Theoretical Optimal Specifications of Memcapacitors for Charge-Based In-Memory Computing

Zichen Qian
Columbia University
New York, New York, USA
zq2189@columbia.edu

Rentao Wan
Columbia University
New York, New York, USA
rw2899@columbia.edu

Chin-Hsiang Liao
University of Minnesota
Minneapolis, Minnesota, USA
liao0120@umn.edu

Steven Koester
University of Minnesota
Minneapolis, Minnesota, USA
skoester@umn.edu

Mingoo Seok
Columbia University
New York, New York, USA
ms4415@columbia.edu

ABSTRACT

This paper presents the optimal theoretical specifications of memcapacitors that allow them to surpass the existing state-of-the-art in-memory computing (IMC) prototypes in terms of energy efficiency and weight density. To do so, we build and simulate the SPICE model of an existing memcapacitor device in an IMC macro featuring charge-based computing. We develop the energy efficiency model and weight density model of the memcapacitor-based IMC macro, which are verified against SPICE simulation. Finally, we present the optimal theoretical specifications for memcapacitor devices to obtain a 10x improvement in energy efficiency and/or weight density over the existing IMC prototypes.

CCS CONCEPTS

• **Hardware** → Emerging architectures; **Emerging simulation**; *Memory and dense storage*; *Non-volatile memory*.

KEYWORDS

memcapacitor, in-memory computing, optimal specifications, energy efficiency, weight density

ACM Reference Format:
Zichen Qian, Rentao Wan, Chin-Hsiang Liao, Steven Koester, and Mingoo Seok. 2025. Theoretical Optimal Specifications of Memcapacitors for Charge-Based In-Memory Computing. In *30th Asia and South Pacific Design Automation Conference (ASPDAC '25), January 20–23, 2025, Tokyo, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3658617.3697576

1 INTRODUCTION

Deep convolutional neural networks (DNN) in artificial intelligence (AI) systems have gained significant interest with a wide range of applications [22]. State-of-the-art DNNs are often faced with hardware constraints due to the large amounts of parallel computation required in complex models. Conventional accelerators for DNN require row-by-row memory access to perform computations on data and parameters, posing a bottleneck to computing performance [1].

In-memory computing (IMC) hardware has gathered significant attention recently due to its ability to perform parallel computations where data is stored. Computing in-memory eliminates the need for row-by-row memory access in each computation cycle. This results in an energy efficiency several orders of magnitude higher than conventional hardware [8, 28].

Researchers have developed IMC circuits based on various memory technologies [5, 20, 25]. One approach to IMC involves using novel memory devices, such as memristors and memcapacitors, to store weight parameters. The charge-based memcapacitors allow for negligible static current consumption, achieving significantly higher energy efficiency than the better-known memristor-based designs [14, 26]. The theory of memcapacitors was proposed over a decade ago [7], with promising developments in memcapacitor implementation in recent years [6, 13]. Memcapacitors are commonly implemented with ferroelectric material which exhibits a nonvolatile capacitive memory effect.

However, memcapacitors are a relatively new technology and previous research mostly focus on device level implementation. Prior works on memcapacitor fabrication mainly aim to maximize capacitive range of the device [6, 13]. There has not been a systematic study on a set of target specifications of memcapacitors which can guide device development.

Therefore, this article aims to provide target specifications for memcapacitor fabrication and would not be considering the feasibility of fabricating such devices. Towards this goal, we have developed a SPICE and Verilog-A memcapacitor model based on the leading memcapacitor device in [6]. We constructed a memcapacitor-based IMC circuit netlist. We developed energy efficiency and weight density models for the memcapacitor-based IMC, with energy consumption verified by SPICE simulations. Finally, using those models, we find a set of the optimal specifications of memcapacitor devices to attain 10x higher energy efficiency and/or weight density than state-of-the-art IMC prototypes. We found that an 8bx8b memcapacitor-based IMC with 10x improvement in both energy efficiency and weight density requires a memcapacitor area of $\leq 0.28 \mu m^2$ and maximum capacitance $\leq 1.02 pF$ with 256 stable capacitance states.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ASPDAC '25, January 20–23, 2025, Tokyo, Japan

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0635-6/25/01
https://doi.org/10.1145/3658617.3697576

Table 1: Fabricated memcapacitors by prior works

Reference	Type	Maximum Capacitance / Area (fF/μm^2)	Range of Capacitance (CR)	Fabricated Area ($L_m \cdot W_m$)	Program Voltage (V_{prog})	Capacitive Read-out Voltage (V_{read})
Nature'21 [6]	FeFET	12	1478x	2500 μm^2	−5.5 to 5.5V	0.5V
EDL'23 [13]	FeFET	8 [a]	25x	5540 μm^2 [a]	−3.5 to 3.5V	0.1V
VLSI'22 [29]	FeFET	12	125x	5390 μm^2 [b]	−6 to 6V	0.4V
EDL'23 [17]	FeFET	5	134x	6350 μm^2	−6 to 5V	0.5V
EDL'22 [15]	Organic	0.45 [c]	8.3x	0.25 mm^2 [c]	−30 to +30V	0.03V

[a]Gate area only. [b]Estimated based on 1kb array image. [c]Estimated based on the area of the pentacene layer.

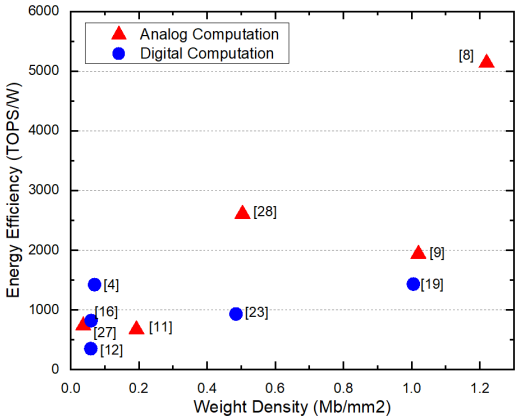


Figure 1: Energy efficiency and weight density of recent SRAM-based IMC macros, normalized to 1-b input, 1-b weight. Normalized energy efficiency = (macro energy efficiency)·(input bits)·(weight bits). Weight density = (macro size)/(macro area).

