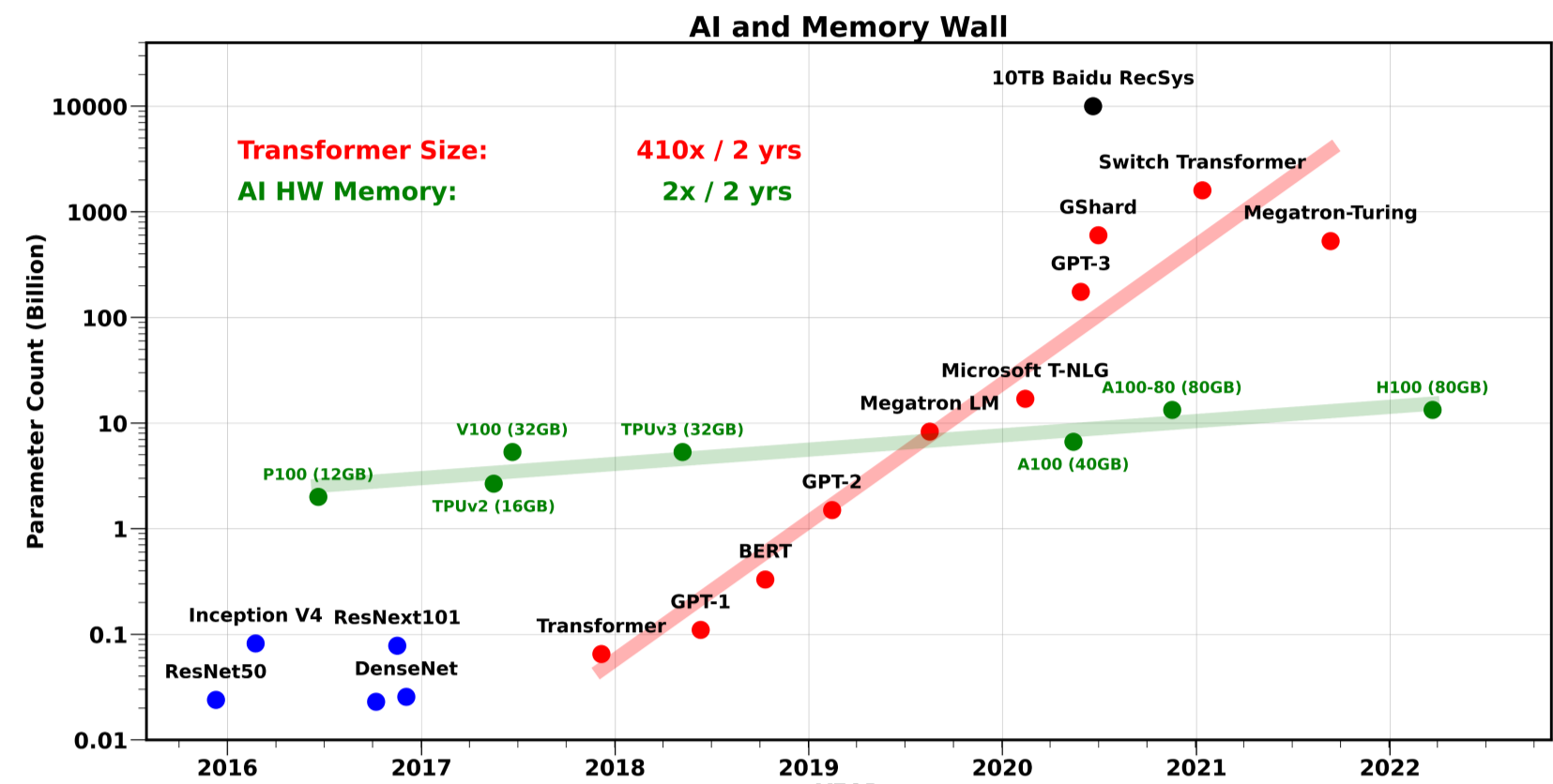


# HW-SW Co-Exploration And Optimization For Next-Generation Learning Machines

Chunyun Chen. Advisor: Mohamed M. Sabry Aly

School of Computer Science and Engineering, Nanyang Technological University, Singapore

## Huge gap between SW and HW

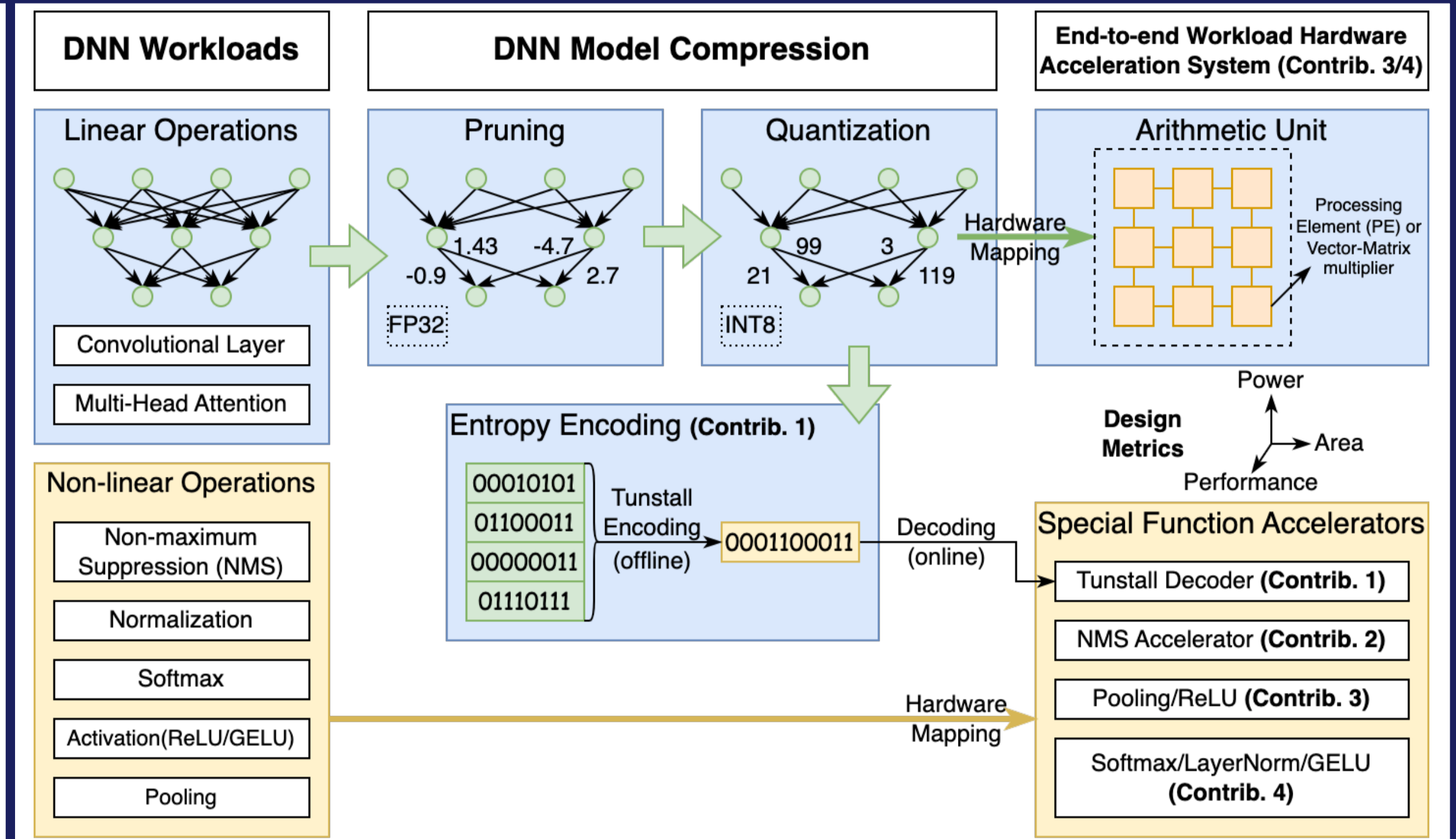


DNNs are both *data-intensive* and *compute-intensive*

- Huge models: **trillions** of params.
  - AI Models size: **240x** every 2 years
  - AI HW Memory: **2x** every 2 years
- Lack special function accelerators
  - # OPs: **2%**
  - Execution time: **73%**
- Lack system-level hardware solutions

