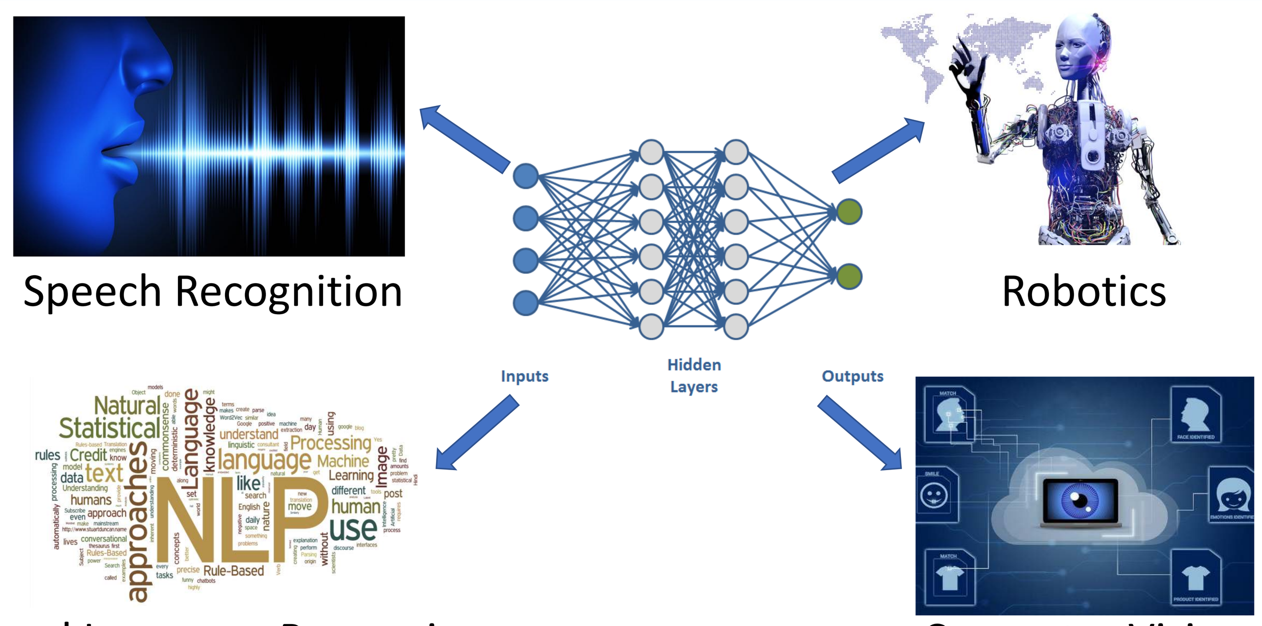
BiHiWe: Mixed-Signal Charge-Domain Acceleration of Deep Neural Networks

Soroush Ghodrati[1], Hardik Sharma[2], Sean Kinzer[1], Amir Yazdanbakhsh[2], Kambiz Samadi[3], Nam Sung Kim[4], Doug Burger[5], Hadi Esmaeilzadeh[1]

