DATE 24
DESIGN, AUTOMATION & TEST IN EUROPE
25 - 27 March 2024 · Valencia, Spain
The European Event for Electronic System Design & Test

An Isotropic Shift-Pointwise Network for Crossbar-Efficient Neural Network Design

Ziyi Guan^[1], Boyu Li[1], Yuan Ren[1], Muqun Niu[1], Hantao Huang[2], Graziano Chesi[1], Hao Yu[2] and Ngai Wong[1]

[1]Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong

[2]*School of Microelectronics, Southern University of Science and Technology, Shenzhen, China*

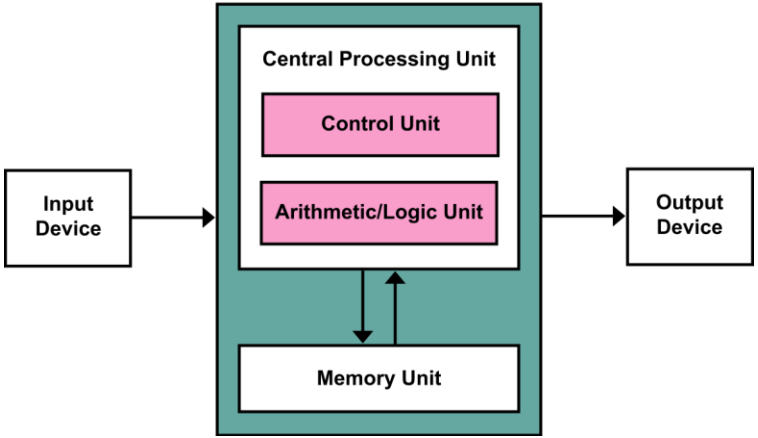
^Presenter

Outline

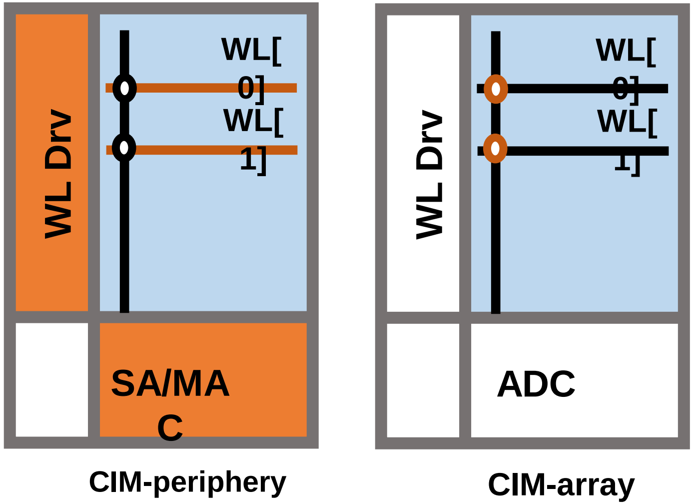
- **Background**
- **Shift Block Design**
- **Isotropic Shift-Pointwise Network Architecture**
- **Hardware Shift Module Design**
- **Experiment Results**
- **Conclusion**

Background: Compute-in-memory (CIM)

