VLSI Architecture Design

Project Presentation

An SRAM-Based Multibit In-Memory Matrix-Vector Multiplier With a Precision That Scales Linearly in Area, Time, and Power

Table of contents

01

Introduction

Introduction

The selected paper describes a novel architecture of an In-Memory Compute Unit (IMCU) for MAC (Multiply-Accumulate) operations.

The IMCU unit allows the analog computation of matrix-vector multiplication which are the base of all machine and deep learning applications.

The described MAC operations can be written in equation as follows:

$$\vec{x}^{\mathsf{T}} \times \boldsymbol{A} = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix}^{\mathsf{T}} \times \begin{pmatrix} w_{1,1} & \cdots & w_{1,M} \\ \vdots & \ddots & \vdots \\ w_{N,1} & \cdots & w_{N,M} \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_M \end{pmatrix}^{\mathsf{T}} = \vec{y}^{\mathsf{T}}$$

Each output element y_m can be represented as a sum of N products of w_n ,m and x_n.

$$y_m = \sum_{n=1}^{N} w_{n,m} \cdot x_n, \quad m = 1, \ldots, M.$$

Introduction

- The conventional Von-Neumann architecture incurs time and energy costs when performing MAC operations by transferring matrix elements stored in memory units to a physically separated digital computation unit (ALU), creating a performance bottleneck.

- Thus, through the user of IMCUs, all computational primitives required are executed within the memory subsystem resulting in energy efficient and faster computation.

- The representation below shows the major differences
between the two architectures.

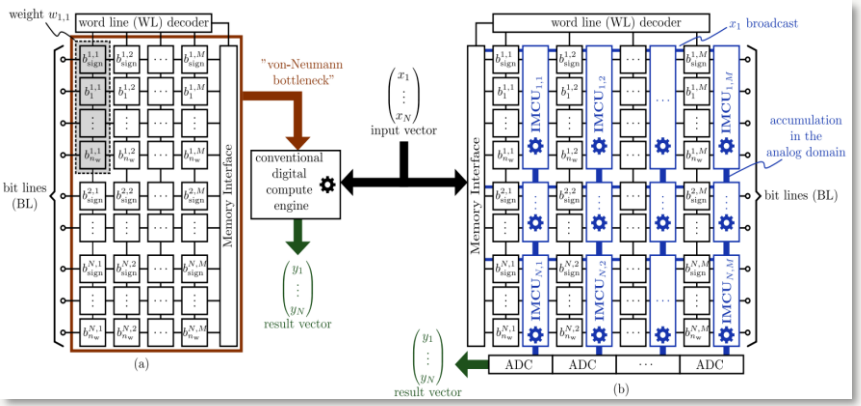


Figure 1: a) Conventional Von-Neumann Architecture. b) In-Memory Computation based Architecture with IMC units

02

Proposed IMCU CIRCUIT

Proposed IMCU Circuit

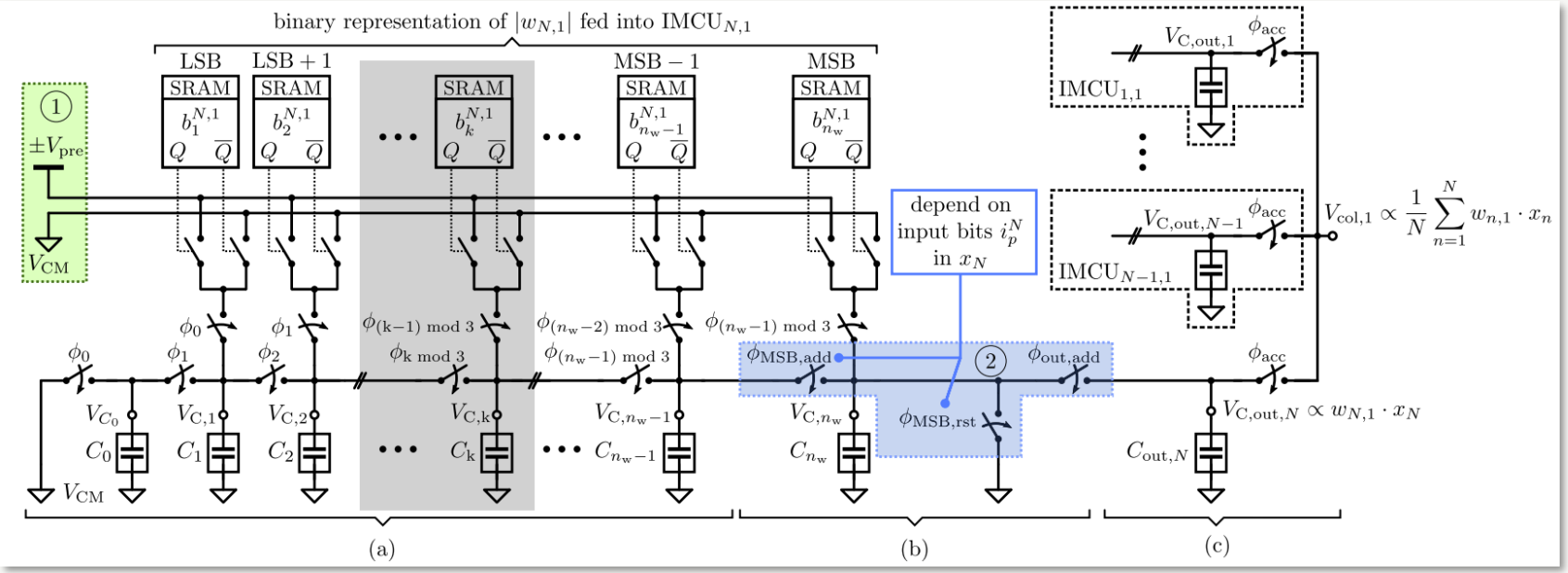


Figure 2: a) DAC Conversion b) Analog Multiplication c) Analog Accumulation

DAC Conversion

- DAC Conversion converts the stored weight bits b_n^k into proportional voltages $V_{w,n}$.
- D/A converter is implemented from equally sized capacitors C_k interconnected through switches.
- To conduct D/A conversion, a set of three non-overlapping digital pulse signals φ_0, φ_1 and φ_2 are used. Each SRAM cell contains one-bit b_n^k of the weights and controls one stage of the pipeline DAC. Based on the value of the corresponding capacitor is either charged to either V_{pre} or to the common node V_{CM} . Next, the top plates of capacitor C_k and C_{k-1} are shorted, which averages their voltages.

$$V_{C,k} = 0.5 \,.\, (b_k^n \,.\, V_{pre} + V_{C,k-1})$$

- This procedure is continued until all the magnitude bits of the weights are processed, yielding the weight-proportional voltage $V_{w,n}$.

$$V_{w,n} = V_{pre} \,.\, \sum_{k=1}^{n_w} b_k^n . 2^{k-n_w-1}$$

Analog Multiplication

- A multi-bit fixed point multiplication of input x_n with weight w_n can be represented as follows:

$$s^n_{\text{result}} \cdot |w_n \cdot x_n| = s^n_{\text{result}} \cdot |w_n| \cdot \sum_{p=1}^{n_x} \left(i^n_p \cdot 2^{-p}\right).$$

- The multiplication is carried out successively in n_x multiply and add steps while going through the input bits one by one.
- Depending on the input bit, in every three cycles, if the input bit is 1, the capacitor produces the weight proportional voltage $V_{w,n}$, else 0. This binary multiplication is then accumulated via charge-sharing on a dedicated capacitor $C_{out,n}$.

$$V^p_{C,out,n} = 0.5.\ (i^n_p.V_{w,n} + V^{p-1}_{C,out,n})$$

- The input bits are traversed LSB to MSB to ensure that added charge corresponds to the respective bit's significance. The valid bit d_{valid} indicates that the correct voltage $V_{w,n}$ is available after which the accumulation is initiated. The final equation for output voltage $V_{C,out,final,n}$ is:

$$V_{C,out,final,n} = s^n_{result} \frac{x_n}{2^{n_x}} \cdot \frac{w_n}{2^{n_w}} . V_{pre} + V_{CM}$$

Analog Accumulator

- After the multiplication of all bits of input vector element with the analog weight values have been executed for all elements of the vector and one column of the weight matrix, results are summed up along each column.
- All output capacitors for one column are shorted to one node V_{col} based on switch controlled by φ_{acc} signal. Final value of this is obtained to be:

$$V_{col} = \frac{1}{N} \cdot \sum_{n=1}^{N} V_{C,out,final,n}$$

- The number of cycles needed for DAC conversion and analog multiplication respectively are as follows:

$$n_{\mathrm{cyc,w}} = n_{\mathrm{w}} + 1. \qquad n_{\mathrm{cyc,i}} = 3 \cdot (n_{\mathrm{x}} - 1) + 1.$$

- The total number of clock cycles needed for the in-memory MAC operation are:

$$\begin{aligned} n_{\mathrm{cyc}} &= n_{\mathrm{cyc,w}} + n_{\mathrm{cyc,i}} + n_{\mathrm{cyc,acc}} + n_{\mathrm{cyc,adc}} + n_{\mathrm{cyc,rst}} \\ &= n_{\mathrm{w}} + 3 \cdot n_{\mathrm{x}} + 2 \end{aligned}$$