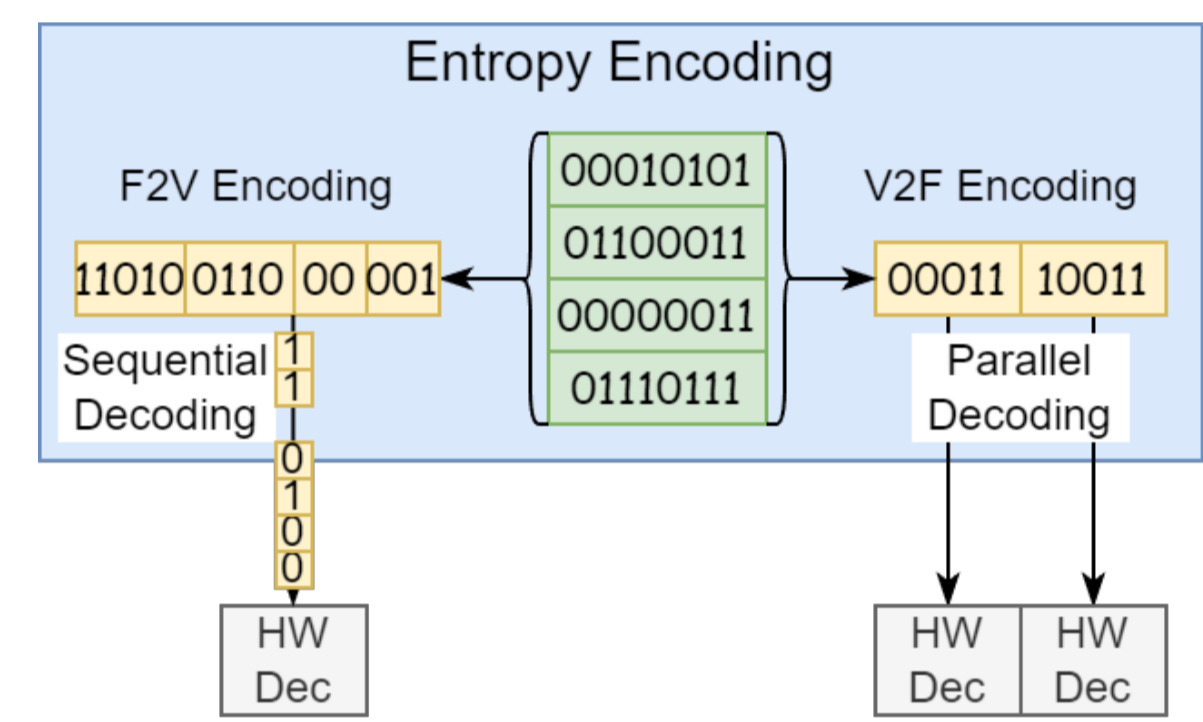
Solutions:

- Models compression
- Special function acceleration hardware
- DNN hardware acceleration systems



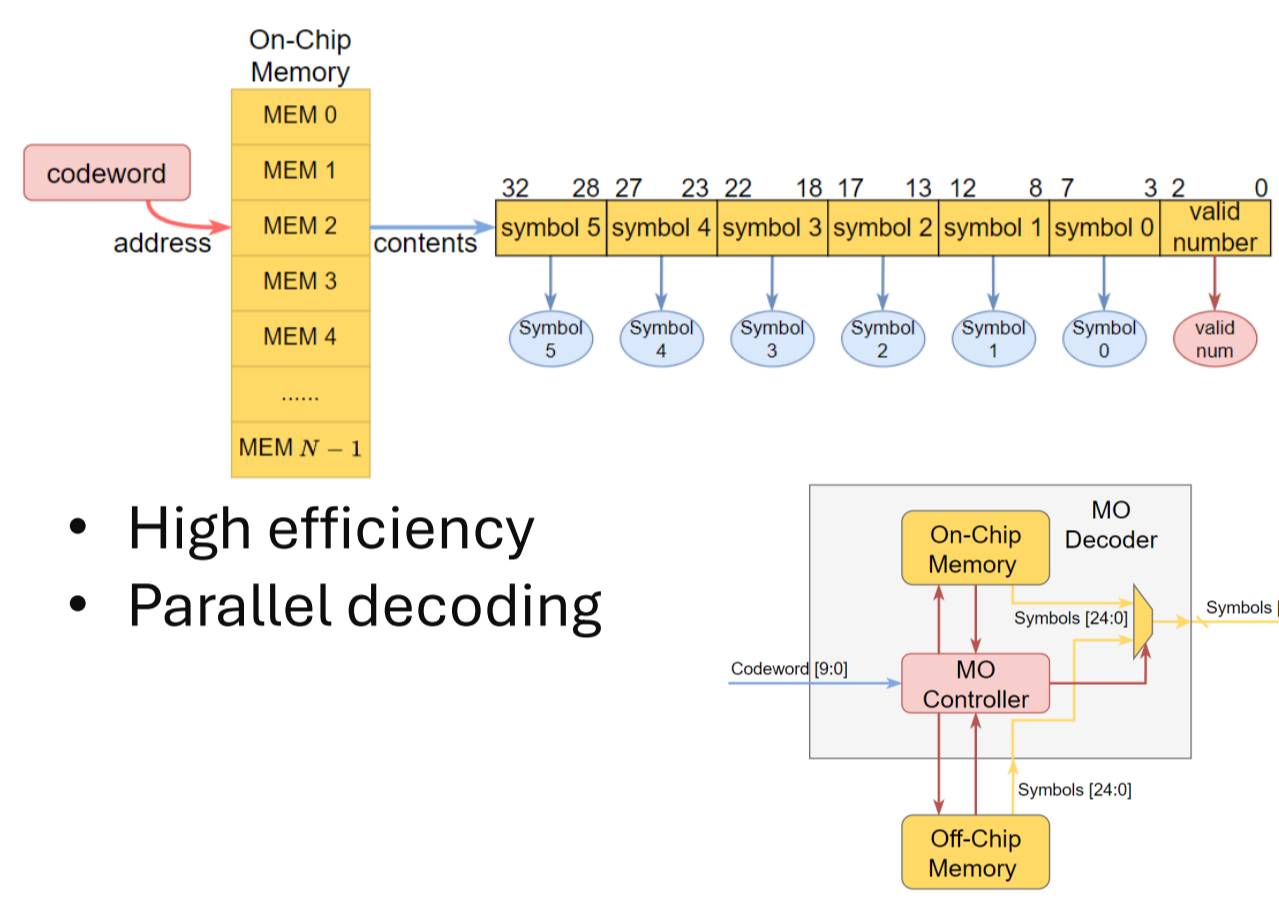
## C1: Efficient Tunstall Decoder for Deep Network Compression