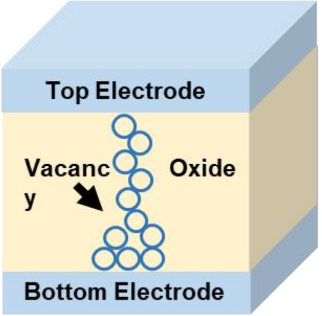
von Neumann architecture



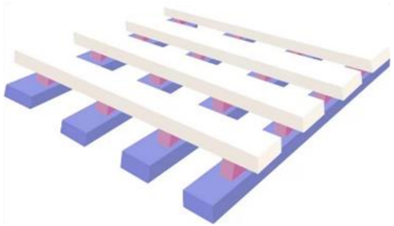
compute-in-memory (CIM) Architecture

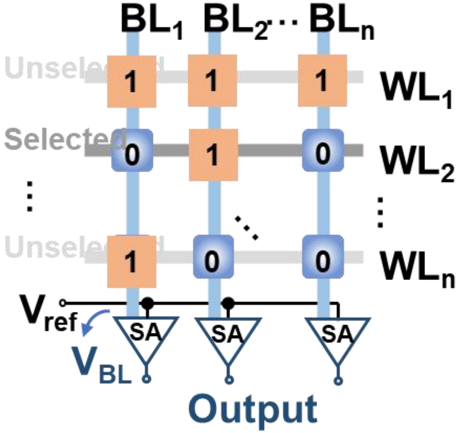


Background: RRAM

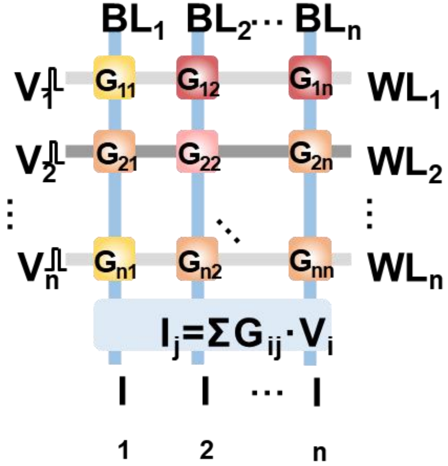


Crossbar Array

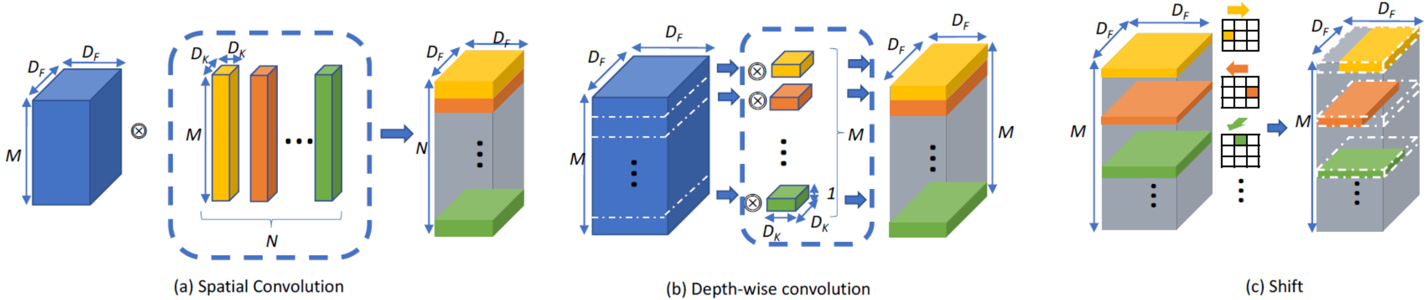


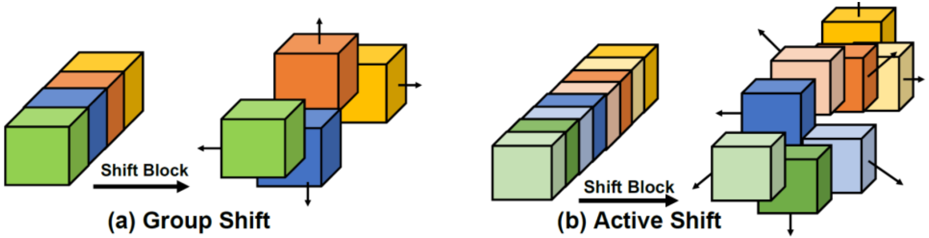


Matrix-Vector Multiplication



Background: Shift Operation

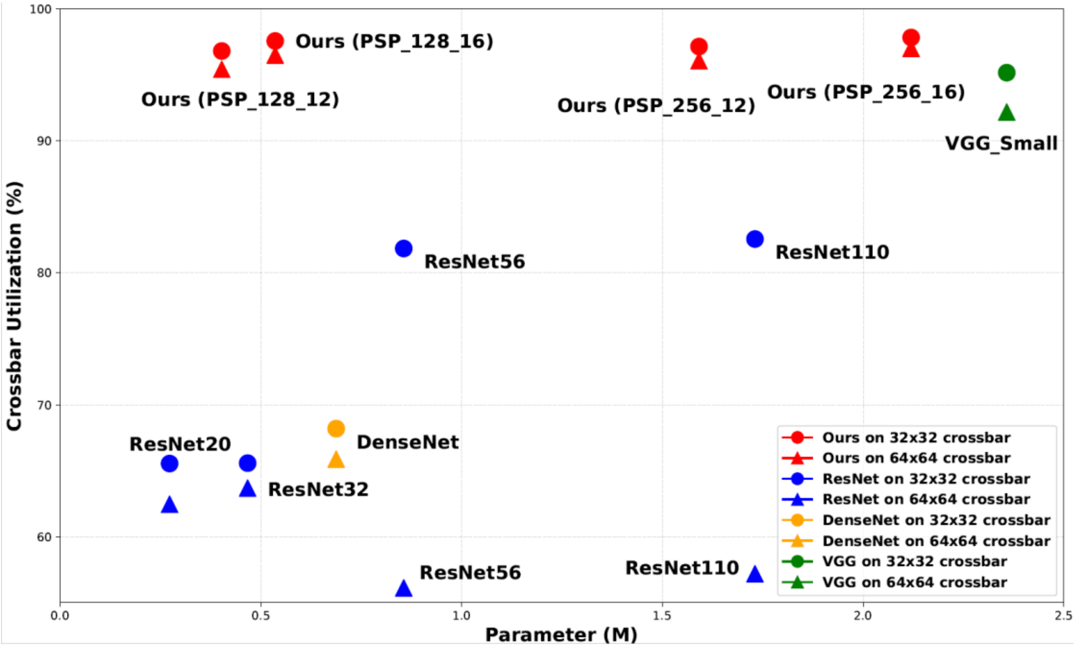




