tinyML® Talks

Enabling Ultra-low Power Machine Learning at the Edge

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

Foroozan Karimzadeh – Postdoctoral Fellow, Georgia Institute of Technology

November 21, 2023

www.tinyML.org

TINY ML TALKS webcast

Thank you, **tinyML Strategic Partners**, for committing to take tinyML to the next Level, together

TINY ML TALKS webcast

Executive Strategic Partners

Qualcomm
AI research

Advancing AI research to make efficient AI ubiquitous

Power efficiency

Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization

Continuous learning, contextual, always-on, privacy-preserved, distributed learning

Efficient learning

Robust learning through minimal data, unsupervised learning, on-device learning

A platform to scale AI across the industry

Perception

Object detection, speech recognition, contextual fusion

Reasoning

Scene understanding, language understanding, behavior prediction

Action

Reinforcement learning for decision making

Edge cloud

Cloud

IoT/IIoT

Automotive

Mobile

Accelerate Your Edge Compute

SYNTIANT

Making Edge AI A Reality

www.syntiant.com

TINY ML TALKS webcast

Platinum Strategic Partner

AITRIOS

DEPLOY VISION AI AT THE EDGE AT SCALE

SONY

TINY ML TALKS webcast

Gold Strategic Partners


Build the
Future of tinyML
on arm

TINY ML TALKS webcast

EDGE IMPULSE

The Leading Development Platform for Edge ML

edgeimpulse.com

Driving decarbonization and digitalization. Together.

Infineon serving all target markets as
Leader in Power Systems and IoT

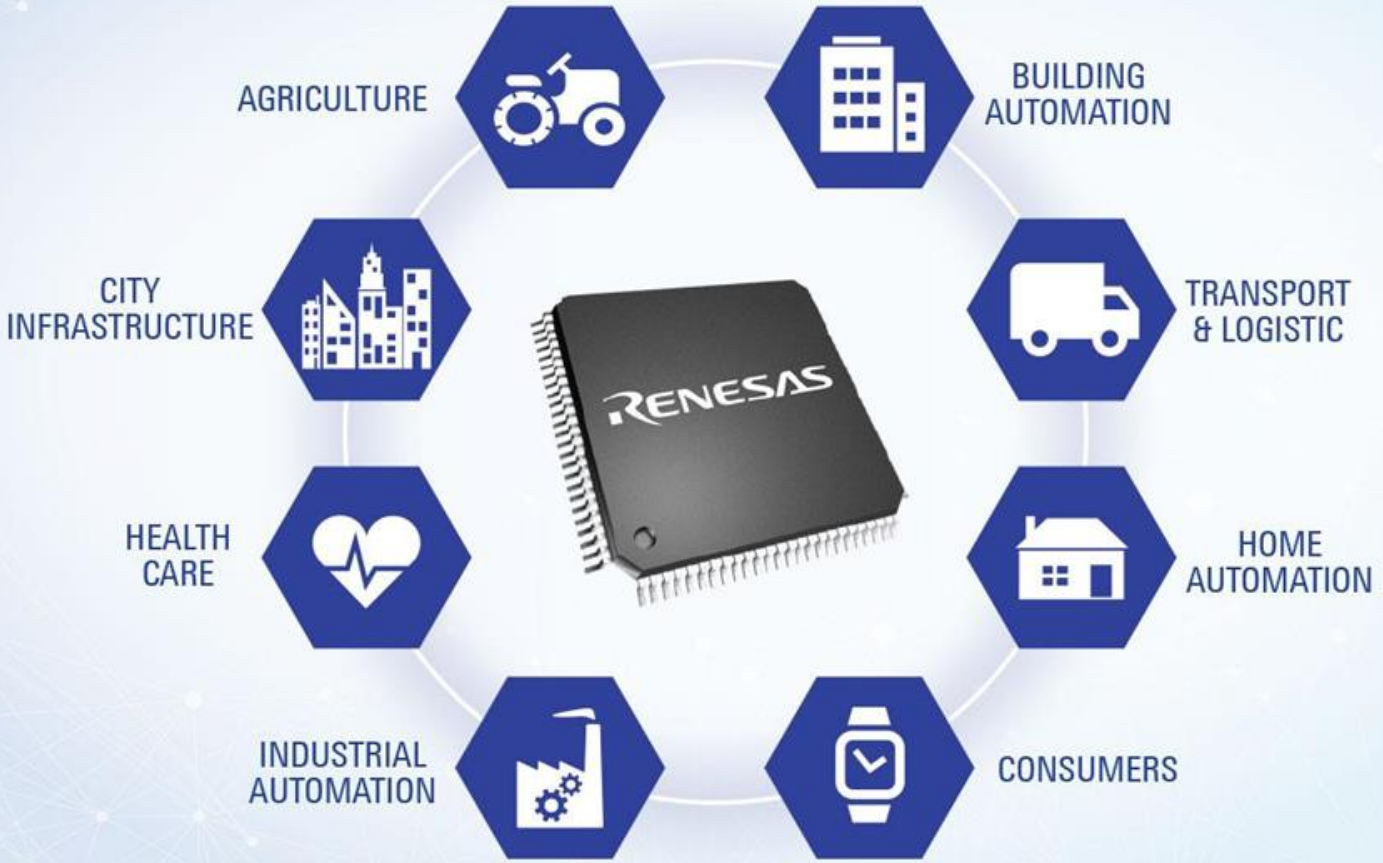
www.infineon.com

innatera

NEUROMORPHIC INTELLIGENCE FOR THE SENSOR-EDGE

www.innatera.com

Renesas is enabling the next generation of AI-powered solutions that will revolutionize every industry sector.



renesas.com

STMicroelectronics provides extensive solutions to make tiny Machine Learning easy

life.augmented

www.st.com/ai

synaptics

ENGINEERING EXCEPTIONAL EXPERIENCES

We engineer exceptional experiences for consumers in the home, at work, in the car, or on the go.

www.synaptics.com

TINY ML TALKS webcast

Silver Strategic Partners

Join Growing tinyML Communities:

17.6k members in
49 Groups in 41 Countries

tinyML - Enabling ultra-low Power ML at the Edge

https://www.meetup.com/tinyML-Enabling-ultra-low-Power-ML-at-the-Edge/

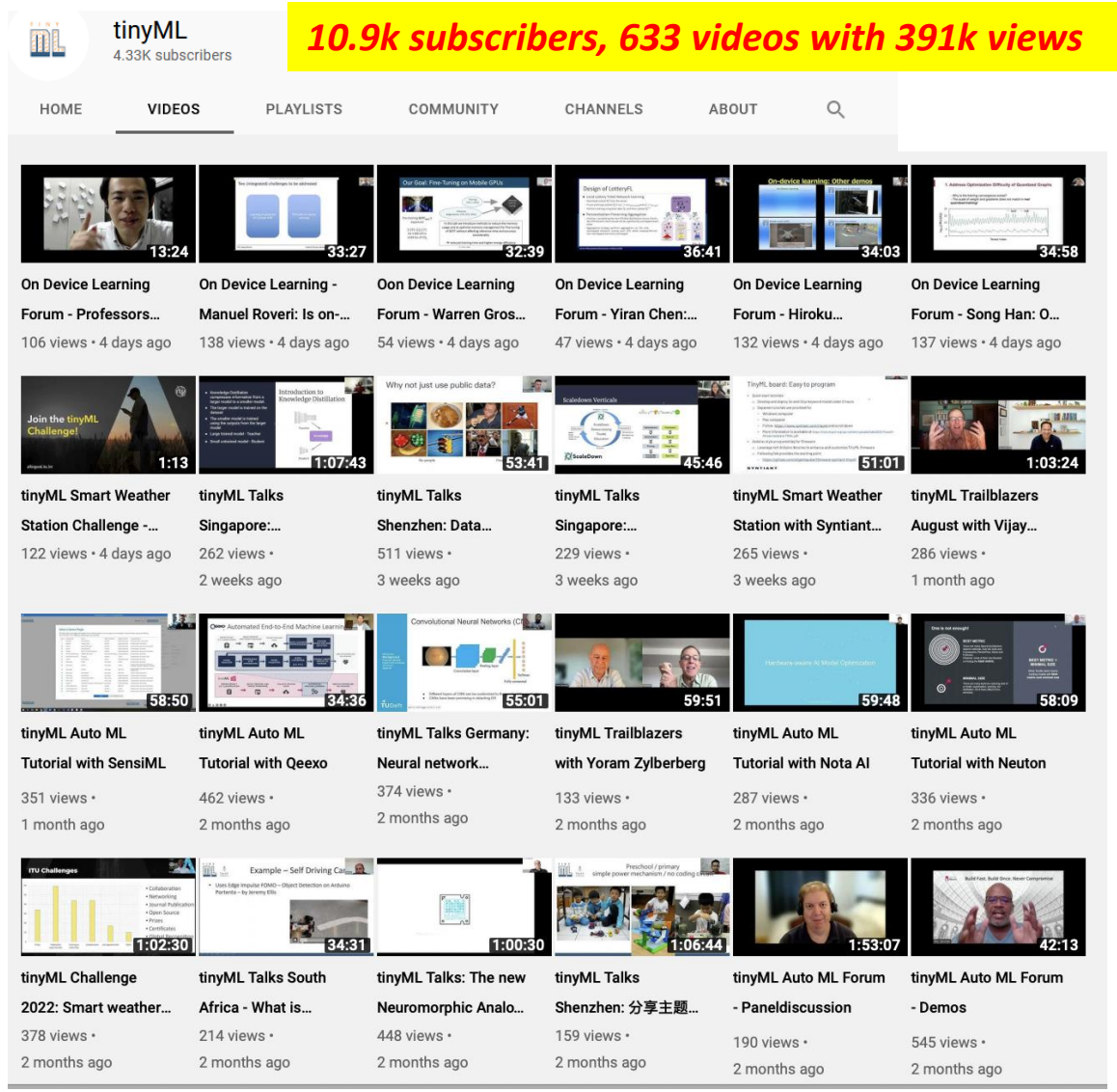
4k members
&
13k followers

The tinyML Community

https://www.linkedin.com/groups/13694488/

TINY ML TALKS webcast

YouTube



Subscribe to tinyML YouTube Channel for updates and notifications

(including this video)