Alternative Computing Technologies (ACT) Lab

[1]UC San Diego, [2]Georgia Tech, [3]Qualcomm, [4]UIUC, [5]Microsoft

Overview



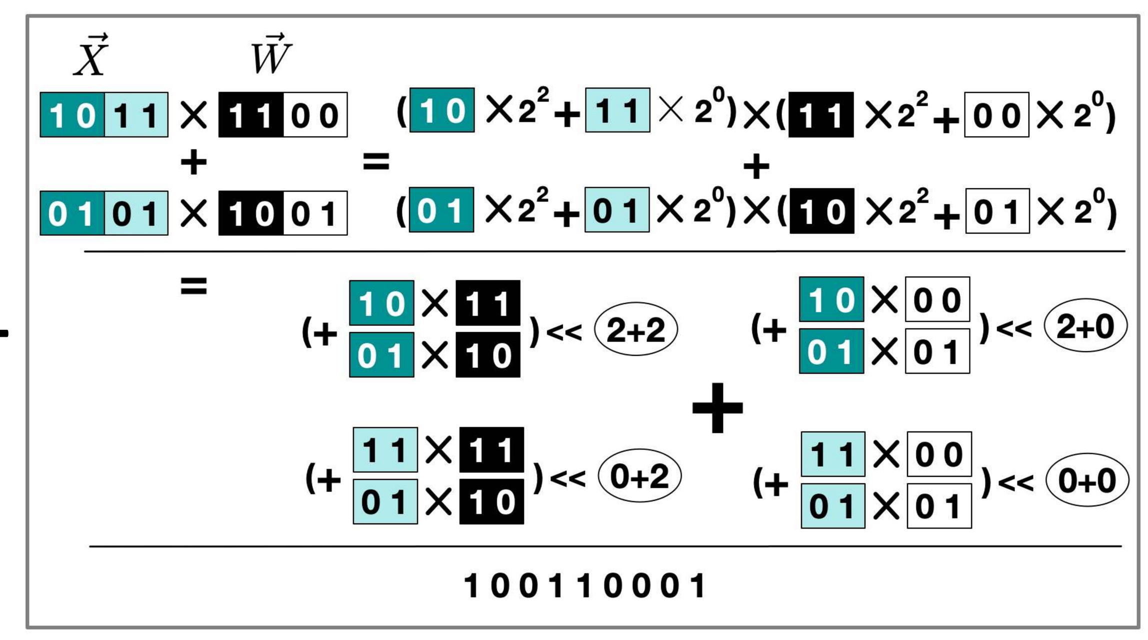
Percentage of operations in different layers

DNN	AlexNet	CIFAR-10	GoogLeNet	ResNet-18	ResNet-50	VGG-16	VGG-19	YOLOv3	PTB-RNN	PTB-LSTM
Convolution Layers	91.8	98.4	99.6	99.4	99.8	99.1	99.3	99.8	—	—
Fully-Connected Layers	8.1	1.5	0.1	0.5	0.1	0.8	0.6	0.1	99.9	99.9
Other Layers	0.1	0.3	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

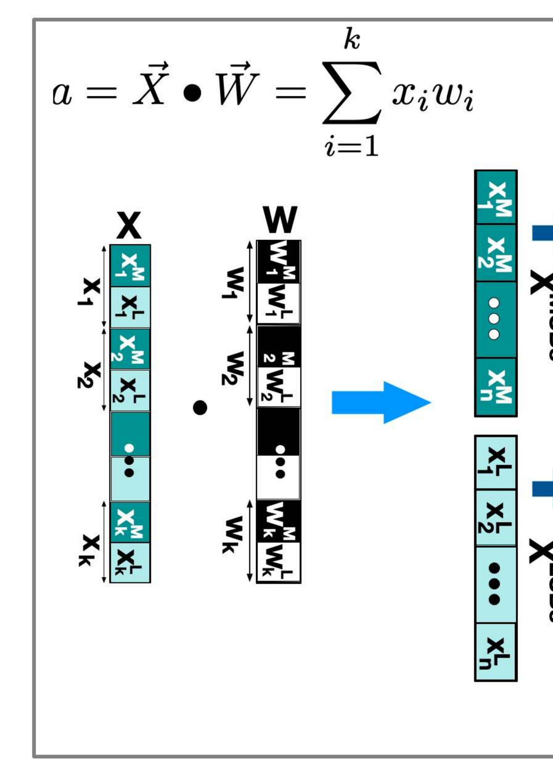
Amdahl's law motivates moving Convolution and Fully-Connected layers to analog domain