3-Bit IMCU

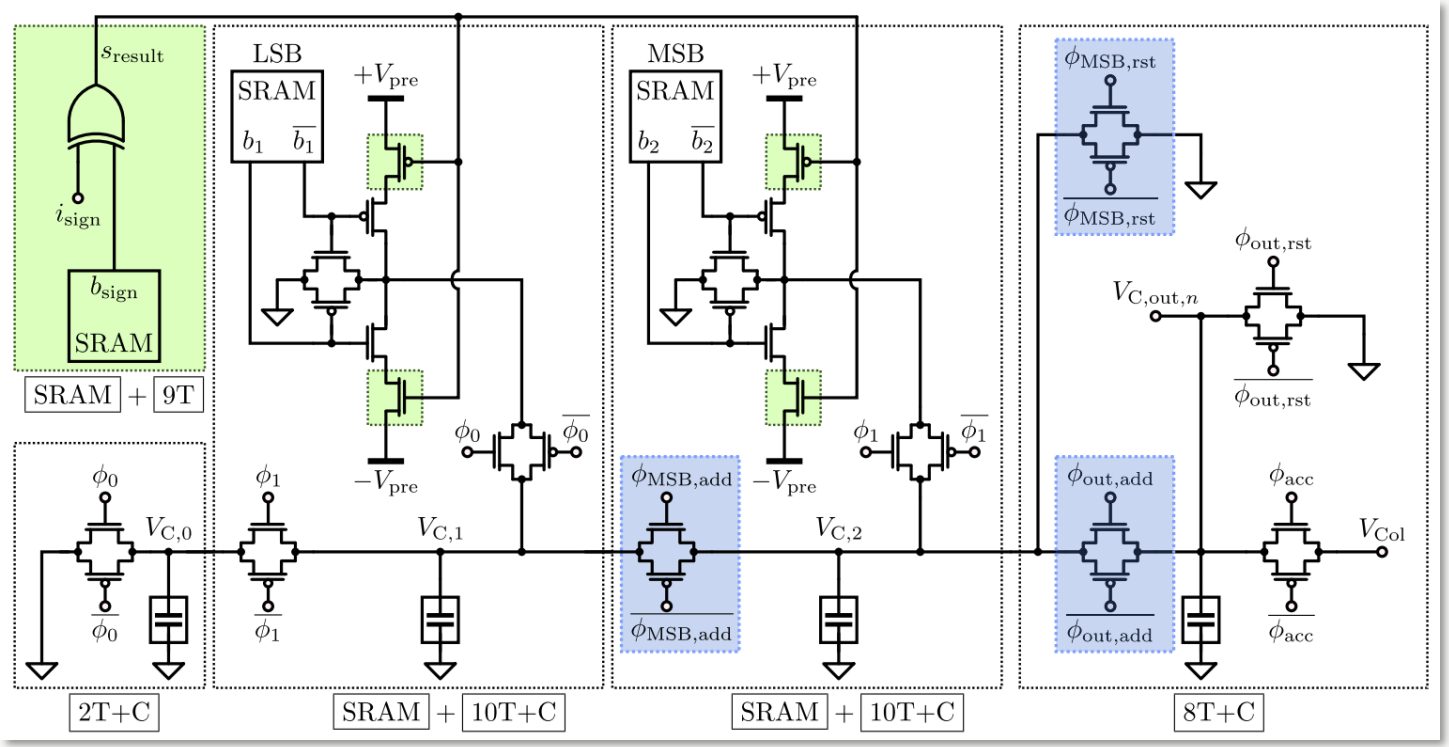


Figure : Given circuit of a 3-bit signed IMCU

3-Bit IMCU

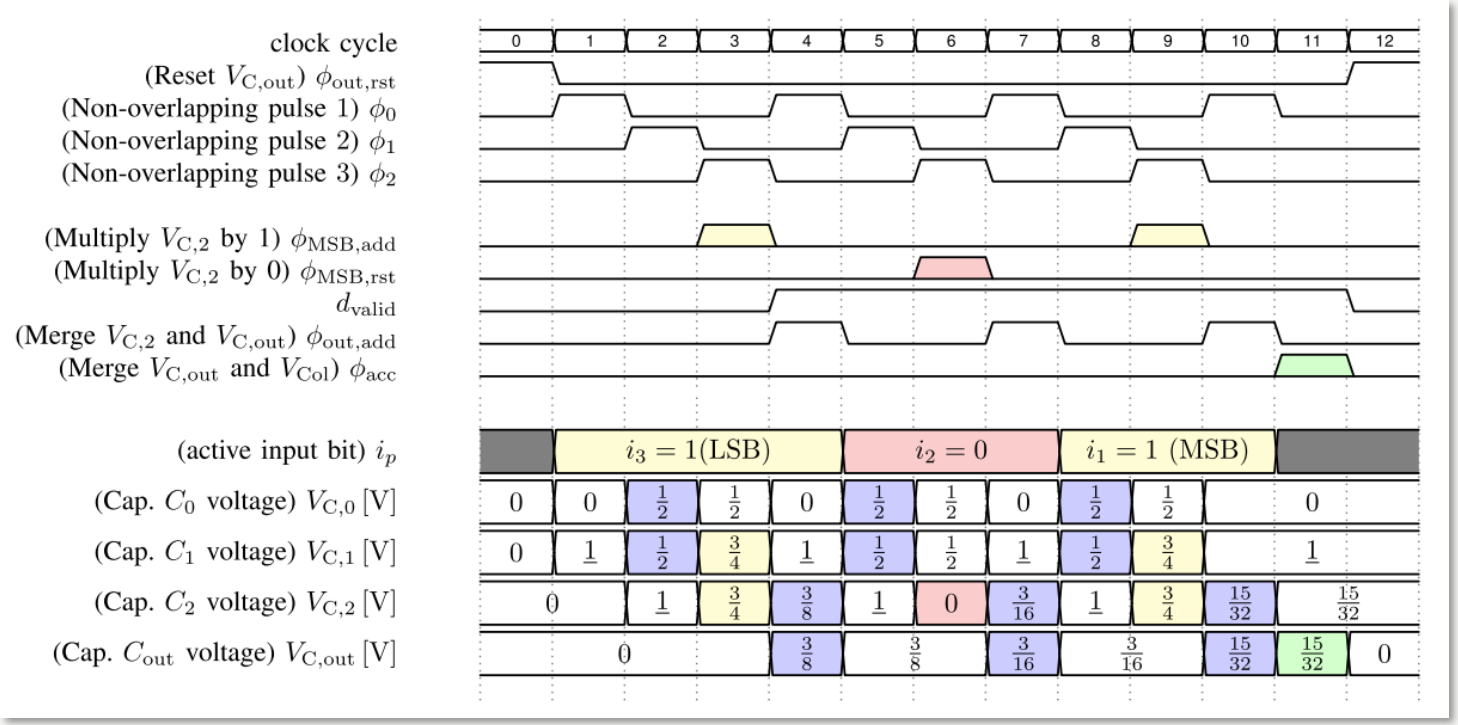


Figure : All signals for inputs and outputs of IMCU

Energy consumption and Power

	clock cycle												
	1	2	3	4	5	6	7	8	$\cdots$	$n_{\mathrm{cyc}}-3$	$n_{\mathrm{cyc}}-2$	$n_{\mathrm{cyc}}-1$	n_{cyc}
$V_{\mathrm{C},0}$	–	■	–	–	■	–	–	■		■	–	–	■
$V_{\mathrm{C},1}$	①	■	■	②	■	■	②	■		■	■	②	■
$V_{\mathrm{C},2}$	■	①	■	■	②	■	■	②		②	■	■	②
$V_{\mathrm{C},3}$	■	■	①	■	■	②	■	■		■	②	■	■
$V_{\mathrm{C},4}$	①	■	■	①	■	■	②	■		■	■	②	■
$V_{\mathrm{C},5}$	■	①	■	■	①	■	■	②		②	■	■	②
$V_{\mathrm{C},6}$	■	■	①	■	■	①	■	■		■	②	■	■
$\vdots$													
$V_{\mathrm{C},n_{\mathrm{w}}-2}$	①	■	■	①	■	■	①	■		■	■	②	■
$V_{\mathrm{C},n_{\mathrm{w}}-1}$	■	①	■	■	①	■	■	①		③	■	■	③
$V_{\mathrm{C},n_{\mathrm{w}}}$	■	–	①	■	–	①	■	–		■	④	■	■
$V_{\mathrm{C,out}}$	–	–	–	–	–	–	–	–		■	–	–	■

■	Charge-sharing
	Precharge cycles (potential capacitive energy consumption)
①	Capacitor precharge event (Initialization)
②	LSB to MSB-1 capacitor precharge event (Steady-state)
③	MSB-1 capacitor precharge event (Steady-state)
④	MSB capacitor precharge event (Steady-state)

$$E_{\mathrm{C},k} = \frac{1}{2} \cdot C_{\mathrm{unit}} \cdot b_k \cdot (V_{\mathrm{pre}} - V_{\mathrm{C,L}})^2$$

