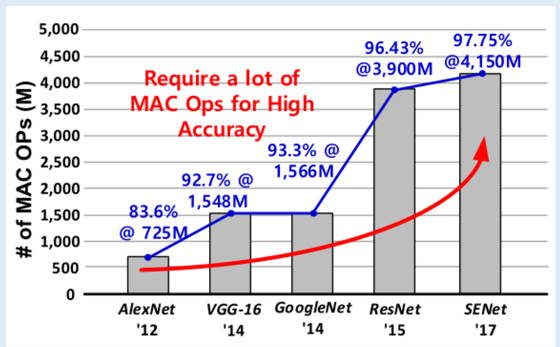
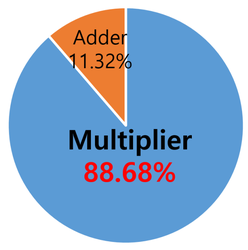
IDEC Chip Design Contest

An Area-efficient Memory-based Multiplier Powering Eight Parallel Multiplications for Convolutional Neural Network Processors

Seongrim Choi, Jeong-Gon Hwang, and Byeong-Gyu Nam

1. Motivation



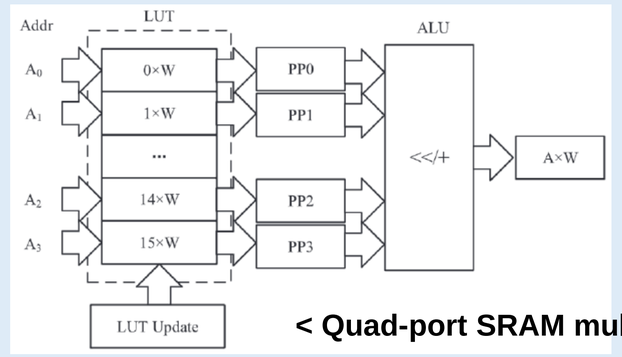


Area breakdown of MAC units[1]

- Challenges for CNN processors
  - Integrating a large number of MAC units
    - Need area-efficient MAC unit
  - Multiplier part is very critical due to its significant contribution to the area of MAC unit

[1] S.-S. Part et al., *electronics*, Jan. 2020.

2. Previous Work

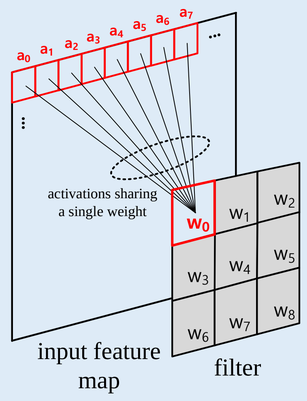


< Quad-port SRAM multiplier[2] >

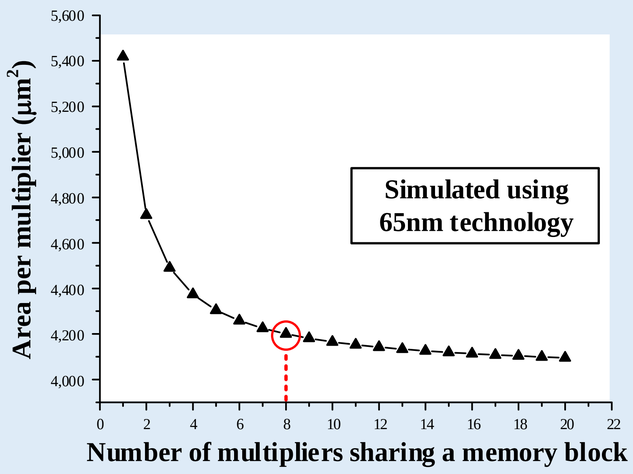
- Memory-based multiplier for improved area
  - Exploiting pre-computed multiplication results
  - Incorporating special quad-port bitcell
- However, CNN requires a more area-efficient design

[2] Y. Gong et al., *Microelectron. J.*, Mar. 2019.

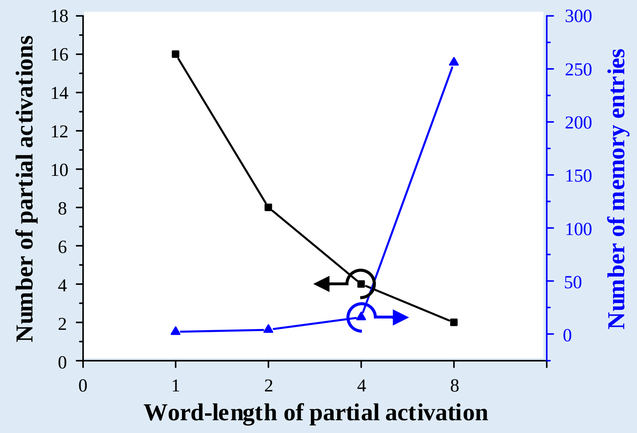
3. Proposed Memory Sharing Multiplier



- Multiple activations sharing single weight in a convolution
  - ➔ Memory-sharing parallel multiplier to share weight