www.youtube.com/tinyML

tinyAI Forum on PdM & Anomaly Detection 2023

Interactive live webinar December 5, 2023 at 8AM Pacific Time
Registration is free of charge

tinyML Research Symposium
April 22, 2023
Call for Papers

tinyML Summit April 23-24, 2024
Call for Presentations and Posters

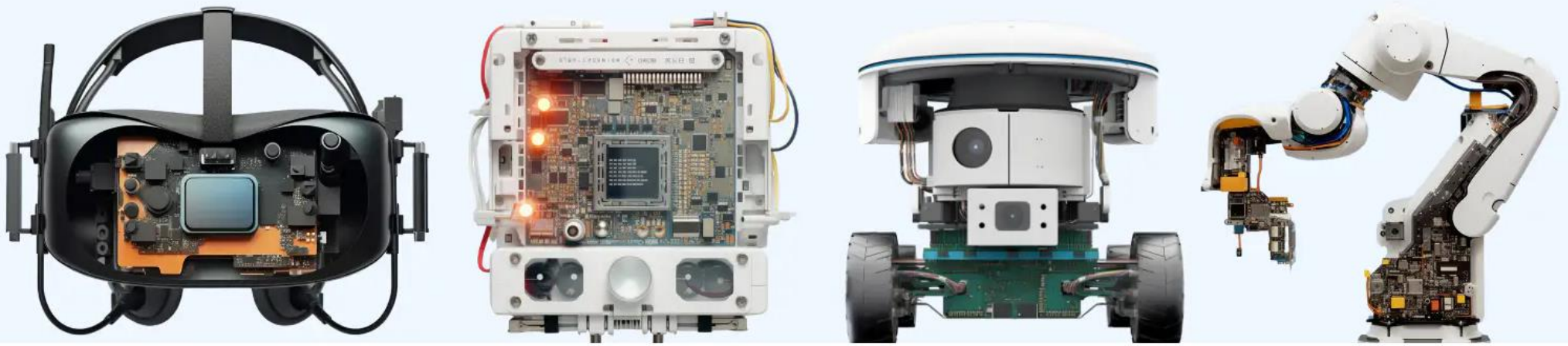
WEVOLVER

https://www.wevolver.com/article/2023-edge-ai-technology-report

TINY ML FOUNDATION

2023 Edge AI Technology Report

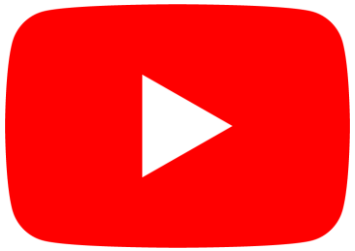
The guide to understanding the state of the art in hardware & software in Edge AI.



TINY ML TALKS webcast

Reminders

Slides & Videos will be posted tomorrow



tinyml.org/forums youtube.com/tinyml

Please use the Q&A window for your questions

TINY ML TALKS webcast

Foroozan Karimzadeh

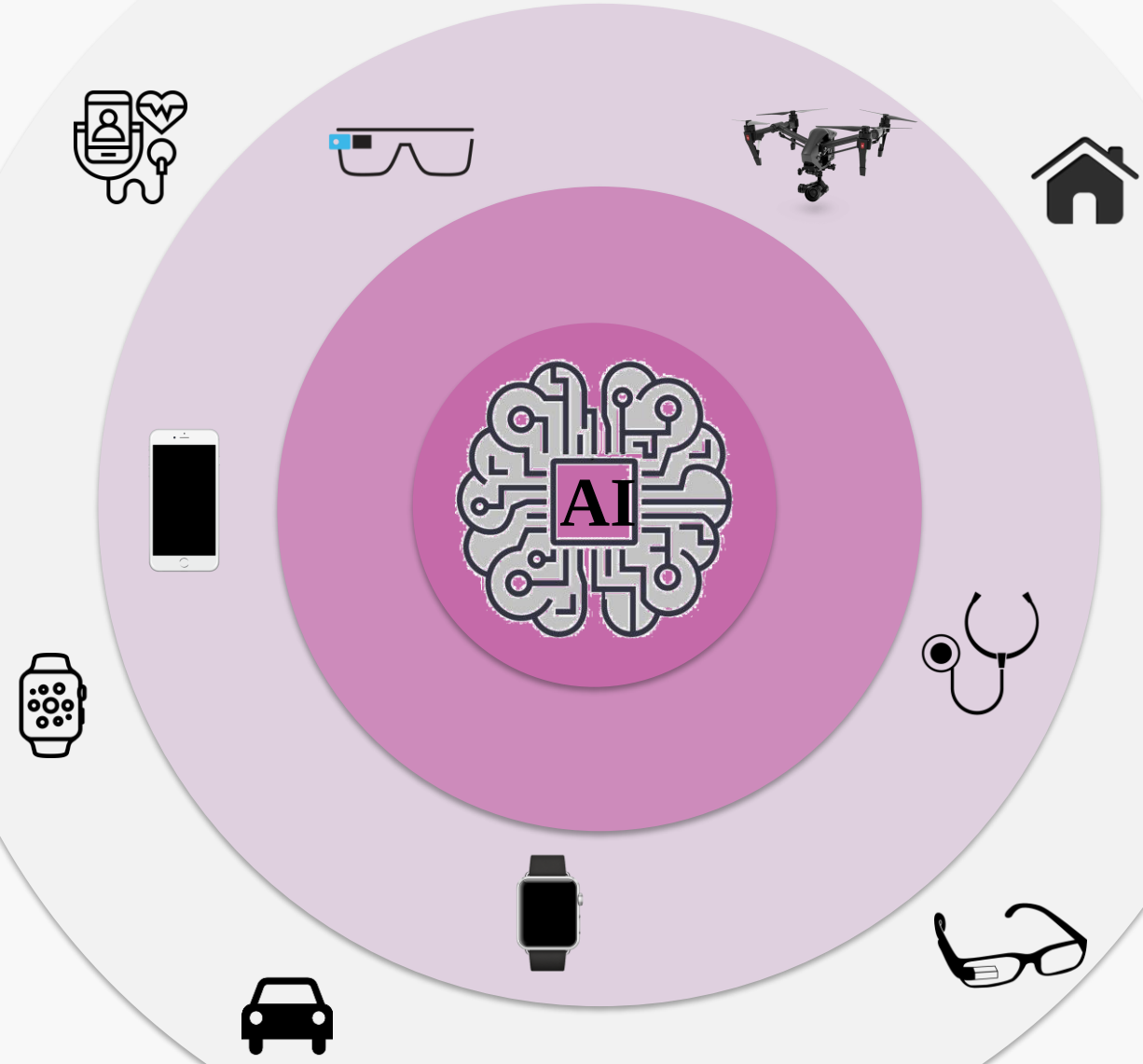
Foroozan Karimzadeh is currently a postdoctoral fellow at Georgia Institute of Technology. She received her PhD degree at Electrical and Computer Engineering department, Georgia Institute of Technology under supervision of Dr. Raychowdhury in 2022. Her research interest mainly includes developing novel algorithms and hardware co-design for energy efficient deep learning and large language models. She was selected as an MIT rising star in EECS, 2023. Foroozan was awarded a prestigious Semiconductor Research Corporation (SRC) Graduate Fellowship, which is awarded in partnership with Texas Instruments. Also, she received DAC Young fellow award at Design and Automation Conference, 2022.

Devices are Getting Smarter

AI Augmentation will create

$3.3 Trillion

of business value by 2025.