Energy consumption and Power

- On analyzing the IMCU during one operational cycle, energy consumption occurs when – initial charge or discharge of capacitors, two charge-sharing procedures: first with previous and then with next capacitor, which consumes no energy expect for switching of TGs.
- The types of energy consumptions can be summarized as follows:

1) Initialization Pre-charge Events: Refers to energy drawn in initial pre-charge cycles, when the capacitors getting charged do not contain LSB information.

$$E_{(1)} = \frac{C_{\text{unit}}}{2} \sum_{r=1}^{n_{\text{r,init}}} \sum_{k_{\text{x}}=3r-2}^{n_{\text{w}}} b_{k_{\text{x}}} (V_{\text{pre}} - V_{\text{C,L,i}}(r, k_{\text{x}}))^2.$$

2) Steady-State Pipeline D/A Pre-charge Events: Refers to all capacitors pre-charging due to switching of φ_0, φ_1 and φ_2, which don't include the MSB-1 and the MSB capacitors.

$$E_{(2)} = \frac{C_{\text{unit}}}{2} \sum_{r=1}^{n_{\text{r,st}}} \sum_{k_{\text{x}}=1}^{n_{\text{d,st}}(r)} b_{k_{\text{x}}} (V_{\text{pre}} - V_{\text{C,L}}(1, k_{\text{x}} + 1))^2.$$

3) MSB-1 Capacitor Pre-charge Events: Refers to energy consumed during pre-charging of MSB-1 capacitor which depends on input bits i_p.

$$E_{(3)} = \frac{C_{\text{unit}}}{2} \sum_{x=1}^{n_{\text{x}}} b_{n_{\text{w}}-1} \cdot (V_{\text{pre}} - V_{\text{C,L,MSB}-1}(i_{\text{x}}))^2.$$

Energy consumption and Power

4) MSB Capacitor Pre-charge Events: Refers to energy consumed during pre-charging of MSB capacitor which depends on current input bit i_p and all previously encountered input bits.

$$E_{(4)} = \frac{C_{\text{unit}}}{2} \cdot b_{n_{\text{w}}-1} \cdot \sum_{p=1}^{n_{\text{x}}} (V_{\text{pre}} - V_{\text{C,out,L}}(p))^2.$$

5) Switching of TGs: Refers to all energy consumed during every transition of transmission gate. This can be found by calculating energy required for single transient of TG and then adding them up for all switching events. Example of this for φ_0 is as follows:

$$n_{\text{TG},\phi_0} = 2 + 2 \cdot \left\lfloor \frac{1}{3} \cdot (n_{\text{w}} - 1) \right\rfloor.$$

$$n_{\text{TG,ev},\phi_0} = \left\lfloor \frac{1}{3} \cdot (n_{\text{cyc}} + 2) \right\rfloor.$$

$$n_{\text{TG,ev},\phi_0} = \left\lfloor \frac{n_{\text{cyc}} + 2}{3} \right\rfloor \cdot \left(2 + 2 \cdot \left\lfloor \frac{n_{\text{w}} - 1}{3} \right\rfloor \right)$$

3-Bit IMCU

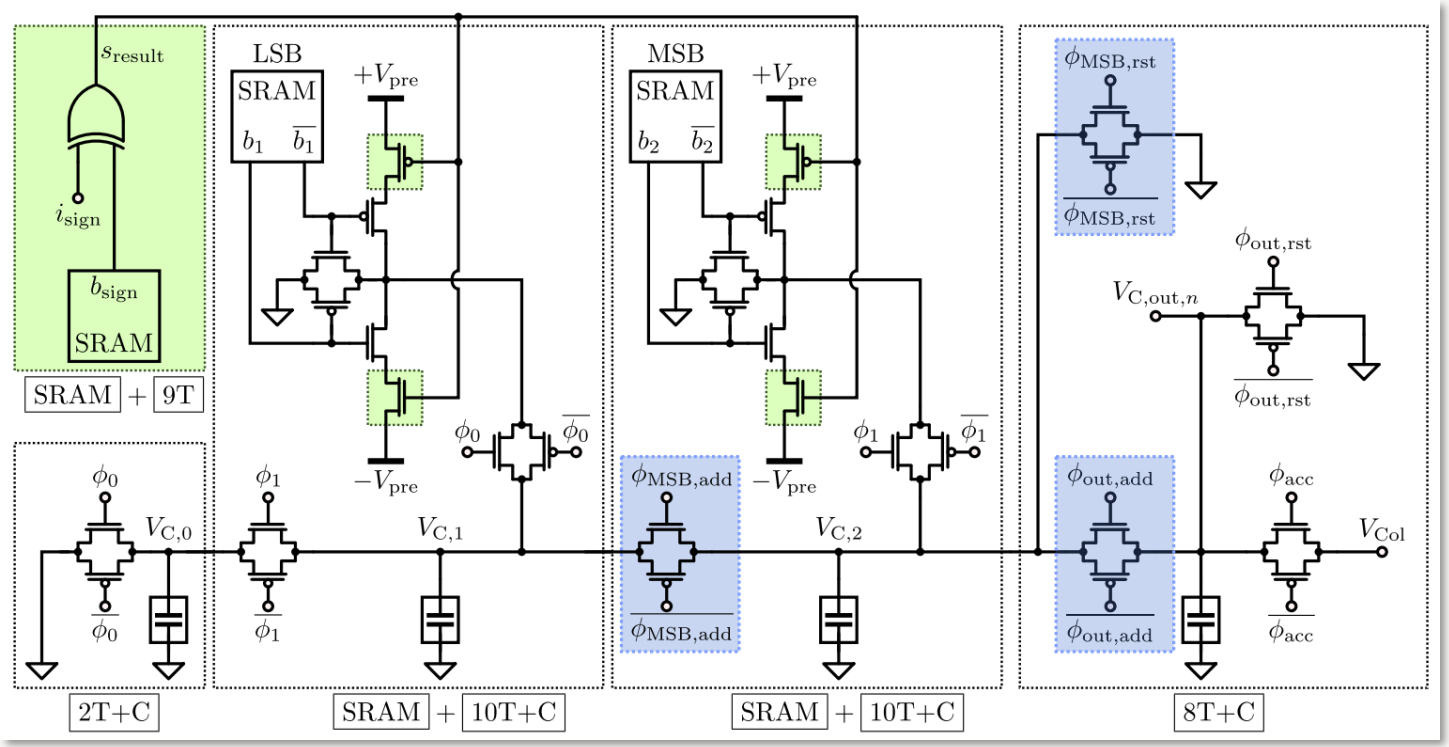
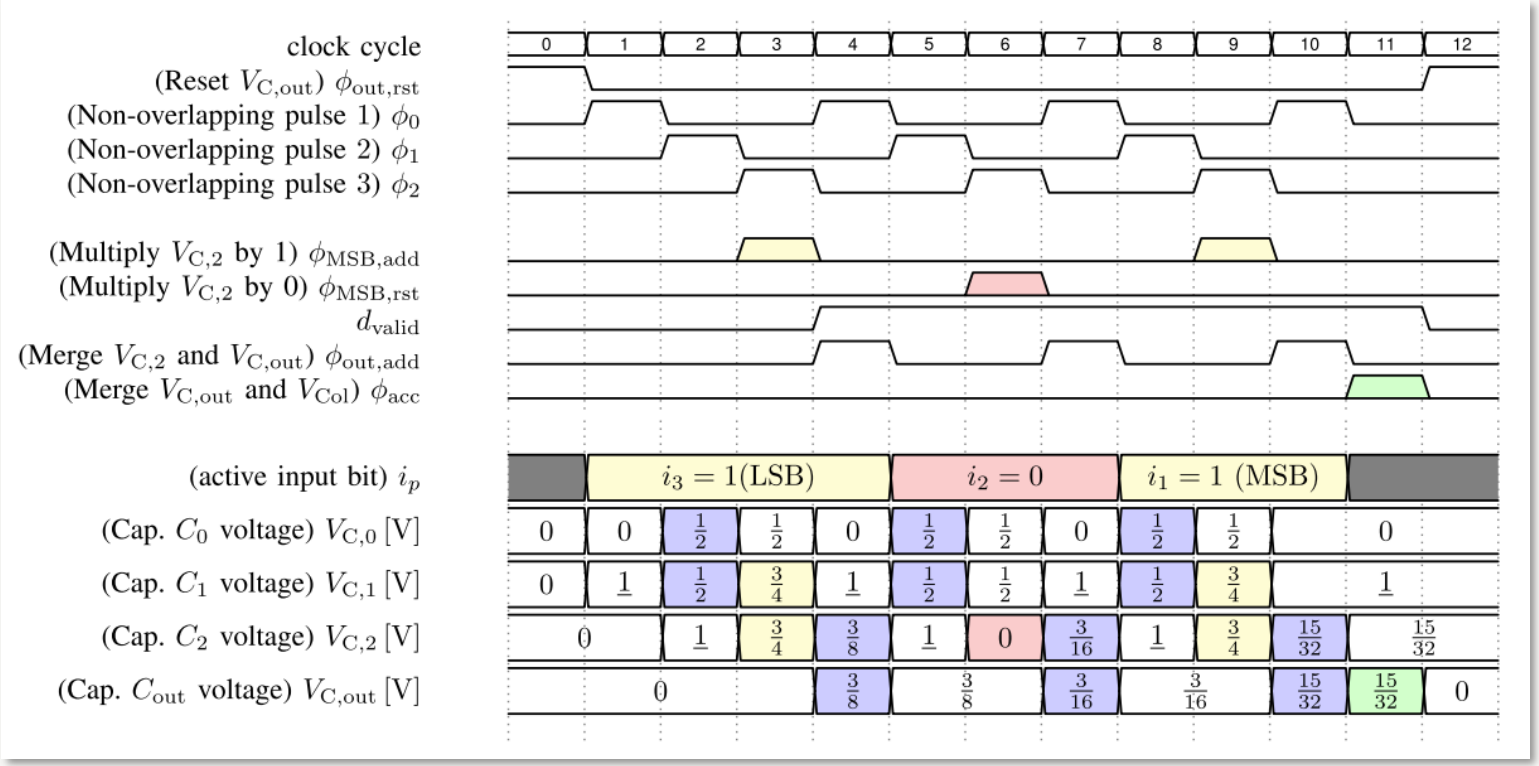