2 PRIOR WORKS

2.1 Recent State-of-the-Art IMC Works

Fig 1. summarizes recent SRAM-based IMC works and their performance in terms of energy efficiency (TOPS/W) and weight density (Mb/mm^2) [4, 8, 9, 11, 12, 16, 19, 23, 27, 28]. SRAM-based IMCs are used as a benchmark as they utilize readily available technology and provide some of the highest energy efficiencies among IMC prototypes [8, 28]. The energy efficiency is normalized to 1-bit input and 1-bit weight. The state-of-the-art SRAM-based IMC prototype achieves an energy efficiency of over 5,000 TOPS/W and a weight density of over 1.2 Mb/mm^2.

SRAM-based IMC can be divided into analog and digital based on the nature of employed arithmetic circuits. Analog IMC macros possess high energy efficiency due to their ability to perform multi-bit multiply-and-accumulate (MAC) operations in a bitcell. Conversely, its disadvantage lies in its analog-digital converters' (ADC) higher energy consumption and accuracy constraints.

2.2 Prior Works on Memcapacitors

Nonvolatile capacitive devices, or memcapacitors, have recently emerged as viable candidates for charge-domain IMC hardware. Memcapacitors have several advantages compared to memristors, including negligible static current consumption, low IR voltage drop and no sneak-path leakage [26]. A memcapacitor can store its "state" as small-signal capacitance, which can be programmed externally. Doped-HfO_2 ferroelectric (FE) material has shown much promise in implementing memcapacitors. Several practical implementations of memcapacitors have been developed recently [6, 13, 17, 29].

Table 1 summarizes the performance of fabricated memcapacitors from prior works. Capacitive range (CR) refers to the range of programmable capacitance, in other words, the ratio of capacitance at high capacitance state to low capacitance state (C_{HCS}/C_{LCS}). Maximum capacitance is formulated to: $C_{max} = C_{min} \times CR$, as is stated in [6, 13, 17, 29]. The range of programming voltage required to achieve the full capacitive range is noted as V_{prog} in Table 1. The capacitive readout voltage V_{read} refers to the small signal voltage used in each work to extract capacitance data without disturbing the programmed capacitive state of the memcapacitor.

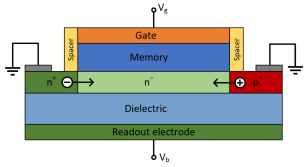
3 SPICE MODEL OF MEMCAPACITOR DEVICE

3.1 Charge Shielding-based Memcapacitor and Model

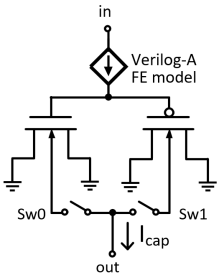
As can be seen in Table 1, the memcapacitor device in [6] has the best capacitive range out of prior works while having a fabricated area similar to the other devices. Capacitive range is important as a higher range can support a greater number of bits, thereby resulting in higher energy efficiency and weight density. Therefore, this article will model the device fabricated in [6] to be used in memcapacitor simulations.

Fig. 2a shows the device structure of the memcapacitor fabricated in [6]. The device consists of a gate electrode, a ferroelectric layer with memory, a shielding layer, a bottom dielectric layer and a readout electrode. A high capacitive ratio is attained depending on the degree of shielding between the gate and readout electrode [6].

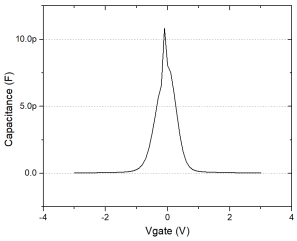
The p^+- and n^+-doped regions in the shielding layer act as reservoirs for electrons and holes, and can inject either carrier type into the n^- region [6]. When enough carriers are injected, a channel forms and the shielding layer is effectively grounded. Changes in gate voltage results in negligible current through the device, and



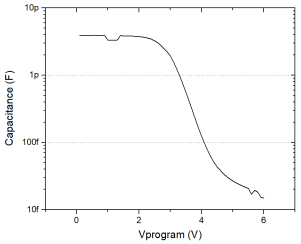
(a) Memcapacitor device structure in [6]



(b) SPICE model of memcapacitor device

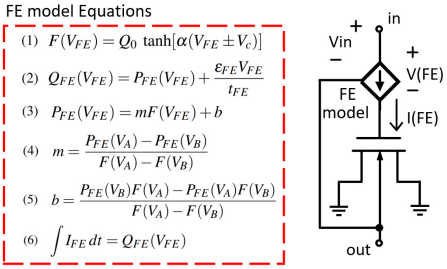


(c) Capacitance-voltage graph of SPICE model simulation without ferroelectric (FE) model

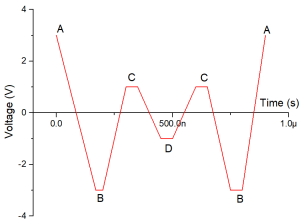


(d) Graph of simulated capacitance against programming voltage

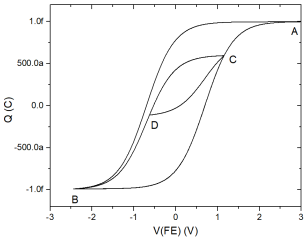
Figure 2: Device model of memcapacitor in [6] and simulation results



(a) Verilog-A implementation of ferroelectric material



(b) Waveform of voltage input to model in Fig. 2b



(c) Q-V curve of FE model with (b) as voltage input

Figure 3: Preisach model-based Verilog-A implementation of ferroelectric material

the device is in a low capacitance state. Without carrier injection, the memory, shielding and dielectric layers act as capacitors in series, resulting in the high capacitance state [24].