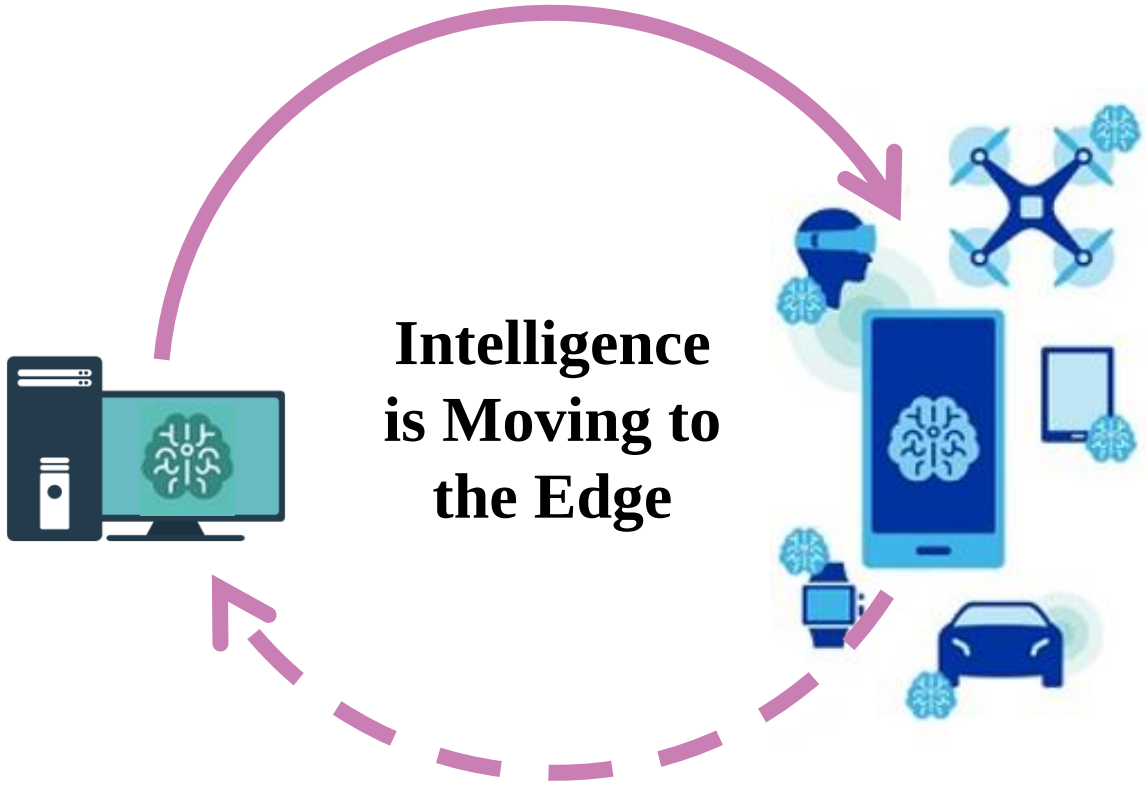
Source Gartner, Qualcomm

Data Explosion and AI

AI at the Edge: Benefits





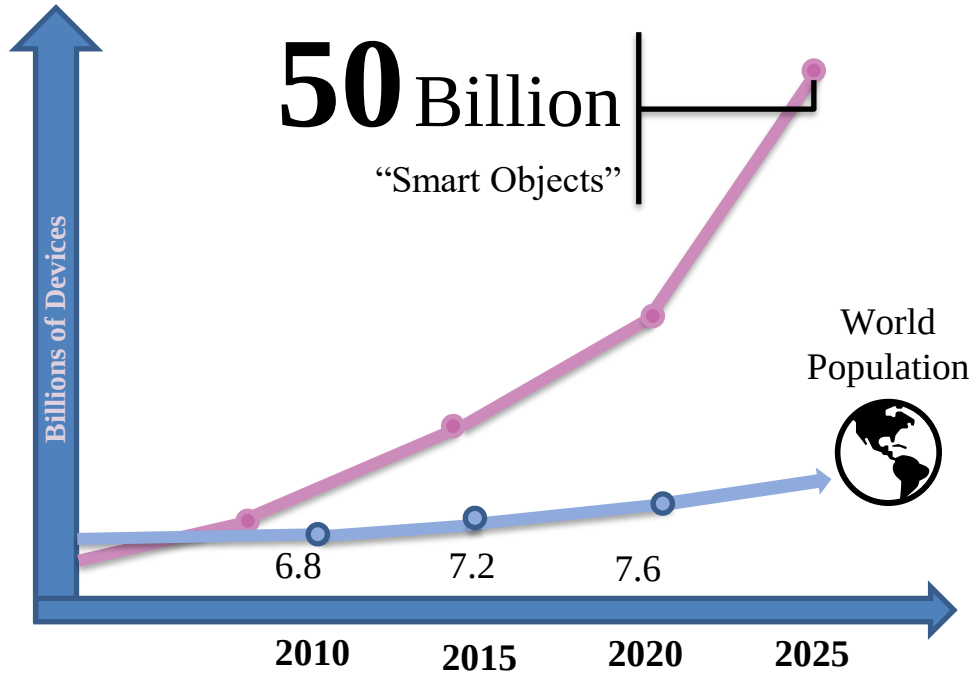



[Privacy](#)

[Reliability](#)

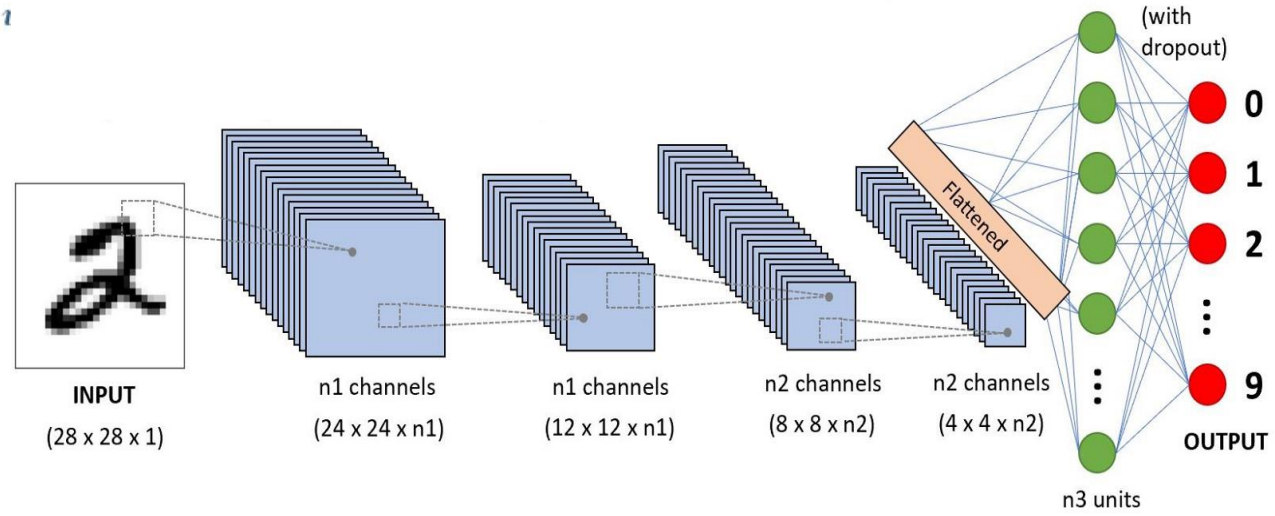
[Low Latency](#)

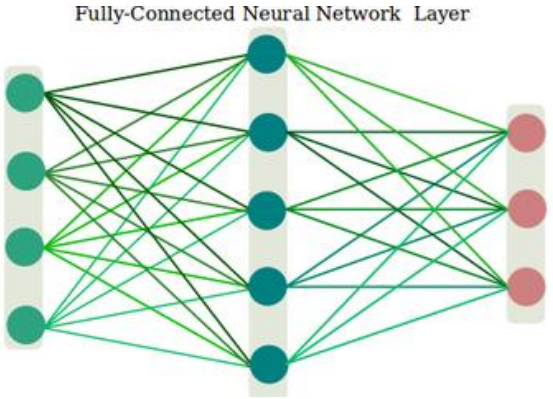
[Low bandwidth](#)



Source Qualcomm, Cisco

AI at the Edge: Benefits





Computer Vision

Health Care Systems

Natural Language Processing

AI at the Edge: Challenges

DNN model workload:

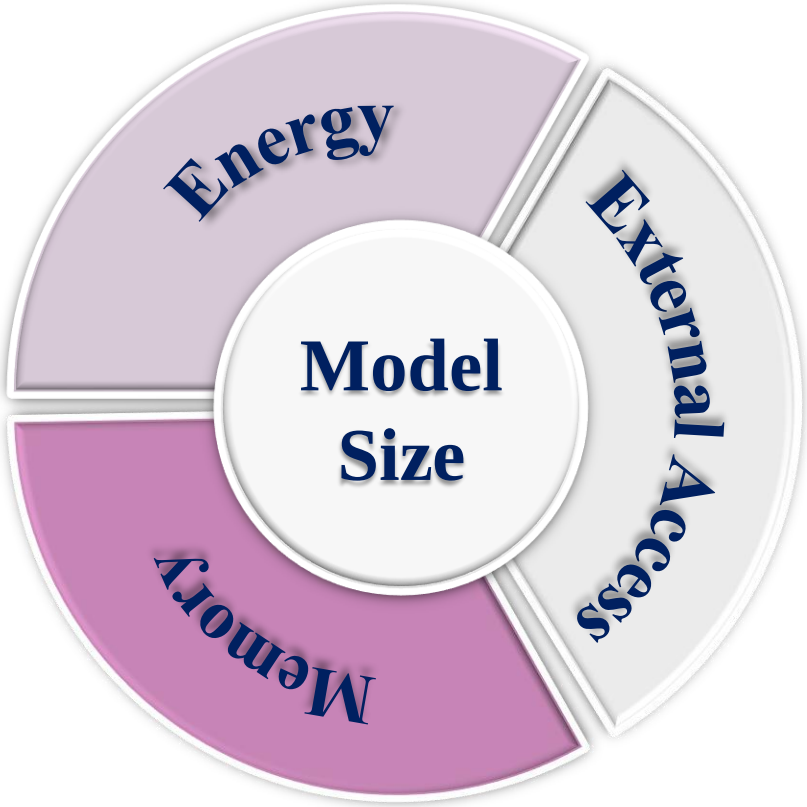
- Large and over-parameterized models
- Computationally intensive
- Always on and real-time processing