Figure : Given circuit of a 3-bit signed IMCU

3-Bit IMCU


clock cycle
0 1 2 3 4 5 6 7 8 9 10 11 12
(Reset $V_{C,out}$) $\phi_{out,rst}$
(Non-overlapping pulse 1) ϕ_0
(Non-overlapping pulse 2) ϕ_1
(Non-overlapping pulse 3) ϕ_2
(Multiply $V_{C,2}$ by 1) $\phi_{MSB,add}$
(Multiply $V_{C,2}$ by 0) $\phi_{MSB,rst}$
d_{valid}
(Merge $V_{C,2}$ and $V_{C,out}$) $\phi_{out,add}$
(Merge $V_{C,out}$ and V_{Col}) ϕ_{acc}
(active input bit) i_p
$i_3 = 1$(LSB)
$i_2 = 0$
$i_1 = 1$ (MSB)
(Cap. C_0 voltage) $V_{C,0}$ [V]
0 0 $\frac{1}{2}$ $\frac{1}{2}$ 0 $\frac{1}{2}$ $\frac{1}{2}$ 0 $\frac{1}{2}$ $\frac{1}{2}$ 0
(Cap. C_1 voltage) $V_{C,1}$ [V]
0 1 $\frac{1}{2}$ $\frac{3}{4}$ 1 $\frac{1}{2}$ $\frac{1}{2}$ 1 $\frac{1}{2}$ $\frac{3}{4}$ 1
(Cap. C_2 voltage) $V_{C,2}$ [V]
0 1 $\frac{3}{4}$ $\frac{3}{8}$ 1 0 $\frac{3}{16}$ 1 $\frac{3}{4}$ $\frac{15}{32}$ $\frac{15}{32}$
(Cap. C_{out} voltage) $V_{C,out}$ [V]
0 $\frac{3}{8}$ $\frac{3}{8}$ $\frac{3}{16}$ $\frac{3}{16}$ $\frac{15}{32}$ $\frac{15}{32}$ 0

Energy consumption and Power

	1	2	3	4	5	6	7	8	9	10	11	12
$V_{c,0}$	–		–	–		–	–		–	–		–
$V_{c,1}$	1			3			3			3		
$V_{c,2}$	–	1			4			4			4	
$V_{c,out}$	–	–	–		–	–		–	–		–	–

Noise and Non-linearity

- Effect of R-TG : We make sure transmission gate $R_{TG\text{-}ON}$'s effect is negligible by designing clock time period to ensure complete voltage settling.

$$R_{\mathrm{TG,on}} \cdot C_{\mathrm{unit}} \cdot n_{\mathrm{acc,mac}} \cdot \ln 2 < T_{\mathrm{cycle}}.$$

- Thermal noise
- Manufacturing differences in capacitances.

03

Implementation

6T-SRAM Block

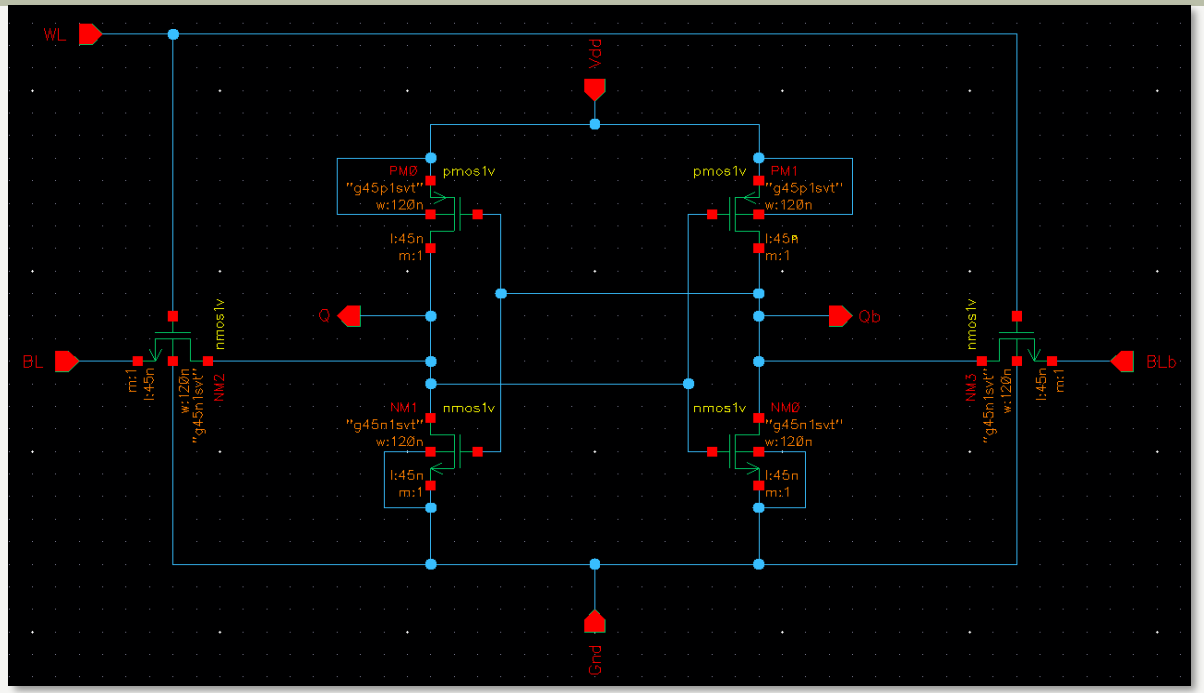


Figure 4: Circuit of a 6T-SRAM memory unit

6T-SRAM Block

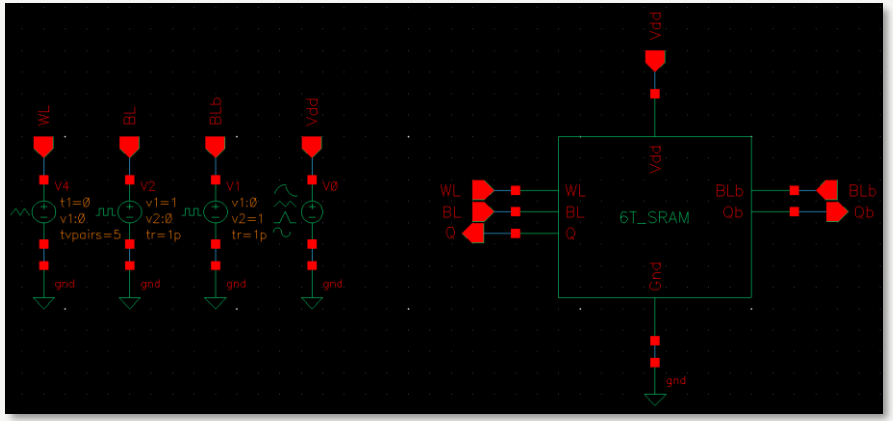


Figure 5: Test circuit for SRAM unit

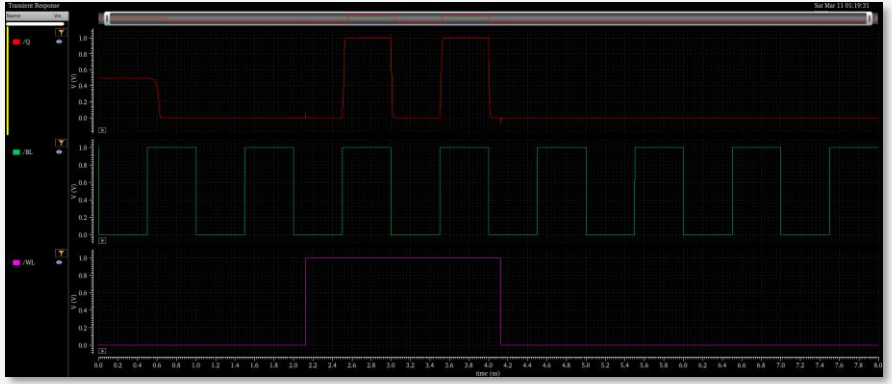


Figure 6: Graphs of inputs and outputs of SRAM test circuit

- This circuit was then tested by giving the a PWL input to the word line and a pulse input to the bit line.