[1] B. Wu, et al, "Shift: A zero flop, zero parameter alternative to spatial convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9127–9135.
[2] Y. Jeon and J. Kim, "Constructing fast network through deconstruction of convolution," Advances in Neural Information Processing Systems, vol. 31, 2018.

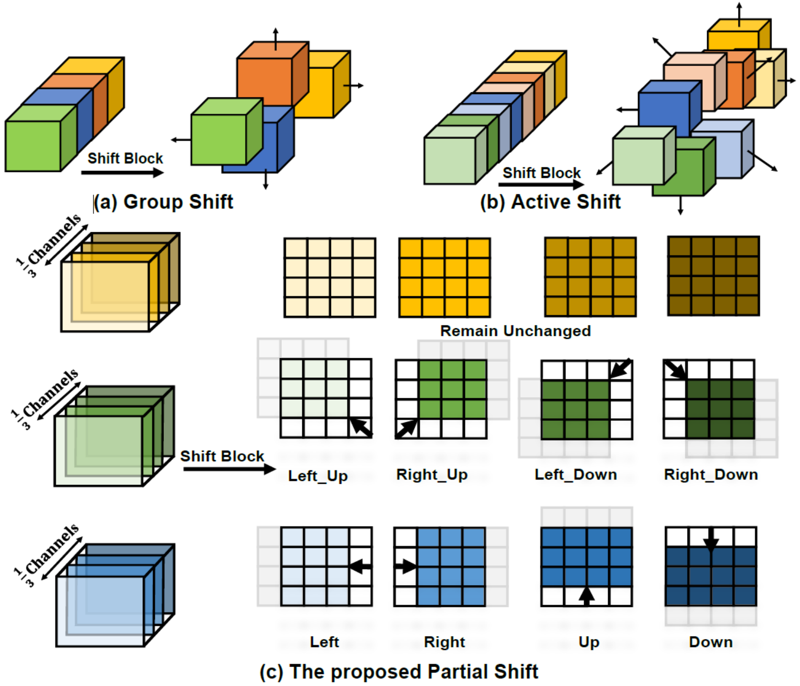
Background: Crossbar Utilization

Comparison of crossbar utilization between the proposed isotropic shift-pointwise networks and some mainstream CNNs.