As there is no existing BSIM or SPICE-compatible model for the $p^+n^-n^+$ shielding layer in Fig. 2a, we use MOSFET transistors to model the shielding layer. The source and drain of NMOS/PMOS act as reservoirs for electrons/holes respectively, and the transistor body models the n^- region in the memcapacitor device. The gate-body capacitance of the MOSFET transistors models the capacitance of the shielding layer in the memcapacitor device.

Fig. 2b shows the SPICE model for the memcapacitor device in [6]. The ferroelectric layer with memory effect is represented by a Verilog-A behavioral model, as explained in the next section. When voltage at gate V_{gate} <0, Sw1 turns on and Sw0 turns off, and the shielding layer of the device is modeled by the PMOS transistor. Conversely, when voltage at gate V_{gate} >0, Sw0 turns on and Sw1 turns off, and the shielding layer is modeled by the NMOS transistor. The switches are implemented with ideal voltage-controlled capacitors.

Fig. 2c shows the simulated capacitance of the SPICE model without considering the ferroelectric layer. The transistors used are on the TSMC 0.18μm technology node, simulated with BSIM level 49 model. The oxide thickness of the BSIM model is increased to reflect the device in [6]. The capacitance of the device is measured by integrating the current I_{cap} divided by the input pulse height. There is some error at the switching point V_{gate}=0 of the model. Fig. 2c resembles the measured CV curve of the fabricated device in [6].

Fig. 2d shows the simulated capacitance against program voltage for the SPICE model including the ferroelectric Verilog-A model. A voltage pulse with pulse height $V_{program}$ and pulse width of 10ns is first given. Subsequently, a voltage pulse of 0.1V is used to extract the capacitance of the device. Fig. 2d gives a capacitive range of around 263x, allowing the memcapacitor model to support 8 bits.

3.2 Verilog-A Model of Ferroelectric Material

Fig. 3a shows the Verilog-A implementation of the ferroelectric layer with memory effect. The behavioral model implements the Preisach model for ferroelectric capacitors [2, 10]. Equation (1) in Fig. 3a is the saturation loop or hysteresis loop, which can be either increasing or decreasing as denoted by the sign of V_c. Q_0 is the maximum charge contribution from ferroelectric switching, V_c is the coercive voltage, and α is derived from experimental data [10]. ε_{FE} and t_{FE} is the permittivity and thickness of the ferroelectric material respectively. The material "remembers" previous alternating pairs of dominant voltage maxima and minima. Within the saturation loop, the Q-V curve always pass through previous turning points stored in memory [10]. The Q-V curve between adjacent turning points is calculated by (3)-(5) in Fig. 3a [18].

By charge conservation, the current through the FE model is equivalent to the current into the gate of the transistor. The FE model acts as a current source, producing a current equivalent to the solution of (6) in Fig. 3a. The Verilog-A model forces I_{FE} such that the equations $V_{in} - V_{FE} = V_{gate}$ and $Q_{FE}(V_{FE}) = Q_{NMOS}(V_{gate})$ holds, by implementing (6) in Fig. 3a. This is modeled after [18] which utilizes a solver to achieve this purpose.

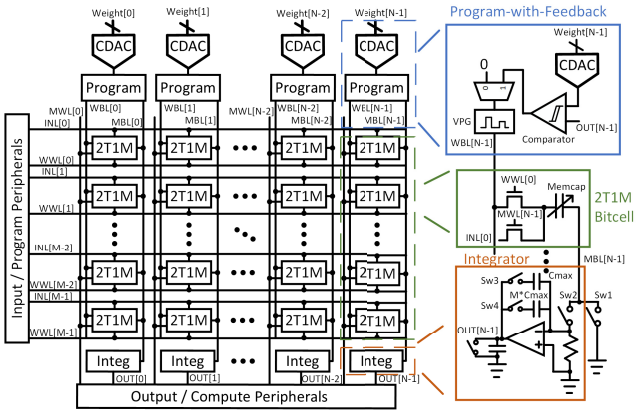


Figure 4: 2T1M analog-in analog-out IMC macro

Fig. 3b and c shows the simulation of minor loops and the resulting Q-V curve of the FE model. Fig. 3c captures the memory effect, turning point and hysteresis properties, exhibiting the same behavior as the verified models in [10, 18].

Fig 2d is obtained with TSMC 0.18μm transistors with oxide thickness changed to 40nm. The FE model parameters are Q_0 = $100pC$, ε_{FE}/t_{FE} = $20pF$, V_c=2.5V and α=2. These parameters are chosen such that the simulation results of the model are closer to the measurement results in [6]. Feasibility of fabrication is not considered as this is a study on theoretical target specifications.

4 MEMCAPACITOR-BASED IMC CIRCUIT

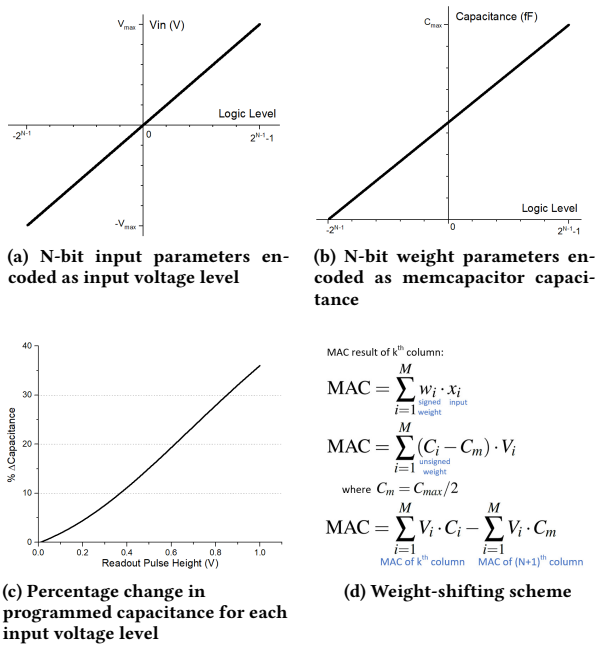
4.1 IMC Architecture

Fig 4 shows a possible memcapacitor-based IMC architecture. It features an analog-input and analog-output interface for higher energy efficiency. Seo. et al. demonstrated such an ADDC processor capable of performing CNN-based model inference in the analog domain without requiring much analog-to-digital conversions [21]. We recognize that the comparison with previous SRAM-IMCs is not completely fair as previous designs include ADCs.