- The output given by the SRAM block is as expected as seen in the graphs above.
- The output follows the bit line whenever word line is active and holds the bit line when the word line turns OFF

Implemented 3-Bit IMCU

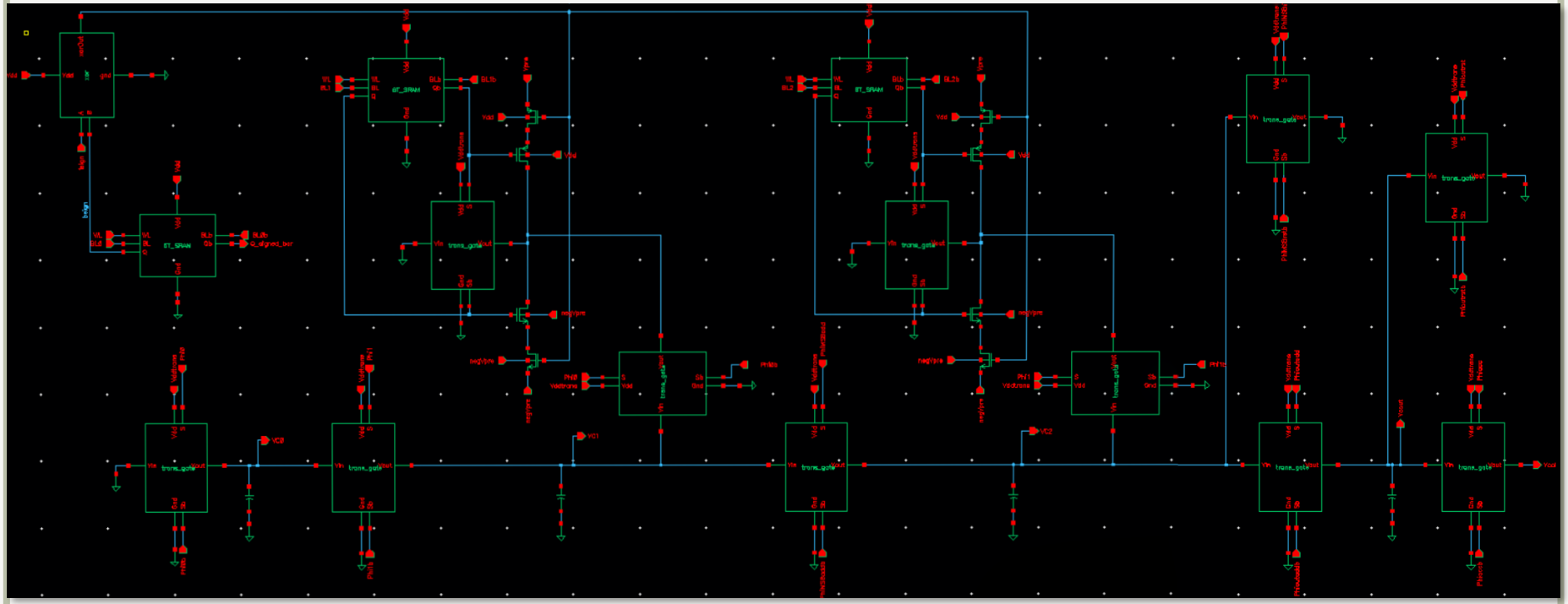


Figure : 3-bit IMCU implemented on Cadence Virtuoso

Extended to 4-bit and 5-bit

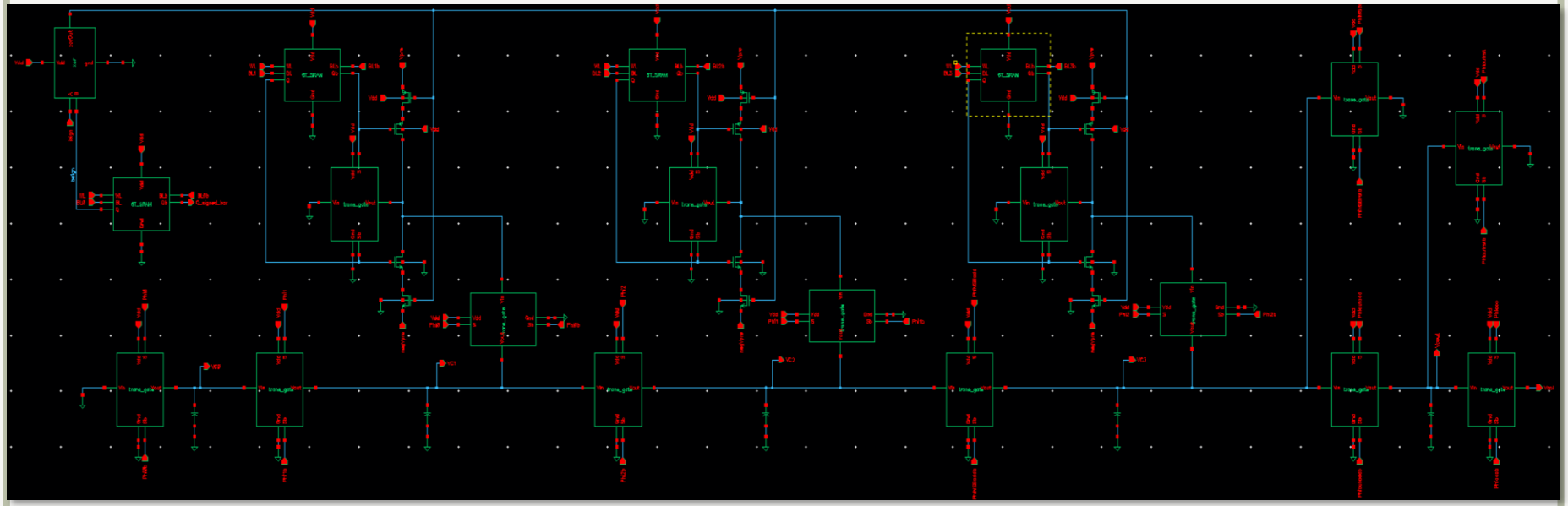


Figure : 4-bit IMCU implemented on Cadence Virtuoso

