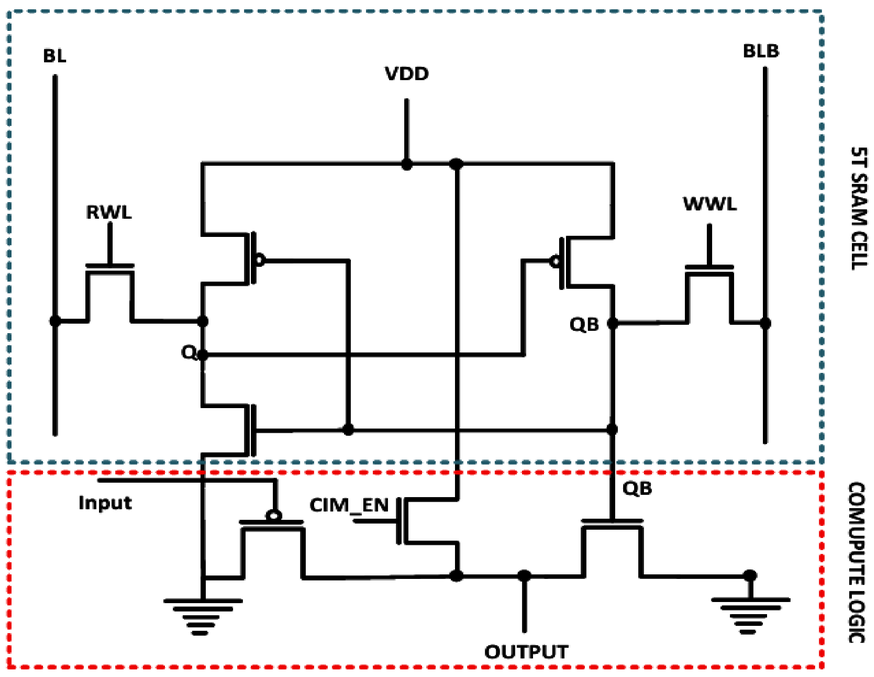
A NOR8T SRAM Digital Compute-in-memory Macro for Sparse and Scalable edge-AI Processing

Pratham Sharma , Mukul Lokhande , Akash Sankhe , Kwok Tai Chui , Brij Bhooshan Gupta , and Santosh Kumar Vishvakarma

Introduction

- The rise of ML algorithms has necessitated a steep growth in the need for computational complexity, highlighting the von Neumann bottlenecks between AI computing hardware and data storage.
- The idea behind compute in-memory is simple: instead of having distinct compartments for memory and processing, the operations are performed in memory itself.

Proposed NOR8T Bit-cell



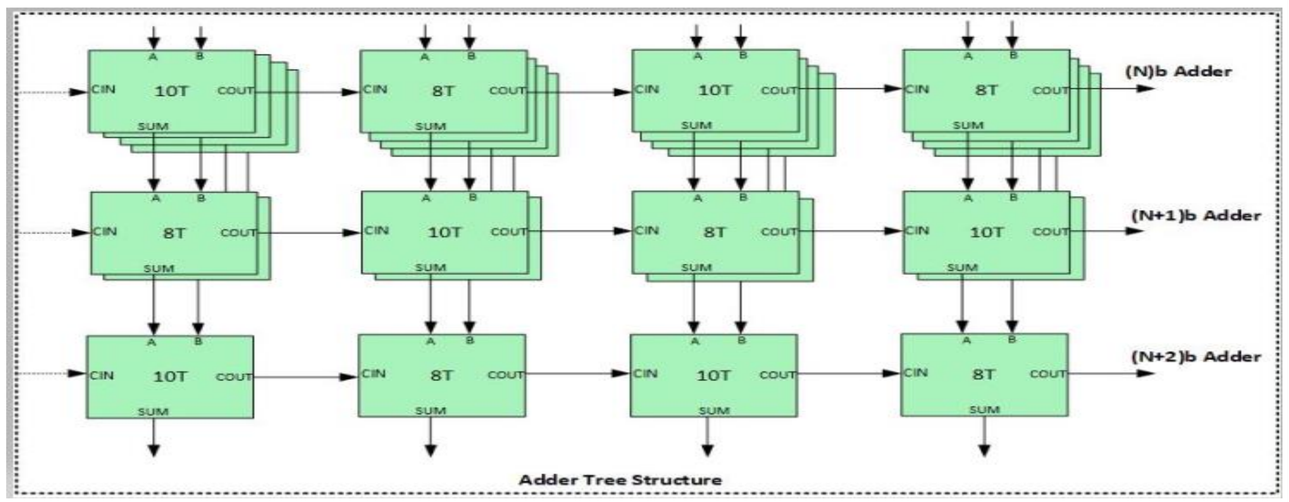
Features of the proposed Bit-cell :

