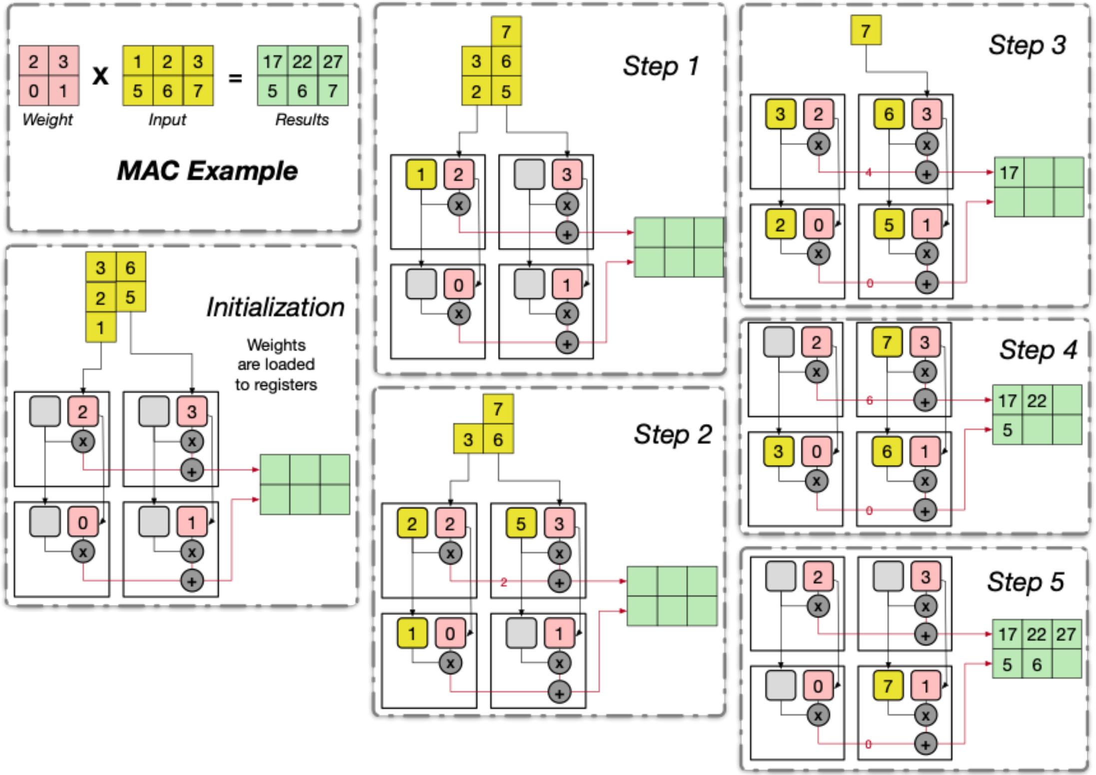
DATE 21
DESIGN, AUTOMATION & TEST IN EUROPE
01 – 05 February 2021 · virtual conference
The European Event for Electronic System Design & Test

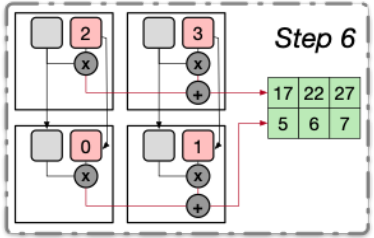
Hardware-Software Codesign of Weight Reshaping and Systolic Array Multiplexing for Efficient CNNs

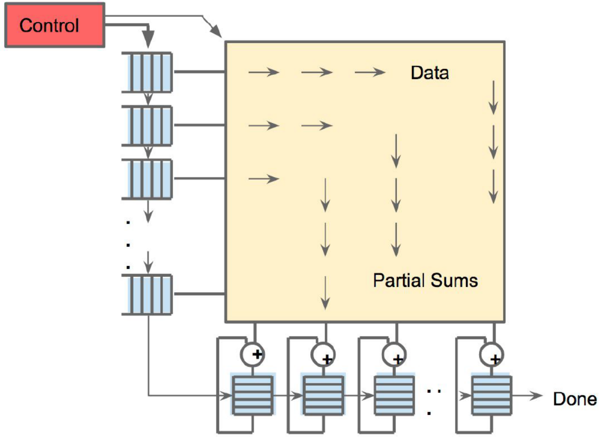
Jingyao Zhang[1], Huaxi Gu[1], Grace Li Zhang[2], Bing Li[2], Ulf Schlichtmann[2]

