Hardware-Software Co-Design of a Collaborative DNN Accelerator for 3D Stacked Memories with Multi-Channel Data

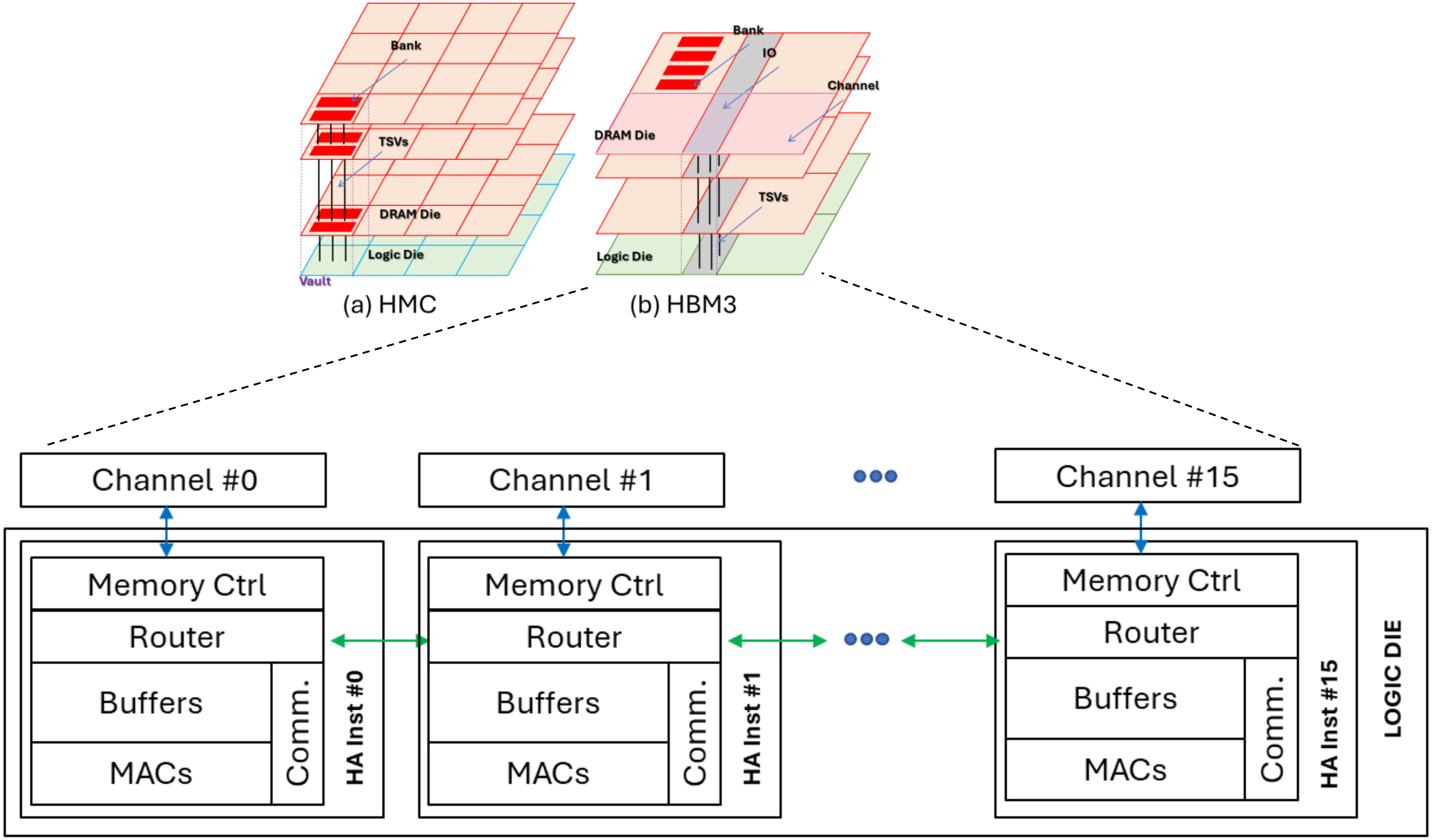
Tom Glint
IIT Gandhinagar
tom.issac@iitgn.ac.in

Manu Awasthi
Ashoka University
manu.awasthi@ashoka.edu.in

Joycee Mekie
IIT Gandhinagar
joycee@iitgn.ac.in

Proposal: 18x FoM

- EnX3D Architecture for ML acceleration
  - 3D memory based architecture with compute in logic layer
  - Hardware-software design and collaborative design paradigm