The IMC architecture employs MxN 2-transistor-1-memcapacitor (2T1M) bitcells. Each column consists of M 2T1M bitcells, a feedback-based programming module, and an output integrator. The 2T1M bitcell contains a write-access transistor, a compute-access transistor, and a memcapacitor. The macro also has several wires, including write bitlines (WBL) and write wordlines (WWL) for programming. It has input lines (INL), MAC wordlines (MWL), and MAC bitlines (MBL) for computation.

The IMC circuit uses pulse height modulation to program the memcapacitors, as is demonstrated in prior works [6]. To ensure robustness across process-voltage-temperature (PVT) variations, it employs the Program-with-Feedback module as shown in Fig. 4, which consists of an 8-bit capacitive digital-to-analog converter (CDAC), a voltage comparator, a 2-to-1 multiplexer (MUX), and a variable-amplitude pulse generator (VPG).

During programming, weight parameters are received as digital inputs and are converted to analog values first with the CDAC. Then, the MBL is grounded by switch 1 (Sw1), and the bitcell row is



(a) N-bit input parameters encoded as input voltage level

(b) N-bit weight parameters encoded as memcapacitor capacitance

(c) Percentage change in programmed capacitance for each input voltage level

(d) Weight-shifting scheme

Figure 5: Encoding scheme for input and weight parameters

selected by WWL and the write-access transistor. One or more negative voltage pulses are used to reset the state of the memcapacitors [6, 17]. The VPG sends a programming pulse. Then, the capacitance value is read out by the VPG sending a small pulse, with switches 2 and 3 turned on. The MUX achieves this alternate pulse generation with an alternating control signal. The integrator outputs a voltage according to (1). This voltage is compared with the analog output of the CDAC. The process repeats until the desired capacitance is programmed.

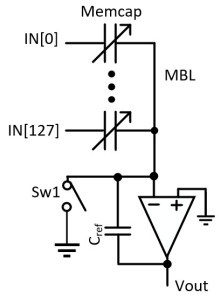
After programming, the macro can perform computing, i.e., an M-by-N vector-matrix multiplication (VMM) with an N target precision. In the compute phase, switches 2 and 4 are turned on. Voltage input is sent and VMM results are read out by charge summing. Assuming a sufficiently large gain of the op-amp (≥40dB), the output voltage can be simplified to:

$$V_{out} = -\frac{\sum_{i=1}^{M} C_i V_i}{C_r}, \tag{1}$$

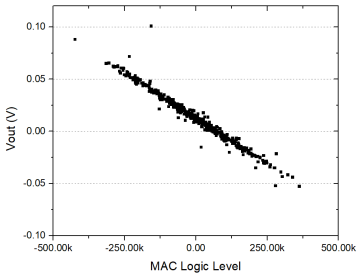
where C_i is the programmed capacitance of the i-th bitcell, V_i is the signed-magnitude of the i-th input ramp signal, and C_r is the capacitor of the integrator. In our simulations, we set $C_r = C_{max} \cdot M$.

4.2 Simulation Setup and Results

Fig. 5a shows the adopted encoding scheme of N-bit inputs in the proposed IMC macro, which linearly maps the logic level of -2^{N-1} to $+2^{N-1} - 1$ onto the voltage swing levels of $-V_{max}$ to $+V_{max}$. V_{max} is the maximum input voltage used in computation. As shown in Fig. 5c, the percentage change in capacitance increases proportionately to the increase in input voltage. This is due to the capacitance characteristic of the transistors as shown in Fig. 2c.

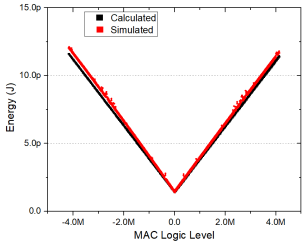


(a) IMC circuit simulated with 128 bitcells

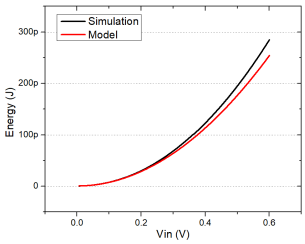


(b) Simulated MAC transfer curve with randomized inputs and weights

Figure 6: Simulation of memcapacitor-based IMC



(a) Simulated and model energy consumption of IMC macro per column



(b) Simulated and model energy consumption for IMC macro per column against V_{in}

Figure 7: Model verification results

Fig. 5c is obtained from capacitance simulations on the SPICE model in Fig. 2b.

Fig. 5b shows the encoding scheme of N-bit weight parameters stored in the memcapacitors. The memcapacitor is assumed to have a dynamic range of 2^N or higher, with 2^N equally spaced stable capacitance states between minimum and maximum programmable capacitance. Fig. 5d shows a weight-shifting technique employed to achieve negative weight computation [3]. An additional column will be used with all memcapacitors set to $C_{max}/2$, which is 2^{N-1} in logic value. The results of this column will be subtracted from the results of each column.

