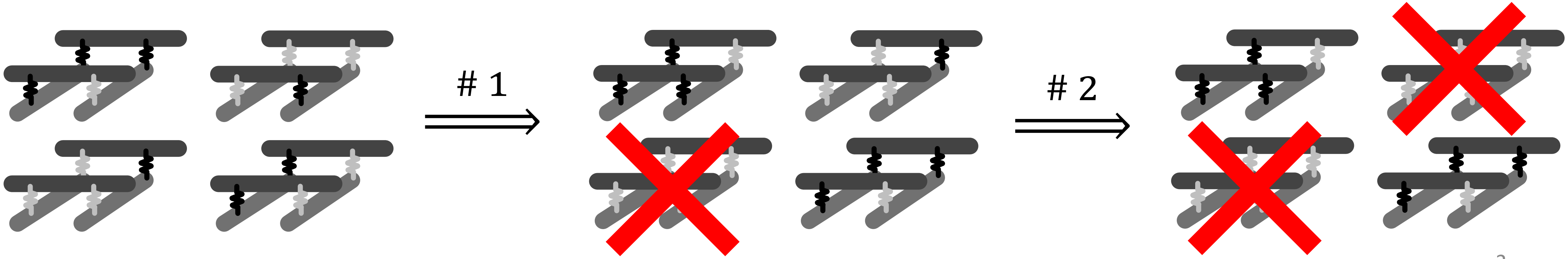
Learning the Sparsity for ReRAM: Mapping and Pruning Sparse Neural Network for ReRAM based Accelerator

Jilan Lin[1,2], Zhenhua Zhu[1], Yu Wang[1] and Yuan Xie[2]

[1]Dept. of ECE, Scalable and Energy-Efficient Architecture Lab (SEAL)
UCSB, Santa Barbara, U.S.
[2]Dept. of EE, Tsinghua National Laboratory for Information Science and Technology (TNList),
Tsinghua University, Beijing, China

Brief Summary

- **Target:** Efficient **sparse** neural network in **ReRAM**-based computing
- **Proposal 1:** Map the huge sparse matrix with column exchanging.
  - Eliminate the unnecessary ReRAM crossbars.
- **Proposal 2:** Prune neural network with grainy of ReRAM crossbar.
  - Further save more ReRAM crossbars.



Outline

- **Background & Motivation**
  - ReRAM based Computing for Neural Networks
  - Sparse Neural Network
- **Proposed Solutions**
  - Sparse Neural Network Mapping
  - Crossbar-Grained Pruning
- **Simulation Results**
- **Conclusion**

Outline

- **Background & Motivation**
  - ReRAM based Computing for Neural Networks
  - Sparse Neural Network
- **Proposed Solution**
  - Sparse Neural Network Mapping
  - Crossbar-Grained Pruning
- **Simulation Results**
- **Conclusions**

Background & Motivation

- Neural networks (NNs) now dominate the field of machine learning.

Neural Networks

- Neural networks (NNs) now dominate the field of machine learning.

Google Translation

Neural Networks

- Neural networks (NNs) now dominate the field of machine learning.

Google Translation

AlphaGo [Silver D_nature_2016]

Neural Networks

- Neural networks (NNs) now dominate the field of machine learning.

Google Translation

AlphaGo

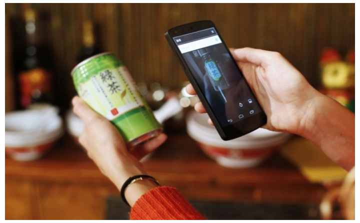
AlphaGo [Silver D_nature_2016]

Pedestrian detection

[Zhang_CVPR_2016]

Neural Networks

- Neural networks (NNs) now dominate the field of machine learning.
  - Key operations: Matrix-Vector/Matrix Multiplication

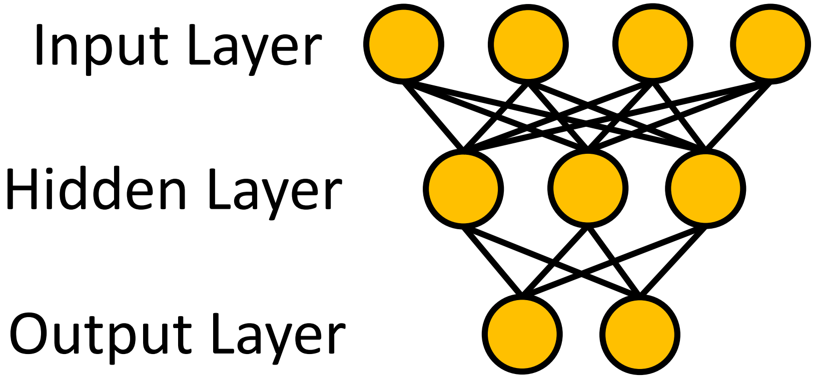


Google Translation

AlphaGo

AlphaGo [Silver D_nature_2016]

Pedestrian detection
[Zhang_CVPR_2016]



Fully Connected NN

Neural Networks

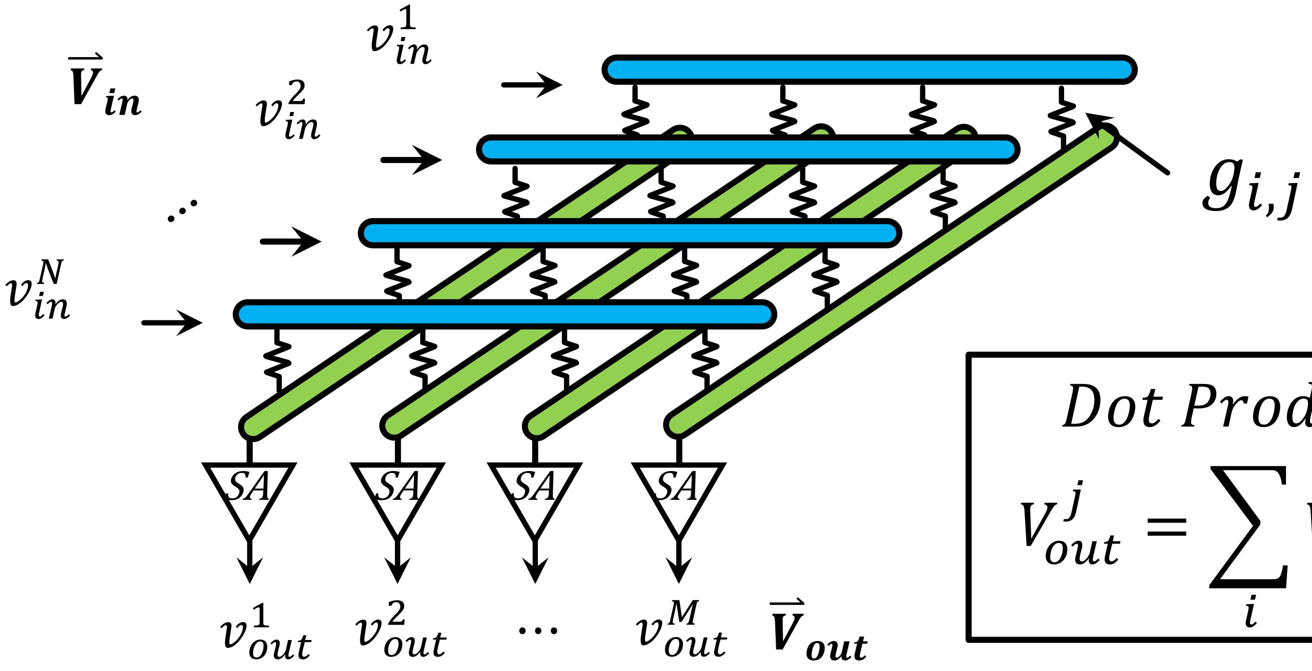
- NNs are hardware-expensive, due to the huge amount of parameters.
  - For VGG-16:
  552 MB paras, 1.6×10^{10} ops (forward), 4×10^{4} iterations (backward) [1][2]