Extended to 4-bit and 5-bit

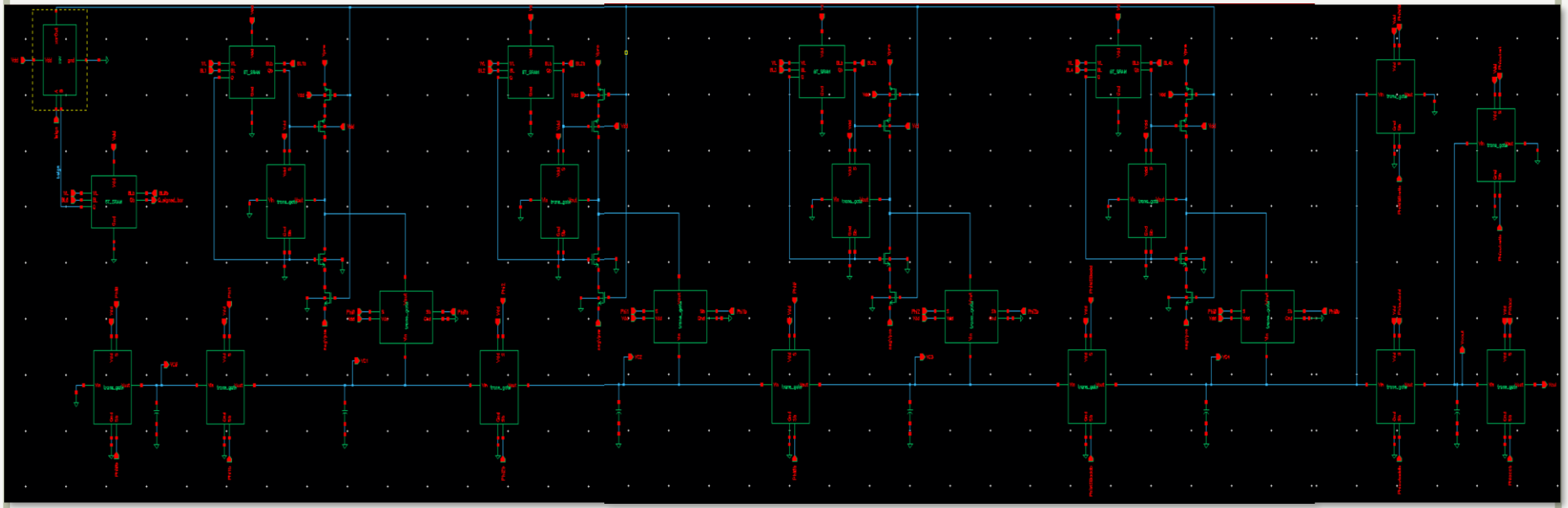


Figure : 5-bit IMCU implemented on Cadence Virtuoso

04

Results

Layout of 3-Bit IMCU

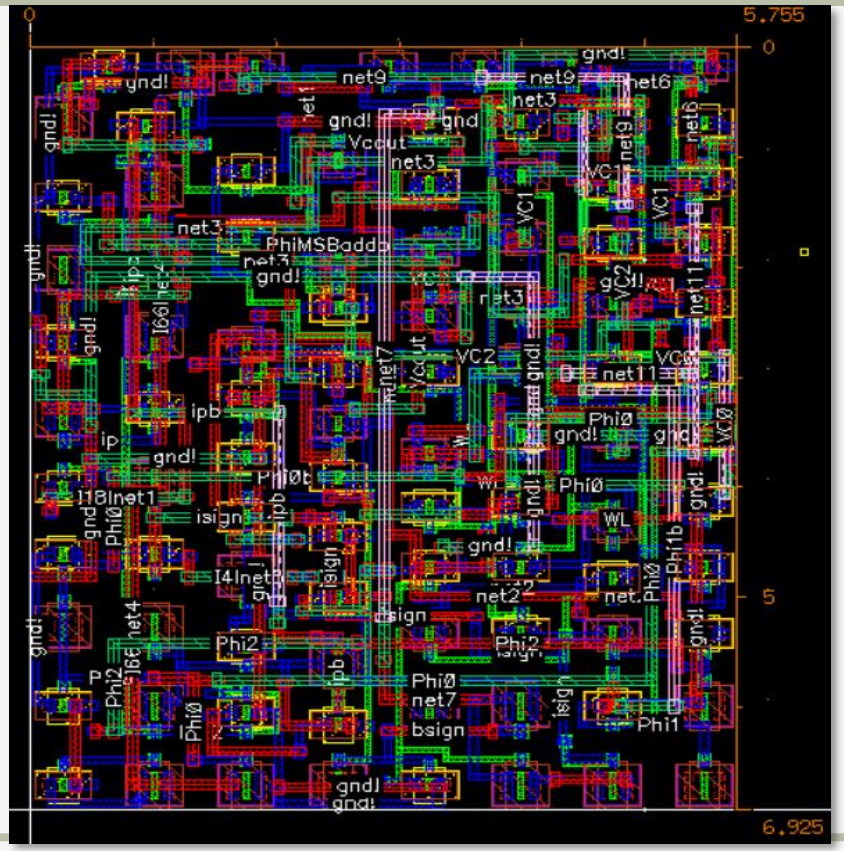
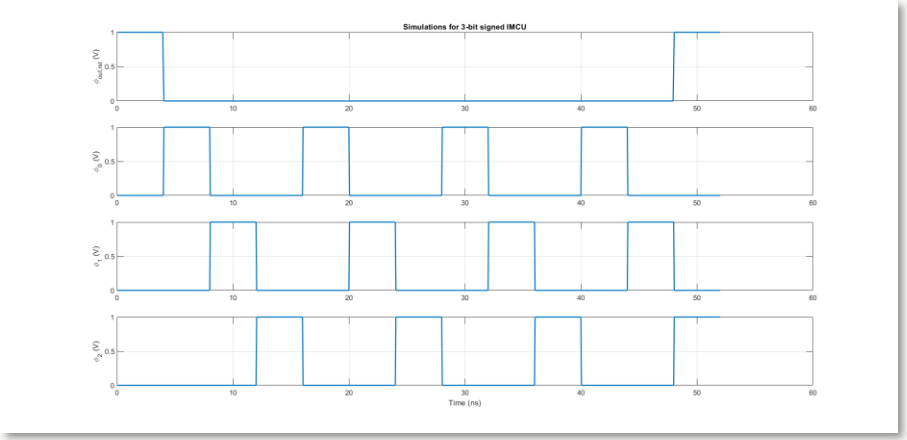
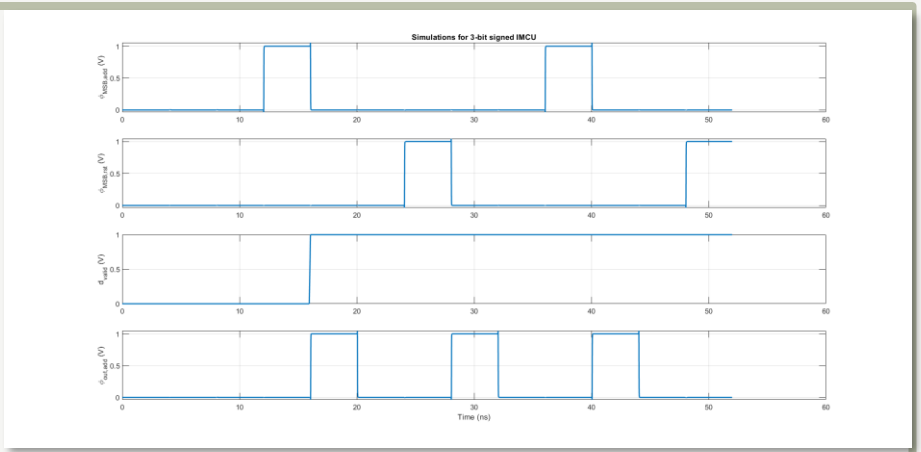
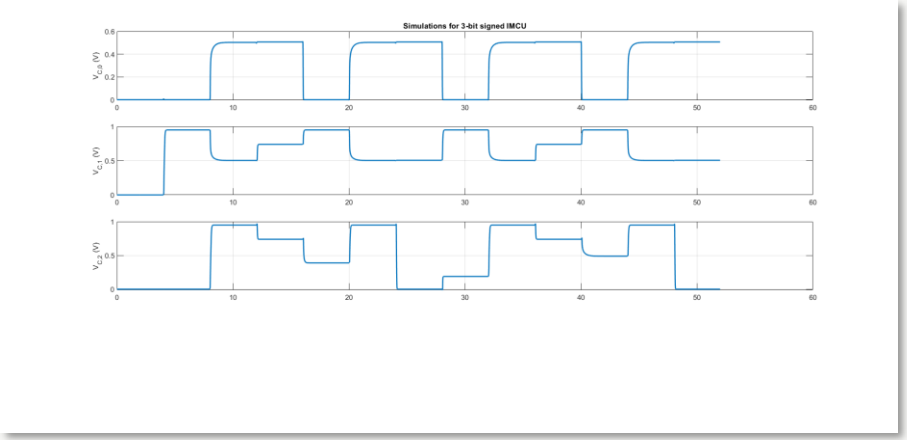


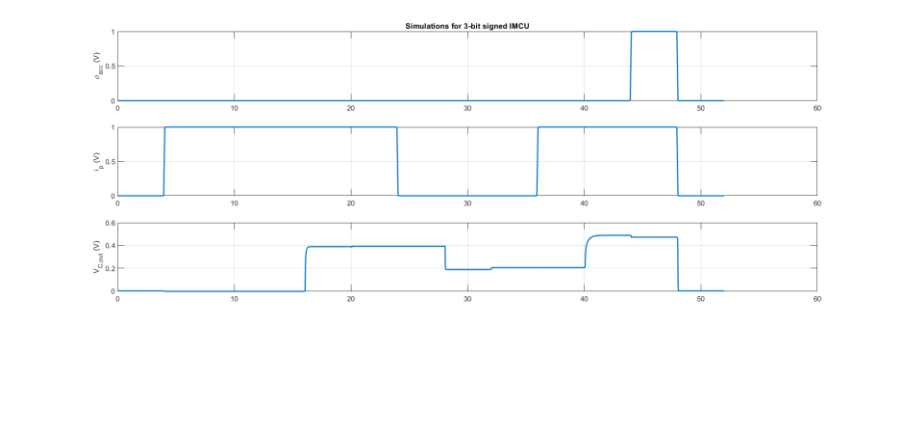
Figure : Layout of 3-bit IMCU implemented on Cadence Virtuoso