Hardware constrained:

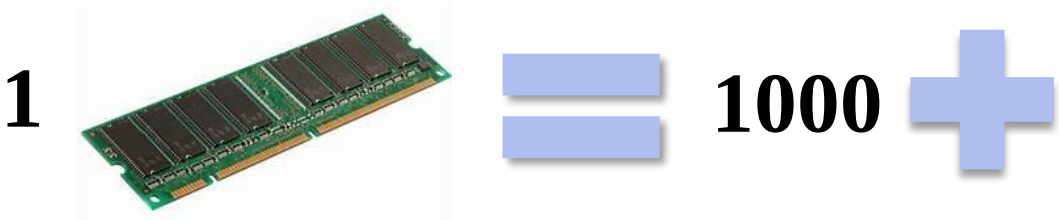
- Memory and bandwidth limitations
- Power/battery constrained

AI at the Edge: Challenges

- Deep Neural Network: Large model size.
- Edge Devices: Resource (Memory, Energy) constrained.



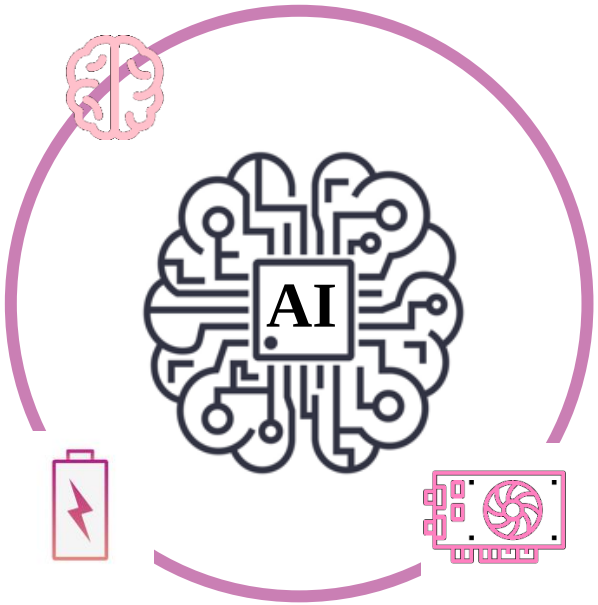
Operation	Energy [pJ]
32-bit float ADD	0.9
32-bit float MULT	3.1
32-bit SRAM Cache	5
32-bit DRAM Memory	640

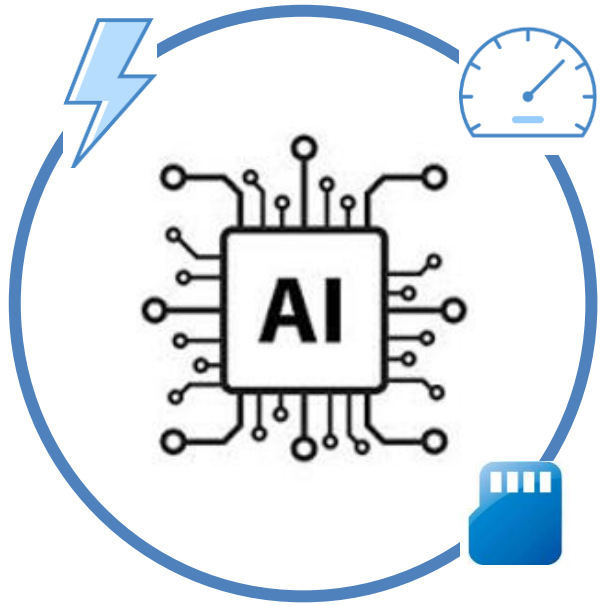


Han, Song, et al. "Learning both weights and connections for efficient neural network." *Advances in neural information processing systems*. 2015.

AI at the Edge: Solutions

Making power efficient AI





Software Advances

Efficient Hardware

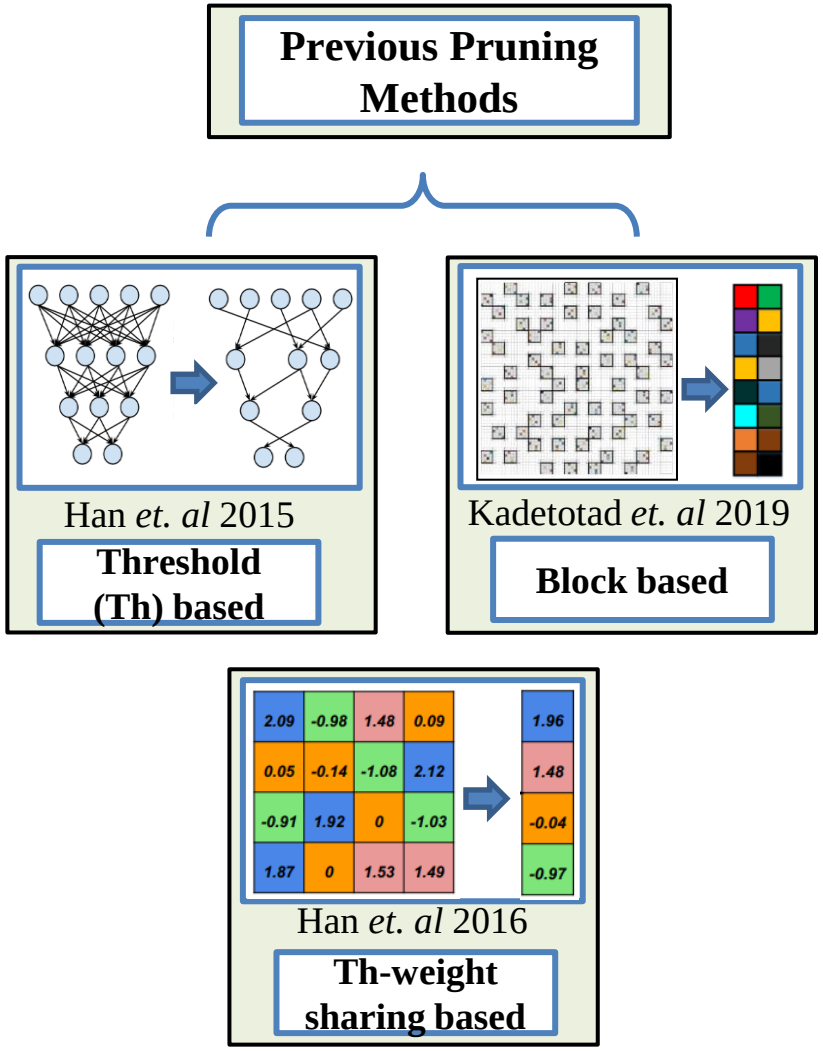
DNN compression: Bit and Network-level Sparsity

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

•**Foroozan Karimzadeh**, R Raychowdhury. "Twofold sparsity: Joint Bit Network- level Sparse Deep Neural Network for Energy-efficient RRAM Based Compute-In-Memory". IEEE transaction of Circuit and Systems (**IEEE-TCAS I**)

Deep Learning on Mobile: Prior Works

Network compression via pruning techniques.



- Not designed to address the real hardware bottlenecks.
- Sparsity matrices add extra levels of irregularity to the weights' matrices.
- Extra memory to save the addresses.

Hardware-aware Pruning of DNNs using LFSR-Generated Pseudo-Random Indices

Why Compute-In-Memory (CIM)?