- Fully connected (FC) layer: frequently used but extremely large.
  - For FC1 in VGG-16: Size of 25088×4096
  - Memory-bound with limited bandwidth.

1. Cheng, Ming, et al. "Time: A training-in-memory architecture for memristor-based deep neural networks." DAC 2017. ACM, 2017.
2. Chi, Ping, et al. "Prime: A novel processing-in-memory architecture for neural network computation in ReRAM-based main memory." ACM SIGARCH Computer Architecture News. Vol. 44. No. 3. IEEE Press, 2016.

Resistive Random Access Memory (ReRAM)

- ReRAM provides a promising solution to compute matrix efficiently.
  - Storing information with resistive cell.
  - Reducing the complexity with crossbar array: $O(n^2) \rightarrow O(n^0)$

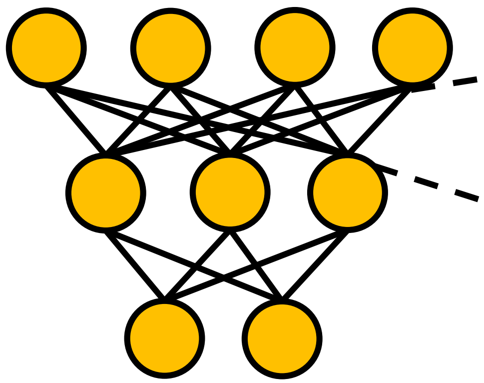


Dot Product:

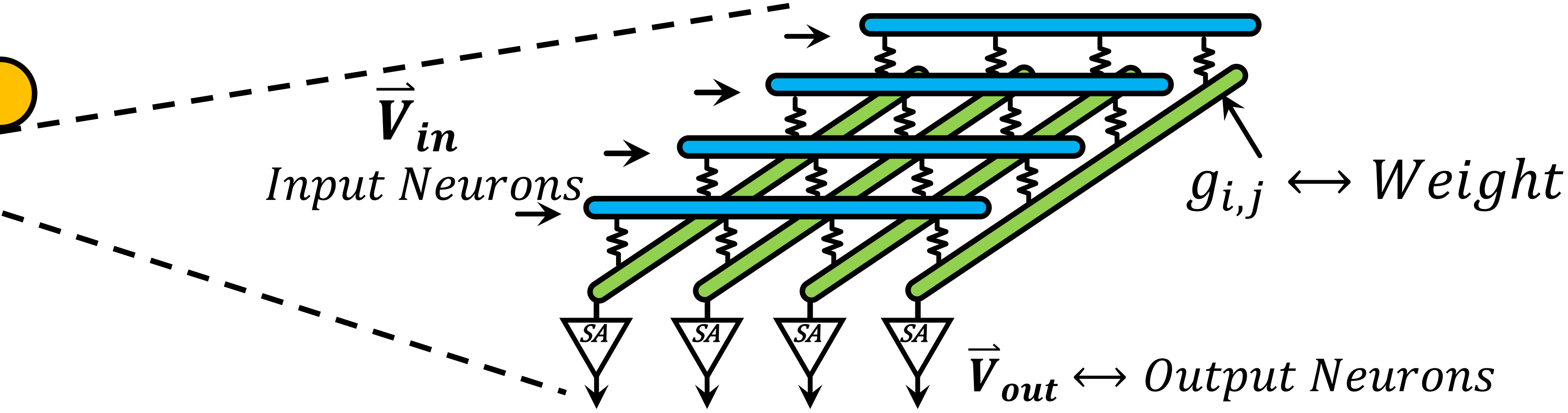
$$V_{out}^j = \sum_i V_{in}^i \, g_{i,j}$$

ReRAM based NN Acceleration

- ReRAM based NN acceleration is attractive.
  - In-memory computing/Low power/Scalable …
  - PRIME [ISCA 2016], ISAAC [ISCA 2016], and PipeLayer [HPCA 2017].



