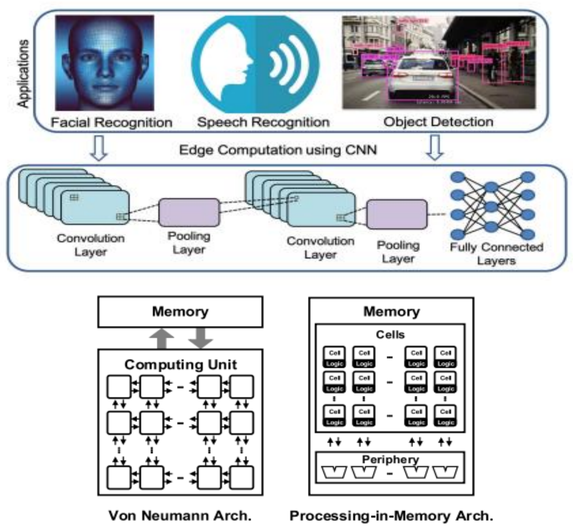
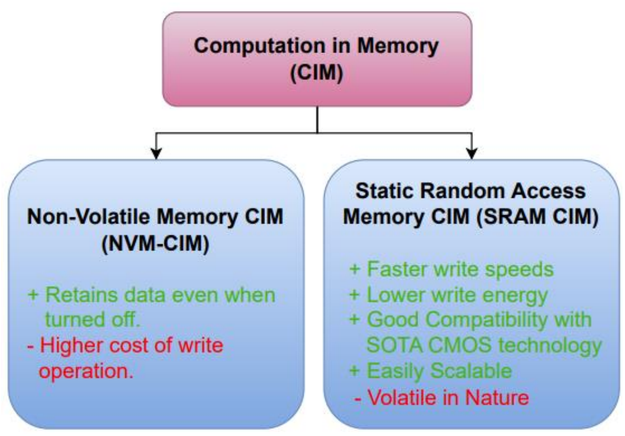
An SRAM-Based Hybrid Computation-in-Memory Macro for CNN Engines

Daniel Giftson E, Joycee Mekie
Department of Electrical Engineering, Indian Institute of Technology Gandhinagar, 382355, India
Email: {daniel.giftson, joycee}@iitgn.ac.in

1.Motivation



In Memory Computing



Types of CIM

2. Challenges

- **Analog-domain computing** performs energy-efficient and fast MAC operations but disadvantages from accuracy degradation due to low SNR and high overhead in power and area due to usage of Analog-to-Digital Converters (ADCs).
- **Digital-domain computing** performs highly accurate MAC operations but disadvantages from lesser energy-efficiency as compared to its analog counterpart.
- Therefore, the main challenges are:
  1) Robust In-Memory-Array Computing (IMAC) techniques
  2) Lightweight ADC Design

3. Proposed DIMAC-PNMAC architecture

- **DIMAC - Digital In-Memory-Array Computing**
  - Performs energy-efficient in-memory bit-wise multiplications and generates pulse-width-modulated (PWM) voltage signals.
  - This is achieved by **LDCAs** (Local Dual Column Arrays),
    - The **4-bit inputs** are stored in two columns according to the even and odd bit positions.
    - Performs bit-wise multiplications energy efficiently by reducing the parasitic capacitance of the LBLs and reusing the charge in the GBLs.

