y_{j,k} \in Y$ is computed as
91 follows:

$$y^{j,k}_{quant} = \sum_{i=0}^{n} x^{j,i}_{quant} w^{i,k}_{quant} - \sum_{i=0}^{n} x^{j,i}_{quant} z_w - \sum_{i=0}^{n} w^{i,k}_{quant} z_x + \sum_{i=0}^{n} z_x z_w \tag{8}$$

92 The output is then re-scaled to FP precision by multiplying Y_{quant} with the scaling factor
93 $\Delta_x \times \Delta_w$. Thanks to the presence of ReLU at the output of all layers, z_x is zero for all layers
94 except the first one that receives the input data.