Compute Model

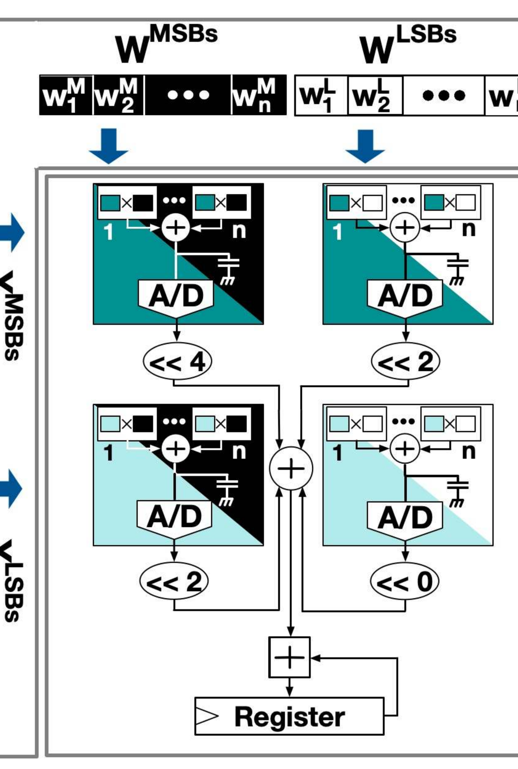
Our Approach: Enabling analog computing via wide, interleaved, and bit-partitioned arithmetic