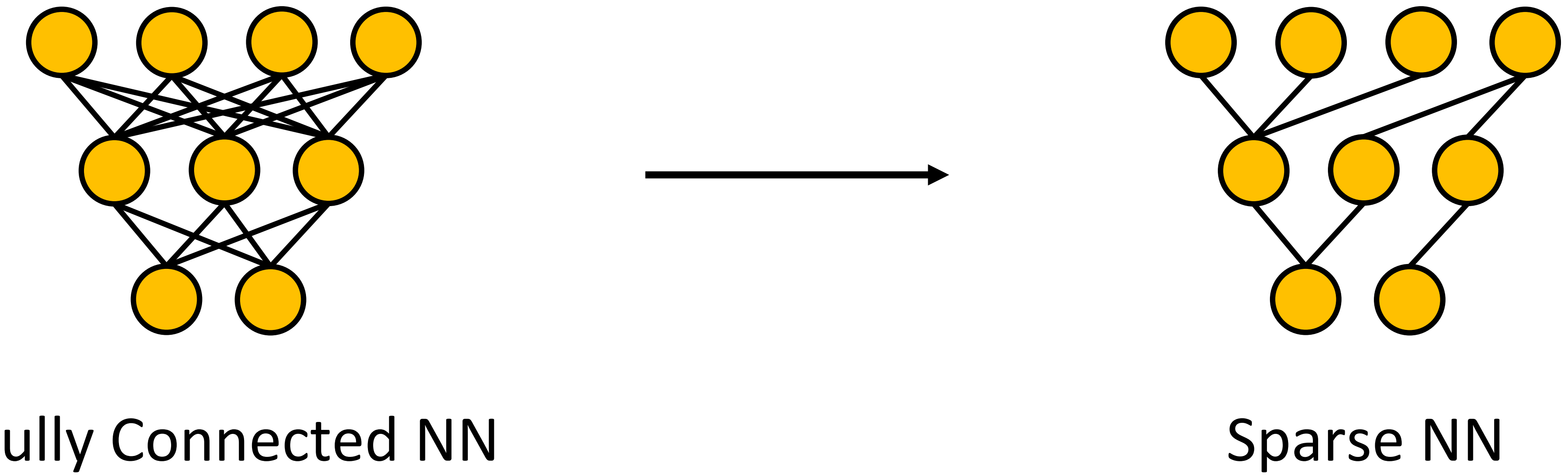
Fully Connected NN



ReRAM based Computing

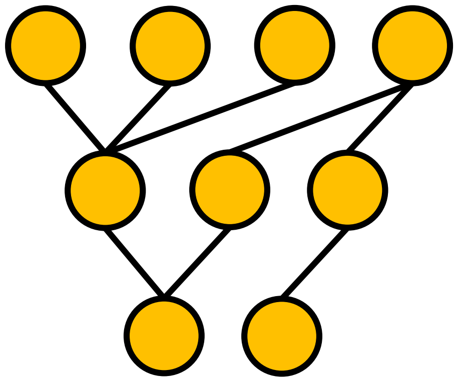
Sparse NN

- Learning the Sparsity for NN brings significant advantages.
  - Compressing the model ~10X
  - Avoiding overfitting.

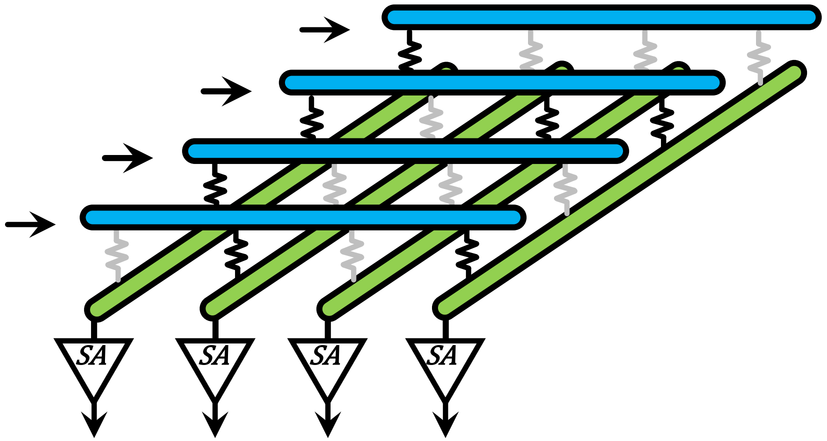


Sparse NN V.S. ReRAM Crossbar

- The crossbar structure is contradictory with sparse matrix.



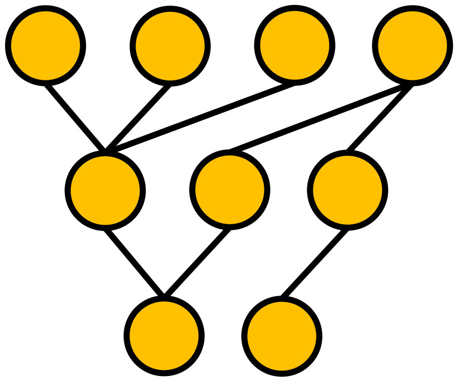
Sparse NN



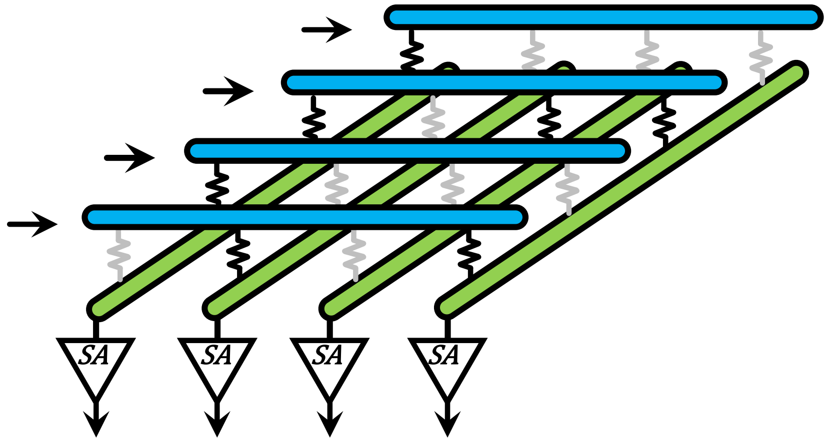
ReRAM based Computing

Sparse NN vs. ReRAM Crossbar

- The crossbar structure is contradictory with sparse matrix.
  - Matrix must be stored in dense way for O(1) computing.
  - No benefits from sparsity.



Sparse NN



ReRAM based Computing

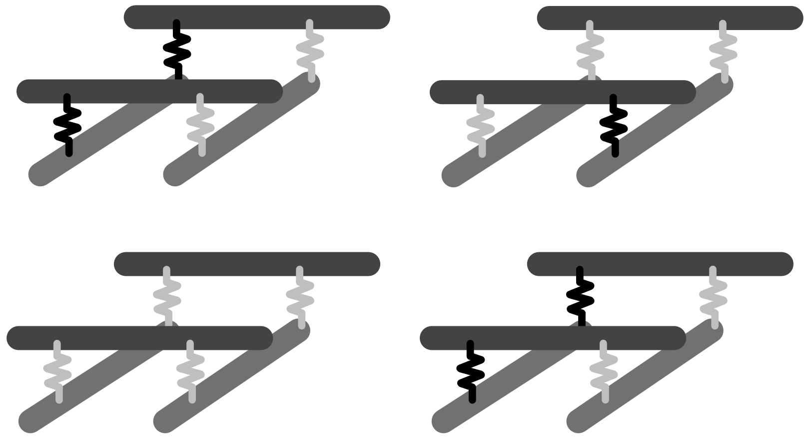
Outline

- Background & Motivation
  - ReRAM based Computing for Neural Networks
  - Sparse Neural Network
- **Proposed Solution**
  - Sparse Neural Network Mapping
  - Crossbar-Grained Pruning
- Simulation Results
- Conclusions