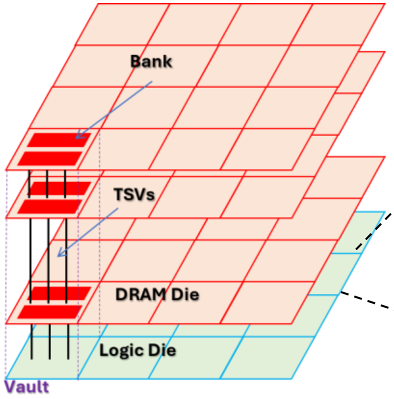


(a) HMC (b) HBM3

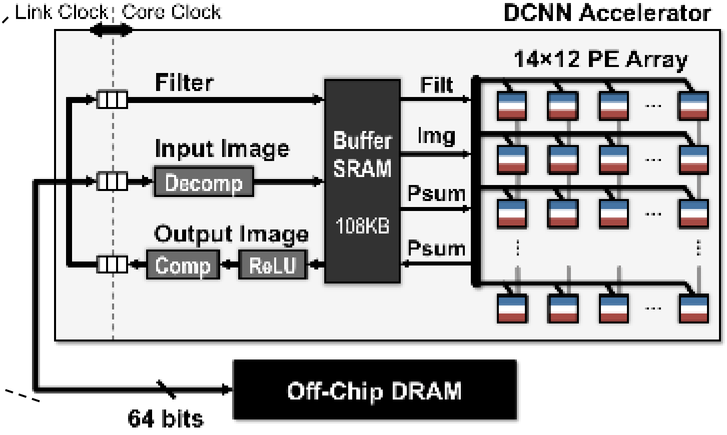
(3D stacked memory based NDP Accelerator)

Motivation

- 3D memories
  - Vacant area in logic die (~50 mm^2)
  - High bandwidth (~10x more than DDR4 DRAM)
  - Low access energy (12x less than DDR4 DRAM)
- Constraints:
  - TDP (9W, 15W), Area
- Current designs are not constraint optimized