[1] Xidian University, [2]Technical University of Munich

Background: Systolic Array





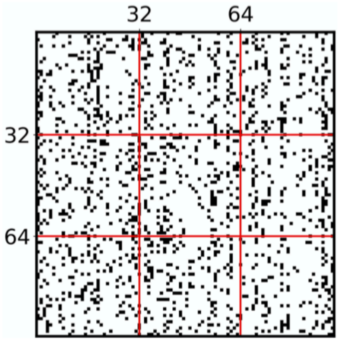


Systolic array in TPU (Google)

N. P. Jouppi et al., "In-datacenter performance analysis of a tensor processing unit," in Proceedings - International Symposium on Computer Architecture, 2017, vol. Part F1286, pp. 1–12.

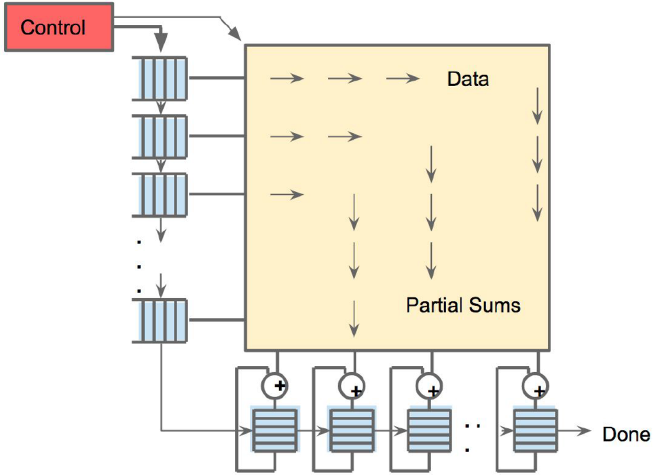
Challenges for Systolic Array

- **Sparse matrix multiplication** is inefficient in **systolic array**
- Disadvantages of systolic array
  - Not good at exploiting irregular parallelism
  - Relatively special purpose → need software, programmer support to be a general purpose model



Sparse Weight Matrix

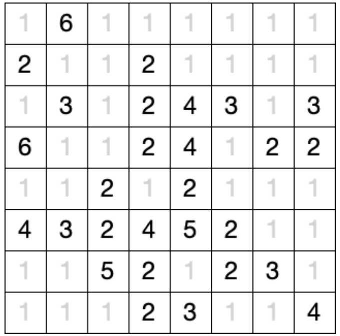
Underutilization



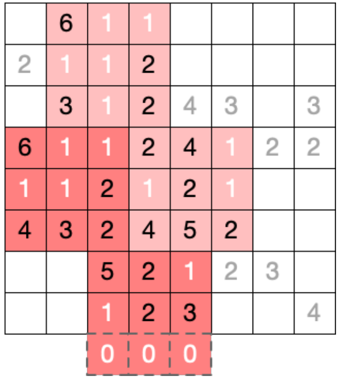
https://safari.ethz.ch/architecture/fall2020/lib/exe/fetch.php?media=onur-comparch-fall2020-lecture27-systolicarrays-afterlecture.pptx

H. T. Kung, B. McDanel, and S. Q. Zhang, "Packing Sparse Convolutional Neural Networks for Efficient Systolic Array Implementations: Column Combining Under Joint Optimization," in Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems, 2019, pp. 821–834.

Previous work: Weight Pruning for Systolic Array



Unstructured Pruning



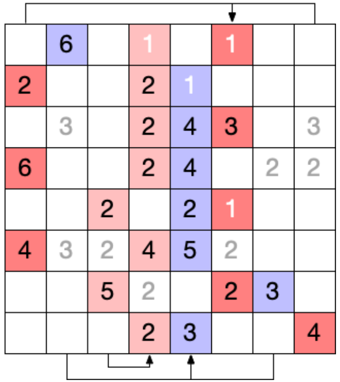
Block-wise Pruning (DAC '19)

Pros:

- *Less computation*
- *No input modification*

Cons:

- *Low compression ratio*



Column Combining (ASPLOS '19)

Pros:

- *Less computation*
- *Relative high accuracy (or low compression ratio)*

Cons:

- *Hard to find optimal group*
- *Rearrange input before computation*
- *MUX for each PE*

B. Asgari et al., "LODESTAR: Creating Locally-Dense CNNs for Efficient Inference on Systolic Arrays," in Proceedings of the 56th Annual Design Automation Conference 2019, 2019, pp. 233:1--233:2.