1: Mapping

- Observation 1: The matrix can be quite large but quite sparse.
  - FC1 in VGG16 (25088 × 4096): Cannot map to a single crossbar
  - 90% paras vs. 96% sparsity after pruning. [Han, NIPS 2016].
  - ReRAM can only be positive: Even more sparse. **Density: 4%→2%**

$$\begin{matrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{matrix}$$



Solution 1: Column Exchanging based Mapping

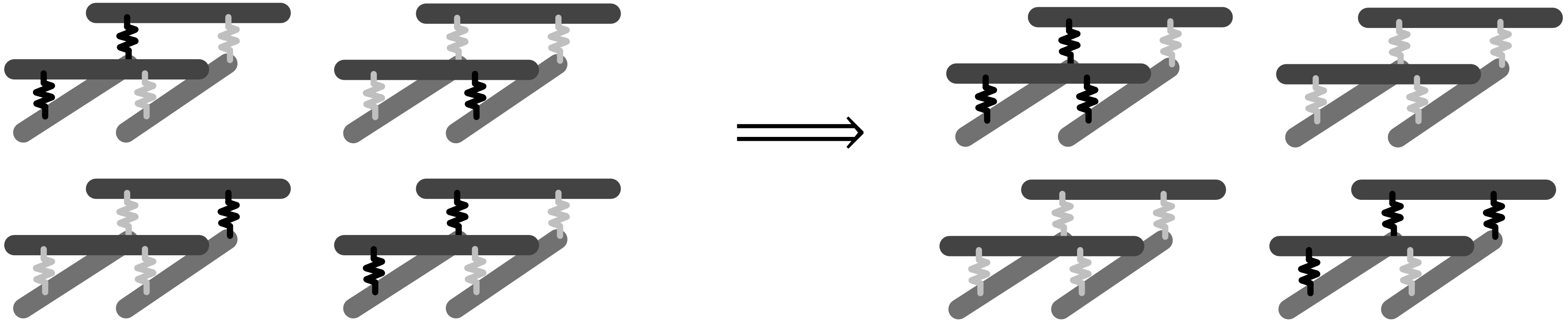
- Key idea: Exchange the column to make non-zero element gathered.

Original Matrix

$$\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{array}$$

Column Exchanging

$$\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \end{array}$$



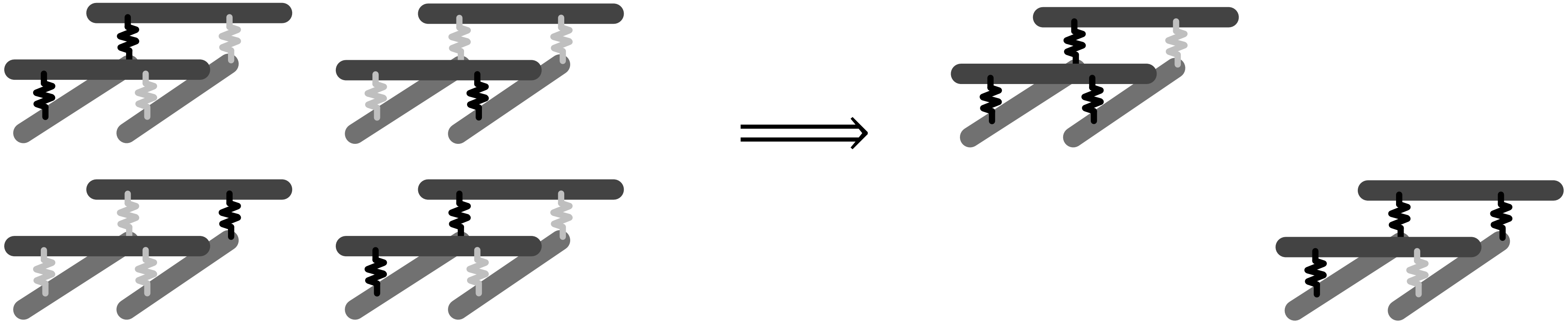
Solution 1: Column Exchanging based Mapping

- Key idea: Exchange the column to make non-zero element gathered.

Original Matrix

$$\begin{matrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{matrix} \Longrightarrow \begin{matrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \end{matrix}$$

Column Exchanging



Solution 1: Column Exchanging based Mapping

- Key idea: Exchange the column to make non-zero element gathered.
- Proposed method: Exchanging the column based on *k-means* clustering.
  - Comparing the similarity of columns based on Hamming distance.
  - Clustering into n categories ($n \sim$ # crossbars)