(a) HMC

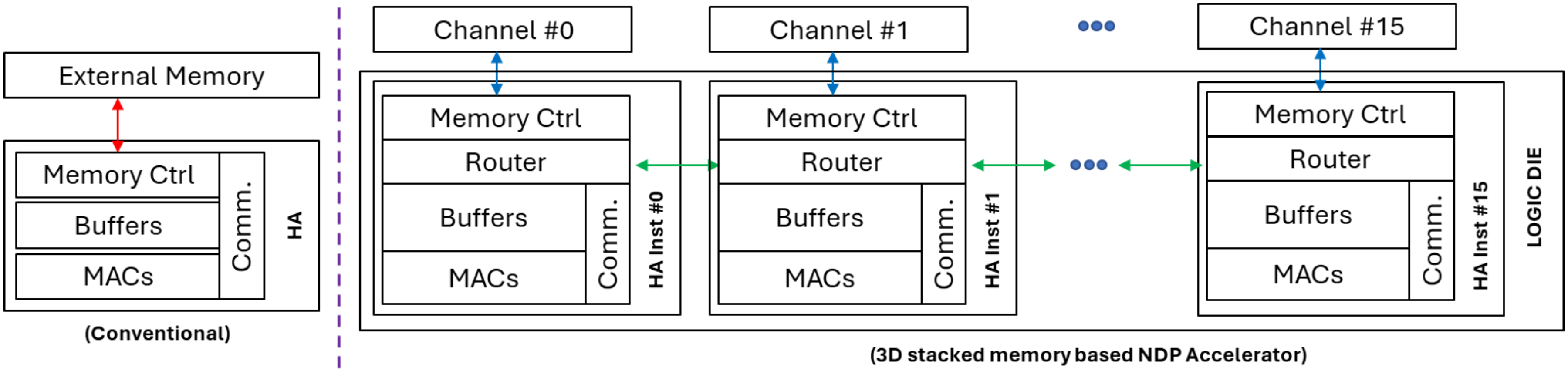


Eyeriss Accelerator

Image courtesy: https://eyeriss.mit.edu/

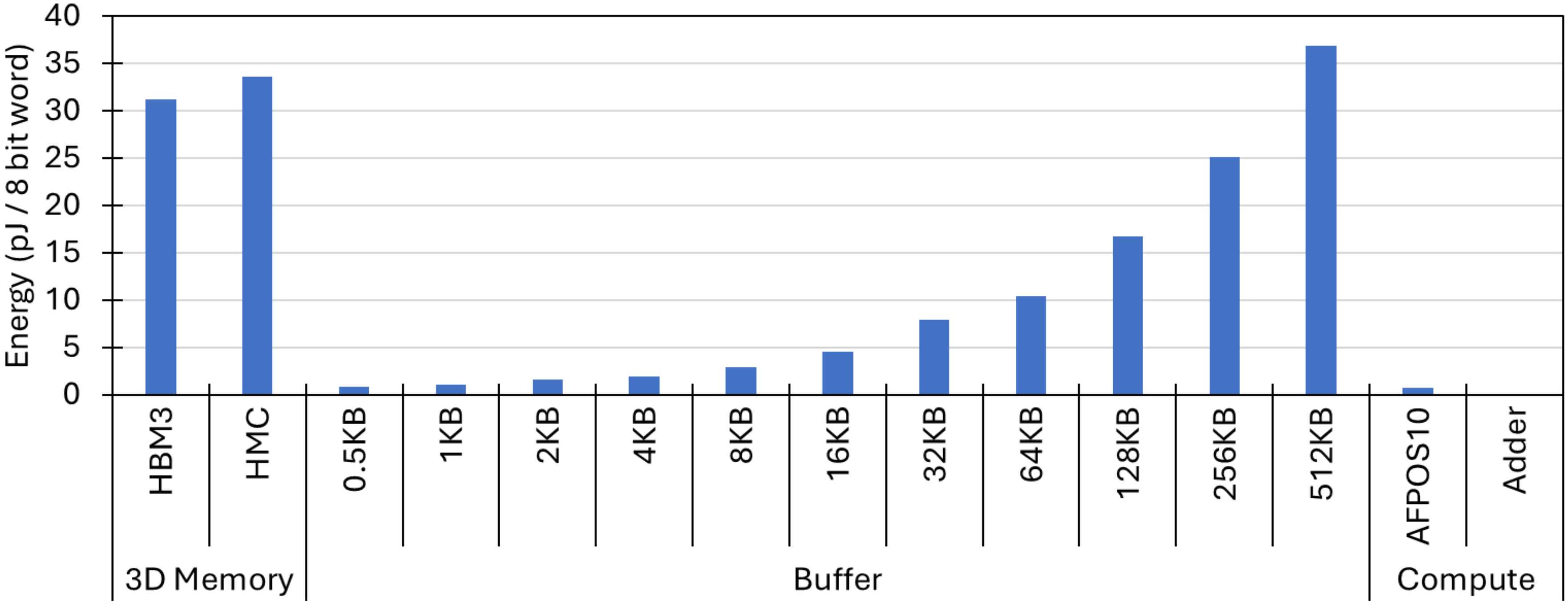
Collaborative Design

- Channel Aware
- Distributed Instances
- Optimized Mapping
  - Timeloop + Accelergy Framework



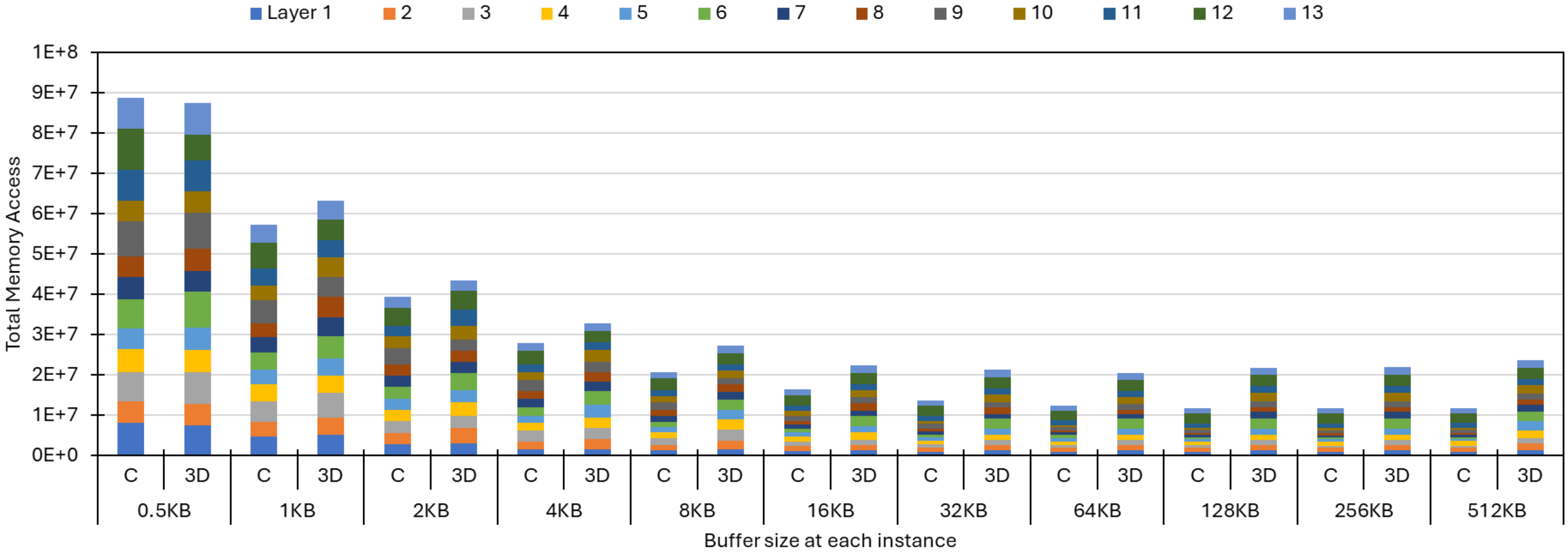
Collaborative Design: Components

- Approximate Fixed Posit Number system
  - No loss of accuracy at 8bit precision (without retraining)
  - No loss of accuracy at 6bit precision (with retraining)