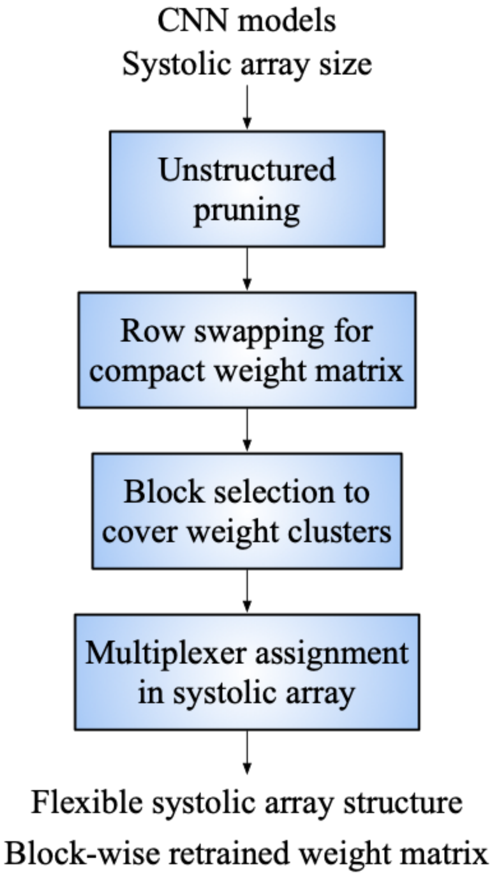
H. T. Kung, B. McDanel, and S. Q. Zhang, "Packing Sparse Convolutional Neural Networks for Efficient Systolic Array Implementations: Column Combining Under Joint Optimization," in Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems, 2019, pp. 821–834.

Solution

To **fully** utilize the systolic array with **high compression ratio of CNN models**, we propose a **hardware-software codesign framework**.

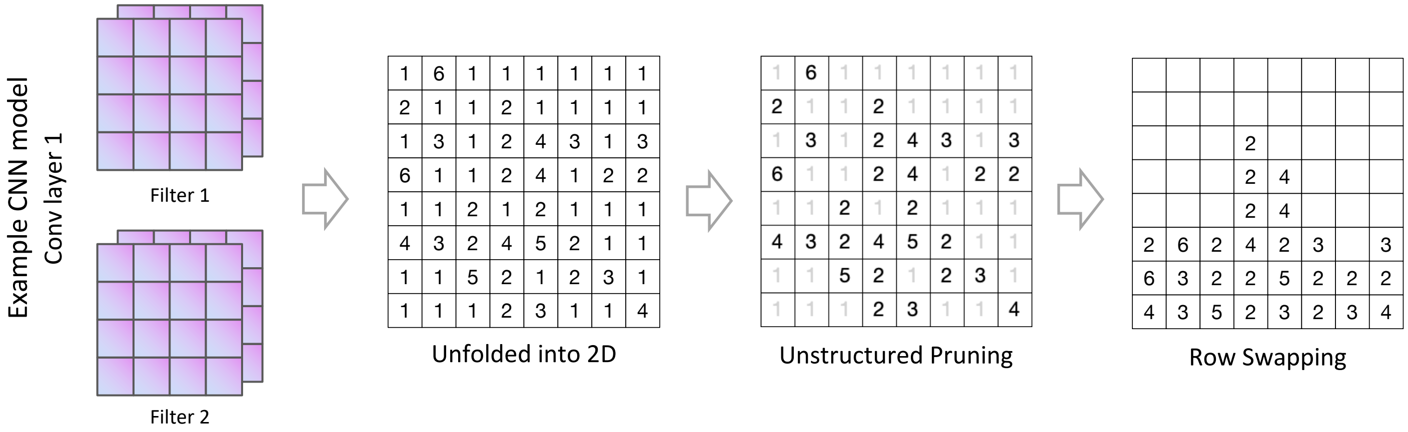
The framework outputs a **flexible** systolic array structure that sustains a balance between **latency** and **hardware cost** by:

- S Row swapping for compact storing of weight matrix
- S Block selection to cover the dense cluster of weights
- H Systolic array multiplexing
- H Genetic searching for flexible structure



Unstructured Pruning and Row Swapping

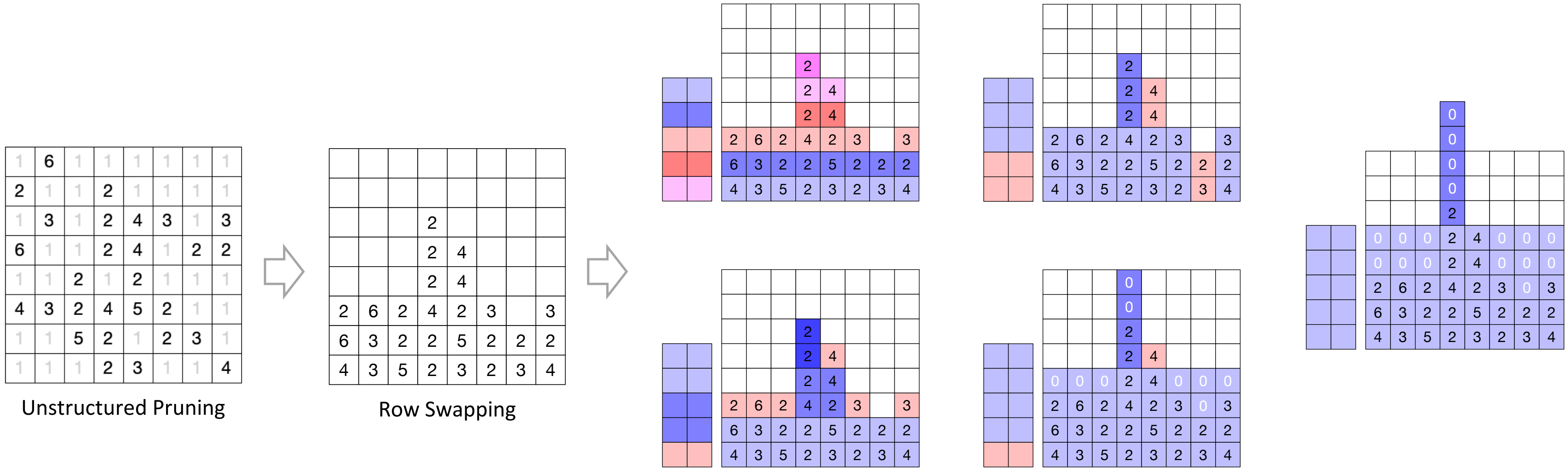
- Unstructured pruning [1] is chosen to achieve a **high compression ratio**
- Row swapping is applied for compact weight matrix
  - Only rows with non-zero weights should be indexed when each column has weights
  - A small number of columns should be indexed only when some columns are empty



[1] S. Han, H. Mao, and W. J. Dally, "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding," arXiv e-prints, p. arXiv:1510.00149, Oct. 2015.