Fig. 6a shows the circuit simulated. One column of the IMC architecture in Fig. 4 is simulated with the programming and computing pulses given by a voltage source. The memcapacitor is implemented with the model in Fig. 2b, and the op-amp is implemented with an ideal voltage controlled voltage source. A 30ns pulse is given to program the memcapacitor, and an input pulse is given for computing with 10ns rise time. Fig. 6b shows the multiply-and-accumulate (MAC) output transfer curve with randomized inputs and weights for 128 bitcells. 8-bit input and 8-bit weights are used with the encoding scheme in Fig. 5, with V_{max} of $0.128V$ and C_{max} of $3.84pF$.

5 ENERGY EFFICIENCY AND WEIGHT DENSITY MODEL

5.1 Model Development

In this section, an energy efficiency model and a weight density model will be developed for the architecture in Section IV. Energy efficiency is formulated in terms of tera-operations per second per watt (TOPS/W), and weight density is calculated by the number of bits in macro divided by macro area.

In formulating the model, we assume that interconnects overlay the memcapacitor. We assume that many cycles of computation occur for each program operation, making the energy consumption of the programming phase negligible. We also assume no peripherals other than the op-amp consume significant power. The model is idealized and actual energy consumption may be higher, thus the aim of this article is to give approximate target specifications for fabricating memcapacitors for IMC.

We then formulate the energy consumption of one column over one cycle of t_c to:

$$E_{col} = E_{OTA} + \frac{C_{MWL} \cdot V_g^2}{2} + \sum_{i=1}^{M} \frac{C_i \cdot V_i^2}{2} \tag{2}$$

where E_{OTA} is the energy consumption of the op-amp implemented with operational transconductance amplifier (OTA) and C_{MWL} is the total capacitance on the MWL. V_g refers to the gate voltage of access transistors and the voltage on MWL, and C_i refers to the memcapacitor capacitance on row i. V_i refers to input voltage in row i.

We further develop E_{col} as a function of memcapacitor device-level parameters as follows:

$$\begin{aligned} E_{col} =& P_{OTA} \cdot t_c + \frac{V_{in}^2 \cdot M}{2} [C_{mem} + C_{wire} \cdot (W_m + W_{tran})] \\ &+ \frac{M \cdot V_g^2}{2} (C_g + L_m \cdot C_{wire}), \end{aligned} \tag{3}$$

where V_g is the voltage at the gate of the access transistors, and M is the number of rows in the macro. C_{wire} refers to the unit-length interconnect capacitance. W_m and L_m refer to the width and length of the memcapacitor, and W_{tran} refers to the width of access transistors. C_{mem} and V_{in} is the average memcapacitor capacitance and average input voltage, and C_g refers to gate capacitance of access transistors. P_{OTA} is the power consumption of OTA in the integrator.

Finally, using (3), we develop the energy efficiency model in TOPS/W as follows.

$$EE = \frac{M \cdot n_{in} \cdot n_{mem}}{E_{col}}, \tag{4}$$

where M is the number of rows, n_{mem} is the bit width one memcapacitor can store, and n_{in} is the bit width of analog voltage input. The energy efficiency model is normalized to 1b×1b operations.

We also develop the weight density model, formulated to:

$$D_w = \frac{n_{mem} \cdot M \cdot N}{M \cdot N \cdot L_m \cdot (W_m + W_{tran}) + A_{peri}}, \tag{5}$$

where A_{peri} refers to the total area of the peripheral circuits of the macro, with M and N being the number of rows and columns of the macro, respectively.

Table 2: Maximum specifications of memcapacitors to achieve 10x improvements over SRAM-based IMC [8]

Performance Goals	Memcapacitor Bit Width n_{mem}	Input Bit Width n_{in}	Area of Memcapacitor (μm^2)	Weight Resolution C_{res} (fF)	Maximum Capacitance C_{max} (pF)	C_{max} / Area (pF/μm^2)	V_{max} (V)
10x Energy Efficiency and 1x Weight Density	4	4	≤1.6	≤5.5	≤0.088	≤0.055	≤0.128
	8	8	≤3.2	≤3.8	≤0.973	≤0.304	
1x Energy Efficiency and 10x Weight Density	4	4	≤0.12	≤179	≤2.86	≤23.9	
	8	8	≤0.28	≤46	≤11.8	≤42.1	
10x Energy Efficiency and 10x Weight Density	4	4	≤0.12	≤8.4	≤0.134	≤1.12	
	8	8	≤0.28	≤4	≤1.02	≤3.66	

5.2 Energy Consumption Model Verification

We verify the energy consumption model against SPICE simulation results. The simulated power consumption is measured by the input voltage multiplied by the current produced by the FE model in Fig. 2b. We set the clock frequency to 100MHz. The OTA energy consumption is not simulated. A_{mem} is set to be $400\mu m^2$, assuming an effective capacitance density of $10fF/\mu m^2$ as is the approximate average in Table 1.

Fig. 7a shows the simulated energy consumption against the MAC logic level of one column with 128 bitcells for 8b×8b operation. We set C_{max} to $3.84pF$ and V_{max} to $0.128V$. Fig. 7a also shows the calculated results from equation (3). As shown in Fig. 2b, the capacitance used in computation is different from total gate capacitance. Thus C_{mem} in (3) is the total gate capacitance of the transistor, measured to be $11pF$ on average.

Fig. 7b compares the simulation result and model calculation while sweeping V_{in} from 0.01V to 0.6V. It targets the same 8b×8b precision. Fig. 7 shows good matching between the simulation and model calculation.