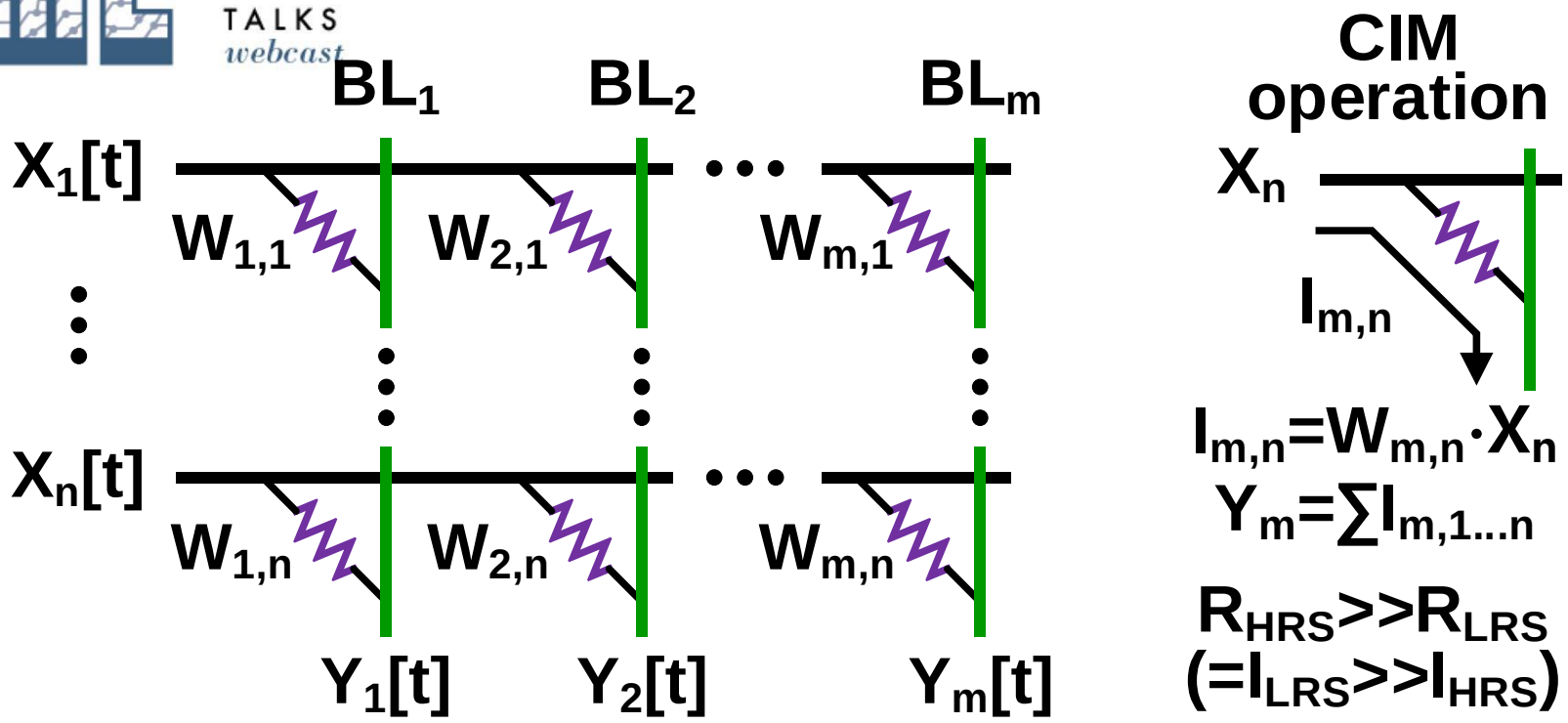
Von Neumann architecture :

- Prohibitive power dissipation
- Massive data transfer between the PEs and memory
- High Latency

CIM architectures :

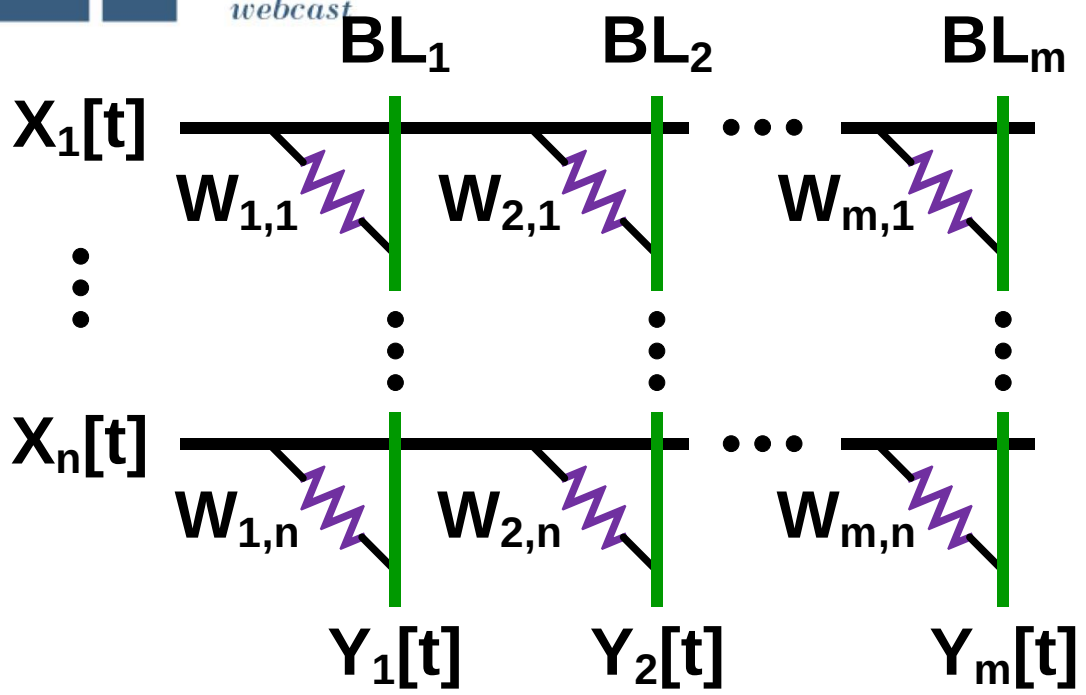
- A memory cell itself serves as a PE and memory
- Low-latency

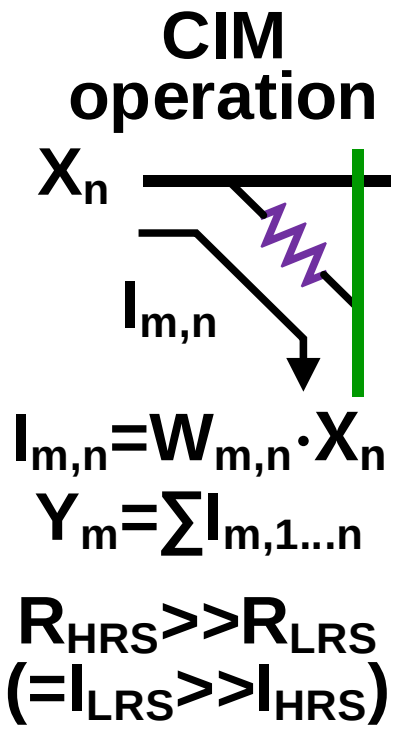
Motivation: Why Compute-In-Memory (CIM)?



Yu, Shimeng. "Neuro-inspired computing with emerging nonvolatile memorys." Proceedings of the IEEE 106.2 (2018): 260-285.

Motivation: Why Compute-In-Memory (CIM)?



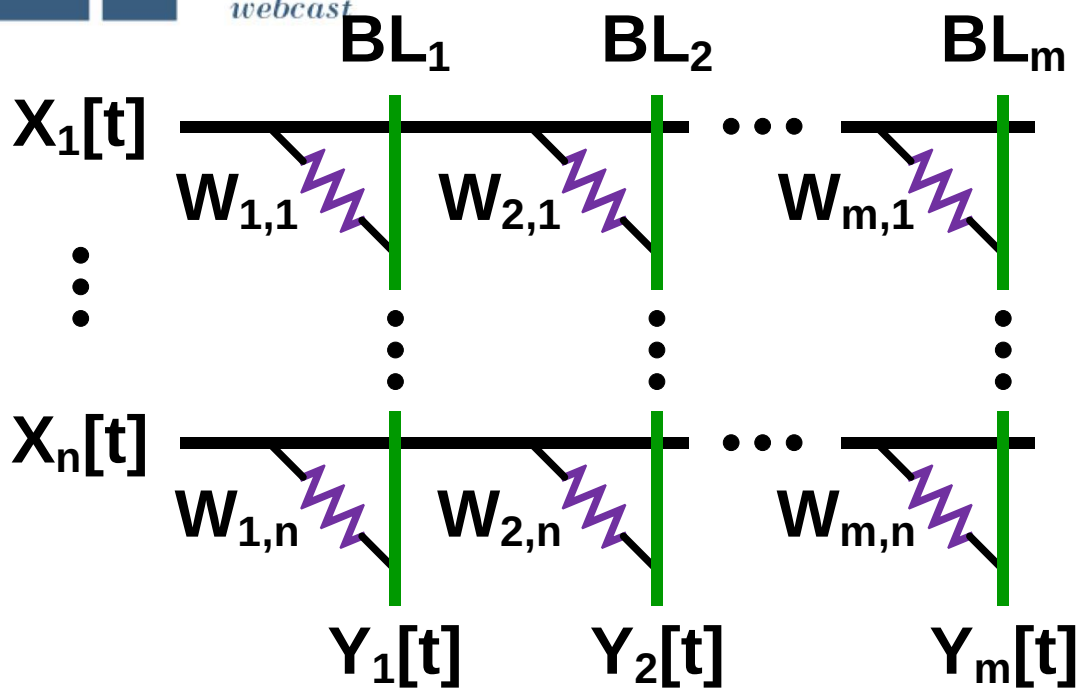


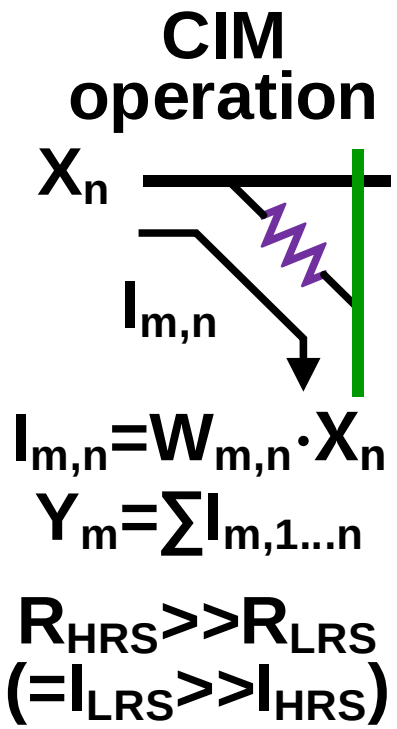
Bitwise multiplication at memory cells

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}

Yu, Shimeng. "Neuro-inspired computing with emerging nonvolatile memorys." Proceedings of the IEEE 106.2 (2018): 260-285.