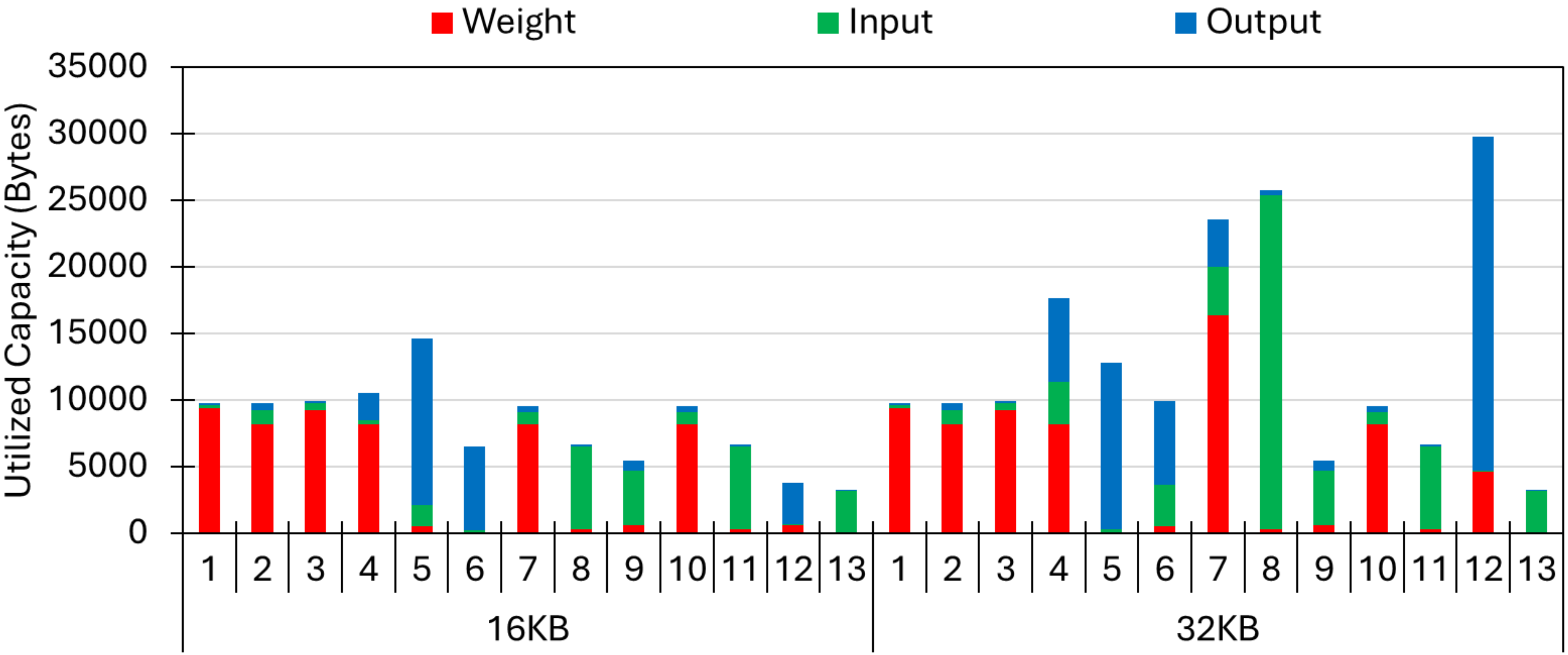
Collaborative Design: Ideal Buffer Capacity