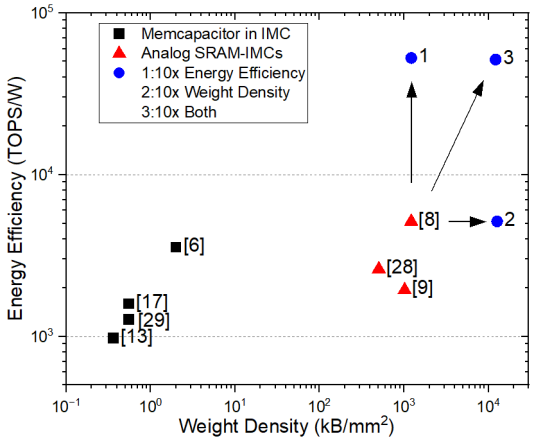


Figure 8: Energy efficiency and weight density of 2T1M IMC using memcapacitors from Table I and SRAM-based analog IMC macros

6 OPTIMAL SPECIFICATIONS OF MEMCAPACITORS

In this section, we investigate the optimal specifications of memcapacitors to obtain a 10x increase in energy efficiency and/or weight density compared to the state-of-the-art SRAM-based IMC.

Fig. 8 shows the energy efficiency and weight density of top-performing SRAM-based IMC macros. It also shows the performance of the previously prototyped memcapacitors in Table 1 when their parameters are substituted into (2)-(5). In estimating the memcapacitors' energy efficiency and weight density, we assume V_{max} to be $0.128V$ with 8-bit precision. n_{mem} is assumed to be the maximum number of bits their capacitance range can support. The OTA's current consumption is assumed to be 1 μA total, clock frequency set to be 100 MHz, V_g to be 1V, and M to be 128. The memcapacitors are assumed to be a square when calculating interconnect capacitances, and A_{peri} is assumed to be half of macro area. It shows the existing memcapacitors are lagging behind significantly in weight density.

However, Fig. 8 also shows the three potential memcapacitor devices in the blue circles that can achieve 10x energy efficiency, 10x weight density, and both, as compared to the state-of-the-art SRAM-IMC in [8]. We summarize the specifications of those memcapacitors in Table 2.

We derived those specifications by the following process. Targeting for 8b×8b operation, we consider two different cases: n_{mem} and n_{in} are either 4 or 8, respectively. Then, we calculate the area of the memcapacitor based on the benchmark value of weight density being 1.22 Mb/mm^2 for 1x improvement (12.20 Mb/mm^2 for 10x improvement) using (5). A_{peri} is assumed to be half of total macro area. M=128 rows are used.

We then calculate the value of capacitance resolution C_{res}, assuming that $C_{res} = C_{min}$ and $C_{max} = 2^{n_{mem}} \cdot C_{min}$. We assume that the capacitance between the gate and n^+/p^+ metal contacts are negligible, according to [6]. Therefore C_{mem} in (3) is equivalent to the capacitance used for computing.

Assuming the memcapacitor layout is a square, we can calculate C_{wire}, L_m, W_m based on the calculated area. Then, using (4) and the benchmark value of 1x energy efficiency 5136 TOPS/W (10x energy efficiency 51,360 TOPS/W) [8], we can calculate C_{res}. Here, we set V_{max} to 0.128V to ensure minimal change in capacitance after programming, as shown in Fig. 5c. We assume the interconnect capacitance to be $0.1fF/\mu m$. V_g is assumed to be 1V, t_c to be 10 ns

and P_{OTA} to be $2\mu W$. Since the energy consumption varies across logic level (Fig. 7a), the average energy consumption assuming $C_{mem} = C_{max}/2$ and $V_{in} = V_{max}/2$ is calculated.

There is a $\leq$ symbol on each metric in Table 2 which represent the maximum possible values of the respective parameters. Achieving the same n_{mem} and n_{in} with smaller capacitance, area, or voltage input will give higher energy efficiency and/or weight density than the performance goals.

A comparison of Table 1 and 2 shows that prior memcapacitor devices can achieve the target maximum capacitance and capacitance density (C_{max}/area). However, the memcapacitor area in prior works are much larger than that required in Table 2, and only [6] has the potential to support 8-bit weights. Thus, the main challenge in fabrication will be obtaining a higher capacitive range while reducing the area by around 1000x to achieve comparable weight density to SRAM-IMCs.

7 CONCLUSION

In this article, we propose a set of target specifications for memcapacitor device fabrication in order for memcapacitor-based IMCs to bring about improvements over current IMC designs. A SPICE and Verilog-A model is built to simulate the leading memcapacitor device [6]. An energy efficiency model and a weight density model are developed based on the memcapacitor-based IMC and model parameters such as input voltage, maximum capacitance, and memcapacitor area. The total energy consumption model is verified with SPICE simulation results. This article suggests a set of maximum values for each critical parameter of memcapacitors to obtain a possible 10x improvement in energy efficiency and/or weight density over the state-of-the-art SRAM-based IMC prototypes.

REFERENCES

[1] Mohd Akhoon, Shahrel Azmin Suandi, Abdullah Alshahrani, Abdul-Malik Saad, Fahad Albogamy, Mohd Abdullah, and Sajad Loan. 2021. High performance accelerators for deep neural networks: A review. *Expert Systems* 39 (10 2021). https://doi.org/10.1111/exsy.12831

[2] Andrei T. Bartic, Dirk J. Wouters, Herman E. Maes, Jurgen T. Rickes, and Rainer Waser. 2001. Preisach model for the simulation of ferroelectric capacitors. *Journal of Applied Physics* 89 (2001), 3420–3425. https://api.semanticscholar.org/CorpusID:52233144

[3] Zhengyu Chen, Xi Chen, and Jie Gu. 2021. 15.3 A 65nm 3T Dynamic Analog RAM-Based Computing-in-Memory Macro and CNN Accelerator with Retention Enhancement, Adaptive Analog Sparsity and 44TOPS/W System Energy Efficiency. In *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, Vol. 64. 240–242. https://doi.org/10.1109/ISSCC42613.2021.9366045