Motivation: Why Compute-In-Memory (CIM)?





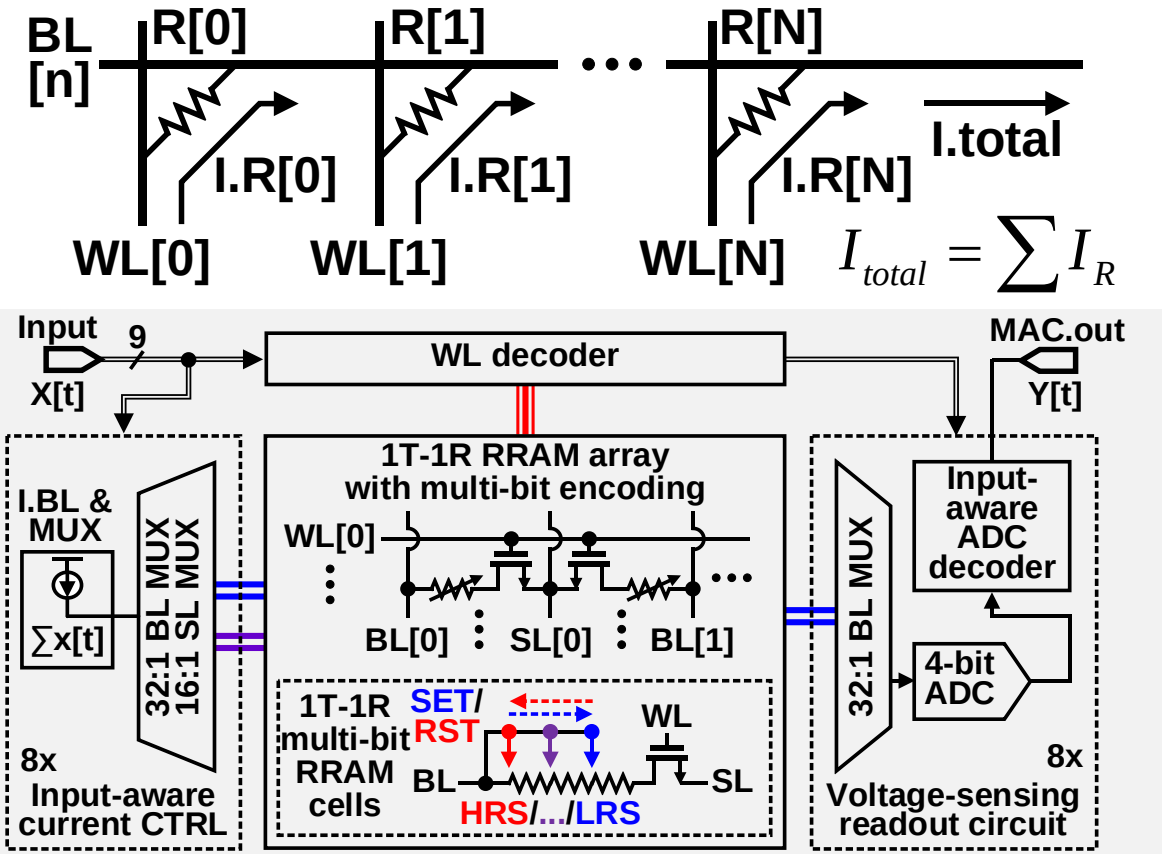
Bitwise multiplication at memory cells

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}

Motivation:
Having more“0”s in the network, as opposed to “1”s reduces the total energy dissipated during inference.

Yu, Shimeng. "Neuro-inspired
memorys." Proceedings of the IEEE 106.2 (2018): 260-285.

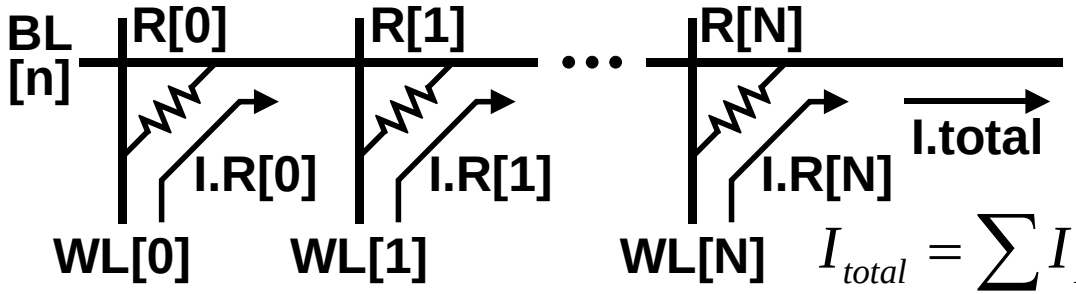
Motivation: 2bit/cell CIM

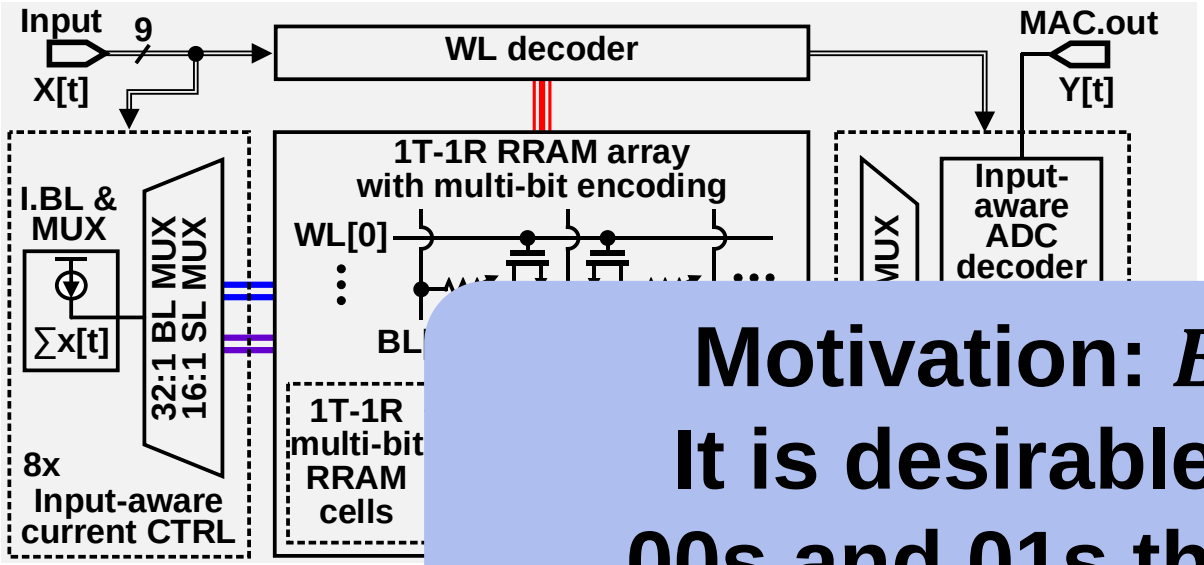


Architecture of voltage-sensing multi-bit RCIM architecture

H. Yoon, M. Chang, W.-S. Khwa, Y.-D. Chih, M.-F. Chang, and A. Raychowdhury, "A 40nm 100kb 118.44 tops/w ternary-weight compute in-memory rram macro with voltage-sensing read and write verification for reliable multi-bit rram operation," CICC. IEEE, 2021.