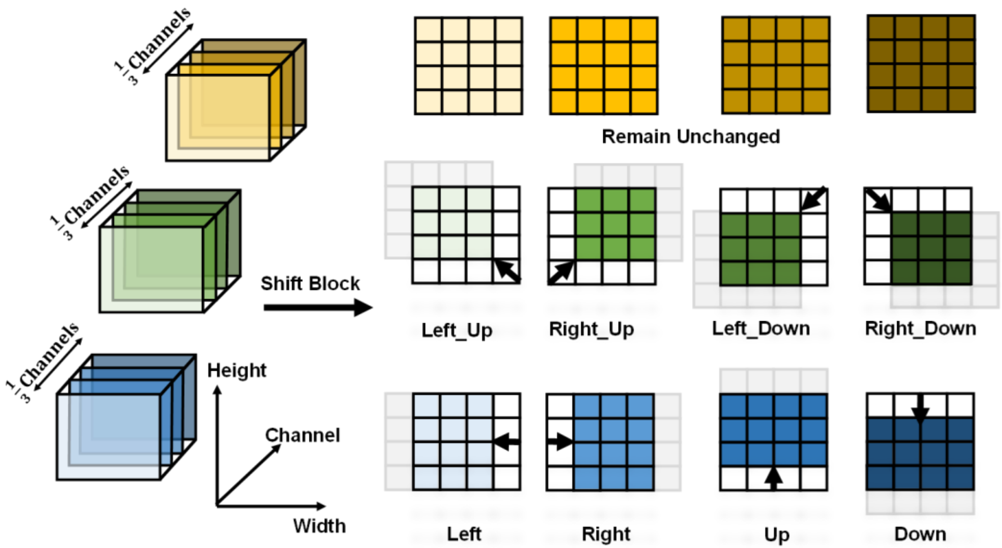
Shift Block Design

Comparison of existing shift operations (a)&(b) and the proposed (c) which comprises 8 directions. 2/3 of all channels are shifted in 8 directions, limiting data movement energy consumption while up-keeping output accuracy.



Shift Block Design

The dataflow and algorithm of the proposed shift block.