[4] Yu-Der Chih, Po-Hao Lee, Hidehiro Fujiwara, Yi-Chun Shih, Chia-Fu Lee, Rawan Naous, Yu-Lin Chen, Chieh-Pu Lo, Cheng-Han Lu, Haruki Mori, Wei-Chang Zhao, Dar Sun, Mahmut Sinangil, Yen-Huei Chen, Tan-Li Chou, Kerem Akarvardar, Hung-Jen Liao, Yih Wang, Meng-Fan Chang, and Tsung-Yung Chang. 2021. 16.4 An 89TOPS/W and 16.3TOPS/mm 2 All-Digital SRAM-Based Full-Precision Compute-In Memory Macro in 22nm for Machine-Learning Edge Applications. In *2021 IEEE International Solid-State Circuits Conference (ISSCC)*. 252–254. https://doi.org/10.1109/ISSCC42613.2021.9365766

[5] Peter Deaville, Bonan Zhang, Lung-Yen Chen, and Naveen Verma. 2021. A Maximally Row-Parallel MRAM In-Memory-Computing Macro Addressing Readout Circuit Sensitivity and Area. In *ESSCIRC 2021 - IEEE 47th European Solid State Circuits Conference (ESSCIRC)*. 75–78. https://doi.org/10.1109/ESSCIRC53450.2021.9567807

[6] Kai-Uwe Demasius, Aron Kirschen, and Stuart Parkin. 2021. Energy-efficient memcapacitor devices for neuromorphic computing. *Nature Electronics* 4 (10 2021), 1–9. https://doi.org/10.1038/s41928-021-00649-y

[7] Massimiliano Di Ventra, Yuriy V. Pershin, and Leon O. Chua. 2009. Circuit Elements With Memory: Memristors, Memcapacitors, and Meminductors. *Proc. IEEE* 97, 10 (2009), 1717–1724. https://doi.org/10.1109/JPROC.2009.2021077

[8] Qing Dong, Mahmut E. Sinangil, Burak Erbagci, Dar Sun, Win-San Khwa, Hung-Jen Liao, Yih Wang, and Jonathan Chang. 2020. 15.3 A 351TOPS/W and 372.4GOPS Compute-in-Memory SRAM Macro in 7nm FinFET CMOS for Machine-Learning Applications. In *2020 IEEE International Solid-State Circuits Conference - (ISSCC)*. 242–244. https://doi.org/10.1109/ISSCC19947.2020.9062985

[9] Hongyang Jia, Murat Ozatay, Yinqi Tang, Hossein Valavi, Rakshit Pathak, Jinseok Lee, and Naveen Verma. 2022. Scalable and Programmable Neural Network Inference Accelerator Based on In-Memory Computing. *IEEE Journal of Solid-State Circuits* 57, 1 (2022), 198–211. https://doi.org/10.1109/JSSC.2021.3119018

[10] Bo Jiang, Zurcher, Jones, Gillespie, and Lee. 1997. Computationally Efficient Ferroelectric Capacitor Model For Circuit Simulation. In *1997 Symposium on VLSI Technology*. 141–142. https://doi.org/10.1109/VLSIT.1997.623738

[11] Zhewei Jiang, Shihui Yin, Jae-Sun Seo, and Mingoo Seok. 2020. C3SRAM: An In-Memory-Computing SRAM Macro Based on Robust Capacitive Coupling Computing Mechanism. *IEEE Journal of Solid-State Circuits* 55, 7 (2020), 1888–1897. https://doi.org/10.1109/JSSC.2020.2992886

[12] Hyunjoon Kim, Taegeun Yoo, Tony Tae-Hyoung Kim, and Bongjin Kim. 2021. Colonnade: A Reconfigurable SRAM-Based Digital Bit-Serial Compute-In-Memory Macro for Processing Neural Networks. *IEEE Journal of Solid-State Circuits* 56, 7 (2021), 2221–2233. https://doi.org/10.1109/JSSC.2021.3061508

[13] Tae-Hyeon Kim, Omkar Phadke, Yuan-Chun Luo, Halid Mulaosmanovic, Johannes Mueller, Stefan Duenkel, Sven Beyer, Asif Islam Khan, Suman Datta, and Shimeng Yu. 2023. Tunable Non-Volatile Gate-to-Source/Drain Capacitance of FeFET for Capacitive Synapse. *IEEE Electron Device Letters* 44, 10 (2023), 1628–1631. https://doi.org/10.1109/LED.2023.3311344

[14] Can Li, Zhongrui Wang, Mingyi Rao, Daniel Belkin, Wenhao Song, Hao Jiang, Peng Yan, Yunning Li, Peng Lin, Miao Hu, Ning Ge, John Paul Strachan, Mark Barnell, Qing wu, Stan Williams, Jianhua Joshua Yang, and Qiangfei Xia. 2019. Long short-term memory networks in memristor crossbar arrays. *Nature Machine Intelligence* 1 (01 2019). https://doi.org/10.1038/s42256-018-0001-4

[15] Li-Xing Li, Jia-Wei Cai, Ya-Nan Zhong, Xu Gao, Jian-Long Xu, and Sui-Dong Wang. 2022. Organic Thin-Film Memcapacitive Device With Analog and Nonvolatile Memory Effect. *IEEE Electron Device Letters* 43, 9 (2022), 1539–1542. https://doi.org/10.1109/LED.2022.3195237

[16] Zhiting Lin, Shaoying Zhang, Qian Jin, Jianping Xia, Yunwei Liu, Kefeng Yu, Jian Zheng, Xiaoming Xu, Xing Fan, Ke Li, Zhongzhen Tong, Xiulong Wu, Wenjuan Lu, Chunyu Peng, and Qiang Zhao. 2023. A Fully Digital SRAM-Based Four-Layer In-Memory Computing Unit Achieving Multiplication Operations and Results Store. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 31, 6 (2023), 776–788. https://doi.org/10.1109/TVLSI.2023.3266651