Graphs for 3-bit IMCU





Layout of 4-Bit IMCU

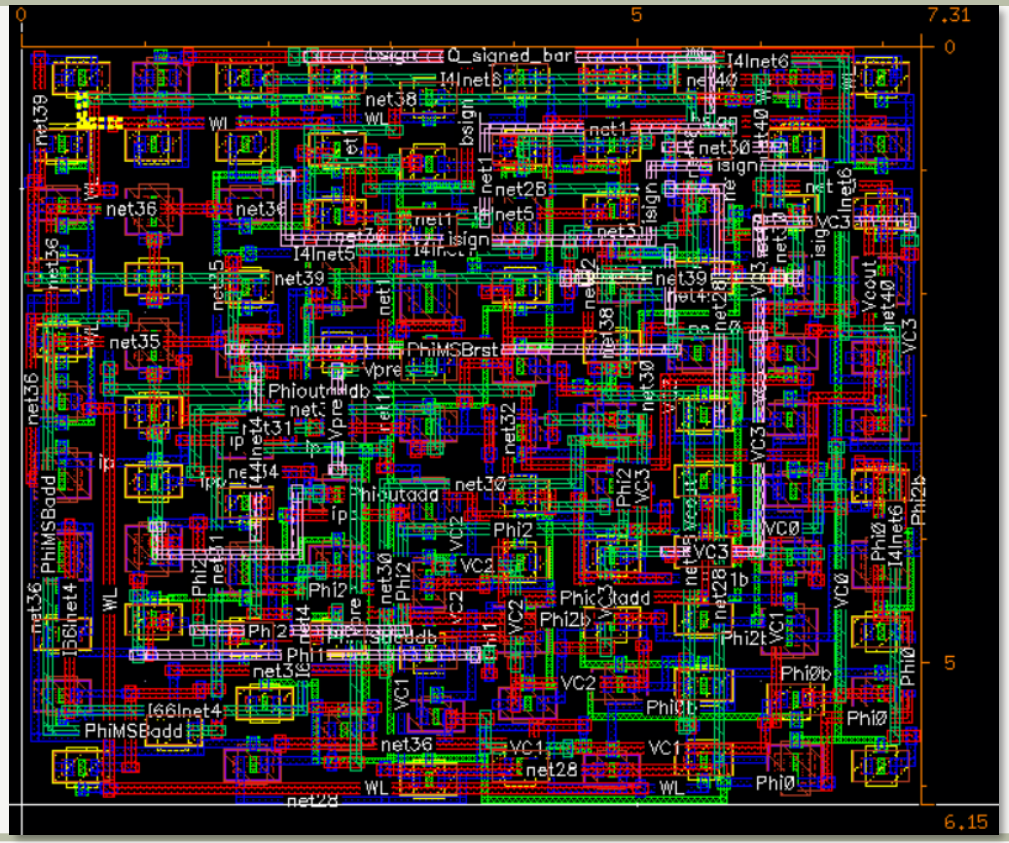
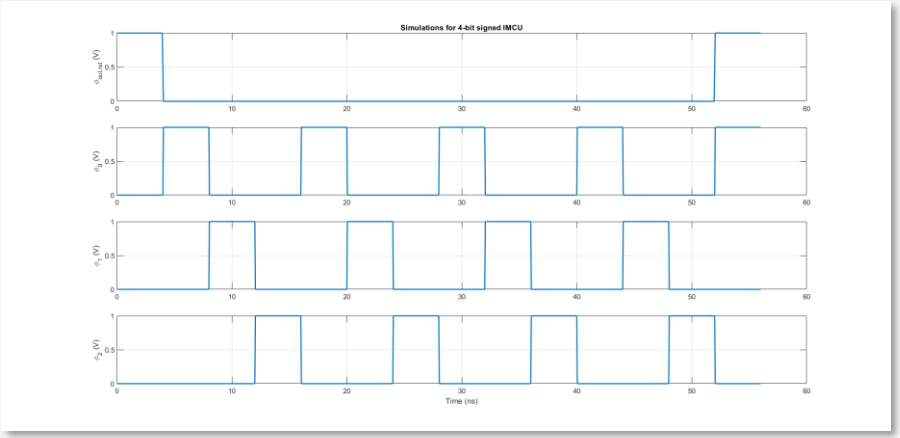
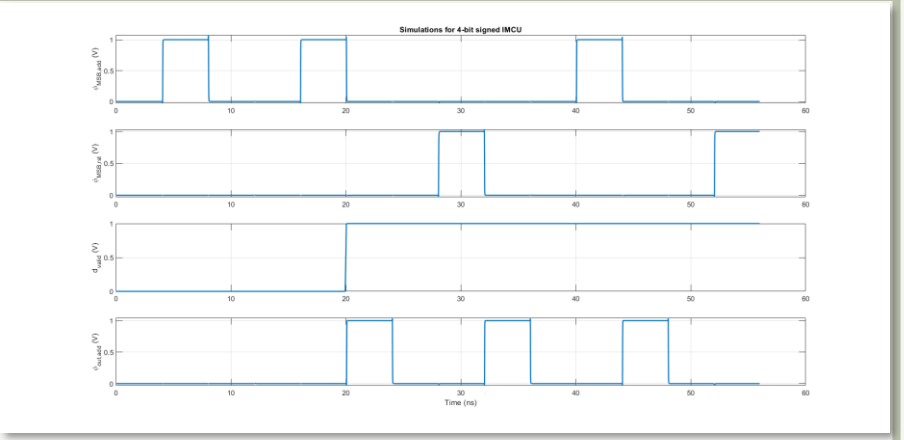
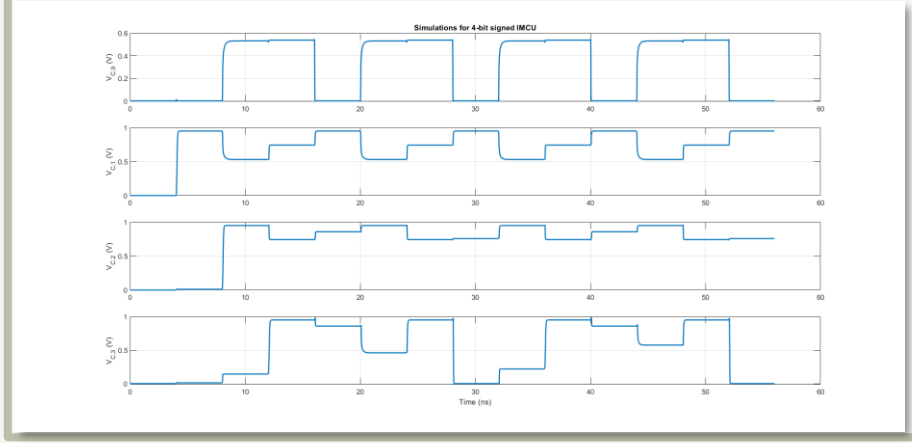


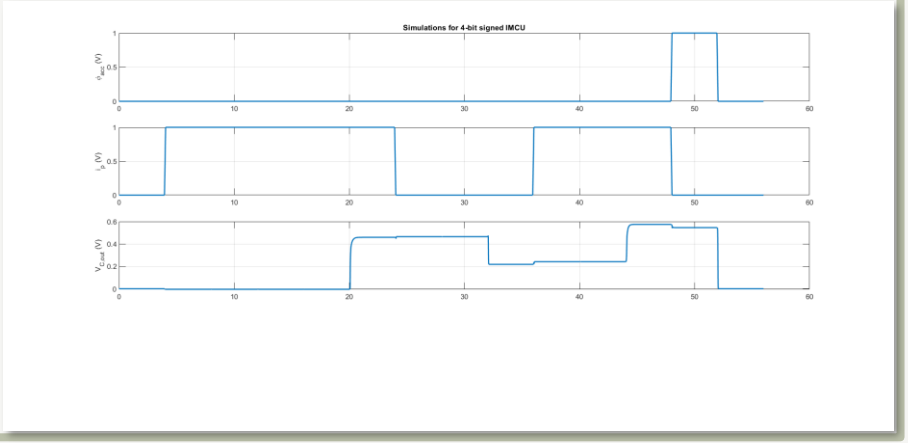
Figure : Layout of 4-Bit IMCU implemented on Cadence Virtuoso