(c) The proposed Partial Shift

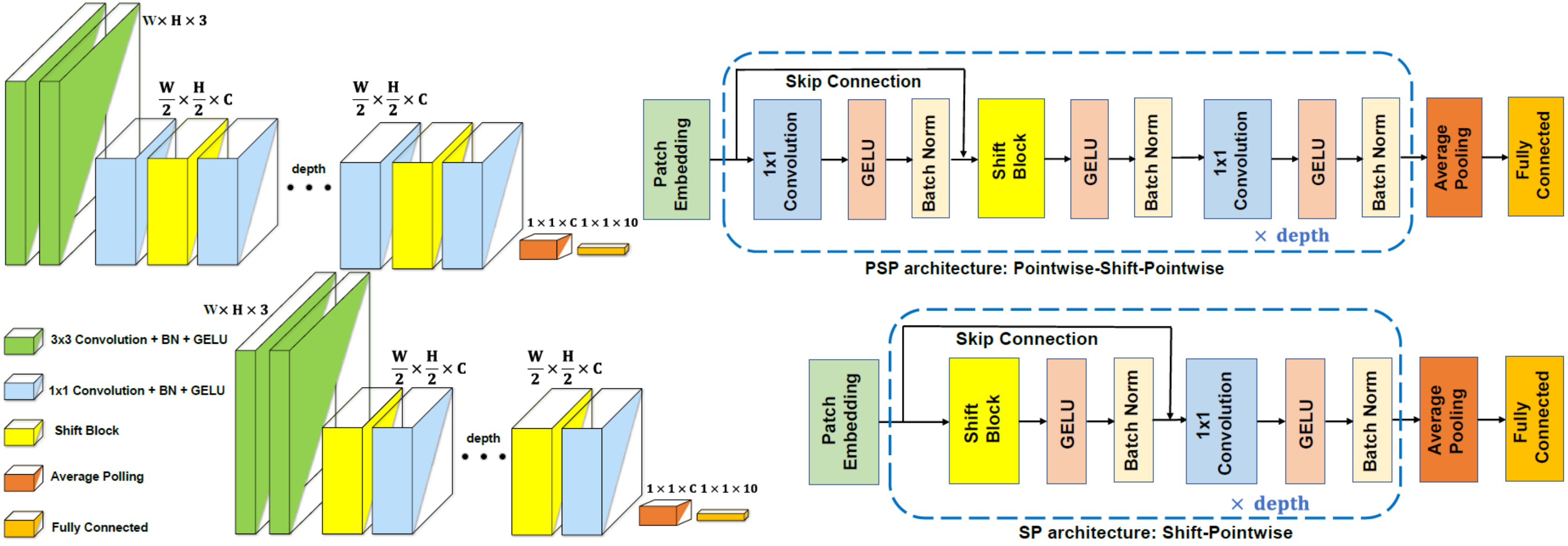
Algorithm 1 Pytorch-based Pseudo code for Shift Block

Input: Input feature tensor x, with a shape of [Batch, Channel, Height, Width], γ a divider to divide the input feature map into nine parts.
Output: The feature map information after the shift block.

def shift($x, g = 1/12$):
  out = torch.zeros_like(x);
  # initial 1/3 of channels with vertical and horizontal shifts
  out$[:, g*0 : g*1, :, : -1] = x[:, g*0 : g*1, :, 1 :]$
  out$[:, g*1 : g*2, :, 1 :] = x[:, g*1 : g*2, :, : -1]$
  out$[:, g*2 : g*3, : -1, :] = x[:, g*2 : g*3, 1 :, :]$
  out$[:, g*3 : g*4, 1 :, :] = x[:, g*3 : g*4, : -1, :]$
  # central 1/3 of channels with diagonal shifts
  out$[:, g*4 : g*5, : -1, : -1] = x[:, g*4 : g*5, 1 :, 1 :]$
  out$[:, g*5 : g*6, :, 1 :] = x[:, g*5 : g*6, :, : -1]$
  out$[:, g*6 : g*7, 1 :, : -1] = x[:, g*6 : g*7, : -1, 1 :]$
  out$[:, g*7 : g*8, 1 :, 1 :] = x[:, g*7 : g*8, : -1, : -1]$
  # final 1/3 of channels with zero shifts
  out$[:, g*8 :, :, :] = x[:, g*8 :, :, :]$
  return out

Isotropic Shift-Pointwise Network Architecture

Isotropic Shift-Pointwise Network Architecture: (Upper). Pointwise-shift-pointwise (PSP) and (Lower) Shift-pointwise (SP). Except for the stem and head, the number of channels in all layers remains unchanged to form an isotropic architecture.