### F2V Encoding: Slow Sequential Decoding



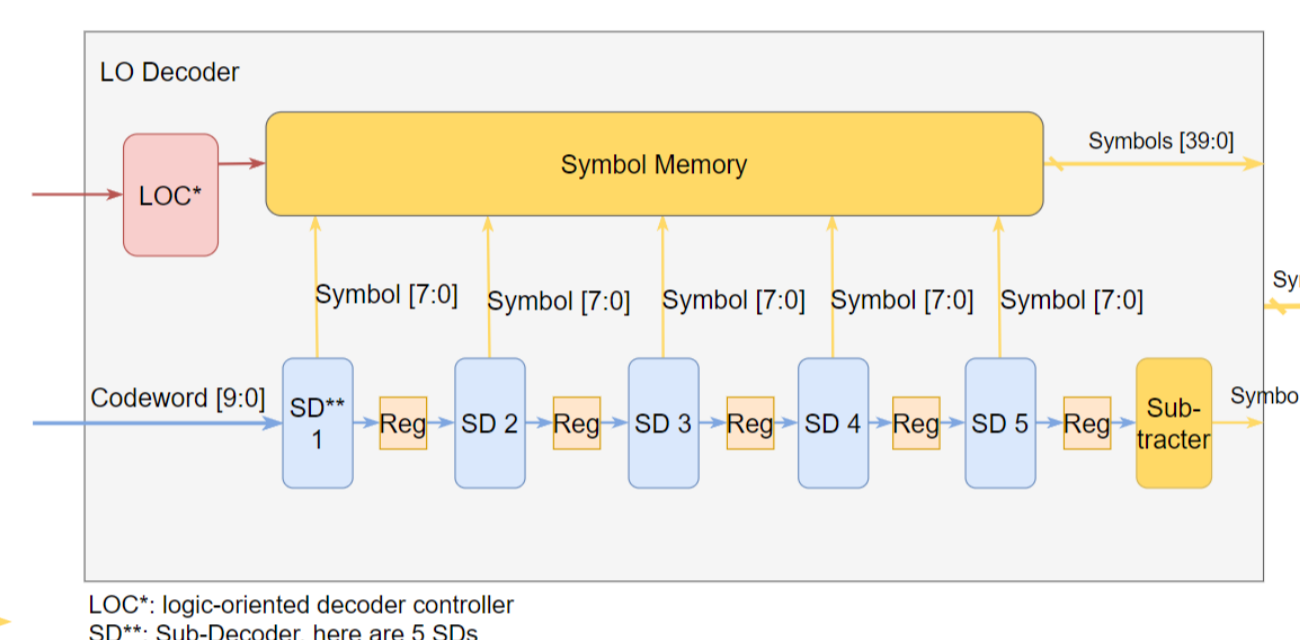
- Solution:
- V2F - Tunstall parallel decoding

### Memory-Oriented (MO) Method



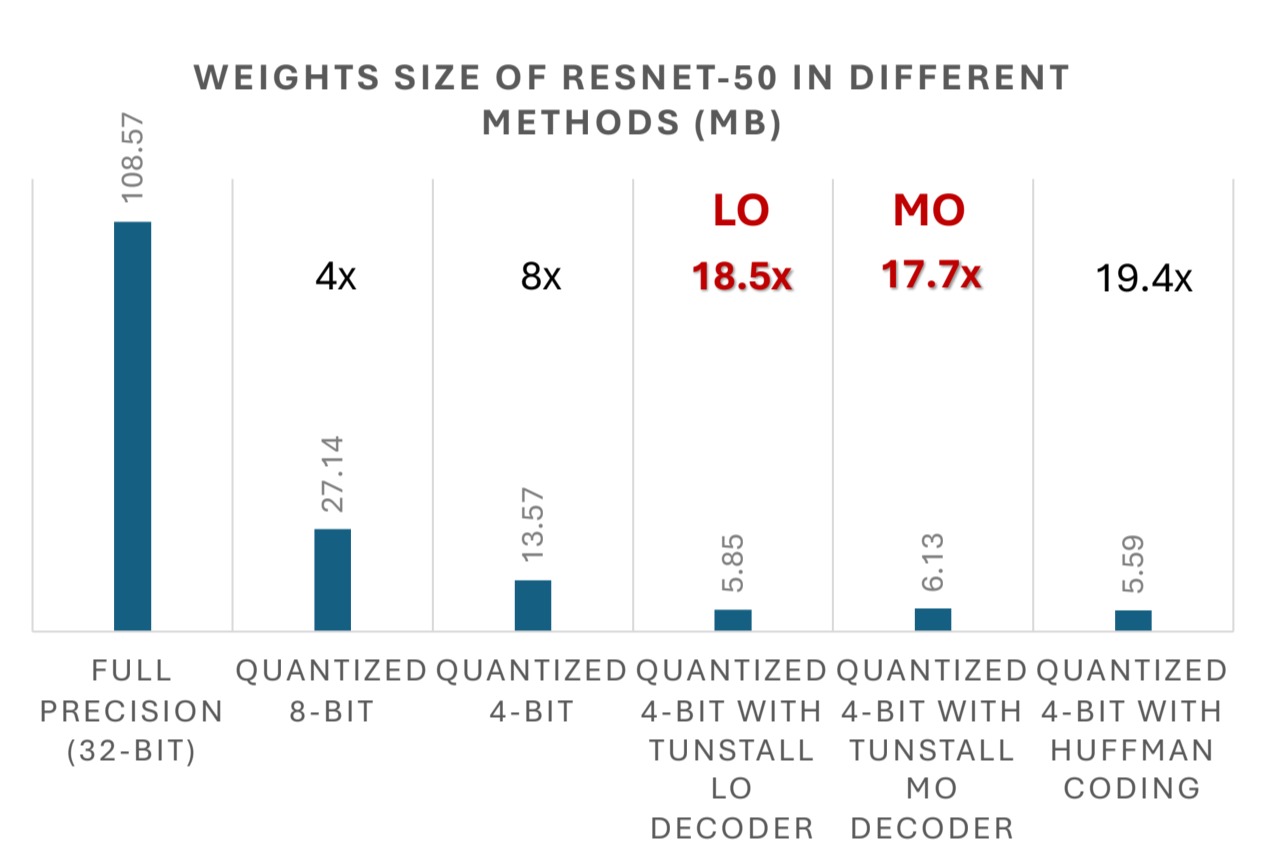
- High efficiency
- Parallel decoding

### Logic-Oriented (LO) Method



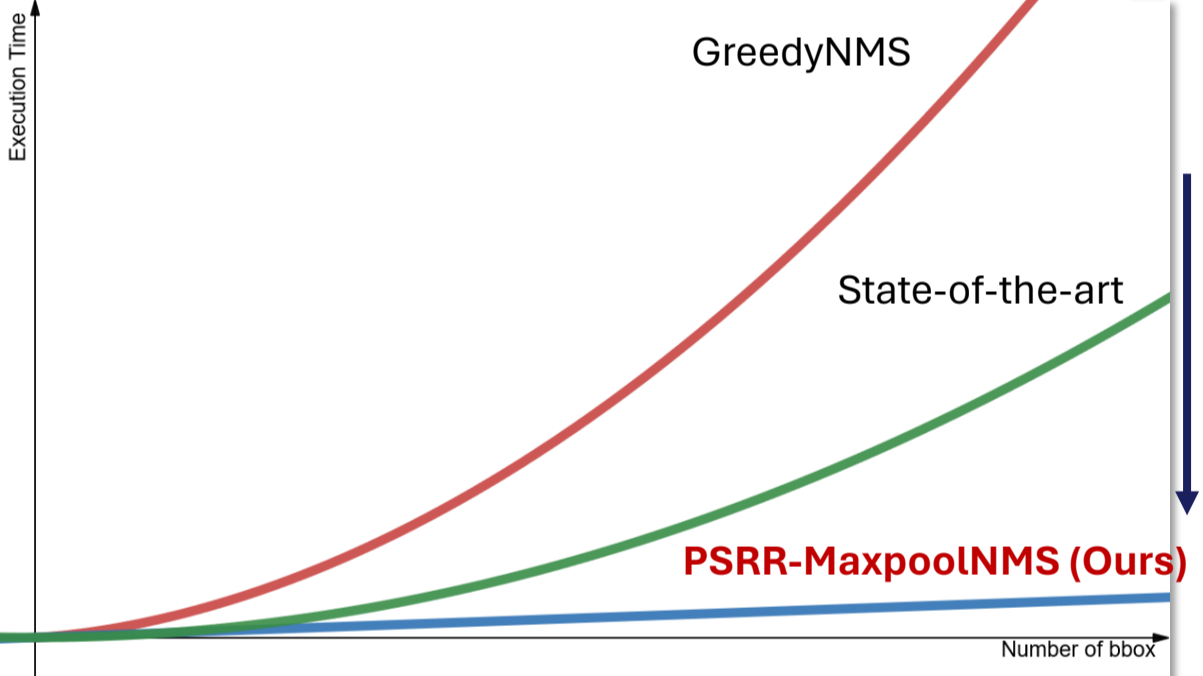
- Pipelined architecture
- Parallel decoding