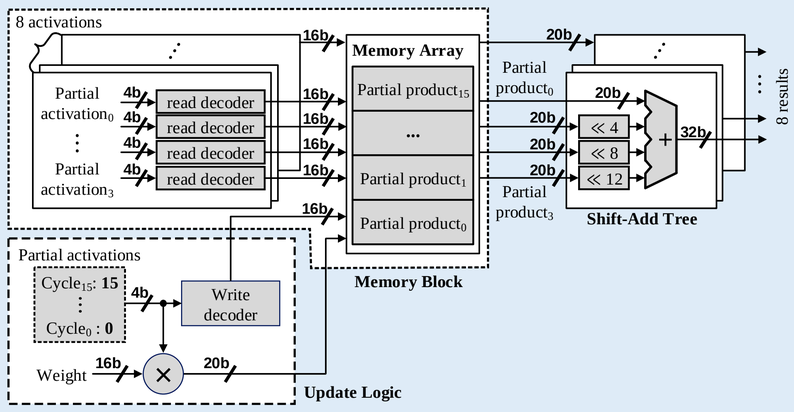


- # of multipliers sharing single memory block to maximize area-efficiency
  - Benefit diminishes at the number of 8



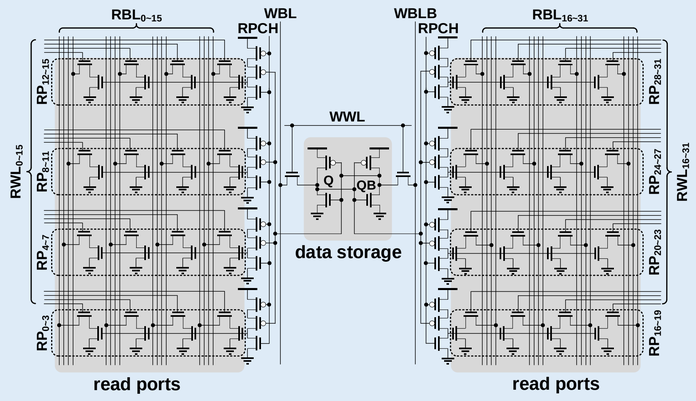
- The number of memory entries is exponential to activation word-length
  - Optimal 4-bit partial activation and 4 partial activations

4. Overall Architecture



- Memory block to support **32 read-out ports**
  - 8 parallel multiplications × 4 partial activations

5. Proposed 32-port Bitcell

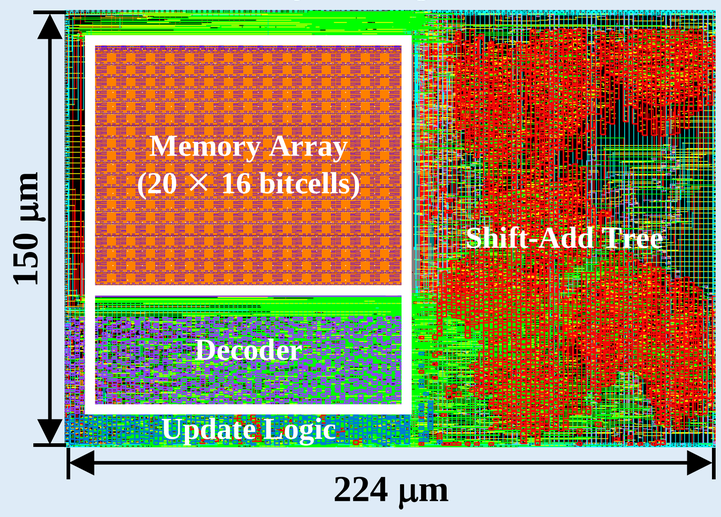


- 32 read ports to simultaneously get 32 partial products
  - 2-stage of FO4 buffering structure to drive large capacitance
  - Domino buffers for fast switching and lightweight driving of output ports

6. Implementation Result

- Achieves **~18.4% area-reduction** from the state-of-the-art memory-based multiplier

<Layout Diagram>



<Comparison with Previous Works>

	Booth Multiplier	LUT Multiplier[3]	Quad-port SRAM Multiplier[2]	This Work
Technology, nm	65	65*	65*	65
Area, μm²	53,630	113,678	41,181	33,600
Supply Voltage, V	1.2	1.2	1.2	1.2
Input Frequency, MHz	200	200	200	200
Power, mW	2.63	1.63	1.43	1.01
Propagation Delay, ns	2.11	2.33	1.94	1.88

[3] D. Shin et al., *ISSCC*, 2017.

※ Acknowledgments: The EDA tool was supported by the IC Design Education Center(IDEC), Korea.

반도체설계교육센터 IC DESIGN EDUCATION CENTER