- Distribution create redundancy

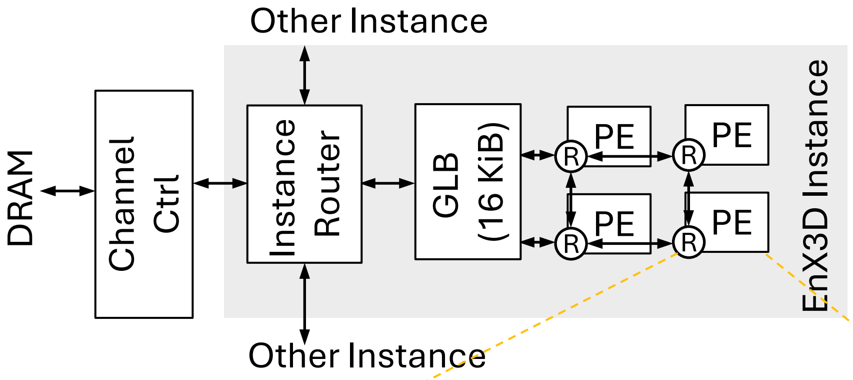


Collaborative Design: Ideal Buffer Capacity

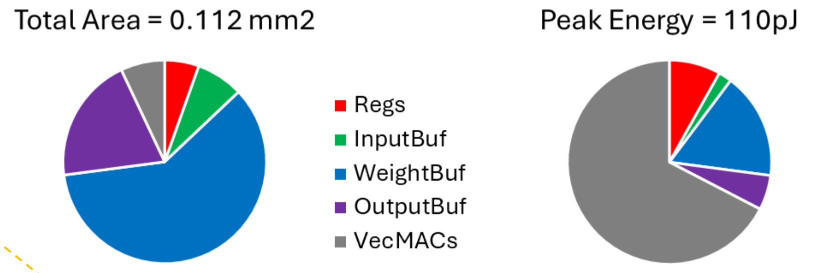
- 16 KB Capacity per instance handles most cases.



