Embedded Systems Week

CASES: International Conference on Compilers, Architectures, and Synthesis for Embedded Systems

Accelerating Large-Scale Graph Neural Network Training on Crossbar Diet

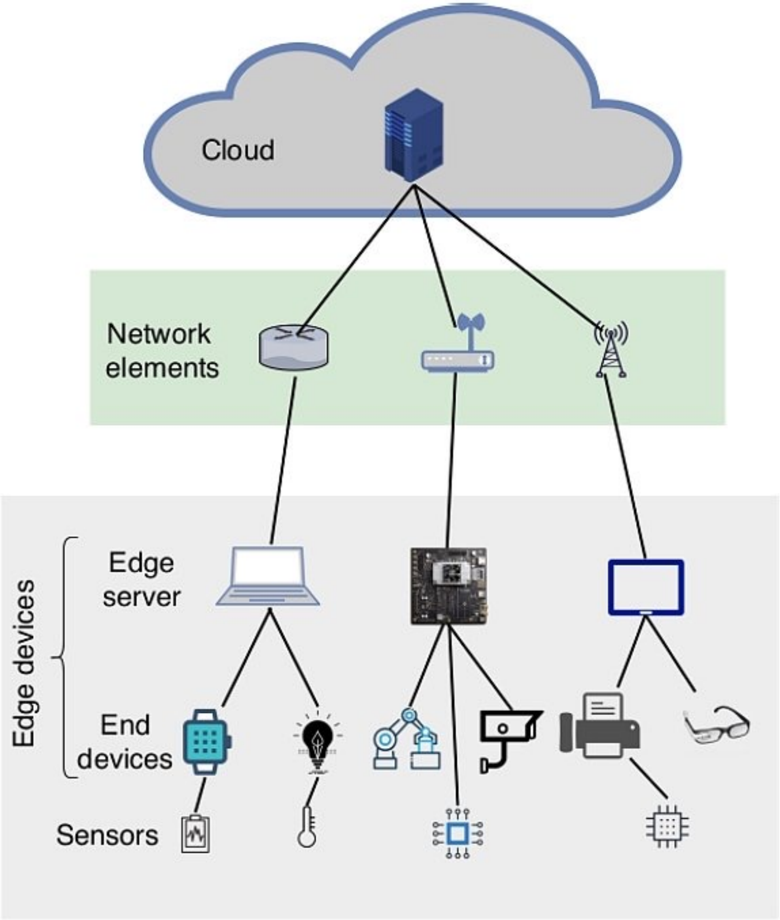
Chukwufumnanya Ogbogu†, Aqeeb Iqbal Arka†, Biresh Kumar Joardar*, Janardhan Rao Doppa†, Hai (Helen) Li*, Krishnendu Chakrabarty*, Partha Pratim Pande†
Washington State University†, Duke University*

Outline

- **Introduction:**
  - **ML models: Graph Neural Networks (GNN)**
  - **Architectures for GNN training & Inference**
- **Motivation:**
  - **ReRAM-based Process-in-memory (PIM) Computing**
- **Background & Overview:**
  - **GNN Training**
  - **3D PIM for GNNs**
- **Methodology & Related work:**
  - **Lottery Ticket Pruning (LTP)**
  - **Crossbar-aware Pruning (CAP)**
  - **DietGNN Technique.**
- **Results**
- **Conclusion**