### Experimental Results



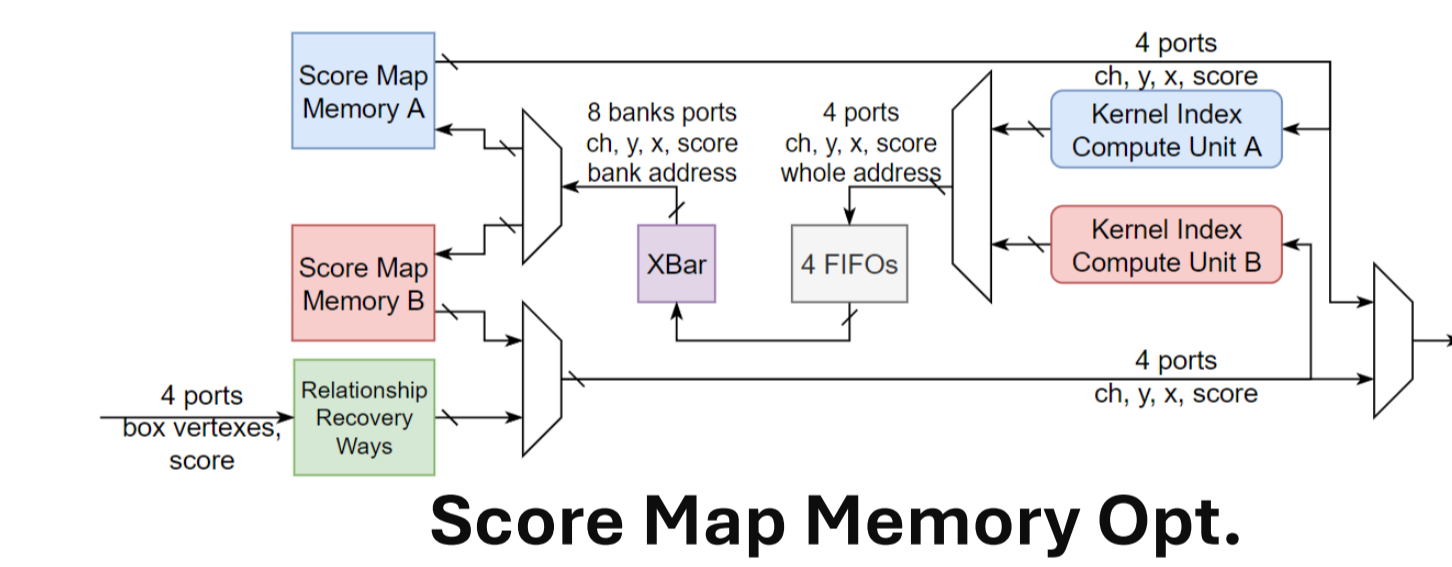
## C2: Scalable Hardware Acceleration of Non-Maximum Suppression

### GreedyNMS: $O(n \log(n)) + O(nm)$



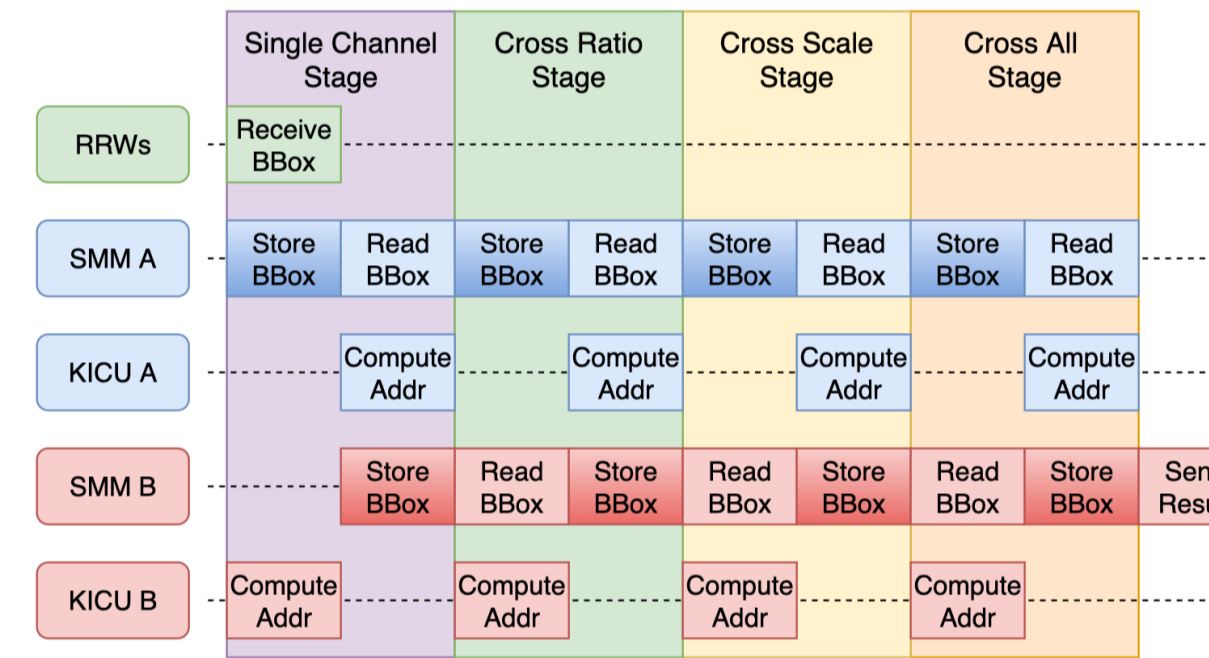
- Reduce the time complexity to  $O(n)$
- Parallelizable
  - No loops