(a) Bit-Partitioning Multiply-Accumulation

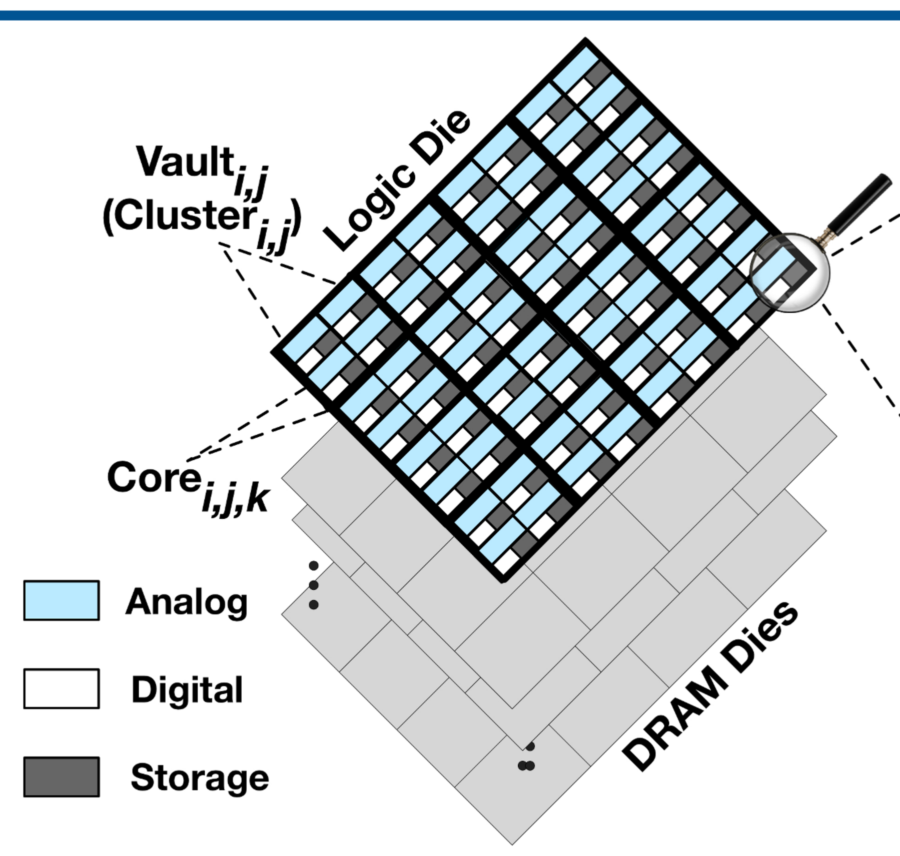


(b) Bit-Partitioned Vector Rearrangement

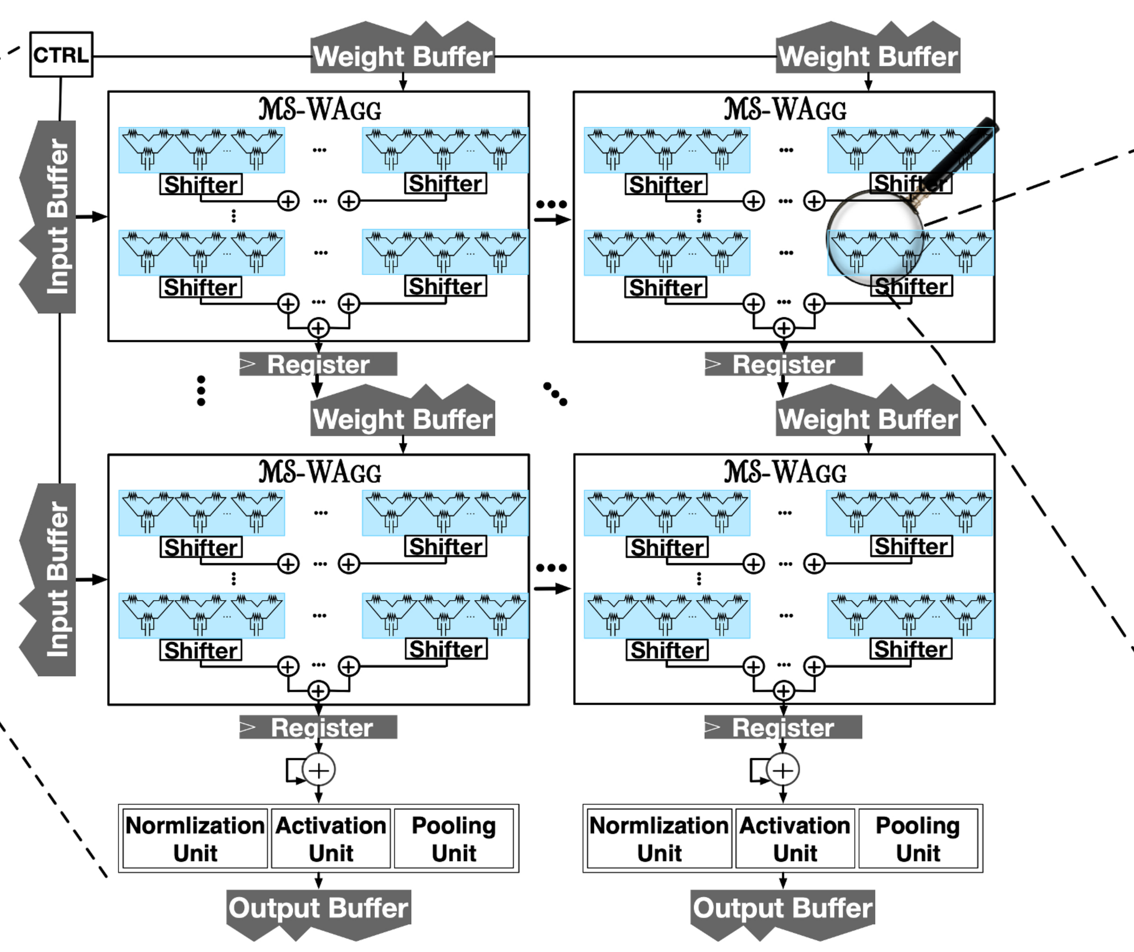