Motivation: 2bit/cell CIM





Architecture of voltage-sens

Energy for multiplication

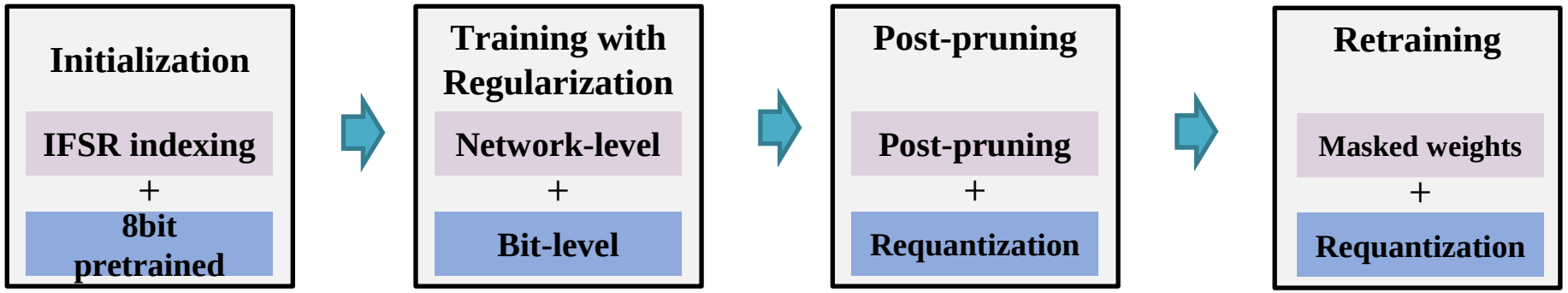
X	W	$E_{(pJ/2bits)}$
1	00	0.079
1	01	0.36
1	10	0.73
		1.49

Motivation: $E_{11} = 20 \times E_{00}$
It is desirable to have more 00s and 01s than 10s and 11s

H. Yoon, M. Chang, W.-S. Khwa, Y.-D. Chih, M.-F. Chang, and A. Raychowdhury, "A 40nm 100kb 118.44 tops/w ternary-weight compute in-memory rram macro with voltage-sensing read and write verification for reliable multi-bit rram operation," CICC. IEEE, 2021.

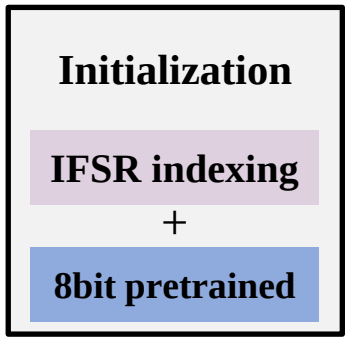
Method: Twofold Sparsity

Joint Bit- and Network-level Sparsity



Method: Twofold Sparsity

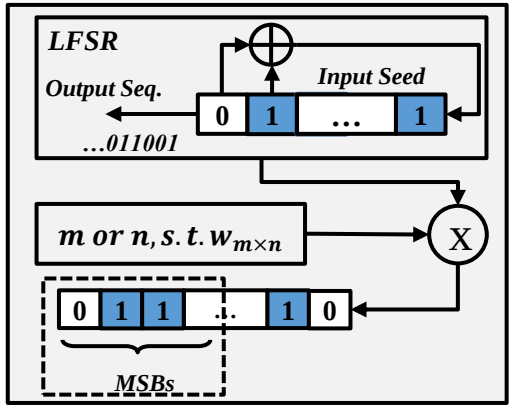
Joint Bit- and Network-level Sparsity



Mask

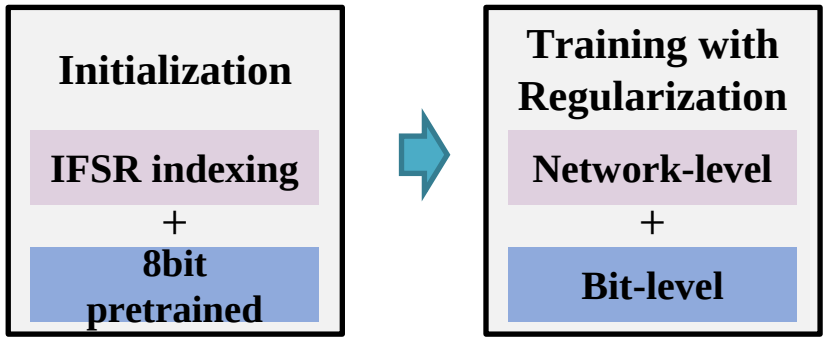
1	0	0	1
1	1	0	0
0	1	0	1
0	0	1	0

Indices from LFSR



Method: Twofold Sparsity

Joint Bit- and Network-level Sparsity



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

$$\lambda \sum_{l=0}^{L} \left|\left| W^{[l]} \odot M^{[l]} \right|\right|_2$$

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

$\odot$

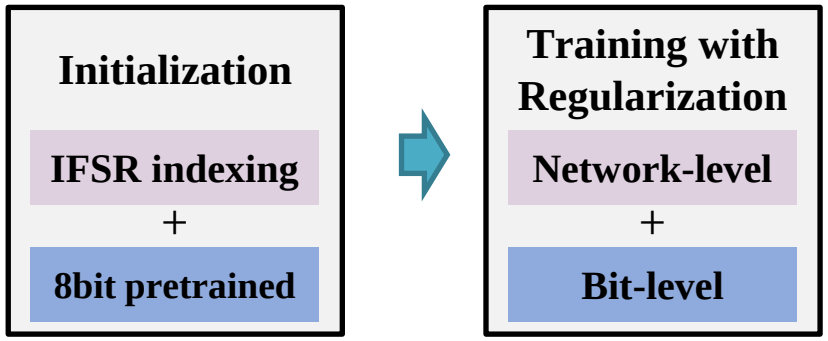
Mask

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

Indices from LFSR

Method: Twofold Sparsity

Joint Bit- and Network-level Sparsity



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

Bit Sparsity: DNN Training under Bit Representation

- **Scaling**:

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

- **Quantization**

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

- **Binary conversion (2's complement)**

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

- **Example**

$$W = \begin{bmatrix} 1.4 \\ 2.0 \\ 0.2 \\ -1.8 \end{bmatrix} \xrightarrow{s = 2.0} W_s = \begin{bmatrix} 0.7 \\ 1.0 \\ 0.1 \\ -0.9 \end{bmatrix}$$

Scaling

$$W_q = \frac{1}{7}\begin{bmatrix} 5 \\ 7 \\ 1 \\ -6 \end{bmatrix}$$

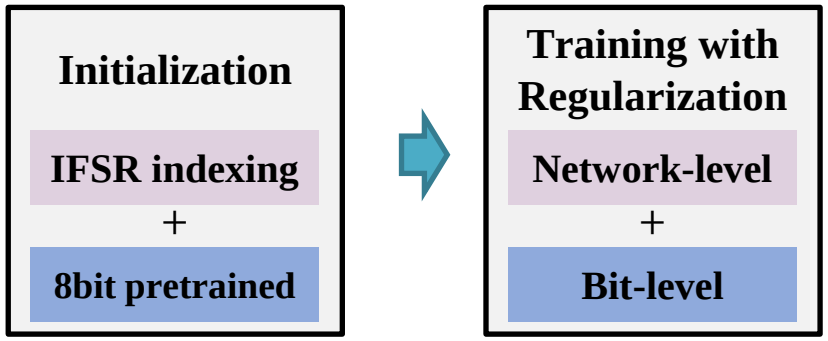
Quantization

$$W_s^{(3:0)} = \frac{1}{7}\begin{bmatrix} 0\,1\,0\,1 \\ 0\,1\,1\,1 \\ 0\,0\,0\,1 \\ 1\,0\,1\,0 \end{bmatrix}$$

Binary conversion
(2′s complement format)

