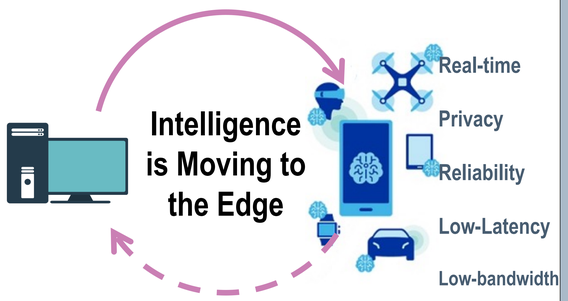
Twofold Sparsity: Joint Bit- and Network-level Sparse Deep Neural Network for Energy-efficient RRAM Based CIM

Foroozan Karimzadeh
Georgia Institute of Technology, Atlanta, GA, USA

Introduction: AI at the Edge



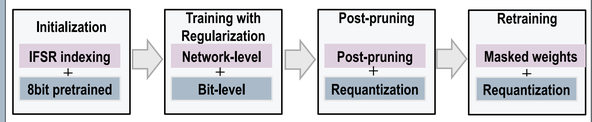
Source: Qualcomm

Challenges and Motivations

- **Challenges:**
  - **DNN model workload:**
    - Large and over-parameterized models
    - Computationally intensive
    - Always on and real-time processing
  - **Hardware constrained:**
    - Memory and bandwidth limitations
    - Power/battery constrained
- **Motivation:**
  - **Von Neumann architecture :**
    - Prohibitive power dissipation
    - Massive data transfer between the PEs and memory
    - High Latency
  - **CIM architectures :**
    - A memory cell itself serves as a PE and memory
    - Low-latency

Twofold Sparsity

Joint Bit- and Network-level Sparsity



$$J = L_{CE} + Reg_{Net} + Reg_{Bit}$$

$$\lambda \sum_{l=0}^{L} \left\| W^{[l]} \odot M^{[l]} \right\|_2 + \beta \sum_{l=0}^{L} \frac{\#param(W^{[l]})}{\#param(W^{[1:L]})} \left\| \sum_{n=0}^{b} W_q^{(b)} \right\|_2$$

Network-level Sparsity

- We use Linear Feedback Shift Register (LFSR) to generate pseudo random indices.
- The network is pruned based on the selected indices.

Weight matrix

4.2	0.3	0.1	2.8
3.1	0.02	0.05	0.12
0.06	2.3	0.03	3.6
0.1	0.01	1.9	0.02

$\odot$ Mask

1	0	0	1
1	1	0	0
0	1	0	1
0	0	1	0

Indices from LFSR

Bit-level Sparsity: DNN training under Bit representation

- Scaling:

$$W = s \,.\, W_s$$
$$s = \max|W|$$

- Quantization

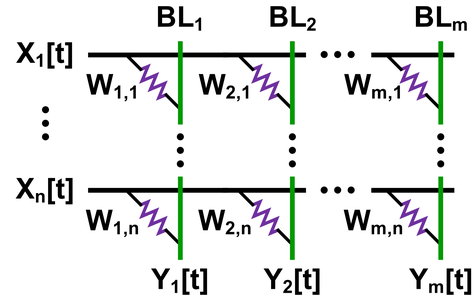
$$W_q = \frac{Round\left[W_s \times \left(2^{b-1} - 1\right)\right]}{2^{b-1} - 1},$$
$$\text{where } w_q \in \left\{0, \pm\frac{1}{2^{b-1}-1}, \pm\frac{2}{2^{b-1}-1}, \ldots, \pm 1\right\}$$

- Binary conversion (2's complement)

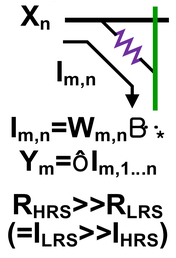
$$W_q = \frac{-W_s^{b-1}2^{b-1} + \sum_{n=0}^{b-2} W_s^{(n)} 2^n}{2^{b-1} - 1}$$

Hardware co-design

- Compute-In-Memory architecture are deployed for the inference.



CIM operation



Bitwise multiplication at memory cells

X	W	I
0	0 (R_{HRS})	0
0	1 (R_{LRS})	0
1	0 (R_{HRS})	I_{HRS}
1	1 (R_{LRS})	I_{LRS}

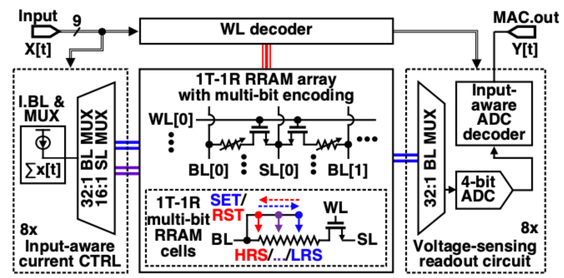
2-bit/cell CIM Energy

- Motivation: $E_{11} \sim 20 \times E_{00}$
- It is desirable to have more 00 and 01 than 10 and 11
- The goal is to sparsify the model in the bit-level to have more 00 in their bit representation.

$$E_{11} \sim 20 \times E_{00}$$

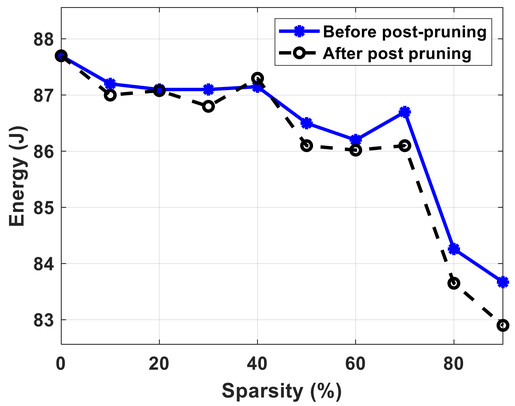
Energy(pJ/2bits)	
00	0.079
01	0.36
10	0.73
11	1.49
ADC	0.208

Architecture of multi-bit RCIM



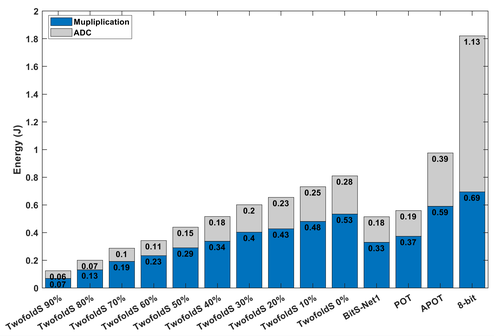
Result: Accuracy

- ResNet-20 using CIFAR-10 dataset.
- Accuracy of the network in different sparsity (%).
- Accuracy before post training is slightly higher.



Result: Energy

- ResNet-20 using CIFAR-10 dataset.
- Estimated total energy including multiplication, ADC and LFSR



Conclusion

- A joint bit- and network-level sparse DNN for energy-efficient RRAM based CIM.
- The network is sparsified during training by adding two regularizer.
- We demonstrate that TwofoldS improves the energy efficiency by up to 5x for ResNet on Cifar10.