Design of EnX3D Instance

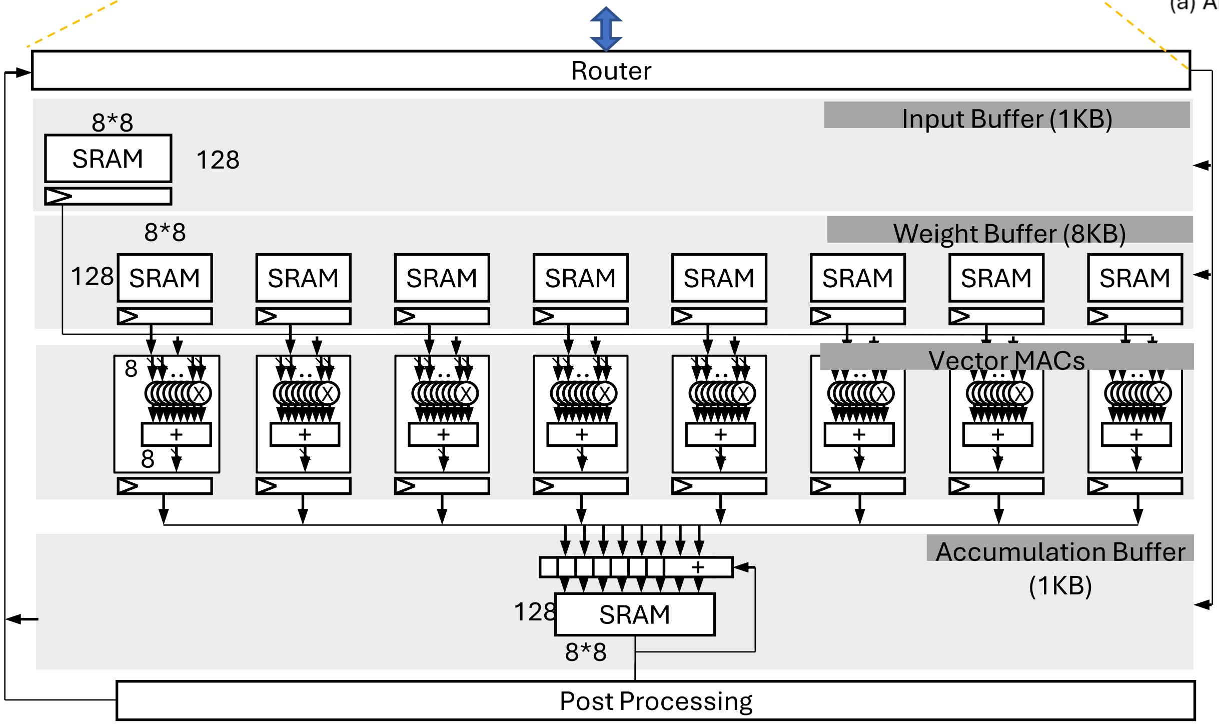


- Vector MAC Architecture
- Shared Input



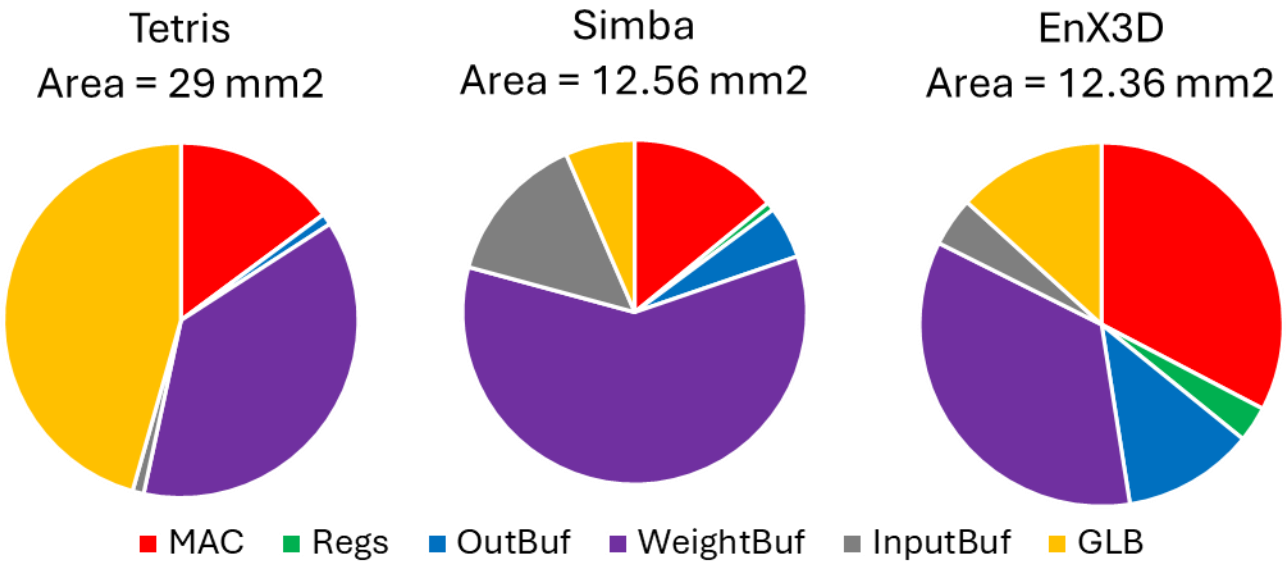
(a) Area

(b) Energy



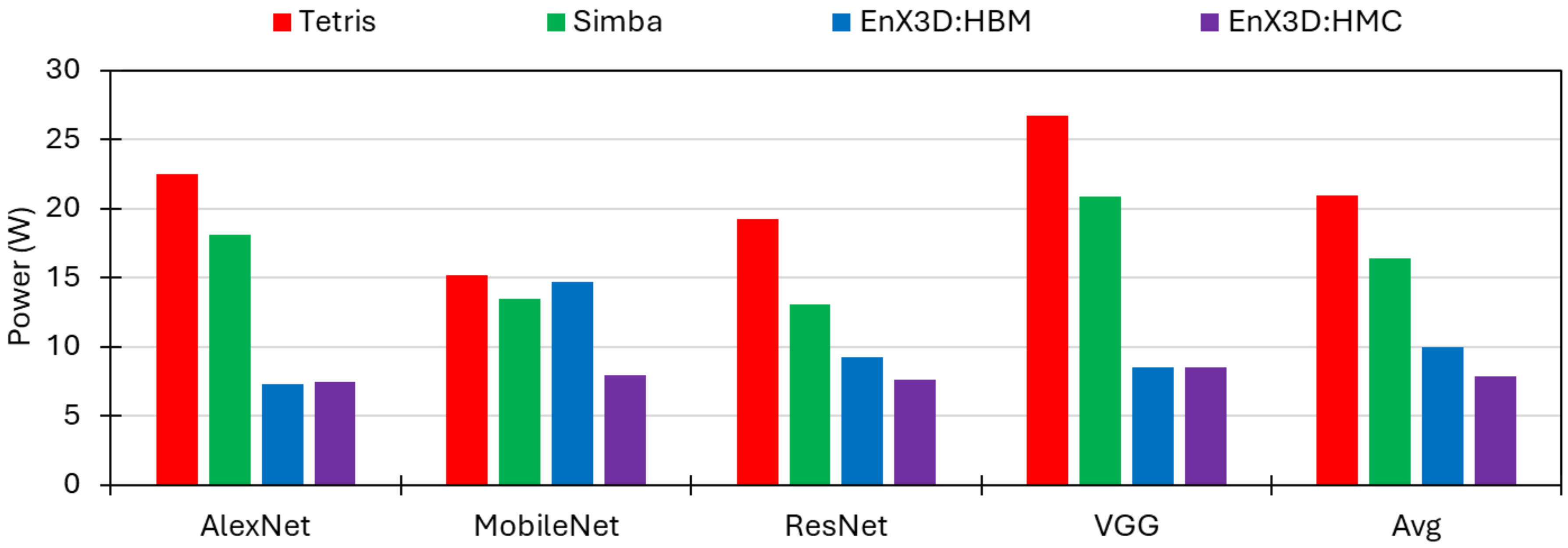
Results: Constraint: Area

- 2.3x reduction in Area



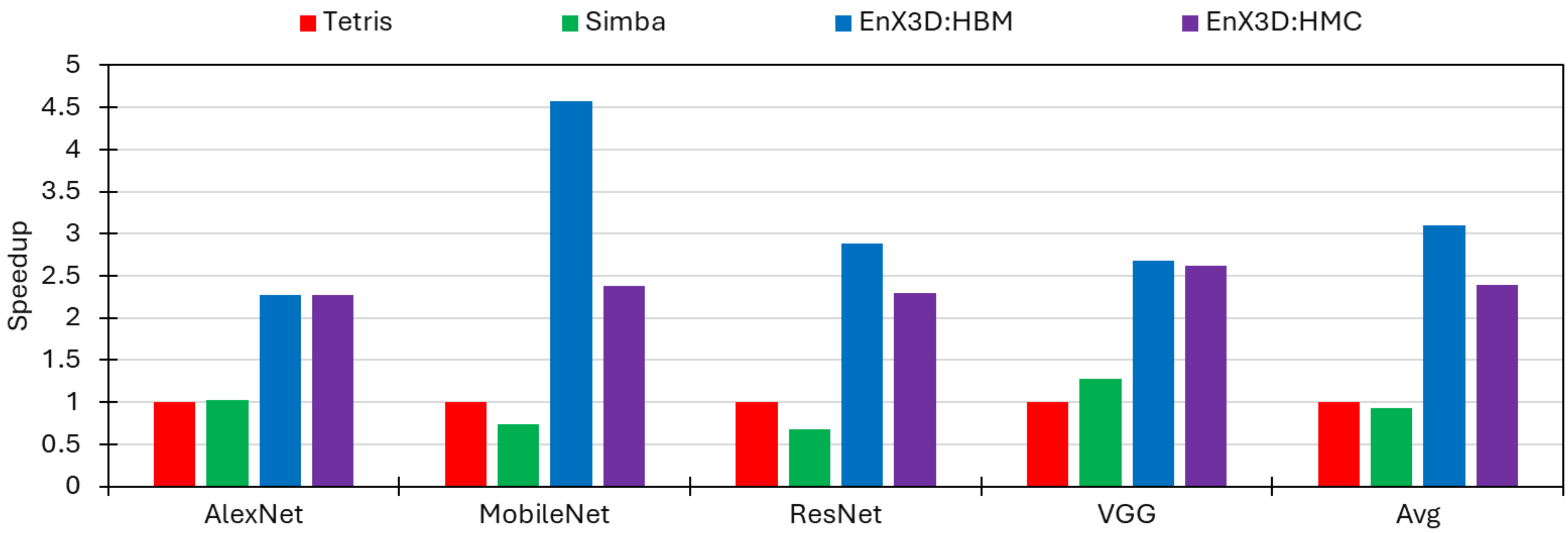
Results: Constraint: Power