### ShapoolNMS Arch

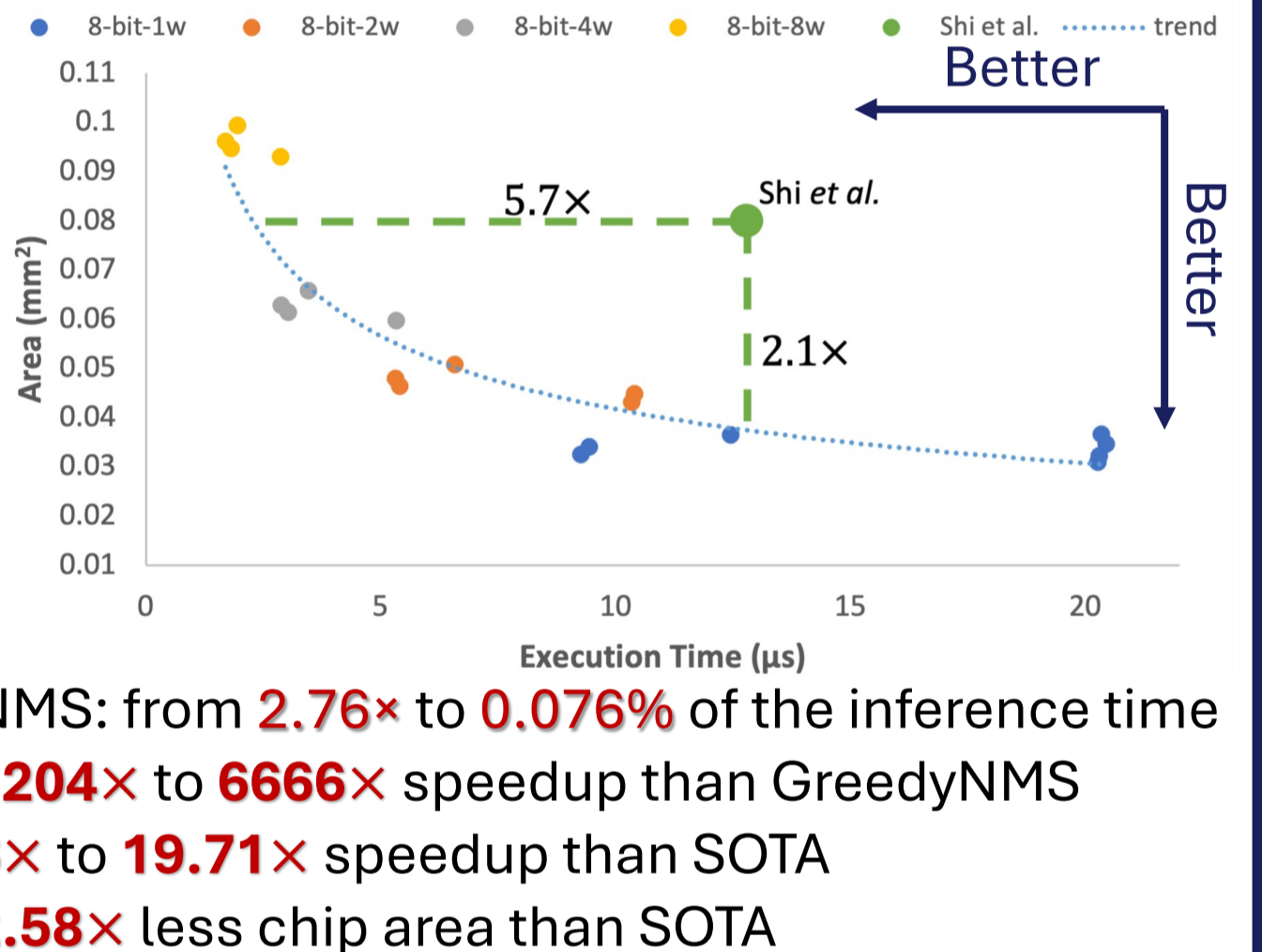


- Utilize the sparsity in the score map
- 2.41x speedup
- 1.05x area overhead

### ShapoolNMS Dataflow



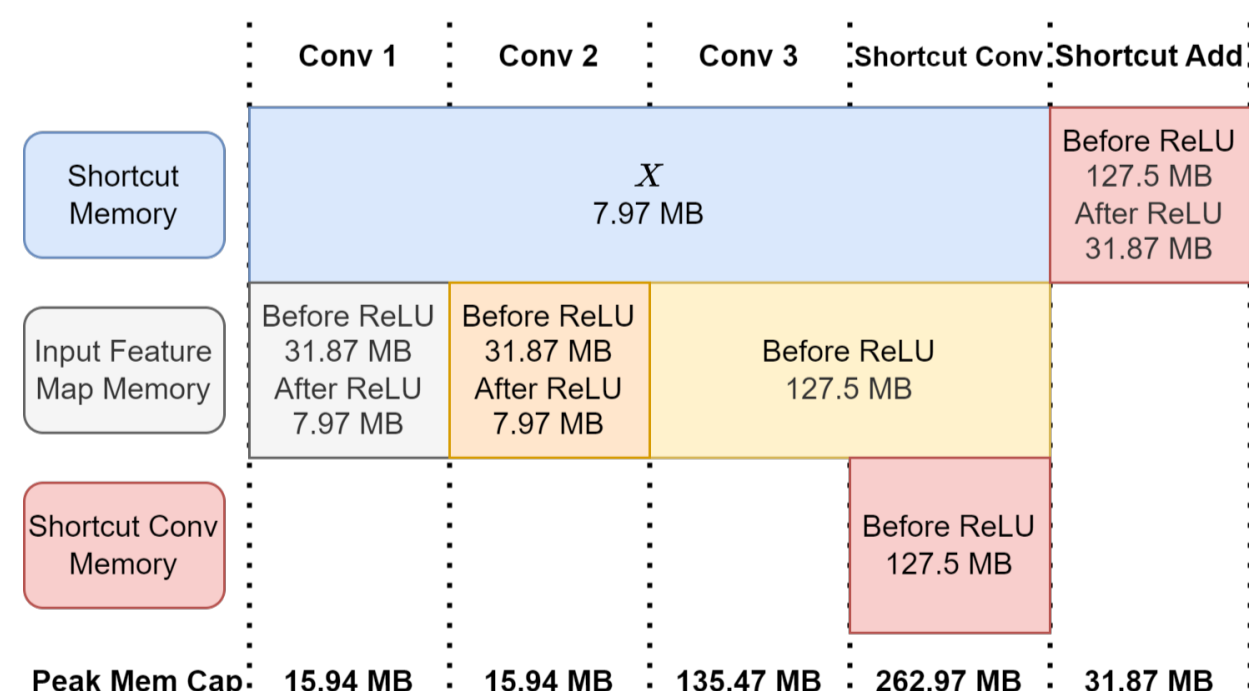
### Experimental Results



- NMS: from 2.76x to 0.076% of the inference time
- 1204x to 6666x speedup than GreedyNMS
- 3x to 19.71x speedup than SOTA
- 2.58x less chip area than SOTA

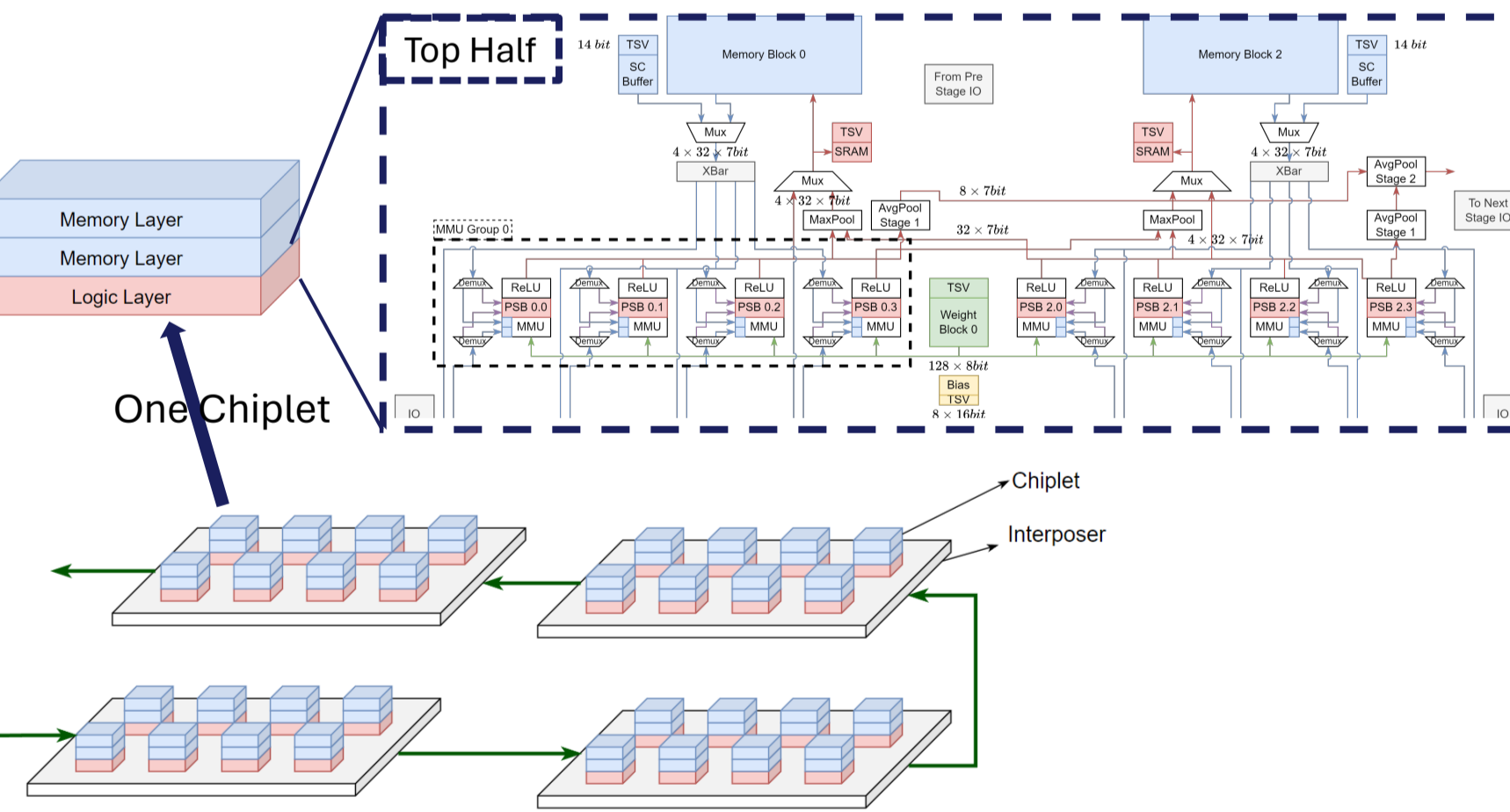
## C3: Chiplet-based Scalable ResNet Accelerating System