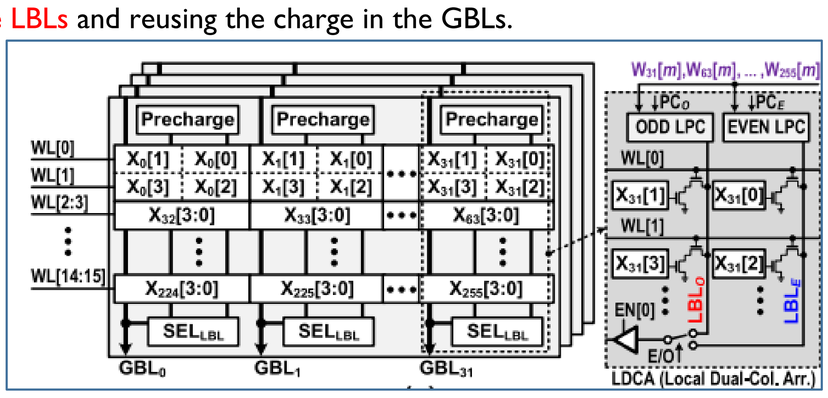


Fig 1. 2x64 8T-SRAM-based CIM Macro containing 32 – 16x2 capacity LDCAs

  - **4-bit weights** are loaded from an external source to the weight buffers
    - Weights are represented in **S&M** (Sign and Magnitude) format.
    - Weight controller performs **Zero-skipping Operation** before forwarding the weight bits serially to the **local precharge circuit**.
  - **In-Memory AND Operation** on the LBLs results in the bit-wise multiplication values.
    - Precharge the LBL with the forwarded weight bit.
    - Once the WL is activated, the ~input is read resulting in bit-wise multiplication.
  - The **E/O signal** ensures the generation of .PWM signals out of the partial MAC results.
    - When E/O signal flips, the respective LBL drives the bit-wise multiplication to the GBL through a **tri-state buffer**.

- **PNMAC – Phase-Domain Near-Memory-Array Computing**

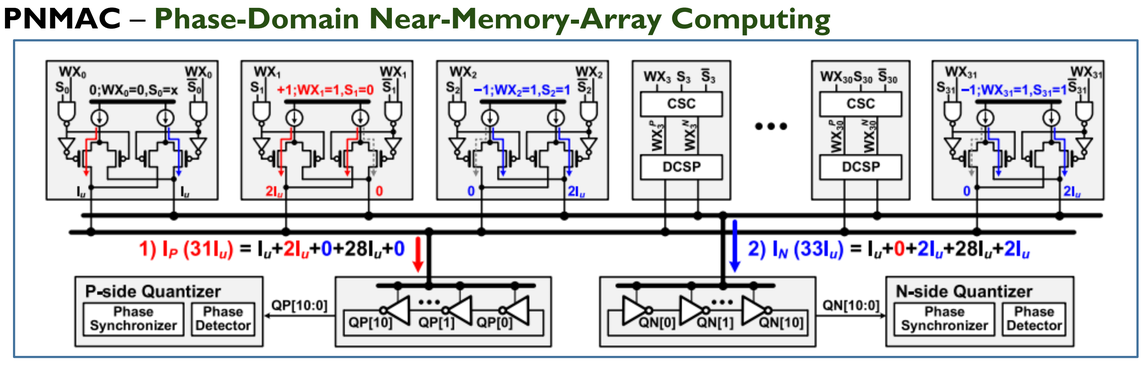


Fig 2. 32 parallel phase-domain bit-wise accumulation

  - Parallely accumulates the partial MAC results continuously and converts the accumulated result to the digital data once at the end with ultra-low power consumption.
  - Each GBL is connected to a **CSC** (current-steering controller) and a **DSCP** (Differential Current Source Pair).
    - CSC generates both positive and negative current-control signals.
    - The DSCP converts the PWM voltage signals to differential PWM current signals using those current-control signals.
  - The output current of the DCSP is translated to the phase by the **DCCO** (Differential Current Controlled Oscillators).
    - CCO is basically a 11-stage ring oscillator.
    - DCCO consists of a P-side CCO and an N-side CCO.