[17] Ning Liu, Jiuren Zhou, Yupeng Yao, Siying Zheng, Wenjing Feng, Mengkuo Cui, Bochang Li, Yan Liu, Yue Hao, and Genquan Han. 2023. HfO2-Based Ferroelectric Optoelectronic Memcapacitors. *IEEE Electron Device Letters* 44, 3 (2023), 524–527. https://doi.org/10.1109/LED.2023.3235909

[18] Kai Ni, Matthew Jerry, Jeffrey A. Smith, and Suman Datta. 2018. A Circuit Compatible Accurate Compact Model for Ferroelectric-FETs. In *2018 IEEE Symposium on VLSI Technology*. 131–132. https://doi.org/10.1109/VLSIT.2018.8510622

[19] Jonghyun Oh, Chuan-Tung Lin, and Mingoo Seok. 2023. D6CIM: 60.4-TOPS/W, 1.46-TOPS/mm2, 1005-Kb/mm2 Digital 6T-SRAM-Based Compute-in-Memory Macro Supporting 1-to-8b Fixed-Point Arithmetic in 28-nm CMOS. In *ESSCIRC 2023- IEEE 49th European Solid State Circuits Conference (ESSCIRC)*. 413–416. https://doi.org/10.1109/ESSCIRC59616.2023.10268725

[20] Giacomo Pedretti and D. Ielmini. 2021. In-Memory Computing with Resistive Memory Circuits: Status and Outlook. *Electronics (Switzerland)* 10 (05 2021). https://doi.org/10.3390/electronics10091063

[21] Jin-O Seo, Mingoo Seok, and SeongHwan Cho. 2022. ARCHON: A 332.7TOPS/W 5b Variation-Tolerant Analog CNN Processor Featuring Analog Neuronal Computation Unit and Analog Memory. In *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, Vol. 65. 258–260. https://doi.org/10.1109/ISSCC42614.2022.9731654

[22] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1–9. https://doi.org/10.1109/CVPR.2015.7298594

[23] Dewei Wang, Chuan-Tung Lin, Gregory K. Chen, Phil Knag, Ram K. Krishnamurthy, and Mingoo Seok. 2022. DIMC: 2219TOPS/W 2569F2/b Digital In-Memory Computing Macro in 28nm Based on Approximate Arithmetic Hardware. In *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, Vol. 65. 266–268. https://doi.org/10.1109/ISSCC42614.2022.9731659

[24] Xuepei Wang, BoYao Cui, Lingling Jing, Xiaolin Wang, Maokun Wu, Yicheng Wen, Yishan Wu, Jinhao Liu, Feilong Zhang, Zidong Lin, Yanan Sun, Pengpeng Ren, Sheng Ye, Runsheng Wang, Zhigang Ji, and Ru Huang. 2024. Enabling Low-Power Charge-Domain Nonvolatile Computing-in-Memory (CIM) With Ferroelectric Memcapacitor. *IEEE Transactions on Electron Devices* 71, 4 (2024), 2404–2410. https://doi.org/10.1109/TED.2024.3367965

[25] Shihui Yin, Yulhwa Kim, Xu Han, Hugh Barnaby, Shimeng Yu, Yandong Luo, Wangxin He, Xiaoyu Sun, Jae-Joon Kim, and Jae-sun Seo. 2019. Monolithically Integrated RRAM- and CMOS-Based In-Memory Computing Optimizations for Efficient Deep Learning. *IEEE Micro* 39, 6 (2019), 54–63. https://doi.org/10.1109/MM.2019.2943047

[26] Shimeng Yu, Yuan-Chun Luo, Tae-Hyeon Kim, and Omkar Phadke. 2023. Nonvolatile Capacitive Synapse: Device Candidates for Charge Domain Compute-In-Memory. *IEEE Electron Devices Magazine* 1, 2 (2023), 23–32. https://doi.org/10.1109/MED.2023.3293060

[27] Jinshan Yue, Xiaoyu Feng, Yifan He, Yuxuan Huang, Yipeng Wang, Zhe Yuan, Mingtao Zhan, Jiaxin Liu, Jian-Wei Su, Yen-Lin Chung, Ping-Chun Wu, Li-Yang Hung, Meng-Fan Chang, Nan Sun, Xueqing Li, Huazhong Yang, and Yongpan Liu. 2021. 15.2 A 2.75-to-75.9TOPS/W Computing-in-Memory NN Processor Supporting Set-Associate Block-Wise Zero Skipping and Ping-Pong CIM with Simultaneous Computation and Weight Updating. In *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, Vol. 64. 238–240. https://doi.org/10.1109/ISSCC42613.2021.9365958

[28] Bo Zhang, Jyotishman Saikia, Jian Meng, Dewei Wang, Soonwan Kwon, Sungmeen Myung, Hyunsoo Kim, Sang Joon Kim, Jae-Sun Seo, and Mingoo Seok. 2024. MACC-SRAM: A Multistep Accumulation Capacitor-Coupling In-Memory Computing SRAM Macro for Deep Convolutional Neural Networks. *IEEE Journal of Solid-State Circuits* 59, 6 (2024), 1938–1949. https://doi.org/10.1109/JSSC.2023.3332017

[29] Zuopu Zhou, Jiao Leming, Jiuren Zhou, Zijie Zheng, Yue Chen, Kaizhen Han, Yuye Kang, and Xiao Gong. 2022. Experimental Demonstration of An Inversion-Type Ferroelectric Capacitive Memory and its 1 kbit Crossbar Array Featuring High CHCS/CLCS, Fast Speed, and Long Retention. In *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*. 357–358. https://doi.org/10.1109/VLSITechnologyandCir46769.2022.9830291

Received 12 July 2024; accepted 8 September 2024