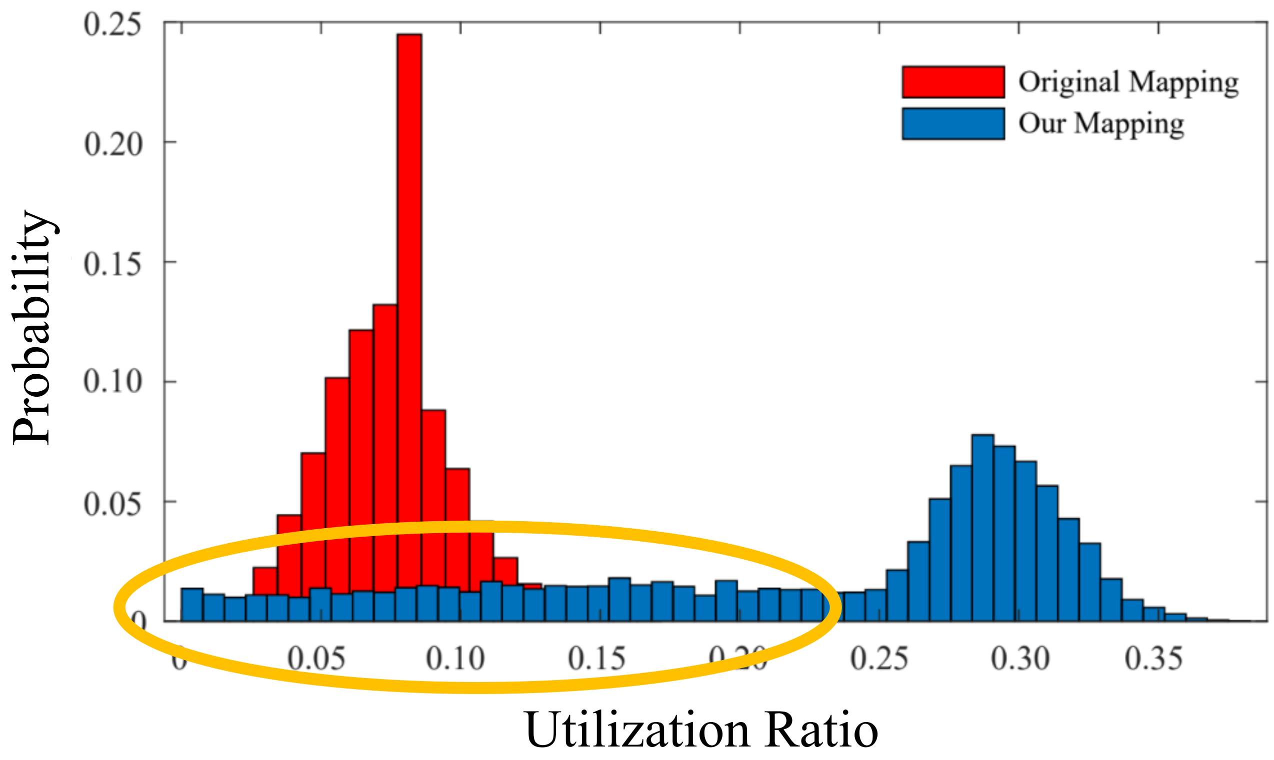
Original Matrix

$$\begin{matrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{matrix} \Longrightarrow \begin{matrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \end{matrix}$$

Column Exchanging

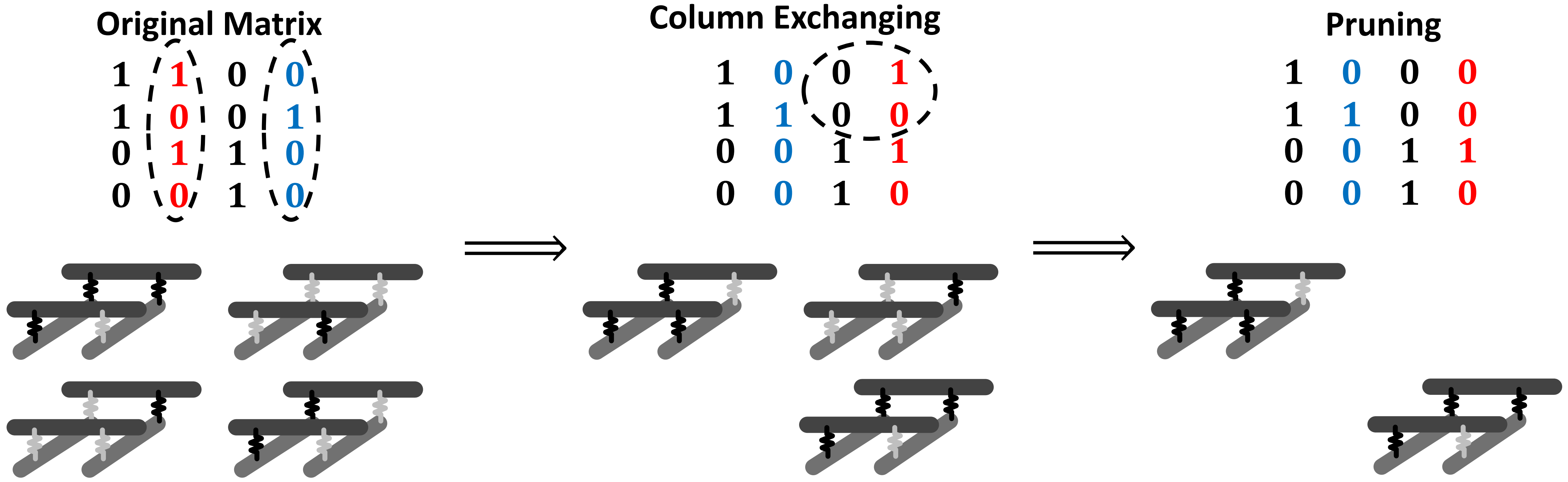
2. Crossbar Utilization

- Observation 2: There still exist crossbars with low utilization.
  - ~ 20% crossbars have less than 20% non-elements for VGG16.