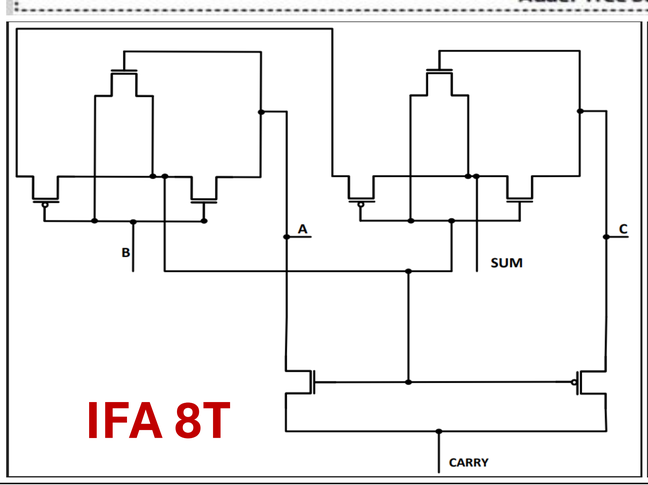
- 5T SRAM Cell is used for storing the data and 3T is used for NOR logic Operation.
- Performs the NOR operation between input and storage (QB) node of the SRAM cell.
- Reduces 4T per bit-cell as compared to conventional NOR based bit-cell as it uses 12T.

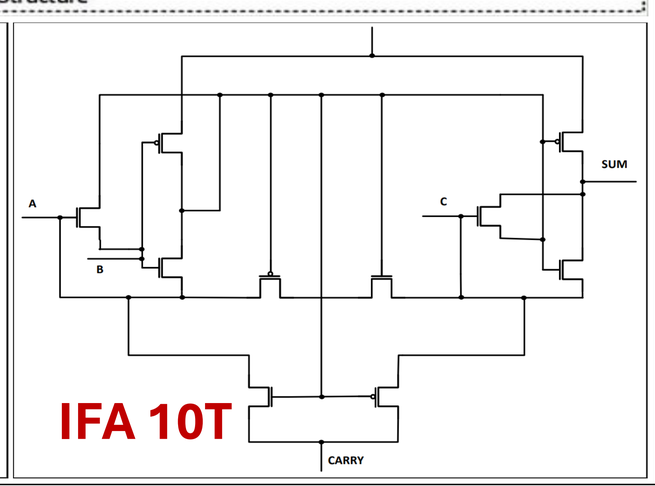
$$\overline{\overline{input} + \overline{weight}} = input.weight$$