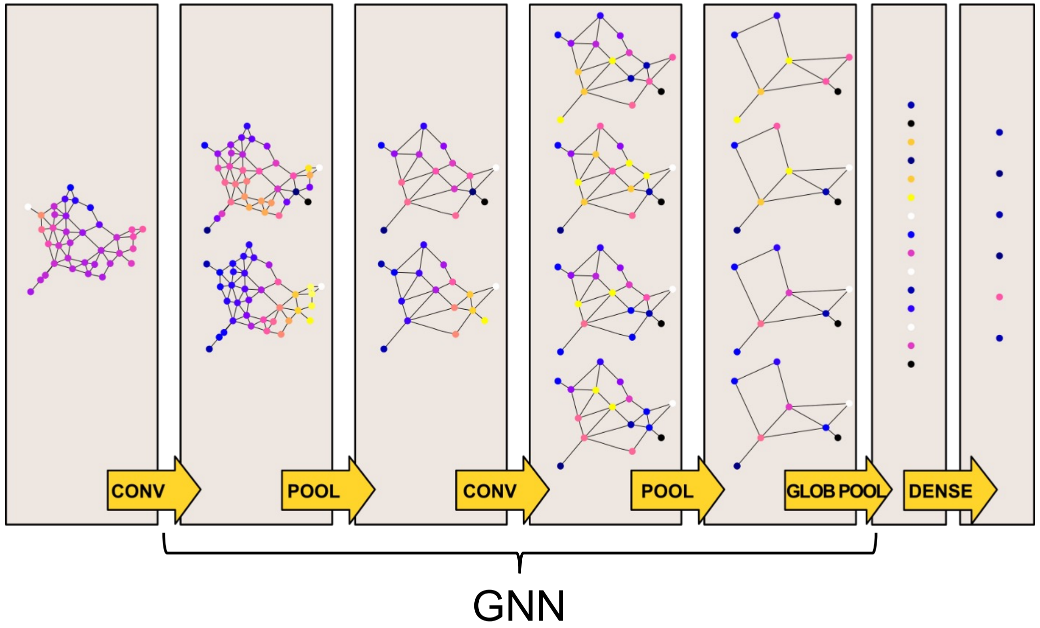
Introduction

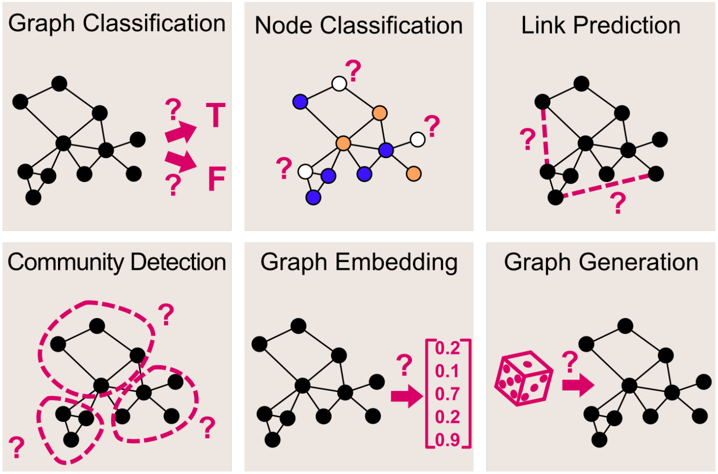
- Training machine learning (ML) models at the edge can address data privacy/security concerns.
  - Federated Learning Applications*
- ML Models are large (Billions of params.)
- Memory bandwidth and power constraints of Edge devices. (‘*memory-wall*’)
- Solutions?
  - Model Compression methods: **Pruning,** Quantization etc.
  - Process-in-memory (**PIM**) computing.



* https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Introduction: Graph Neural Networks (GNN)





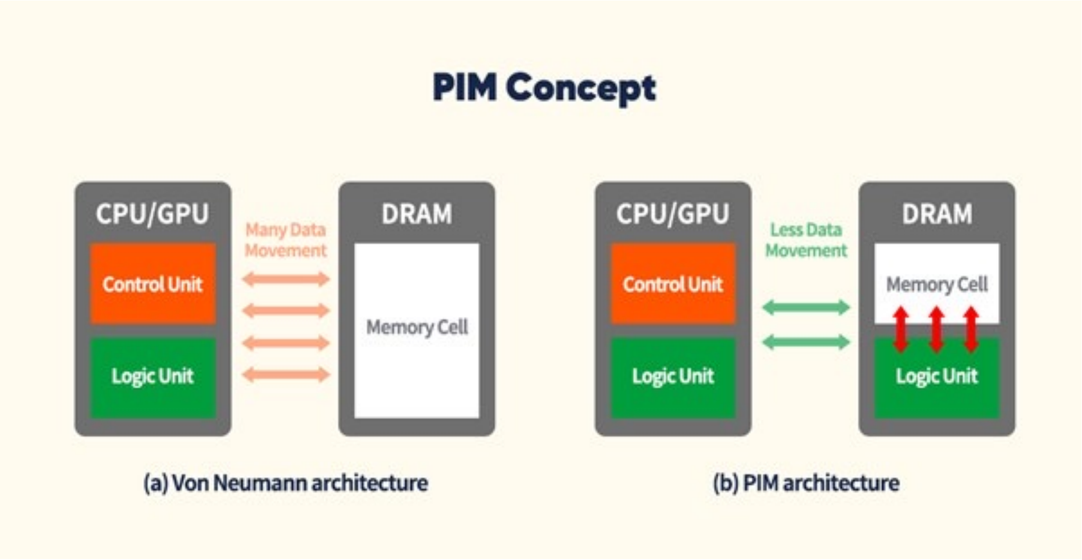
- Non-Euclidean structured data.

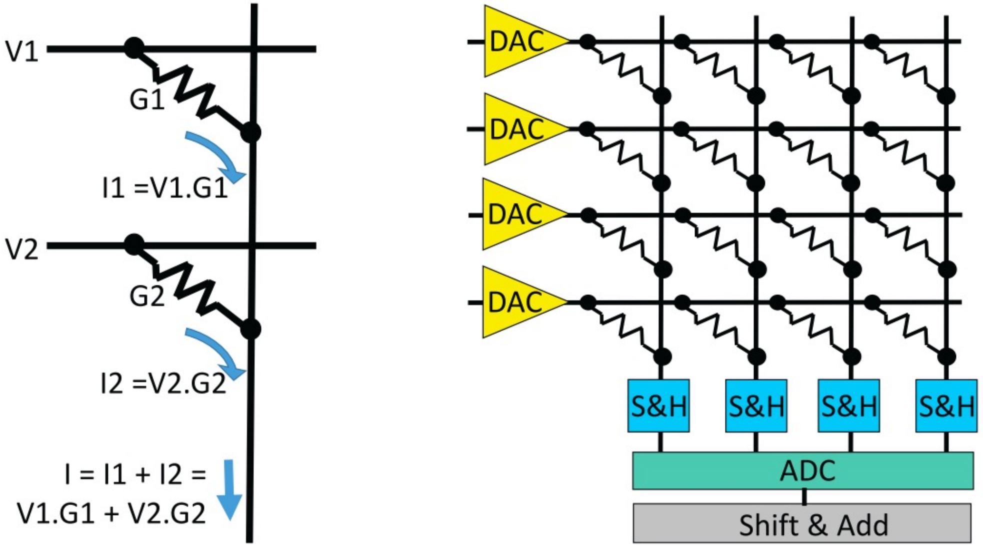
- Graph Convolution Layers
  - Learns features through neighborhood expansion

Why not GPUs?