### ResNet Bottleneck: 262.97 MB Mem

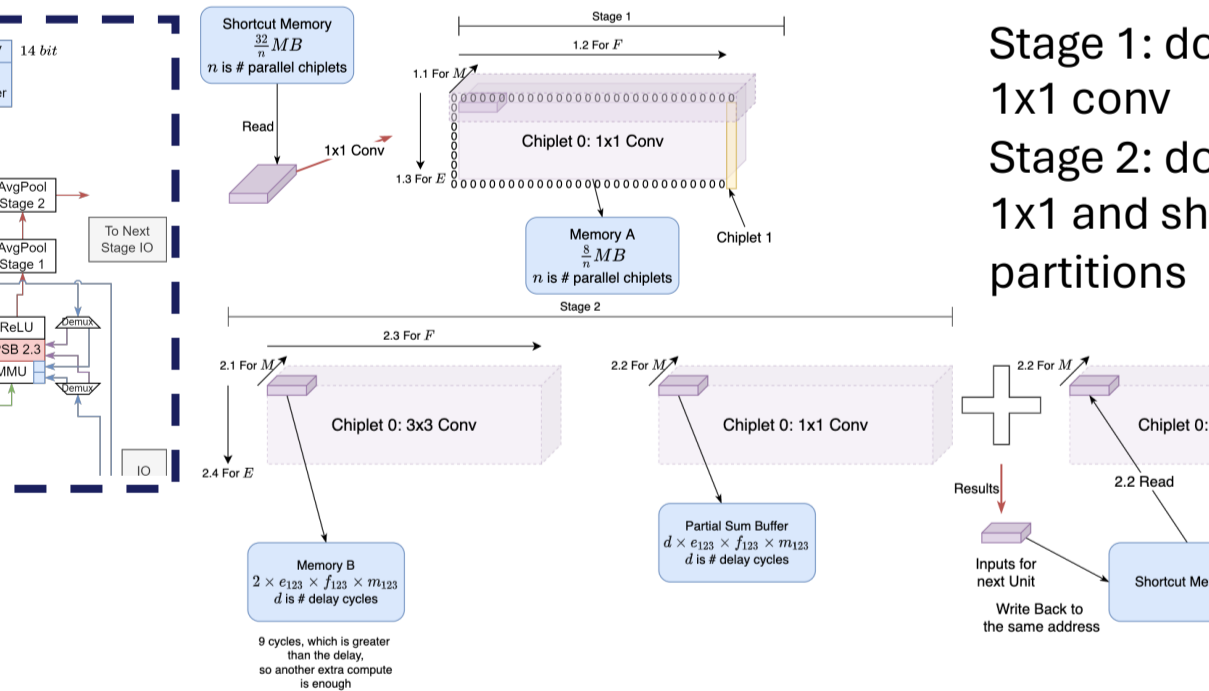


- Solution:
- Optimize the computational dataflow
  - System-level hardware accelerations