Working Operation Of NOR based MAC

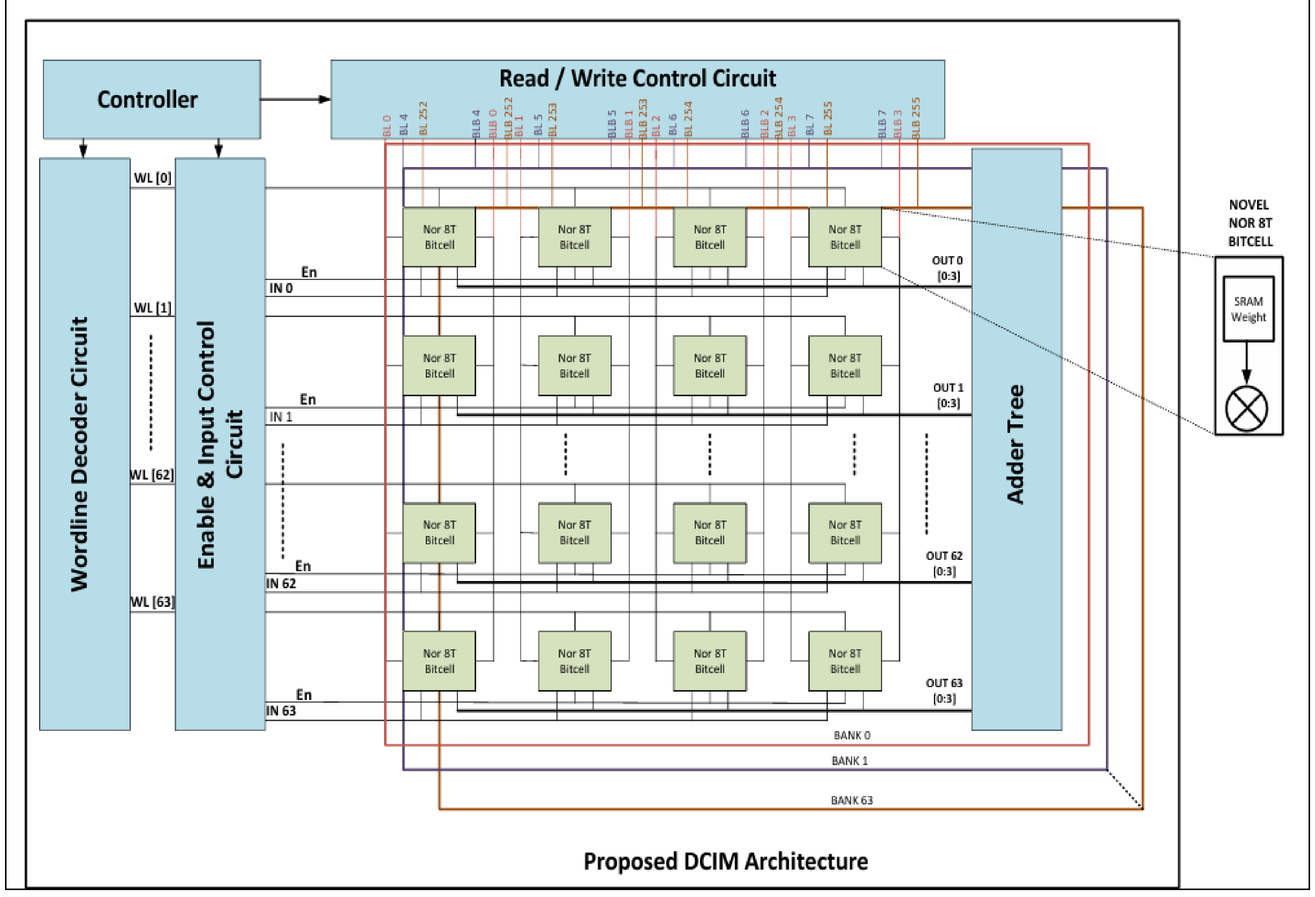
Proposed Adder-Tree Architecture







Proposed DCIM Architecture



Proposed DCIM Architecture

Performance Comparison of Bit-cell with SOTA Works

Parameters	6T	8T	5T [14]	6T+4T [15]	1R1W [16]	MUL8T [18]	XNOR [17]	NOR8T
Area (µm^2)	1.75	2.35	1.55	3.67	4.2	2.52	3.78	**2.43**
Power (µW)	20.1	23.4	17.1	27.4	39.8	34	37.3	**30.92**
Operation	Storage	Storage	Storage	Storage, CIM-NOR	Storage, CIM-NOR	Storage, CIM-AND	Storage XAC/MAC	**Storage, CIM-NOR**
Layout Density	High	High	Irregular	Low	Low	Medium	Medium	**High**
Read Delay (ps)	106.2	102.7	104.6	127.5	158	108.2	134	**121.3**
Write Delay (ps)	169.6	164.3	157.2	216.7	282	173.2	237.6	**173.9**

Overall Performance Comparison

Parameters	JSSC'19 [25]	ISSCC'21 [15]	JETCAS'22 [11]	TNano'23 [17]	TCAS-I'23 [26]	TCAS-I'24 [13]	ISQED'25 [8]	DATE'25 [19]	Proposed
Tech. (nm)	65	22	65	65	65	55	65	65	65
MAC-operation	AMS	Digital	Analog	Digital	Digital	Digital	Digital	Digital	Digital
Cell-Type	10T	6T + 4T	AND8T	10T	7T	8T	MUL8T	6T + 2T	NOR8T
Supply Voltage VDD (V)	0.8-1	0.72	1	1.2	0.8-1.1	1.2	1.2	0.6-1.2	0.8-1.2
Array Size (Kb)	16	64	16	16	80	64	16	4	16
Op. Freq. (MHz)	5	100	100	25	200	200	250	40	66
Bit-cell Area (µm²)	10.53	0.37	2.78	4.5	2.83	4.21	4.1	2.25	2.43
Input Precision (b)	6	1-8	4	4	1-16	4/8/12/16	1-8	4/8	1-16
Weight Precision (b)	1	4/8/12/16	4	1-4	1-16	4/8/12/16	4/8	4/8/12/16	1-16
Model	LeNet-5	NA	VGG-8	CNN-Type	Inception-v4	ResNet-10	LeNet-5	NA	LeNet-5
Accuracy (%)	98.3	NA	96.05	98.67	95.3	93.7	99.1	98.5	99.14
Throughput (TOPS)	0.64	0.52*	0.41	0.82	0.41*	0.43*	2.2	2.52*	1.63
Energy Efficiency (TOPS/W)	50.6	23	180	273	63	67	480	404	321

Conclusion

- This work presents a DCIM macro designed to address the challenges in state-of-the-art works, specifically focusing on reducing the overall area utilisation of the SRAM array and the adder tree.
- We present a novel NOR8T SRAM bitcell for compute-in-memory operation and two full adders, 8T and 10T, used in the 2D interleaving adder tree. which mitigates the threshold voltage drop issue of the bitcell and the voltage drop from the RCAs. The design has 68% area utilisation and 57% power reduction in comparison to conventional adder tree designs.
- The design achieves a maximum throughput of 1.63 TOPS at 66 MHz and at 1.2 V utilising CMOS 65 nm technology.
- The macro is evaluated for 4A4W configurations for the LeNet-5 network using MNIST datasets, achieving 99.14% inference accuracy.

Contact information for Q/A and suggestions:
Pratham Sharma: ms2404102001@iiti.ac.in
Dr. Santosh Kumar Vishvakarma : skvishvakarma@iiti.ac.in