- GNN training is highly compute- and data-intensive.
- GPUs are not optimized for GNN training
  - High Area & Power Requirements.
  - Low performance per watt (Energy efficiency).
  - Limited Memory Bandwidth.
- Alternative Computing Paradigms?

Motivation: Process-in-memory Computing





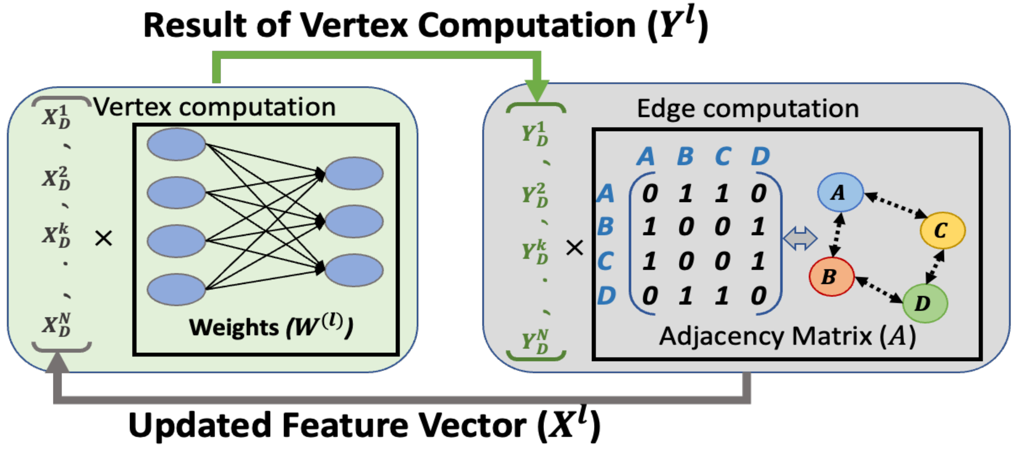
(a) Multiply-Accumulate operation (b) Vector-Matrix Multiplier

- ReRAM-based PIM Architectures

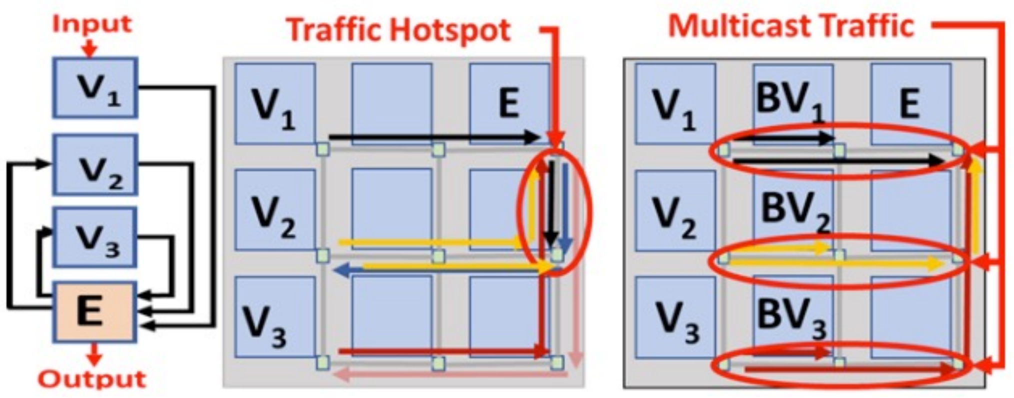
- ReRAM crossbars are natural multipliers
  - Energy Efficient
  - O(1) time
  - > 100 × computation speed-up

Background: GNN Training

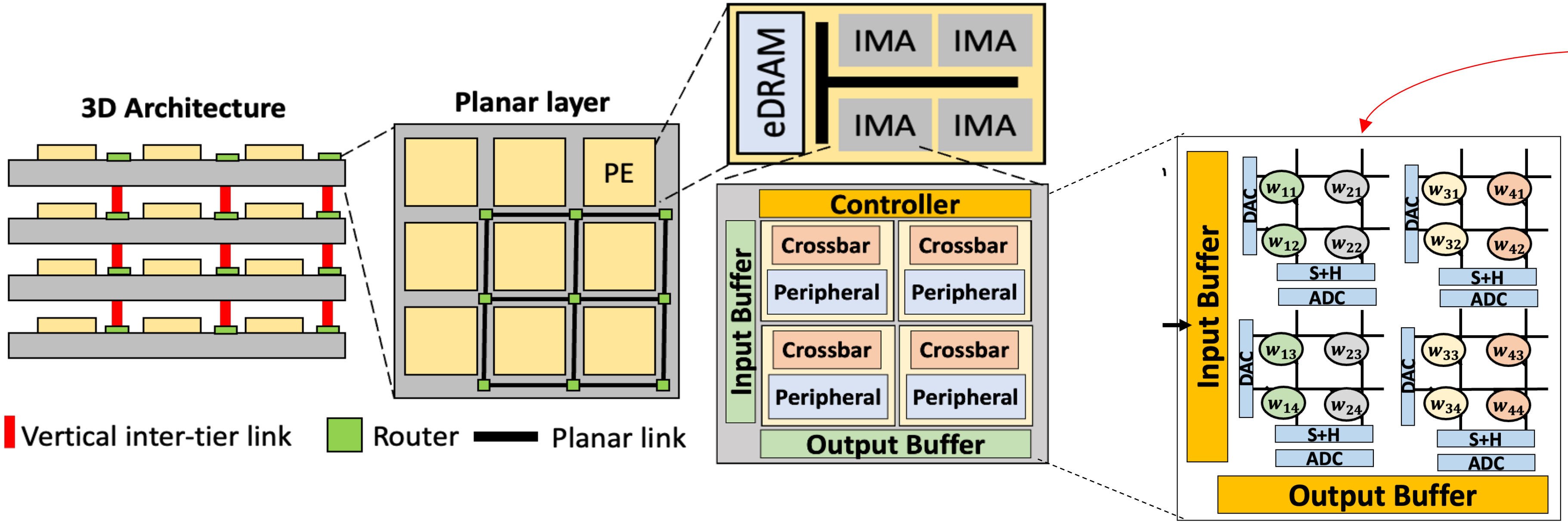
- **Graph Data:**
  - $N \times N$ sparse adjacency Matrix (A)
  - Node-level feature vectors (X)