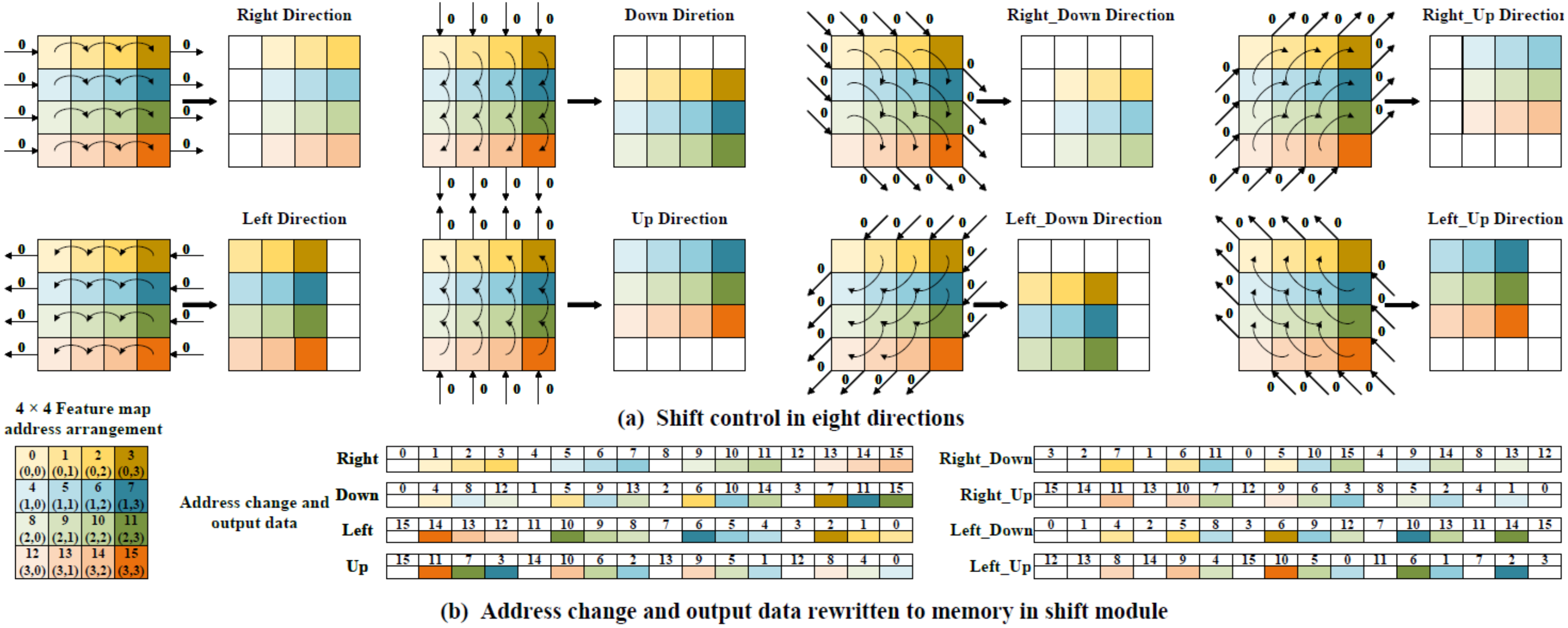


Digital Shift block: Spatial mixing

Analog Pointwise Convolution: Channel mixing

Hardware Shift Module Design

Addressing mechanism and data flow in the proposed shift module: (a) Changes of the address pointer on the feature map in 8 different shift modes; (b) Address change and output data rewritten to memory in shift module in 1-dimension.



(a) Shift control in eight directions

(b) Address change and output data rewritten to memory in shift module

Experiment Results

TABLE I: Comparison of PSP and SP networks vs. mainstream CNNs (ResNet, VGG, and DenseNet40) trained on CIFAR-10 and deployed on 64×64 RRAM crossbars.

Model	Parameters(M)	Top-1 Accuracy(%)	Crossbar Utilization(%)	Latency(ms)	Energy Efficiency(Tops/W)	Chip Area(mm^2)
ResNet110 [1]	1.73	94.52	57.18	9.82	1.89	23.18
DenseNet (40,12) [2]	0.17	91.04	60.53	11.00	3.32	35.50
VGG8 [12]	12.97	90.58	99.39	5.00	4.83	284.18
ShiftResNet110 [6]	0.20	90.67	57.18	10.04	1.88	23.18
PSP_128_8	0.26	92.97	93.40	2.40	6.90	6.84
PSP_256_8	1.06	94.24	94.29	3.57	**7.42**	12.52
PSP_256_12	1.59	94.98	96.07	5.33	7.34	17.91
SP_256_24	1.60	95.21	96.07	6.43	4.68	26.80
PSP_256_16	2.12	**95.64**	97.01	8.13	6.86	23.62

TABLE II: Comparison of PSP architectures against mainstream CNNs such as ResNet, MobileNet w.r.t. Top-1 Accuracy on ImageNet dataset.