- Power within envelope for passive cooling



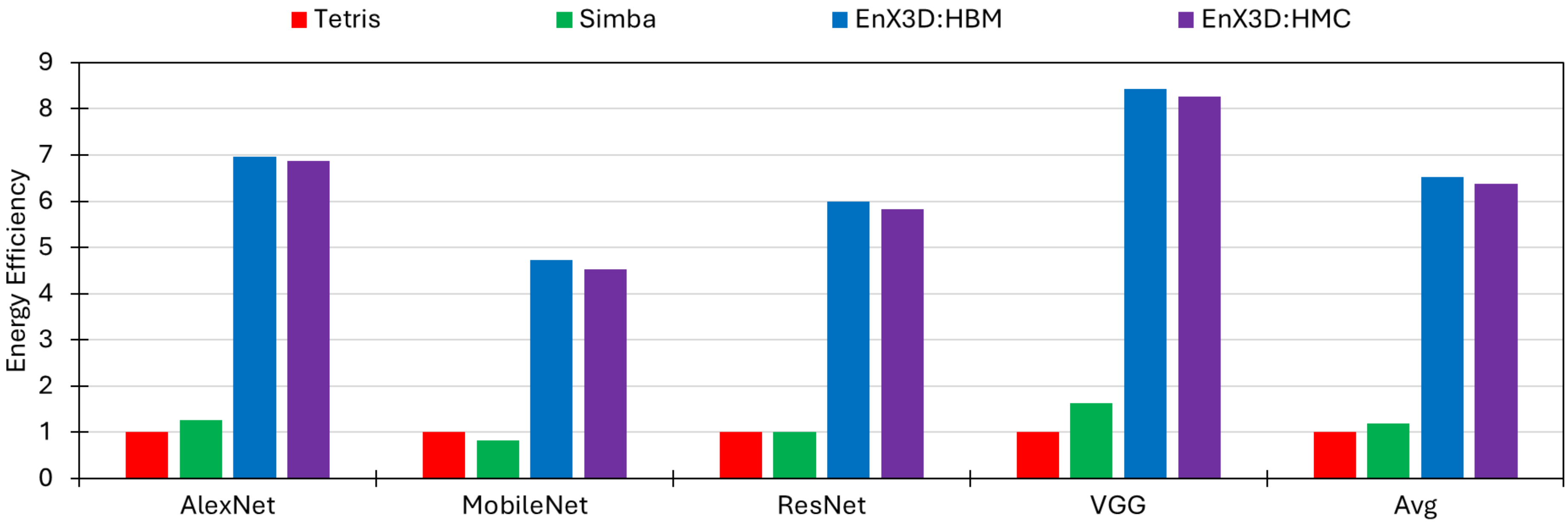
Results: Speedup

- 2.5x to 3x speedup



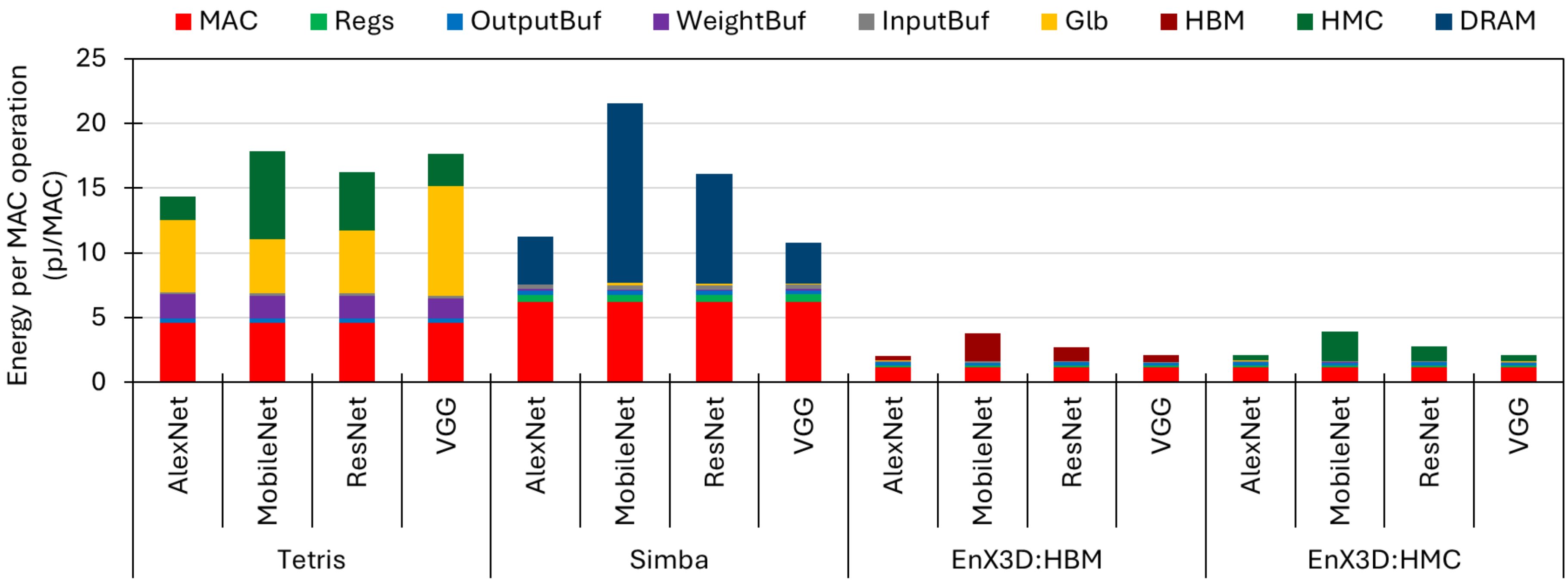
Results: Energy Efficiency

- 6x increase in energy efficiency



Results: RAW energy cost

- Improvements across compute, buffer and RAM access



Key Observations and takeaway

- Multiple Channels of data constraint
    - Collaborative Design
    - DRAM access reduction
- Area and thermal constraint
    - Hardware-software co-design
        - With and without retraining
        - Approximate compute with POSIT number system
- Buffer optimization is critical
- 3x speedup
- 6x reduction in energy
- 2x reduction in area
- No loss of accuracy

QUESTIONS

THANK YOU