- **GNNs:**
  - Multiple Layers (L) with weights $W^{(l)}$
  - High amount of on-chip communication.
  - Communication-intensive Vertex & Edge computation. (many-one)



GNN training on 3D PIM Architecture



$$W^{(l)} = \begin{bmatrix} w_{11} & w_{12} & w_{13} & w_{14} \\ w_{21} & w_{22} & w_{23} & w_{24} \\ w_{31} & w_{32} & w_{33} & w_{34} \\ w_{41} & w_{42} & w_{43} & w_{44} \end{bmatrix}$$

- 3D ReRAM-based PIM Architecture
  - Area Efficency
  - High Communication Bandwidth
  - Shorter interconnect
- High on-chip communication traffic can be addressed.

- GNNs are over-parameterized (millions / billions of weights)
- Model Compression techniques: **Pruning**

Methodology: Lottery Ticket Pruning (LTP)

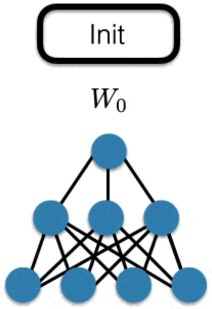
1. Randomly initialize weights (W_i)

2. Train network, to arrive at (W_T)

3. Prune p% of weights in W_T

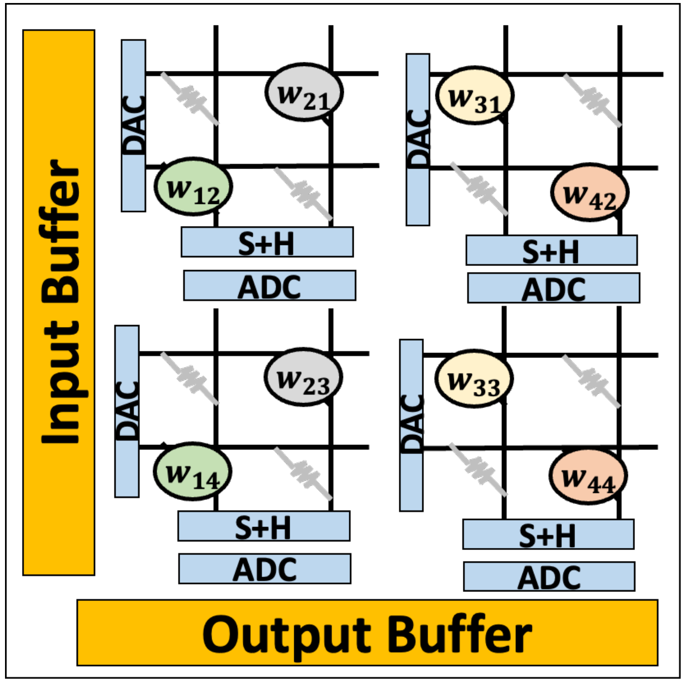
4. Reset remaining weights to initial values in W_i , and Repeat

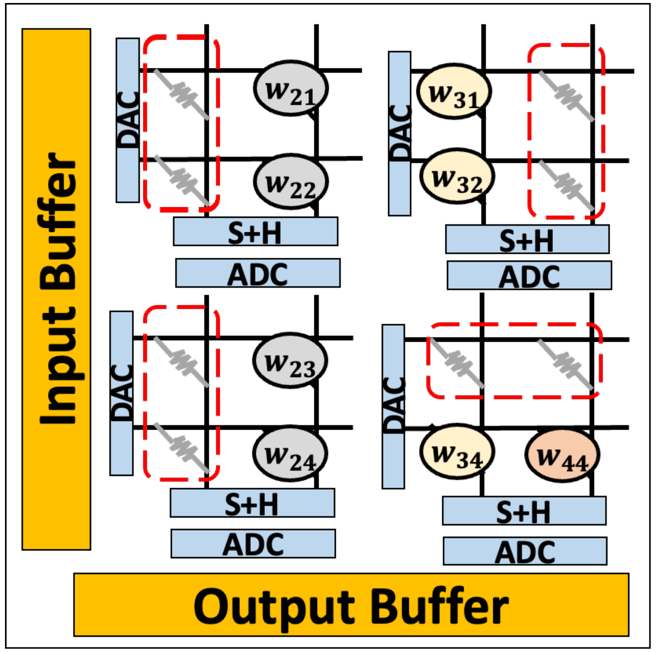
Iterative Magnitude Pruning with Rewinding