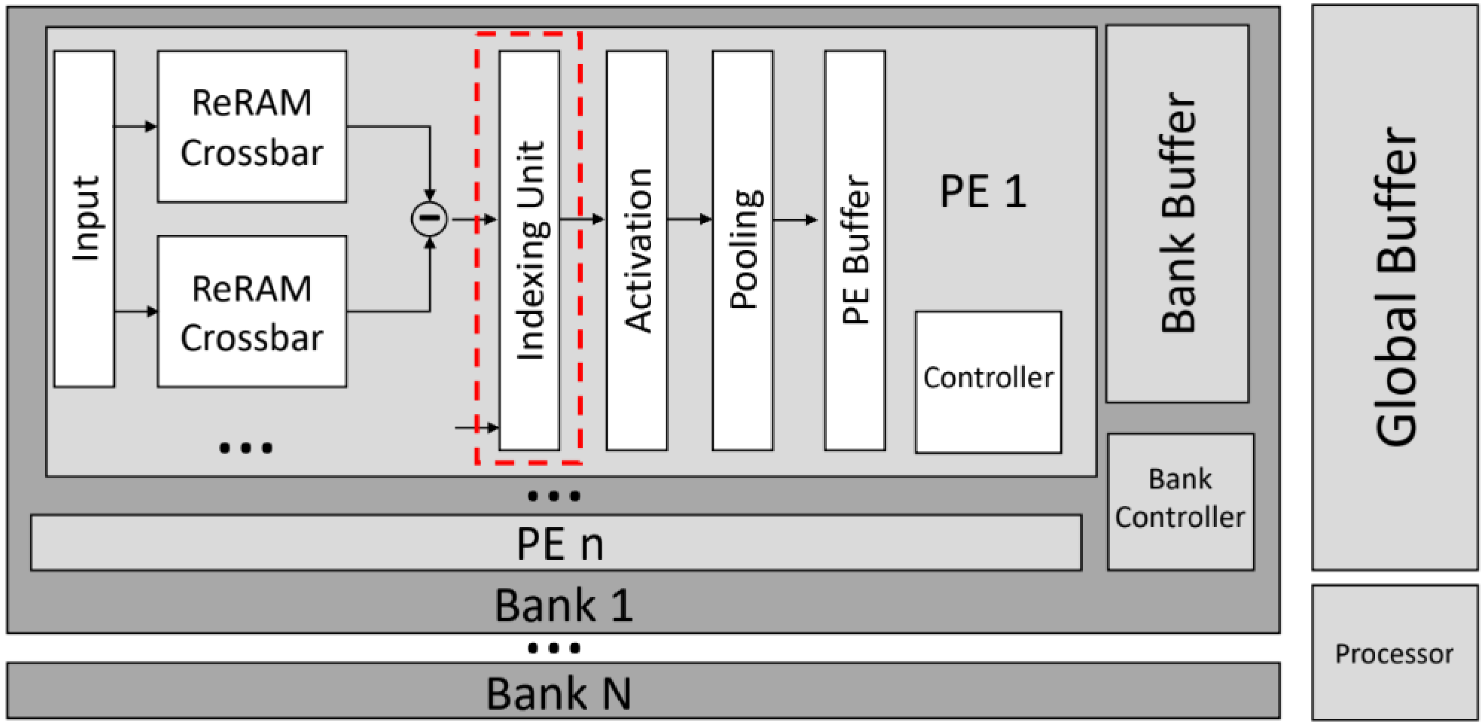
Solution 2: Crossbar-Grained Pruning

- Key idea: Prune the weights in low-utilization crossbars.
  - Finetuning the model after pruning.



Architectural Implementation

- The re-ordered mapping can be implemented in various architectures.
  - Only for outputs and not necessary for inputs.

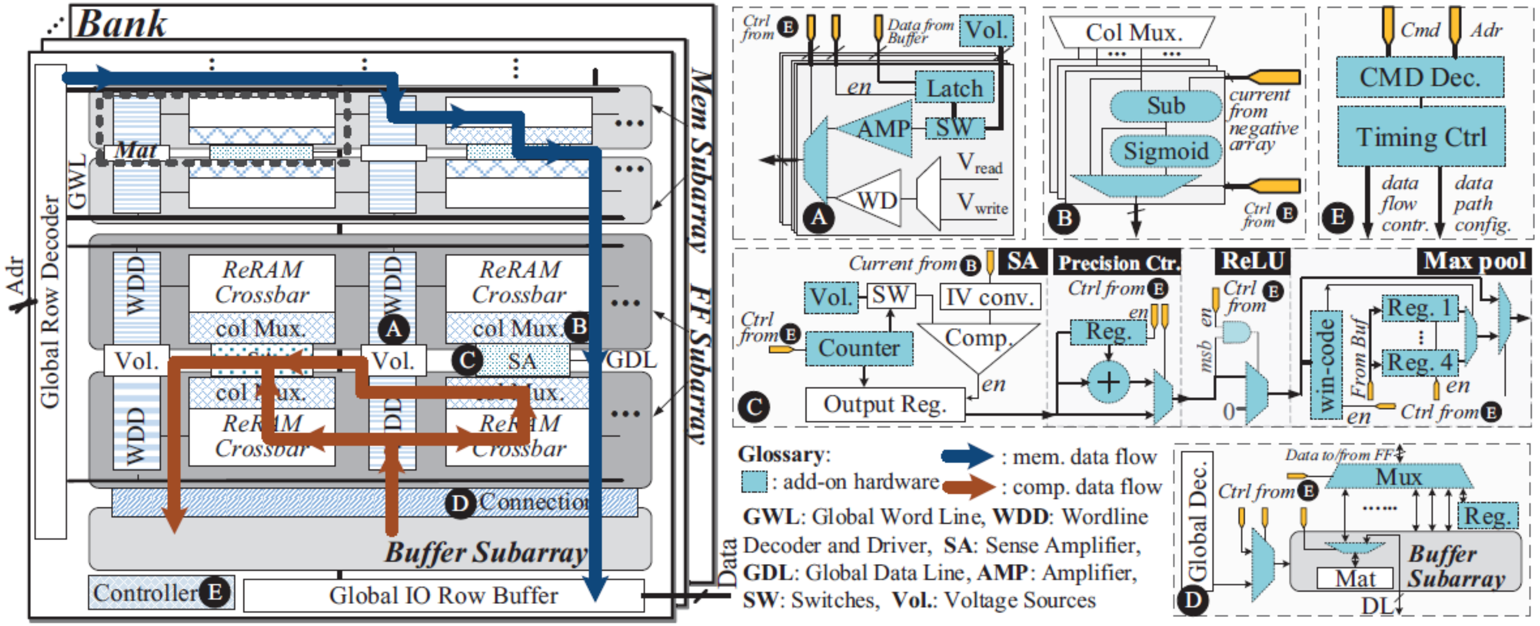


Outline

- Background & Motivation
  - ReRAM based Computing for Neural Networks
  - Sparse Neural Network
- Proposed Solution
  - Sparse Neural Network Mapping
  - Crossbar-Grained Pruning
- **Simulation Results**
- Conclusions

Simulation Setup

- Simulation setup:
  - Implemented on PRIME [ISCA 2016] with 45nm technology.