(c) Wide Bit-Partitioned Vector Dot-Product

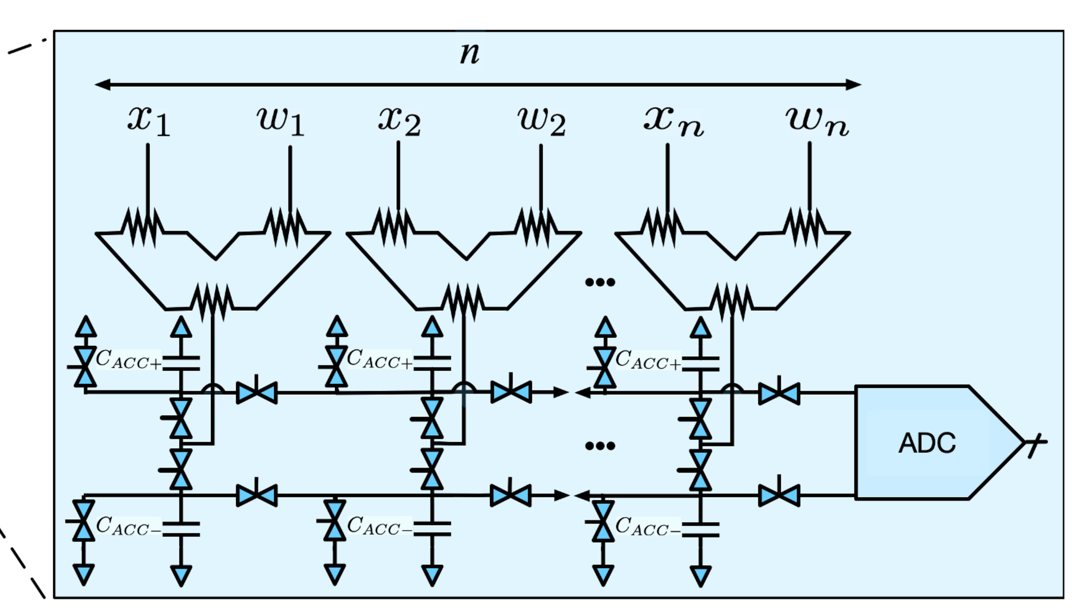
Hierarchically Clustered Architecture of BiHiWe