Frankle et al., 2019
Viz: @RobertTLange

Related Work

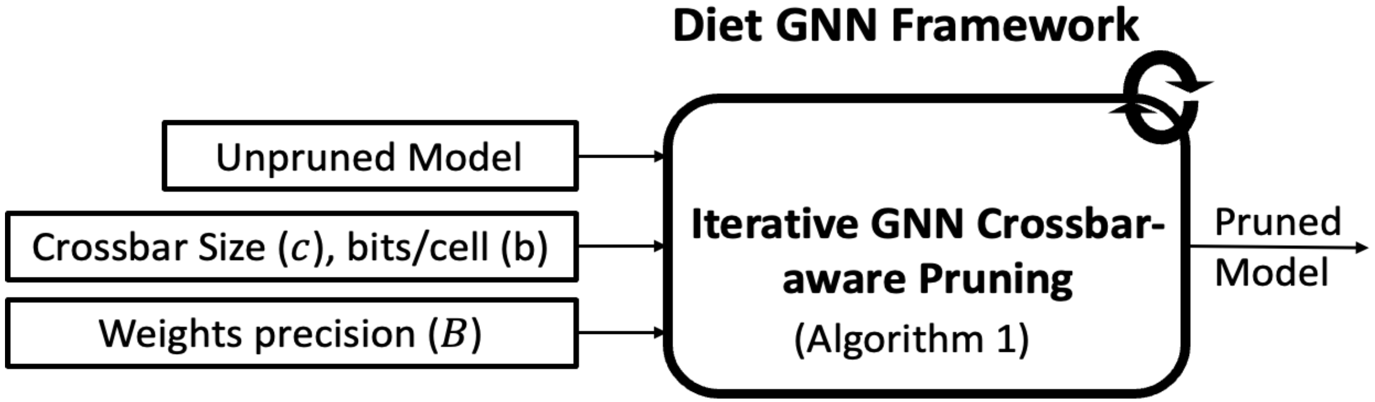


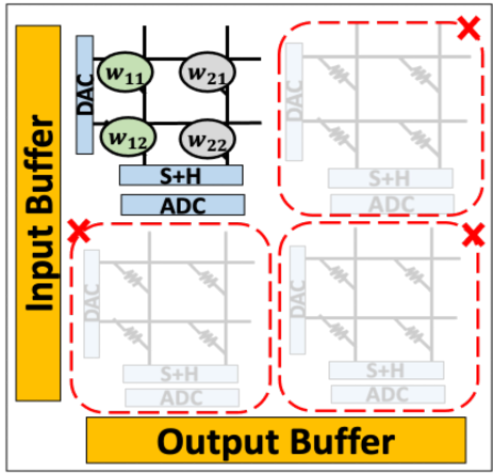


- LTP & Unified Graph Sparsification (UGS)
  - Unstructured pruning, High Sparsity
  - Low Energy & Area Savings

- Existing Crossbar-Aware Pruning (CAP)
  - Marginal reduction in Energy and Area cost.

Methodology: DietGNN Framework





Algorithm 1. Pruning with DietGNN

Input: GNN model, crossbar structure, prune percentage p
Output: Pruned GNN model or winning ticket
Algorithm:

1: **Initialize**: $W^l \leftarrow W_{initial}$;
2: **Partition** W^l into blocks (B^l) of size $c \times \left(c * \frac{b}{B}\right)$
3: **While** *itr* $< n$:
4:     **Train** for E epochs
5:     **Prune** $p\%$ of B^l based on average magnitude
6:     **Reinitialize** remaining **weights** with $W_{initial}$
7: **Return** *Pruned Model (Hardware-friendly winning ticket)*

- DietGNN achieves:
  - Significant reduction in Peripheral Circuit Area & Energy overhead
- Models can be reused multiple times

Experimental Setup

- Five benchmark real-world graph datasets: PPI, Reddit, Amazon2M, Flickr, and Yelp for the performance evaluation

- Graph Convolution Networks (GCNs)