The architecture of PRIME

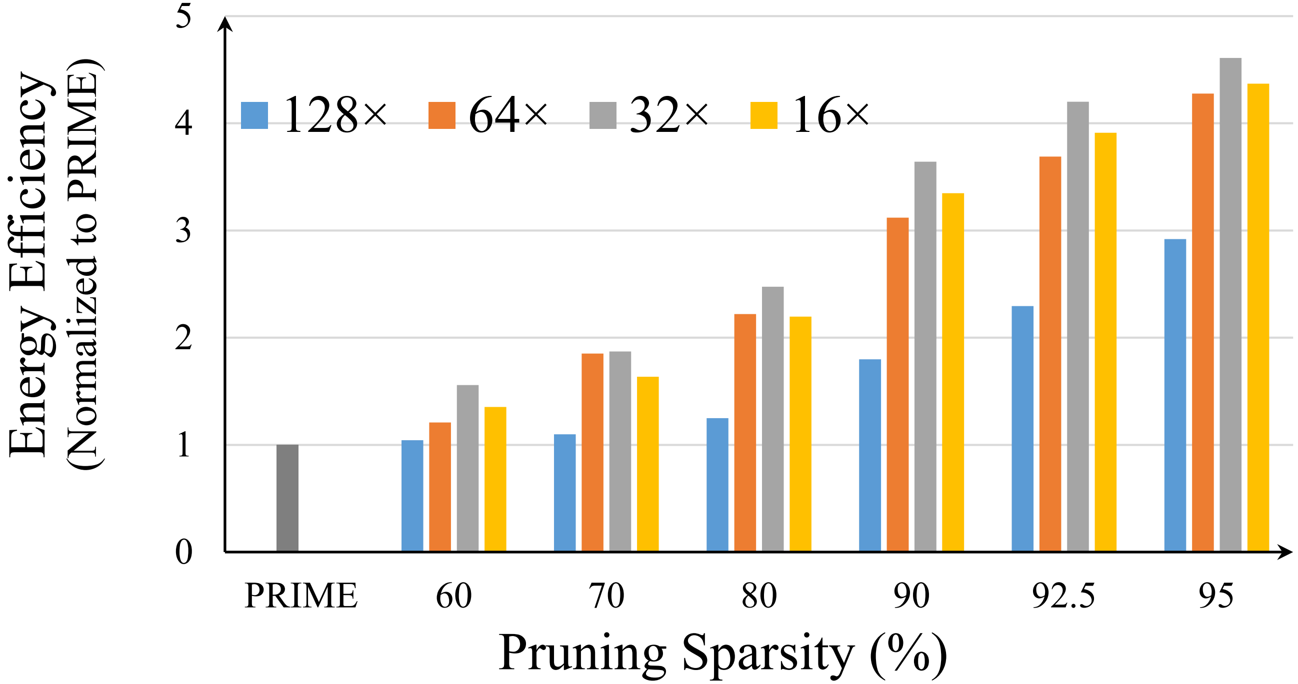
Simulation Setup

- Simulation setup:
  - Implemented on PRIME with 45nm technology.
  - Benchmarks:

NNs	LeNet-5	AlexNet	VGG-16	ResNet-18	LSTM-5
Dataset	MNIST	ImageNet	CIFAR-10	CIFAR-10	LibriSpeech
Sparsity	92%	89%	92.5%	75%	85%

Energy Results – Sparse Mapping

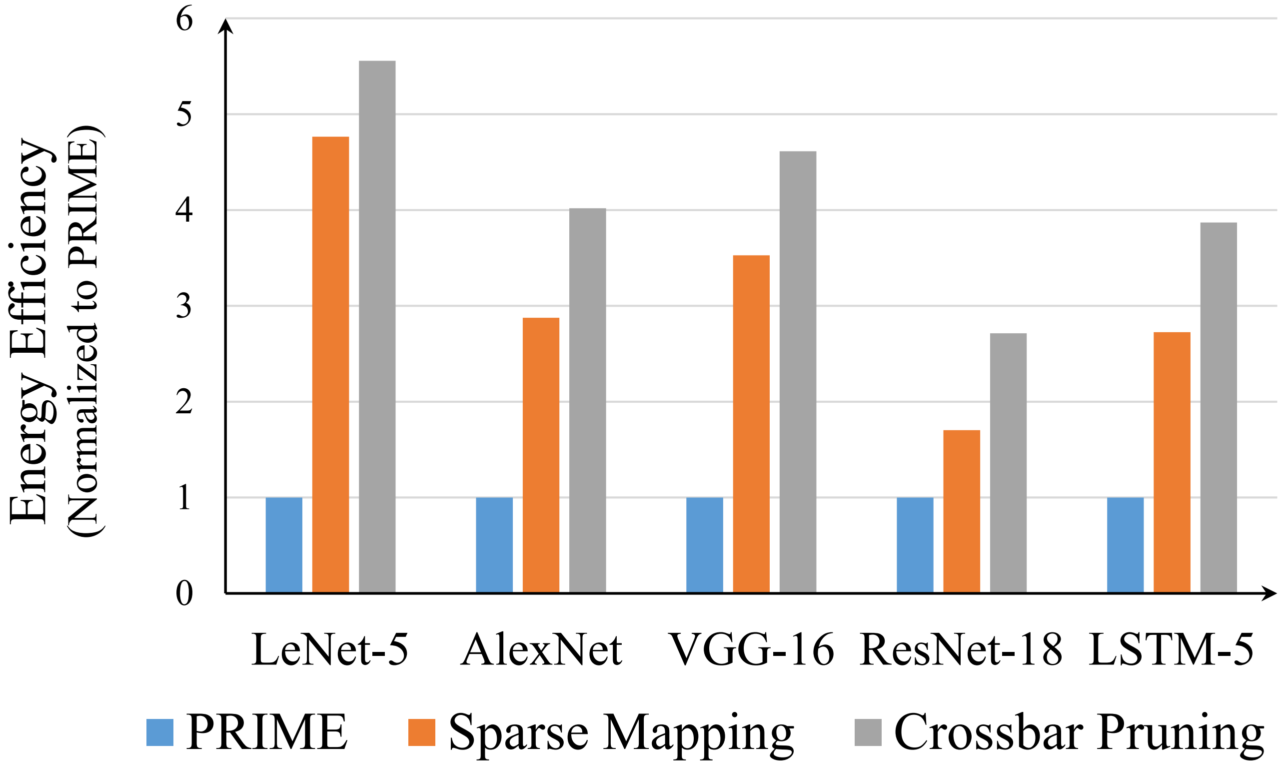
- Energy results among different crossbar sizes:
  - Works better for smaller ReRAM crossbars/more sparse models.
  - ~ 3x boosting on average observed for 90% sparsity.