Graphs for 4-bit IMCU





Layout of 5-Bit IMCU

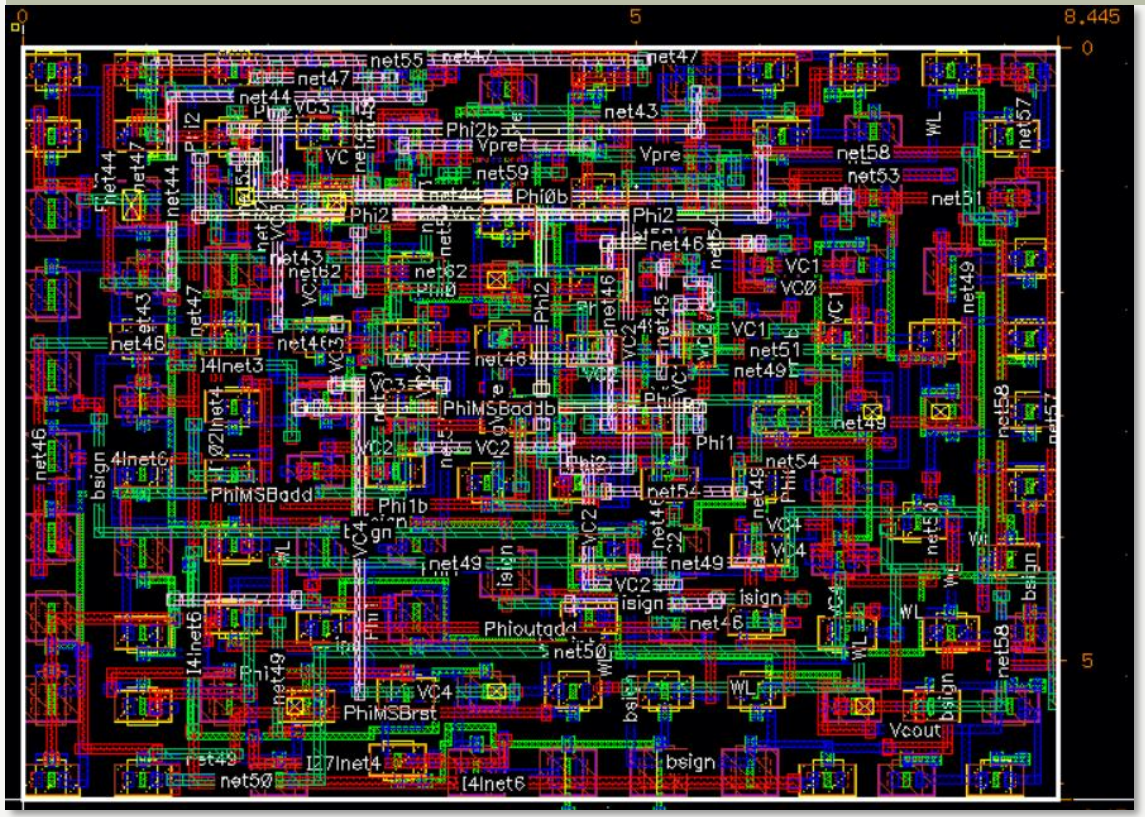
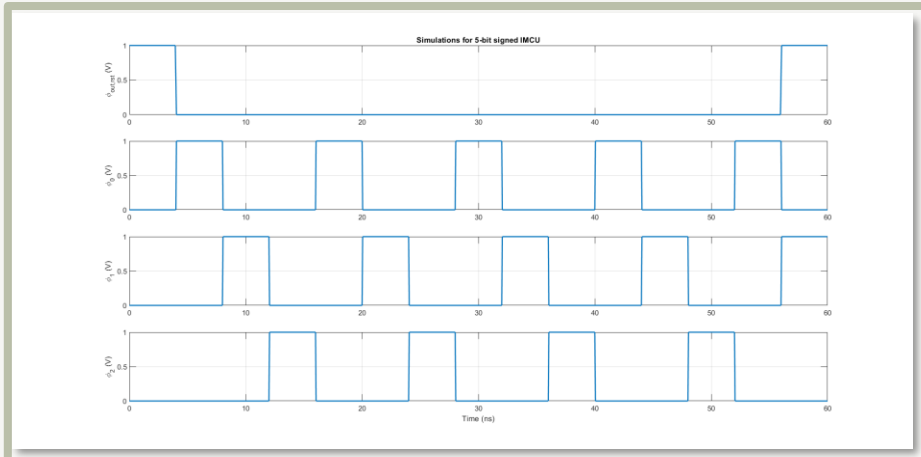
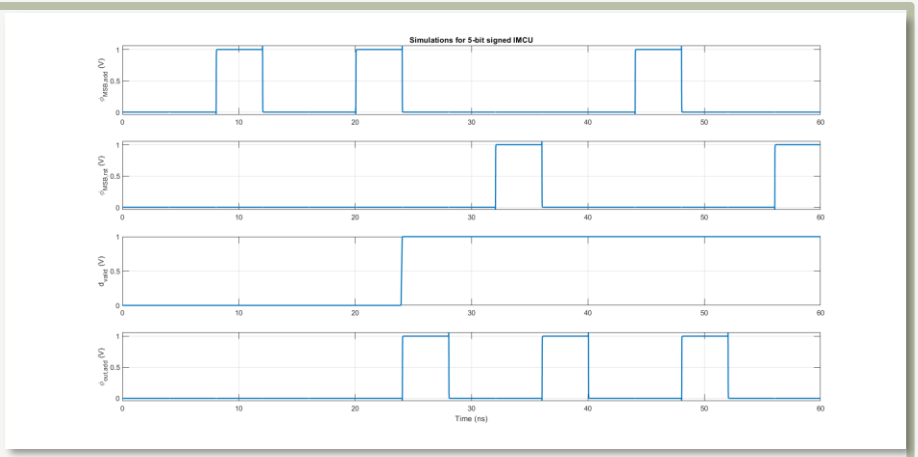
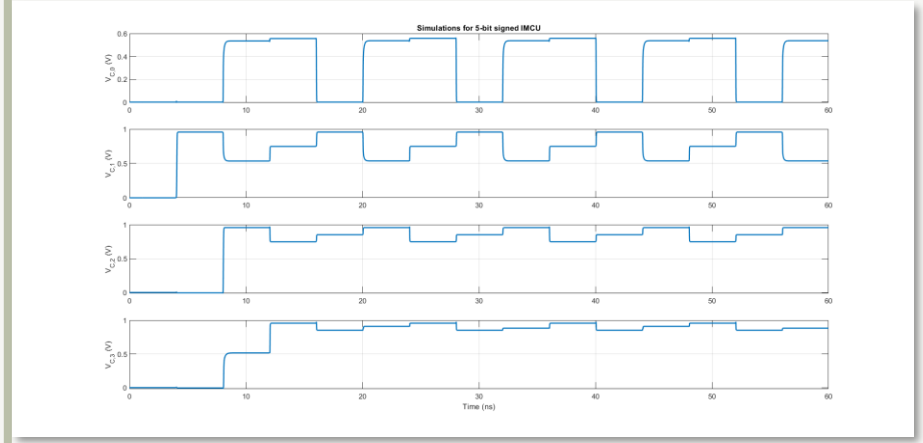


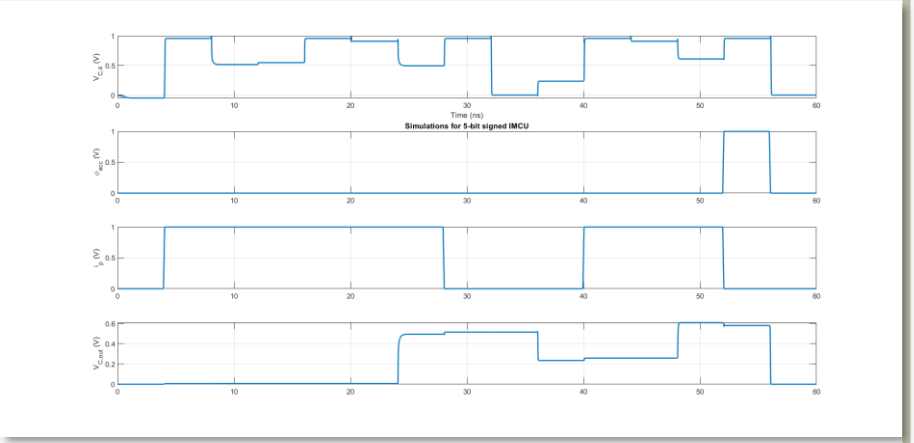
Figure : Layout of 5-Bit IMCU implemented on Cadence Virtuoso





Graphs for 5-bit IMCU





Final Results

Number of weight bits (n_w)	DELAY (I/P – O/P)	ENERGY (fJ)	LAYOUT	
			W x H	Area(μm^2)
3-bit	40 ns	$E_{IMCU,total} = E_{(1)} + E_{(2)} + E_{(3)} + E_{(4)} + E_{TG,total}$ 30.342 = 4.483 + 3.392 + 0 + 5.683 + 16.784	5.755 x 6.925	39.853
4-bit	44 ns	$E_{IMCU,total} = E_{(1)} + E_{(2)} + E_{(3)} + E_{(4)} + E_{TG,total}$ 45.561 = 6.7 + 3.5134 + 1.6479 + 5.282 + 28.418	6.15 x 7.31	44.956
5-bit	48 ns	$E_{IMCU,total} = E_{(1)} + E_{(2)} + E_{(3)} + E_{(4)} + E_{TG,total}$ 61.066 = 10.138 + 3.4231 + 0.7511 + 5.22 + 41.534	6.13 x 8.445	51.767

The results in the above table show that the properties vary linearly with n_w.

05

Conclusions

Conclusions

- The design implemented in this paper succeeds in obtaining linear increase in delay, energy and area with increase in number of bits.
- This was also verified in the simulations performed on 3-bit, 4-bit and 5-bit IMCUs.

06

Future Work

Future Work

- On simulating the circuit we found that for negative analog outputs, the widths for the PMOS and NMOS need to be varied, adjusting the strength of the pull-up and pull-down circuits. This issue can be resolved by using transmission gates in place of pass transistors at the cost of 4 additional transistors per bit.

- The paper also performs a full system implementation of the proposed design using SRAM arrays, an ADC converter and a separate module for the IMCUs. We plan on doing this hardware implementation and try incorporating the earlier proposed change.

References

1. P. Athe and S. Dasgupta, "A comparative study of 6T, 8T and 9T decanano SRAM cell," in Proc. ISIEA, vol. 2, Oct. 2009, pp. 889-894. - Binder1.pdf (ijltet.org)
2. F.-J. Wang, G. C. Temes, and S. Law, “A quasi-passive CMOS pipeline D/A converter,” IEEE J. Solid-State Circuits, vol. 24, no. 6, pp. 1752–1755, Dec. 1989. - A quasi-passive CMOS pipeline D/A converter | IEEE Journals & Magazine | IEEE Xplore
3. E. H. Lee and S. S. Wong, “Analysis and design of a passive switchedcapacitor matrix multiplier for approximate computing,” IEEE J. SolidState Circuits, vol. 52, no. 1, pp. 261–271, Jan. 2017. - Analysis and Design of a Passive Switched-Capacitor Matrix Multiplier for Approximate Computing | IEEE Journals & Magazine | IEEE Xplore
4. L. Kull et al., “A 24-to-72gs/s 8b time-interleaved SAR ADC with 2.0-to-3.3pj/conversion and 30db SNDR at Nyquist in 14nm CMOS FinFET,” in IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers, Feb. 2018, pp. 358–360. - A 24-72-GS/s 8-b Time-Interleaved SAR ADC With 2.0-3.3-pJ/Conversion and >30 dB SNDR at Nyquist in 14-nm CMOS FinFET | Request PDF (researchgate.net)

Q&A