### Res-DLA Chiplet Accelerating System

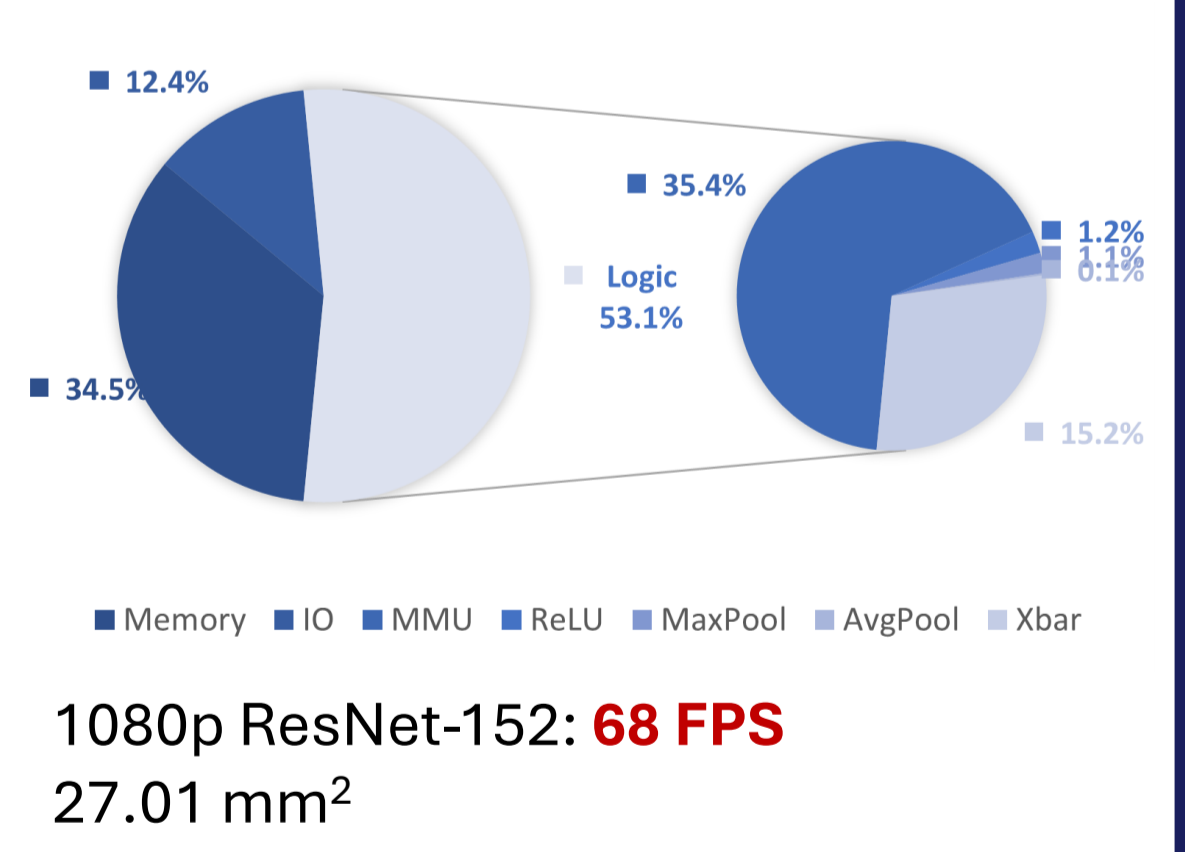


### Cross Layer Optimization Dataflow



- 6.6x less memory for features
- 6.2x less memory for weights

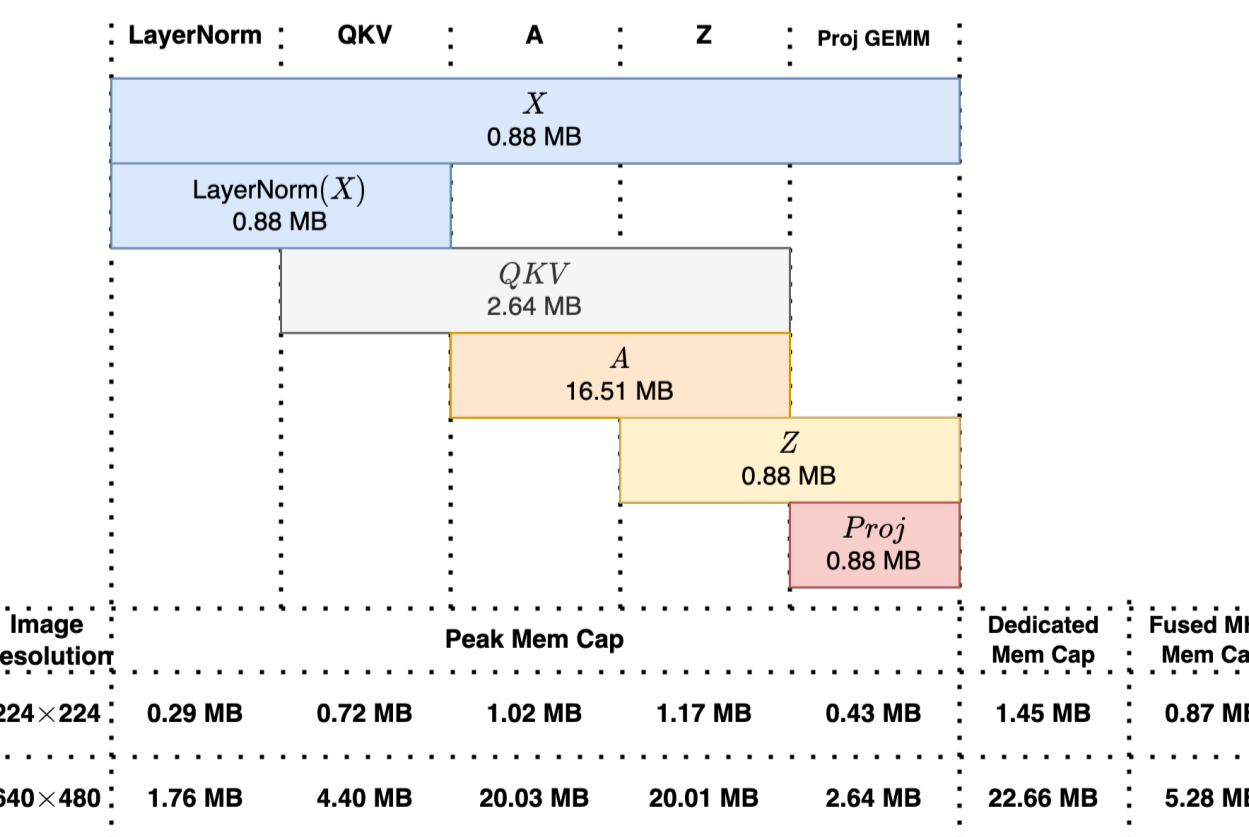
### Experimental Results



- 1080p ResNet-152: 68 FPS
- 27.01 mm²

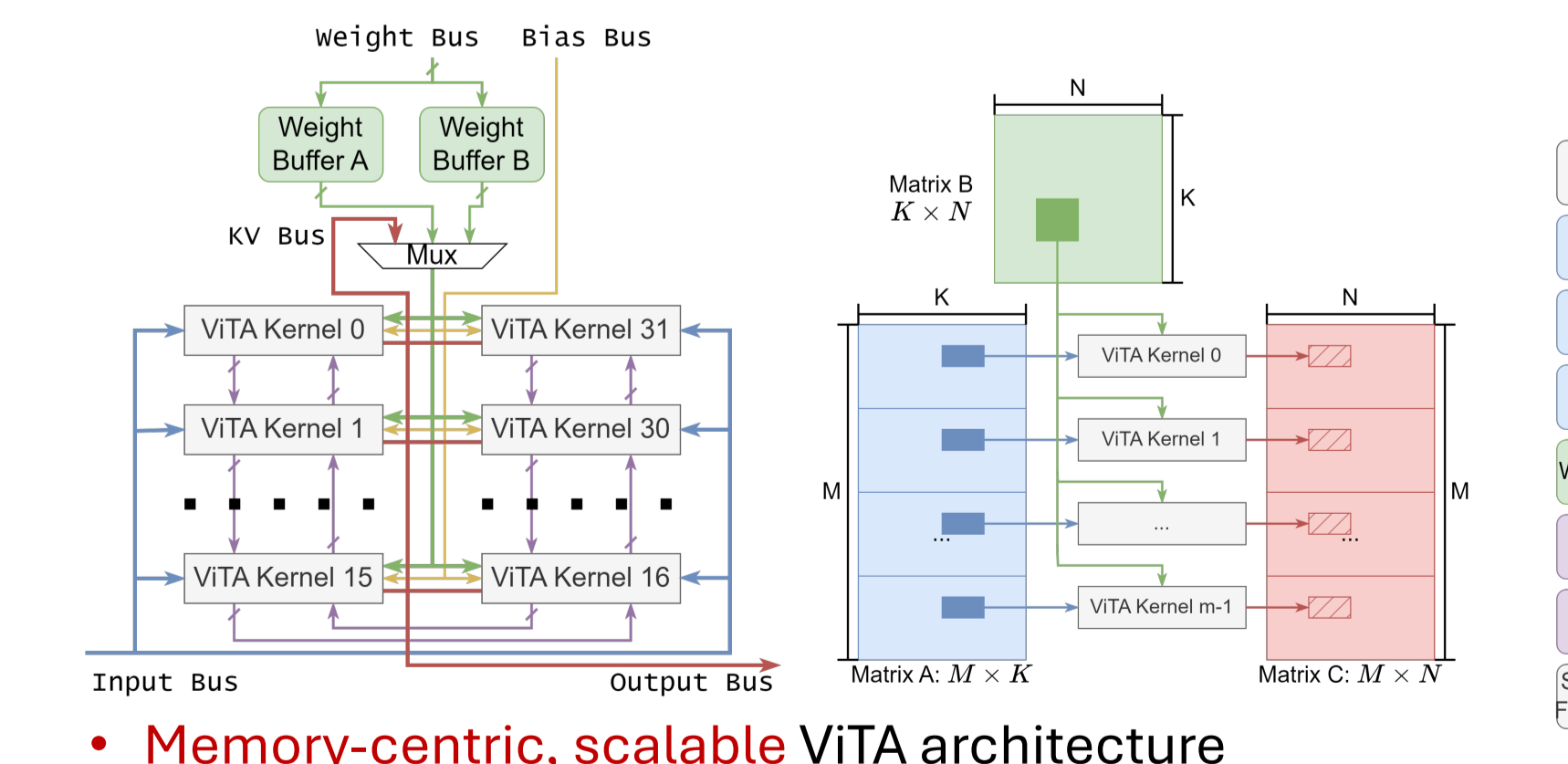
## C4: ViTA: A Highly Efficient Dataflow and Architecture for Vision Transformers