Model	Parameters(M)	Top-1 Accuracy(%)
ResNet18 [1]	11.17	69.15
MobileNetV1 [16]	4.2	70.60
PSP_256_12	1.8	72.20
PSP_512_12	6.32	73.18
PSP_512_16	8.43	**74.87**

TABLE III: Comparison of energy consumption and latency of different models in the shift module.

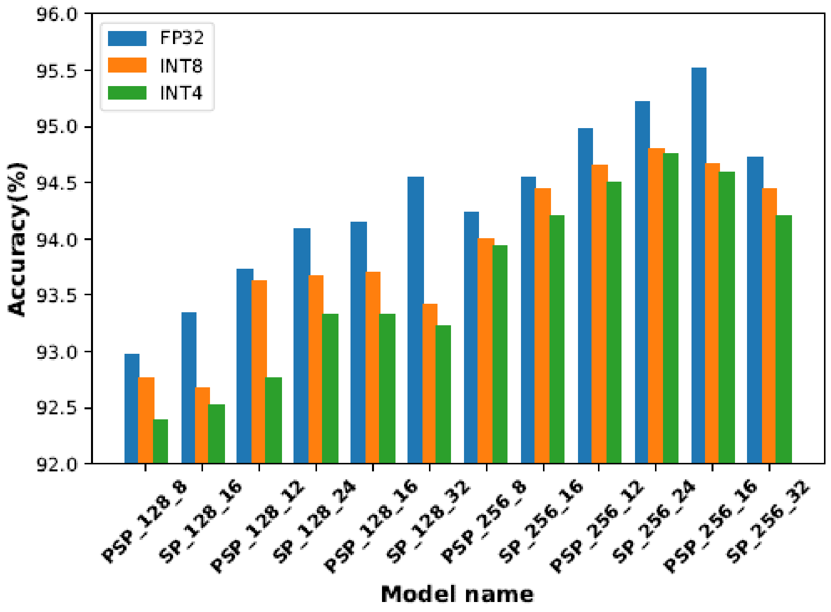
Model	Energy(nJ)	Normalized Energy(nJ)	Latency(ms)
PSP_128_16	66.800	54.177	0.04301
PSP_256_16	215.000	174.371	0.12902
SP_256_32	284.000	230.333	0.17203

Experiment Results

TABLE IV: Comparison of various PSP and SP architectures and ResNet for w.r.t. QAT Top-1 Accuracy for FP32, INT8 and INT4 on CIFAR-10/100 datasets.

Model	Parameters(M)	FP32(%)	INT8(%)	INT4(%)
ResNet32	0.47	92.61/71.05	92.76/70.11	92.39/69.78
ResNet110	1.73	94.52/73.44	92.76/71.44	92.39/71.20
PSP_128_12	0.40	93.73/74.22	93.62/72.05	92.76/70.94
PSP_256_12	1.59	94.98/**77.94**	**94.88/76.55**	94.50/**76.10**
PSP_256_16	2.12	**95.64**/77.86	94.72/76.02	94.59/75.64

Comparison between FP32, INT8 and INT4 among various PSP and SP architectures on CIFAR-10 dataset.



Conclusion

- **We are among the first to design a lightweight isotropic shift-pointwise network with near-100\% RRAM crossbar utilization. The proposed PSP and SP networks outperform standard CNNs in model accuracy and hardware metrics.**
- **A novel reconfigurable and energy-efficient shift module is developed, enabling accurate characterization of the hardware metrics affiliated with the shift operation.**
- **We utilize an algorithm-hardware co-design to exploit shift operation in digital domain for spatial mixing and pointwise operation in analog domain for channel mixing.**

Thank you!
Q&A