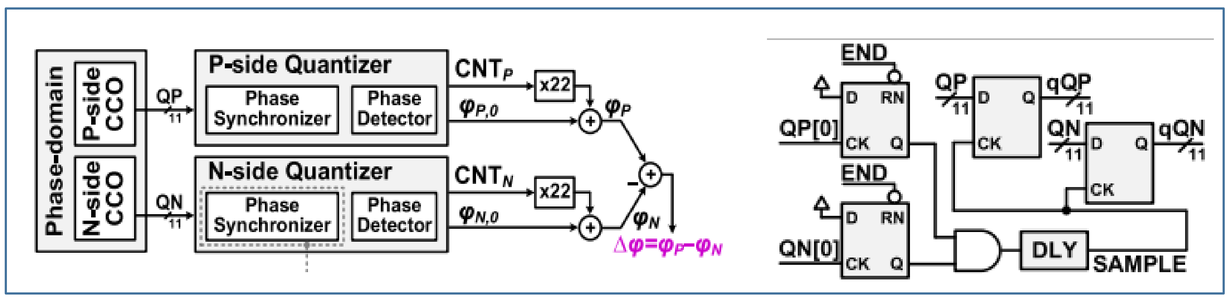


Fig 3. a) Readout Circuit **b)** Phase Synchronizer

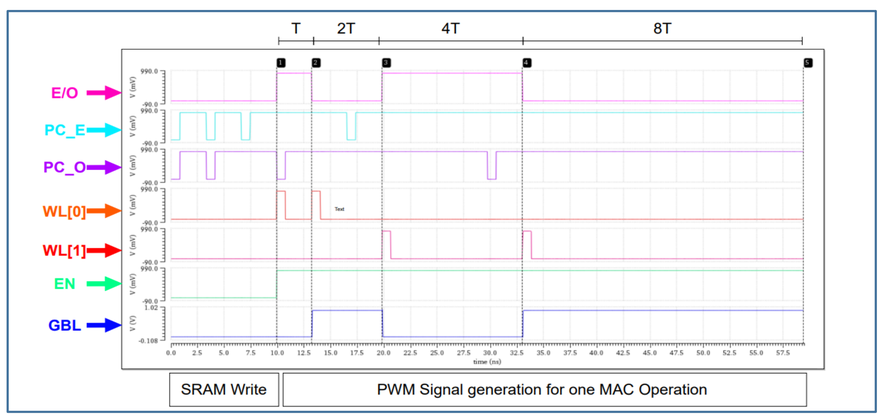
  - The final MAC result is sampled once by the **Readout** circuit after every set of MAC operations

4. Results and Observations

Technology: CMOS 28nm

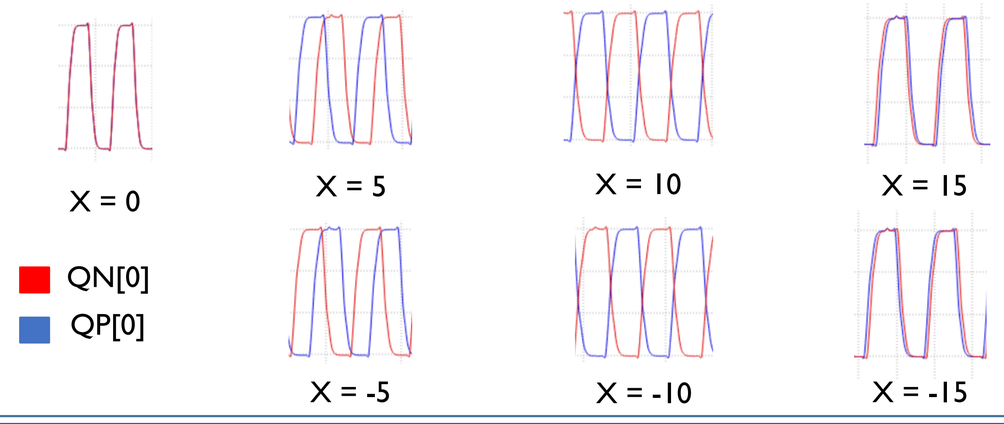
Clock Frequency: 300 MHz

PWM pulse generation timing diagram with overlapping precharging technique



  - The MAC rate of this work (20 MHz) is atleast 6.41 times faster than the existing Analog time-domain computing.

DCCO Waveform Outputs at the end of a MAC operation for various input data



5. Conclusion and Future Work

- This work proposes an ultra-low power consuming Hybrid CIM suitable for resource-constrained edge devices.
- DIMAC is introduce to perform energy-efficient bit-wise multiplications.
- PNMAC is proposed to perform accumulation parallelly with ultra low power consumption and a wide dynamic range analog-to-digital conversions.
- Future works include:
  - PVT analysis of the Hybrid CIM Macro.
  - Proposing an architecture to generate the complete multiplication result (4-b input x 4-b weight) at once, potentially improving the throughput of the CIM.

Acknowledgement

Prateek Sharma, MTech

References

1. Choi et al., "An SRAM-based hybrid computation-in-memory macro using current-reused differential CCO," IEEE J. Emerg. Sel. Topics Circuits Syst., vol. 12, no. 2, pp. 536–546, Jun. 2022.`
2. Y. Toyama, K. Yoshioka, K. Ban, S. Maya, A. Sai, and K. Onizuka, "An 8 bit 12.4 TOPS/W phase-domain MAC circuit for energyconstrained deep learning accelerators," IEEE J. Solid-State Circuits, vol. 54, no. 10, pp. 2730–2742, Oct. 2019.