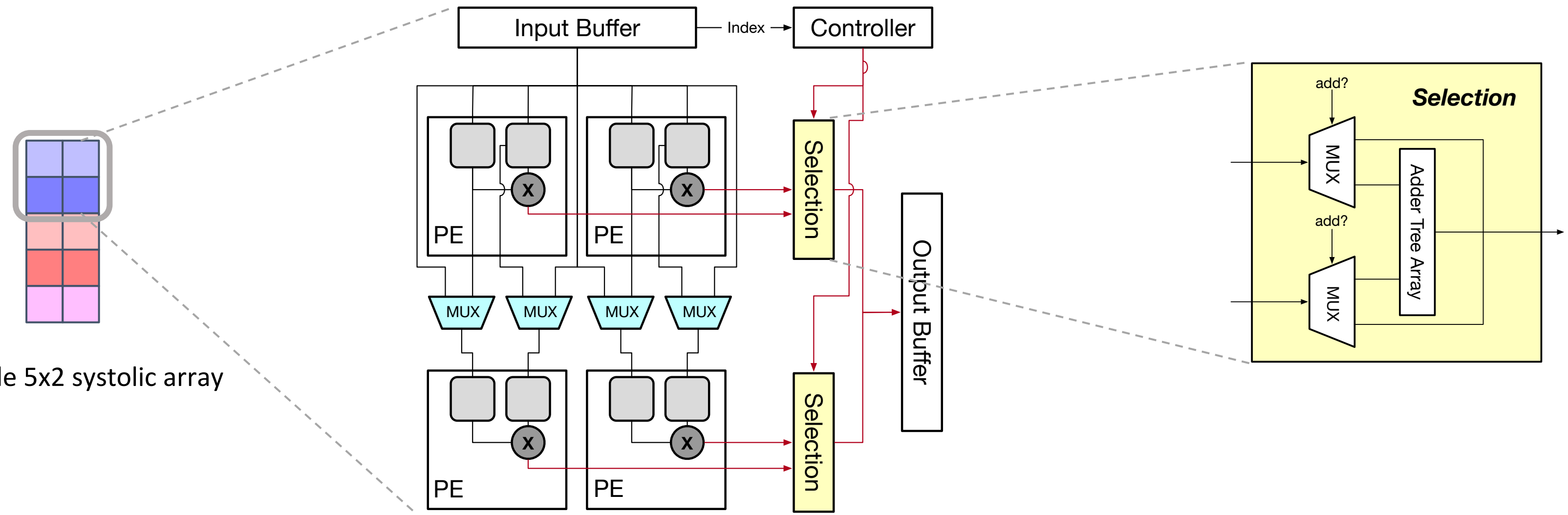
Block Selection

- After row swapping, weights are reorganized into a **dense cluster**
- To find the optimal block set to cover the **weight cluster** seamlessly, we enumerate all the possible block sets according to the hardware constraint



Microarchitecture of Systolic Array

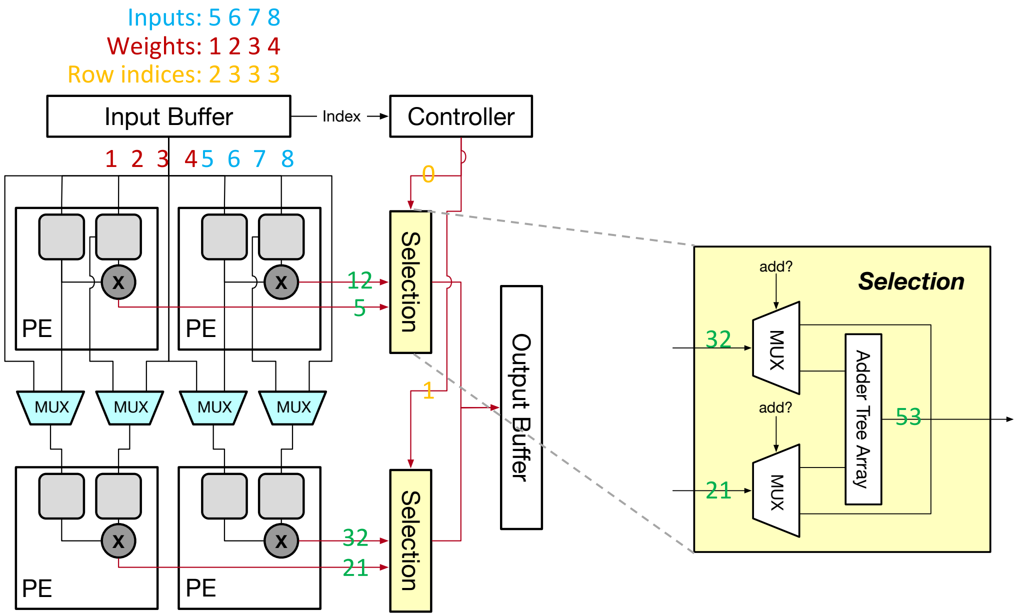
- To support the concurrent computations of various blocks with the corresponding inputs on systolic arrays, multiplexers are inserted into the arrays.
- To support the row swapping, a controller and selection modules are required to schedule the multiplication results.