Energy Results – Pruning

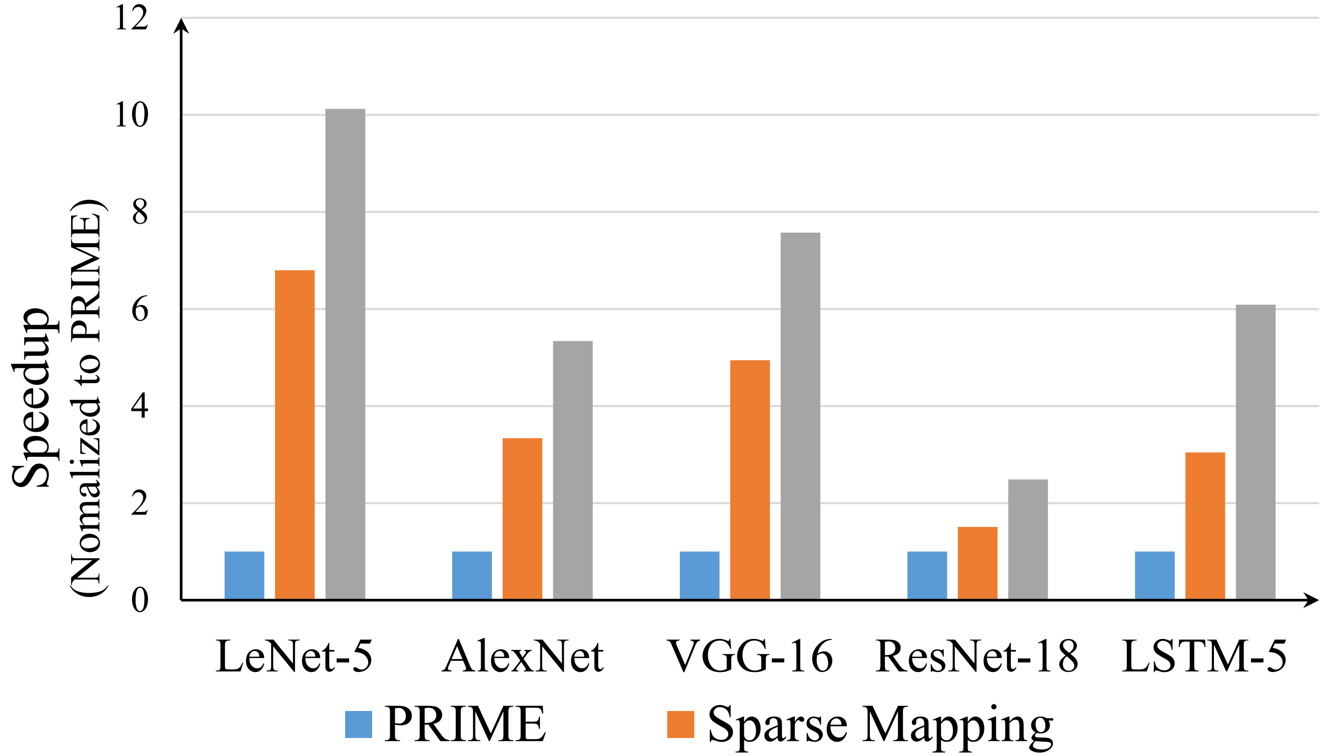
- Energy results among different benchmarks:

  -Works better for those models with large FC layers



Performance Results

- Performance results among different benchmarks:
  - Works better for those models with large FC layers



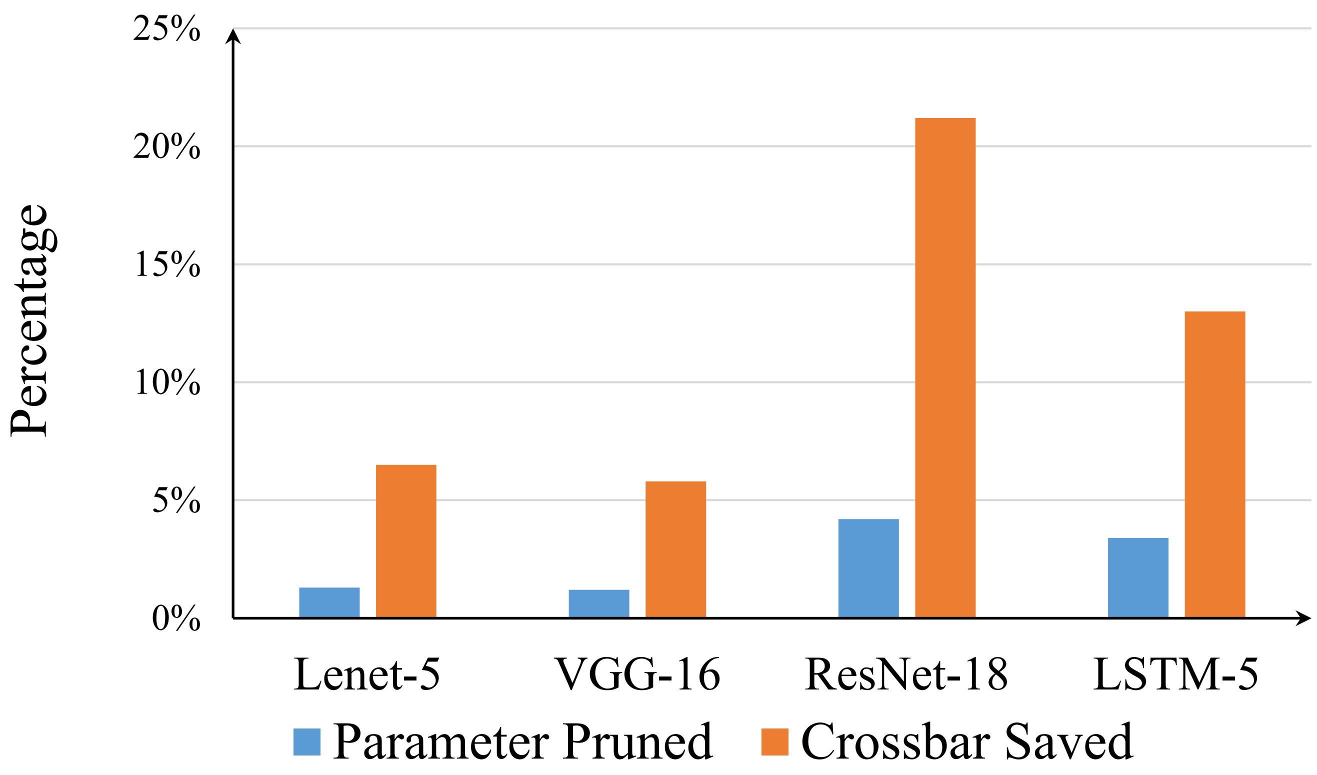
Accuracy Results

- Almost no accuracy loss/acceptable loss.
  - Compared with conventional pruning, < 0.5% accuracy loss.

Nerual Networks	LeNet-5	VGG-16	ResNet-18	LSTM-5
Original	**99.23%**	93.64%	**92.37%**	**89.24%**
Normal Pruning	99.13%	93.62%	92.07%	88.49%
Crossbar Pruning	99.15%	**93.72%**	91.78%	88.01%

Accuracy Results

- Pruned paras vs. saved crossbars:
  - Save 5x crossbars compared to pruned parameters.



Conclusions

- We propose a novel sparse NN mapping scheme based on weight columns clustering, to achieve better ReRAM crossbar utilization.
- We propose crossbar-grained pruning algorithm to reduce the crossbars with low utilization.
- Evaluation results indicate 3–5$\times$ energy efficiency and 2.5–6$\times$ speedup.
- Our pruning algorithm appears to have almost no accuracy loss.

Thanks for your attending!