Method: Twofold Sparsity

Joint Bit- and Network-level Sparsity



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

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

Method: Twofold Sparsity

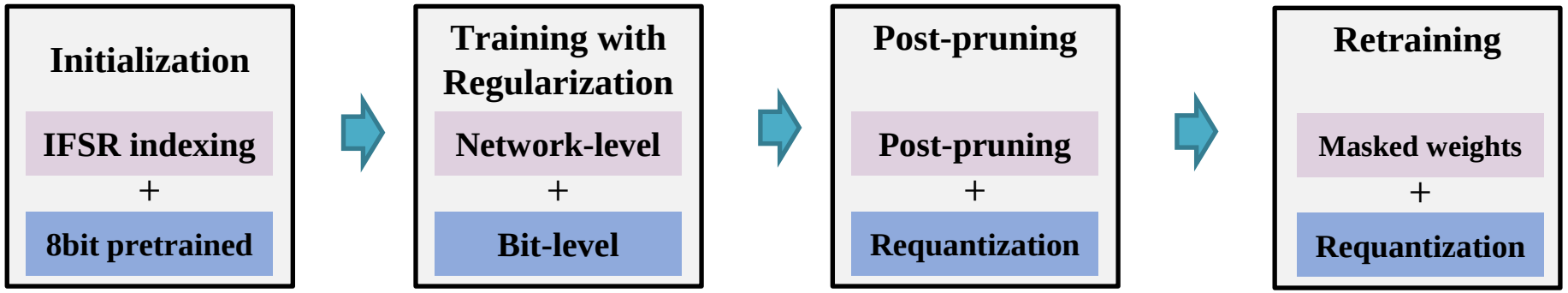
Joint Bit- and Network-level Sparsity

0.71	0	0	0.15
0.34	-0.9	0	0
0	0.28	0	-0.8
0	0	1	0

0.7	0	0	0.1
0.3	-0.9	0	0
0	0.2	0	-0.8
0	0	1	0

Method: Twofold Sparsity

Joint Bit- and Network-level Sparsity



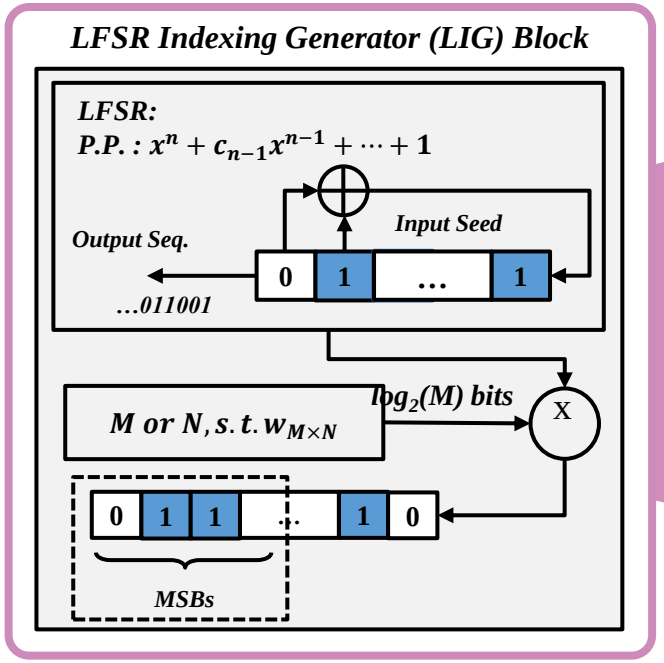
- Masked weights
- Smaller reg. parameters

0.7	0	0	0.1
0.3	-0.9	0	0
0	0.2	0	-0.8
0	0	1	0

0.7	0.1	0.3	-0.9	0.2	-0.8	1

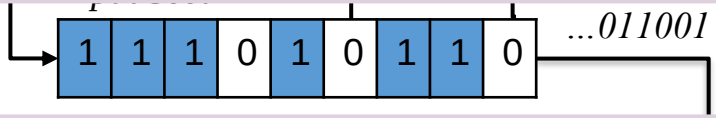
Saved in the memory based on the generated PRS from LFSR

Deep Learning Compression: Hardware



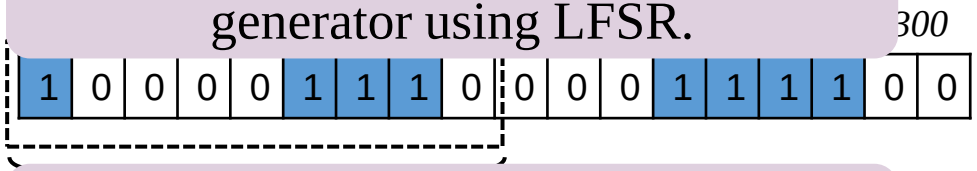
Example: $P.P. = X^9 + X^4 + 1;\ M = 300$

1. LFSRs can be easily be implemented in hardware.

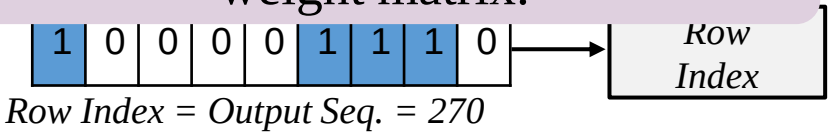


2. Preserves the rank of the generated connectivity matrix.

3. Real-time and automatic address generator using LFSR.



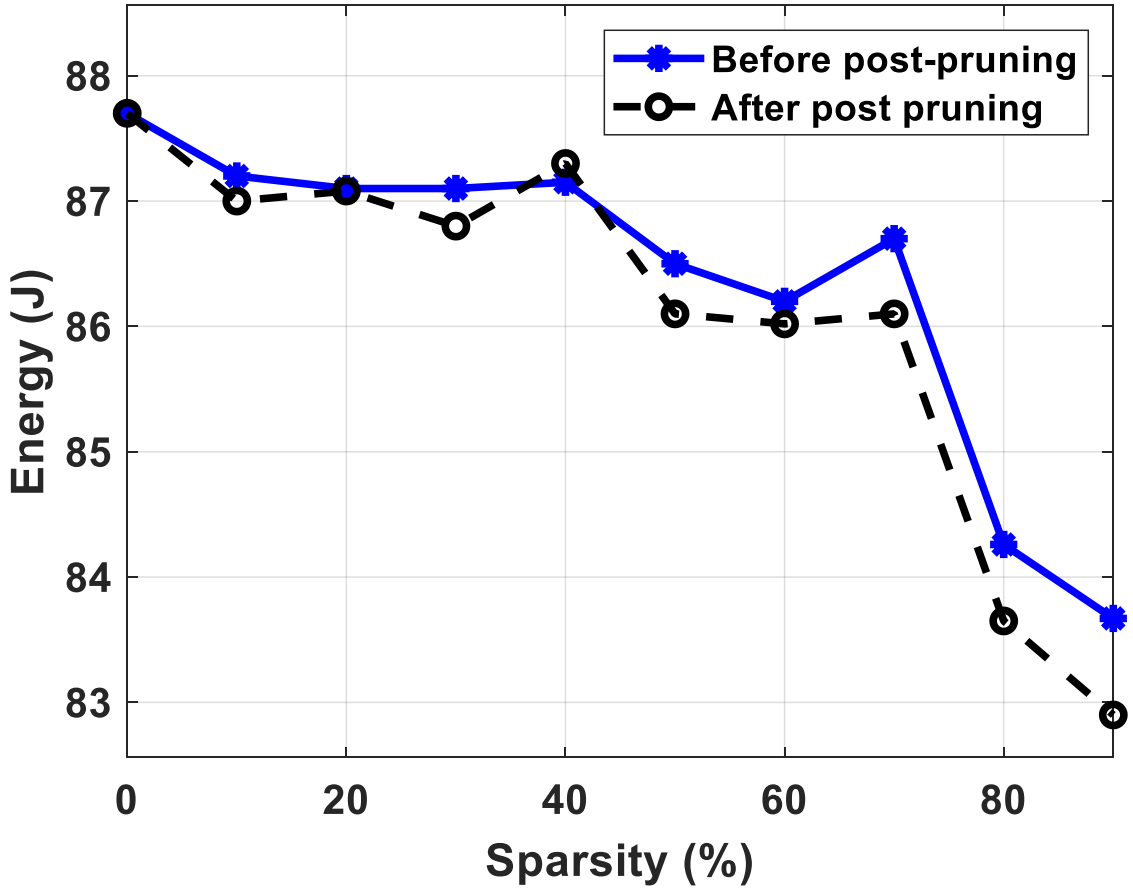
4. No need to store addresses of sparse weight matrix.



Row Index = Output Seq. = 270

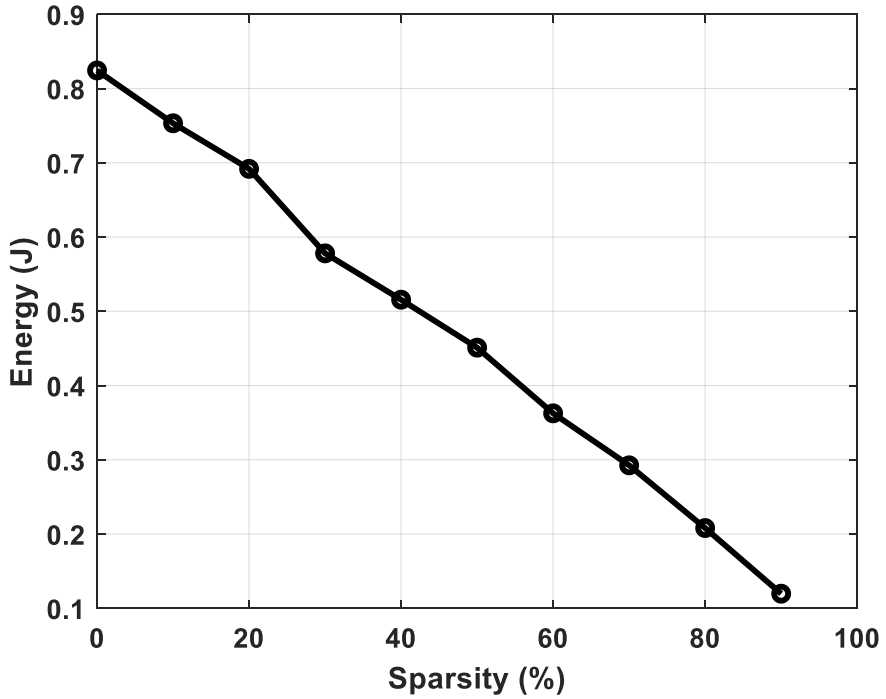
Result: Twofold sparsity, Accuracy

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



Result: Twofold sparsity, Energy

- ResNet-20 using CIFAR-10 dataset.



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

TINY ML TALKS webcast

Thank You!

Email address: fkarimzadeh6@gatech.edu

C-BRIC Center for Brain Inspired Computing

Semiconductor Research Corporation

Copyright Notice

www.tinyml.org

Copyright Notice

www.tinyml.org