Example 5x2 systolic array

Computation of Modified Systolic Array

1. Decode row indices of weights by the **controller**
2. Load weights
3. Load inputs
4. Multiply weights and inputs
5. Transmit products to **selection** modules
6. **Controller** sends signal to **selection** modules
7. **Selection** modules handle received products
   - If weight are in the different row of the weight matrix, stores products respectively
   - If weights are in the same row of the weight matrix, add products and store the addition results



Genetic Algorithm for Multiplexer Assignment

- The locations of multiplexers for **different CNN models** vary significantly
- A **genetic algorithm** is deployed to select where the multiplexers should be inserted considering both latency and hardware cost

Algorithm 1: Genetic algorithm

START
1 Generate the initial population
2 Compute fitness: E
REPEAT
3 Tournament selection
4 Crossover
5 Polynomial Mutation
6 Compute fitness: $E = C_l * L + C_a * A + C_w * W$
UNTIL stopping criteria are satisfied

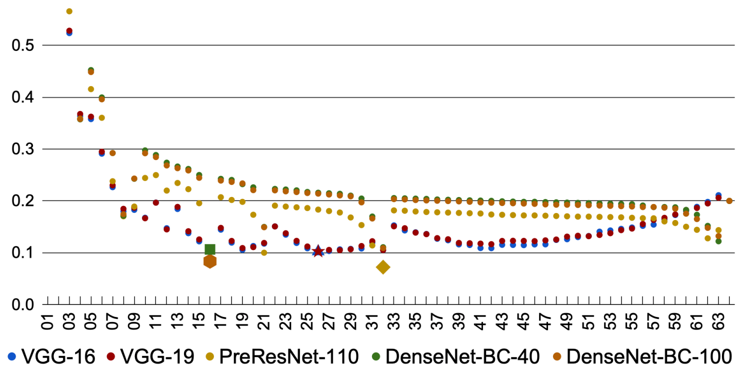
Mutation equation: $x_i^{\{1,t+1\}} = x_i^{\{1,t\}} + \beta_i$

$$\beta_i = \begin{cases} (2u_i)^{\frac{1}{\eta_u+1}} - 1, & u_i < 0.5 \\ 1 - [2(1-u_i)]^{\frac{1}{\eta_u+1}}, & u_i \geq 0.5 \end{cases}$$

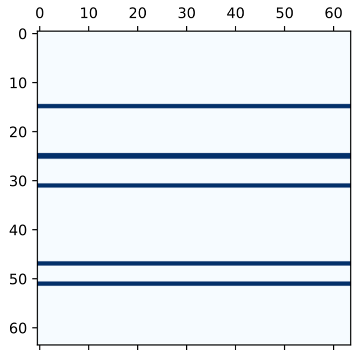
Experimental Results

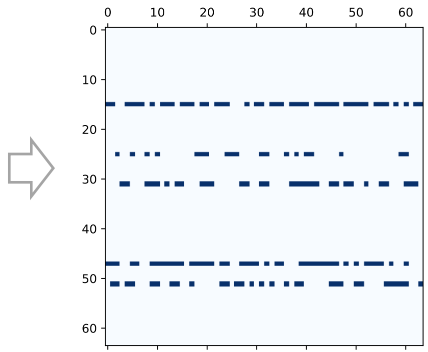
TABLE I
EXPERIMENTAL RESULTS COMPARED WITH THE UNSTRUCTURED PRUNING METHOD

Dataset CIFAR-100	Compression Ratio (%)		Accuracy (%)		Latency		#Multiplexers	
	Baseline	*Proposed*	*Baseline*	*Proposed*	*Baseline*	*Proposed*	*Proposed*	*#Mult./#PEs (%)*
VGG-16	94.67	93.66	71.32	72.07	1	0.324	128	3.13
VGG-19	94.76	94.04	70.81	71.5	1	0.398	128	3.13
PreResNet-110	92.89	67.37	65.2	73.59	1	0.681	64	1.56
DenseNet-BC-40	90.99	75.11	67.27	71.95	1	0.291	192	4.69
DenseNet-BC-100	93.41	75.48	73.87	76.56	1	0.269	192	4.69



Evaluation results for CNN models


Before genetic searching


After genetic searching

Conclusion

- A **hardware-software codesign framework** is proposed to exploit systolic arrays for the computations of various CNNs efficiently

- By row swapping, block selection, systolic array multiplexing and genetic searching for multiplexer assignment, a flexible systolic array structure is developed with accordingly pruned CNN models

- The experimental results show that the **latency** can be reduced significantly with low **hardware cost** and high **inference accuracy**

Q&A

- All the questions and comments are welcomed