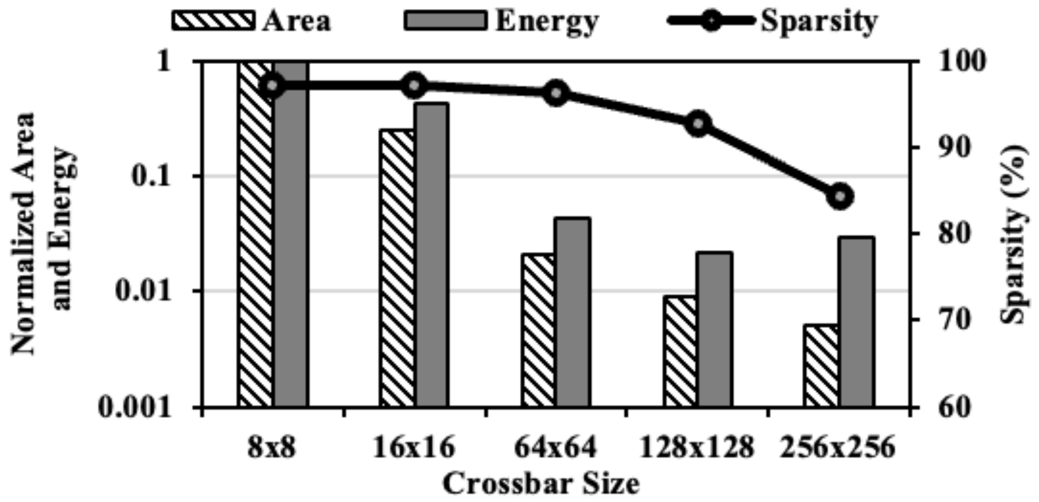
ARCHITECTURAL SPECIFICATIONS

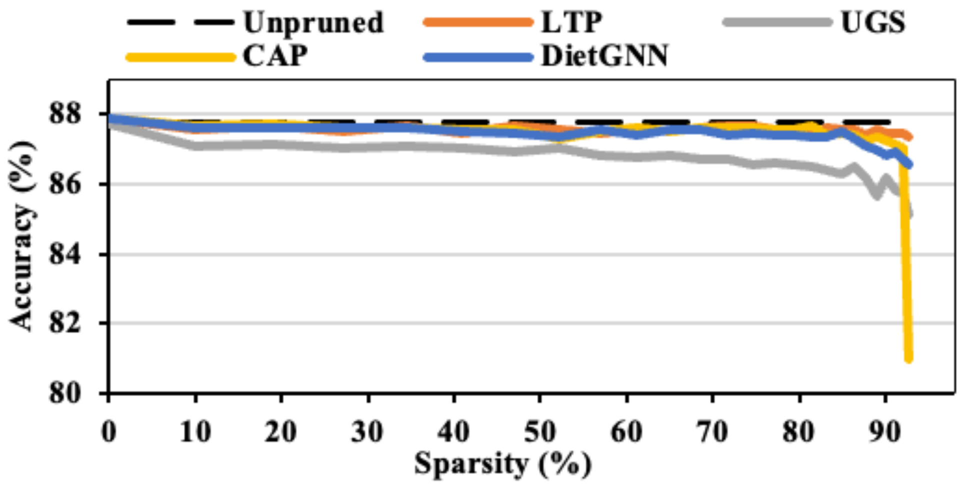
4 planar tiers, 9 cores per tier, 4 tiles per core	
ReRAM Tile	96-ADCs (8-bits), 12x128x8 DACs (1-bit), 96 crossbars, 128x128 crossbar size, 10MHz, 2-bit resolution

GNN DATASET STATISTICS

Dataset	# of Nodes	# of Edges	# of GNN layers	# of Features
PPI	56,944	818,716	5	50
Reddit	232,965	11,606,919	4	602
Amazon2M	2,449,029	61,859,140	4	100
Flickr	89,250	899,756	3	500
Yelp	716,847	13,945819	3	300

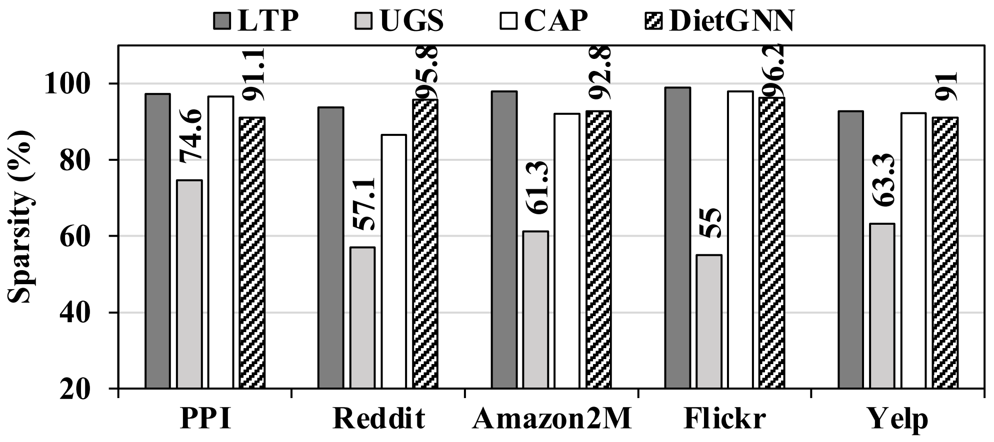
Results: Accuracy vs Sparsity

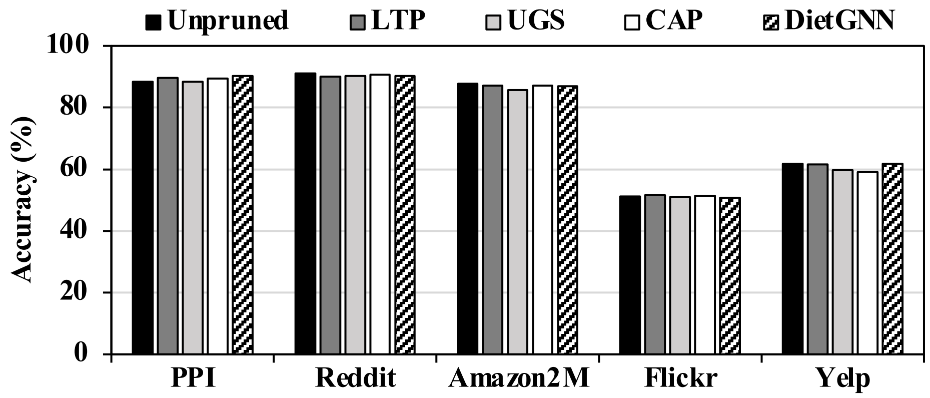




- 128×128 crossbar size is the sweet spot in the sparsity-area-energy trade-off.
- Amazon2M Accuracy vs Sparsity
  - Up to 90% Sparsity with 1% accuracy drop constraint.