### ViT MHA: 22.66 MB Mem



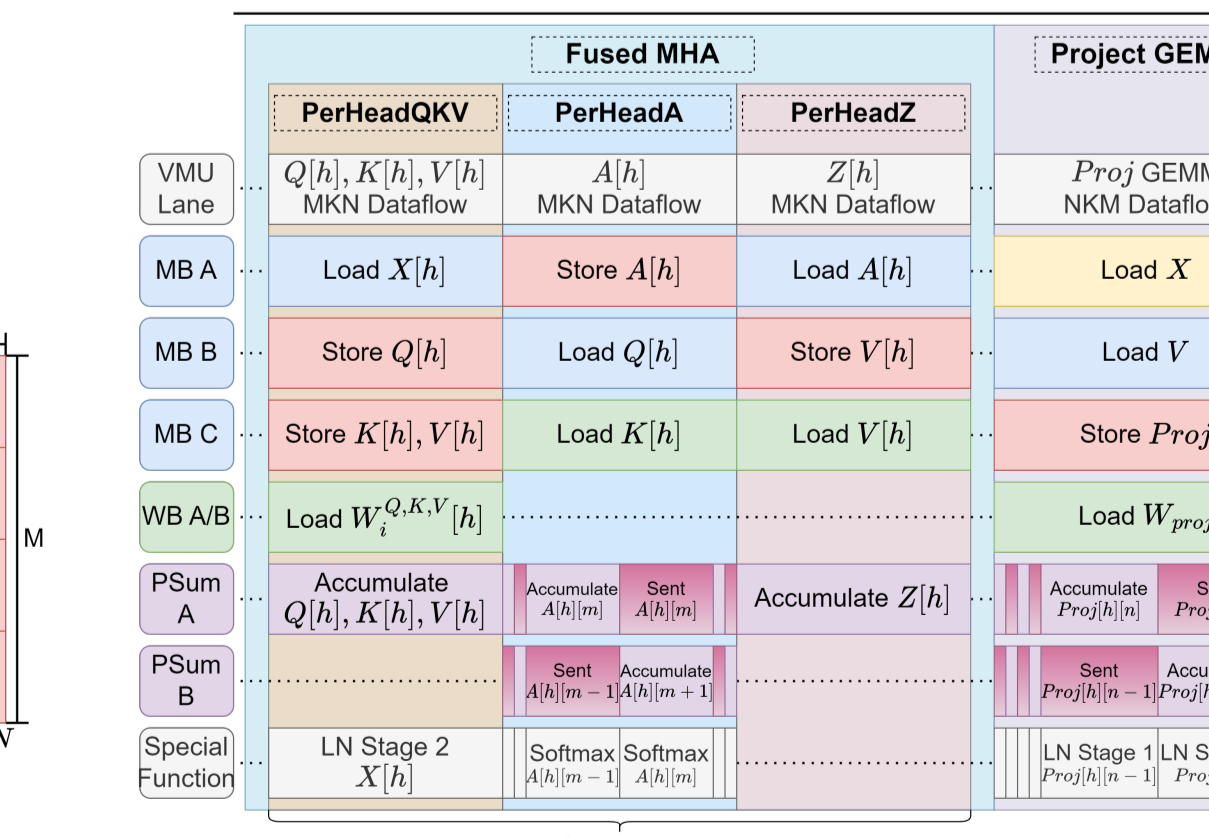
- Solution:
- Optimize the computational dataflow
  - System-level hardware accelerations

### ViTA Top Arch and GEMM Mapping



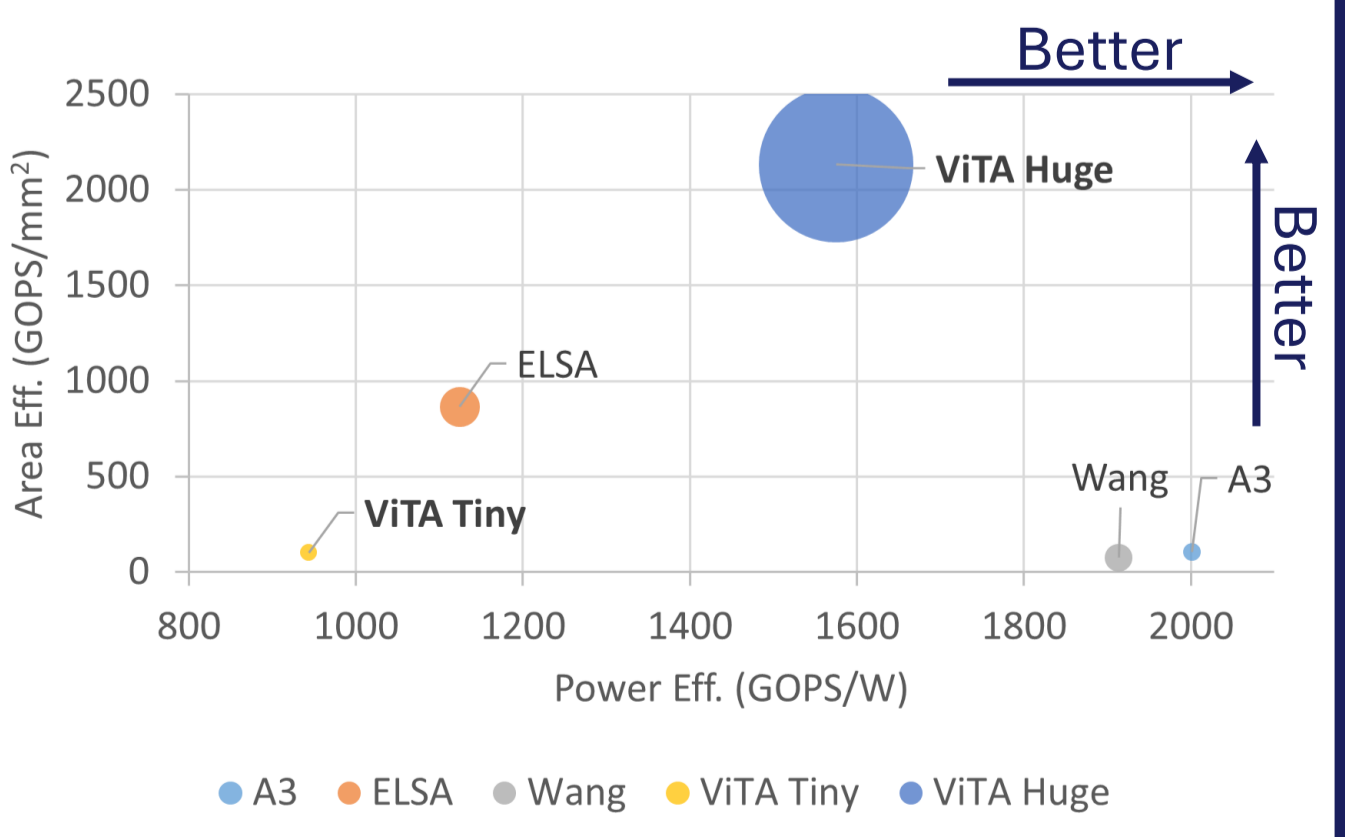
- Memory-centric, scalable ViTA architecture
- A new module for non-linear functions

### ViTA Fused-MHA Dataflow



- With fused MHA: Mem. for VGA: 22.66  $\rightarrow$  5.28 MB
- 4.3x less memory for MHA (VGA img)
- 9.4x smaller bandwidth requirement (VGA img)
- 16.384 TOPS @ 1 GHz
- 2.13 TOPS/mm², 1.57 TOPS/W
- Surpassing SOTA by 27.85x and 1.40x

### Experimental Results



[1] Chunyun Chen, Zhe Wang, Jie Lin and Mohamed M. Sabry Aly, "Efficient Tunstall Decoder for Deep Neural Network Compression", in Proceedings of the 2021 28th ACM/IEEE Design Automation Conference (DAC), 2021, pp. 1021-1026, doi: 10.1109/DAC18074.2021.9586173.

[2] Chunyun Chen, Tianyi Zhang, Zehui Yu, Adithi Raghuraman, Shwetalaxmi Udayan, Jie Lin and Mohamed M. Sabry Aly, "Scalable Hardware Acceleration of Non-Maximum Suppression", in Proceedings of the 2022 Design, Automation & Test in Europe Conference & Exhibition (DATE), 2022, pp. 96-99, doi: 10.23919/DATE54114.2022.9774717.

[3] Chunyun Chen, Lantian Li, and Mohamed M. Sabry Aly, "ViTA: A High Efficient Dataflow and Architecture for Vision Transformers", in Proceedings of the 2024 Design, Automation & Test in Europe Conference & Exhibition (DATE), 2024.