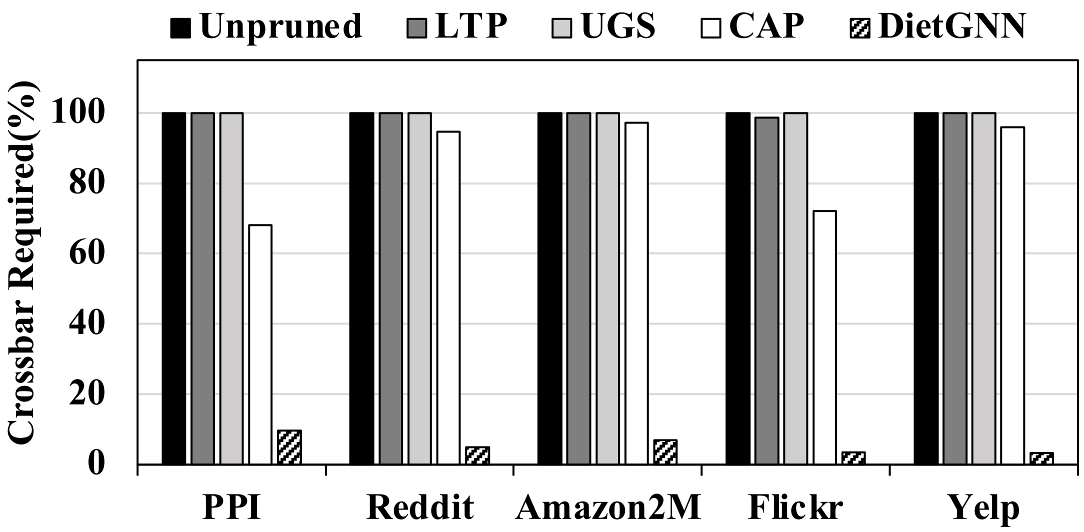
Results: Accuracy and Sparsity

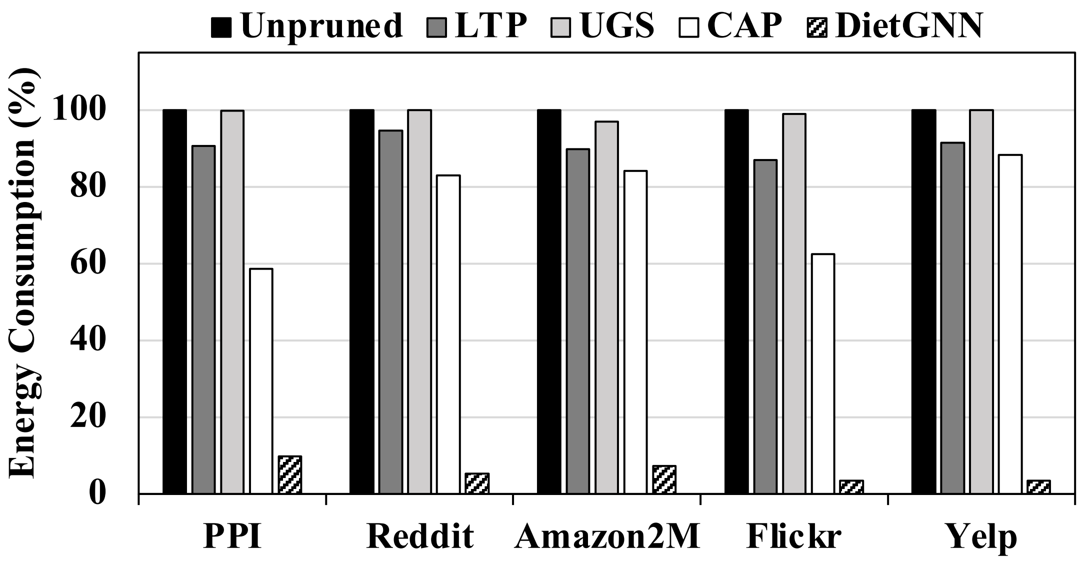




- DietGNN high sparsity like with LTP
- <1% accuracy loss constraint

Results: Area and Energy





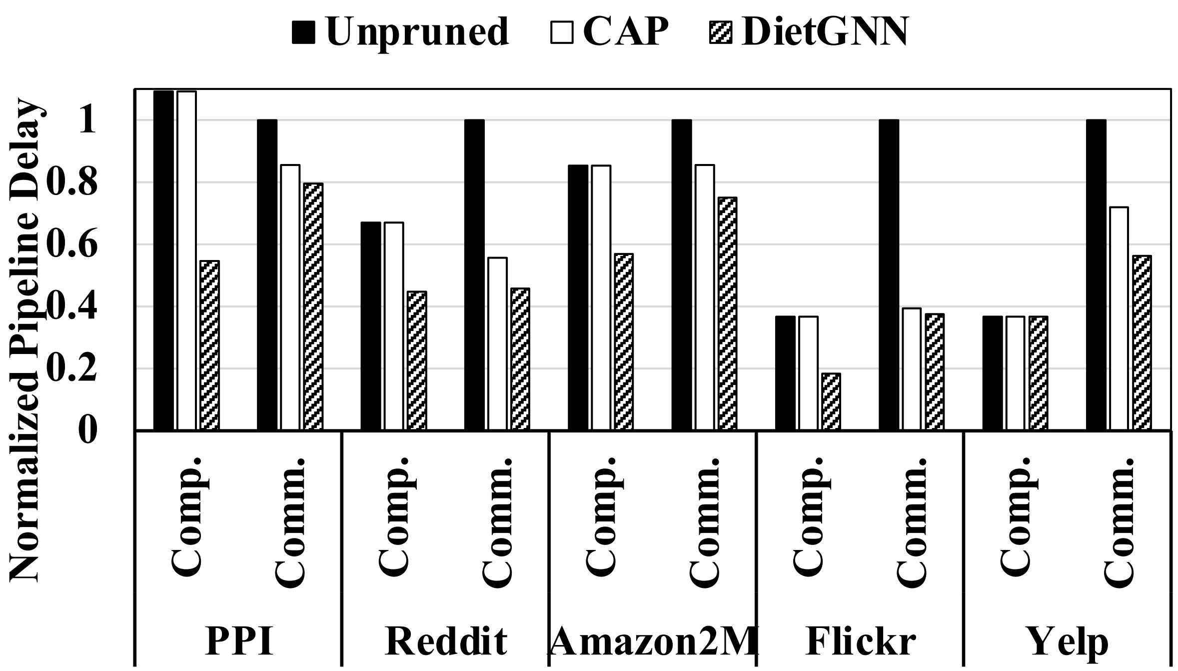
- DietGNN achieves >95% area and Energy reduction.

Duke University

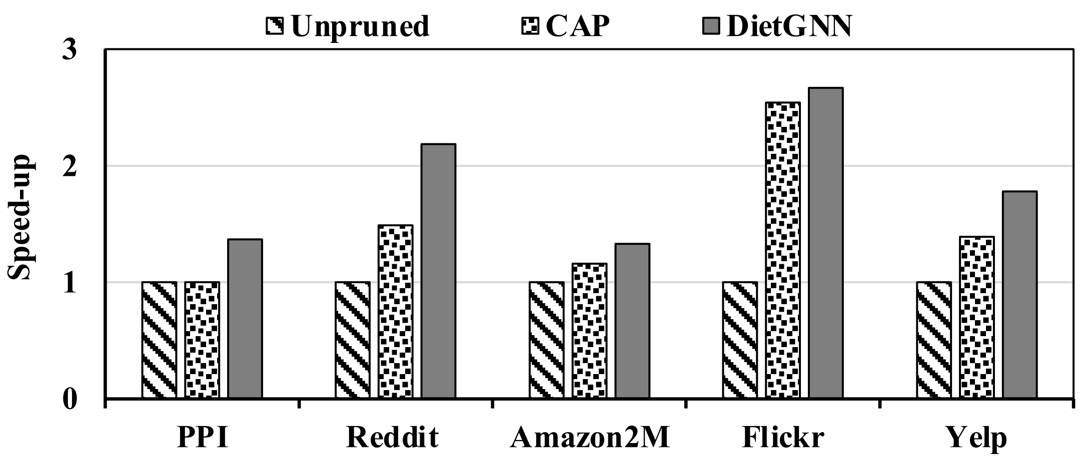
Washington State University

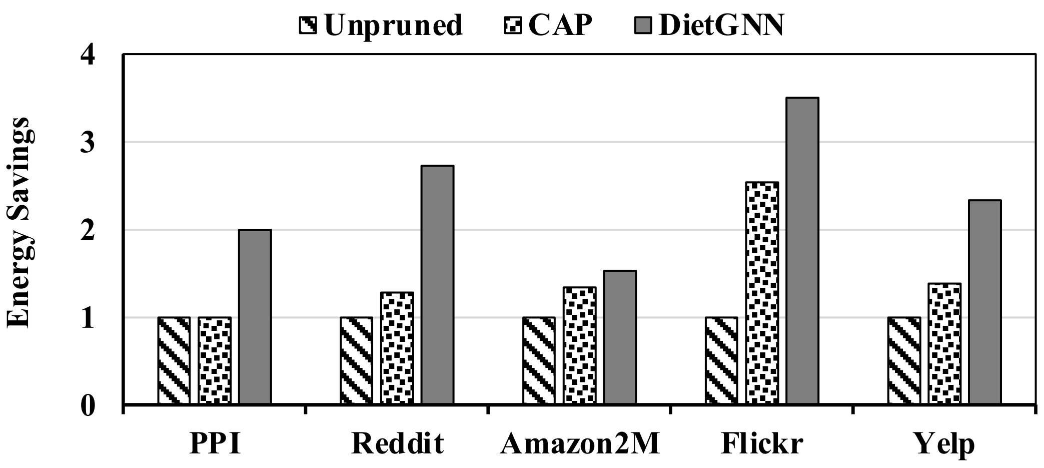
Results: Computation & Communication Delay

- 41.5% communication delay reduction on Average.
- 58% average improvement in computation delay



Results: Overall System Performance





- DietGNN achieves 87% and 52% speed-up in overall execution time on average compared to Unpruned and CAP, respectively

Duke UNIVERSITY   WASHINGTON STATE UNIVERSITY

Conclusion

- ReRAM-based PIM architectures are good candidates for accelerating large-scale GNN training at the edge.
- GNN models contain many parameters and training is compute and communication intensive. The DietGNN pruning method addresses this challenge.
- DietGNN framework achieves $\sim 2.7\times$ speedup and $3.5\times$ energy efficiency for GNN training.

Thank You !

Duke UNIVERSITY

WASHINGTON STATE UNIVERSITY