(a) Clustered Architecture

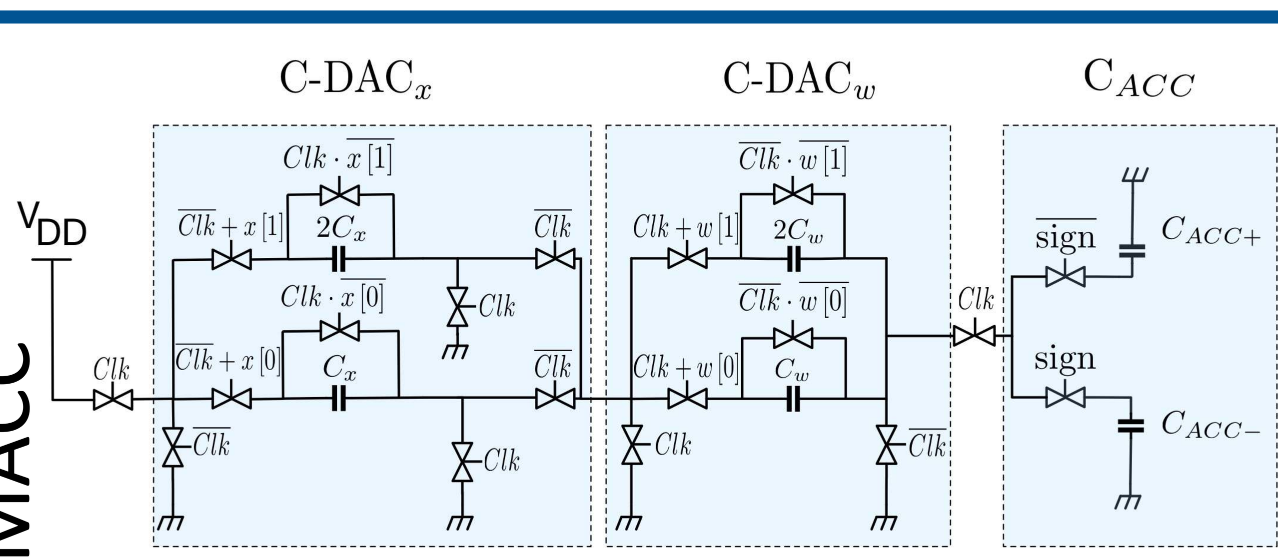


(b) Accelerator Core



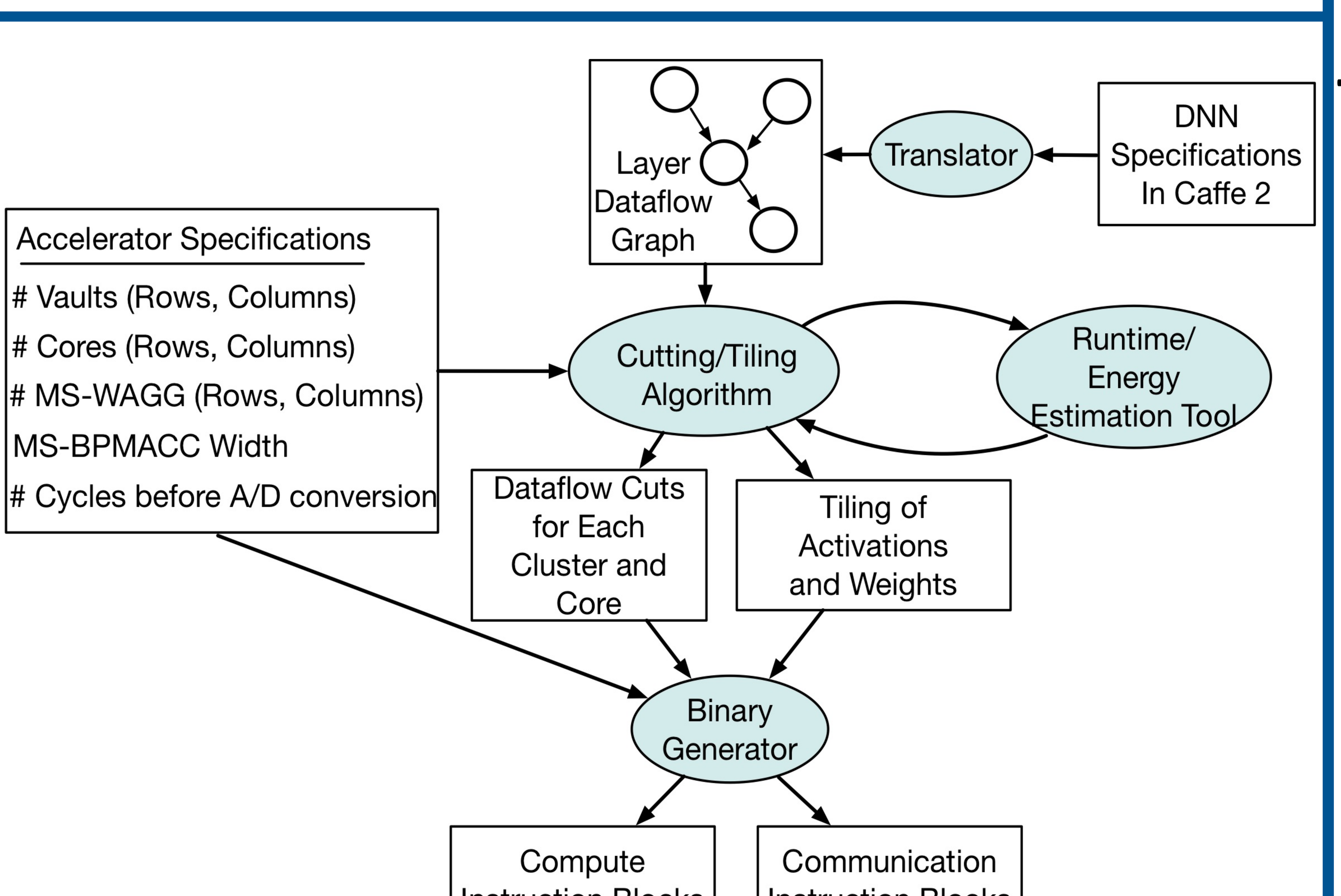
(c) Mixed-Signal Bit-Partitioned MACC

Low-Bitwidth Switched-Capacitor MACC

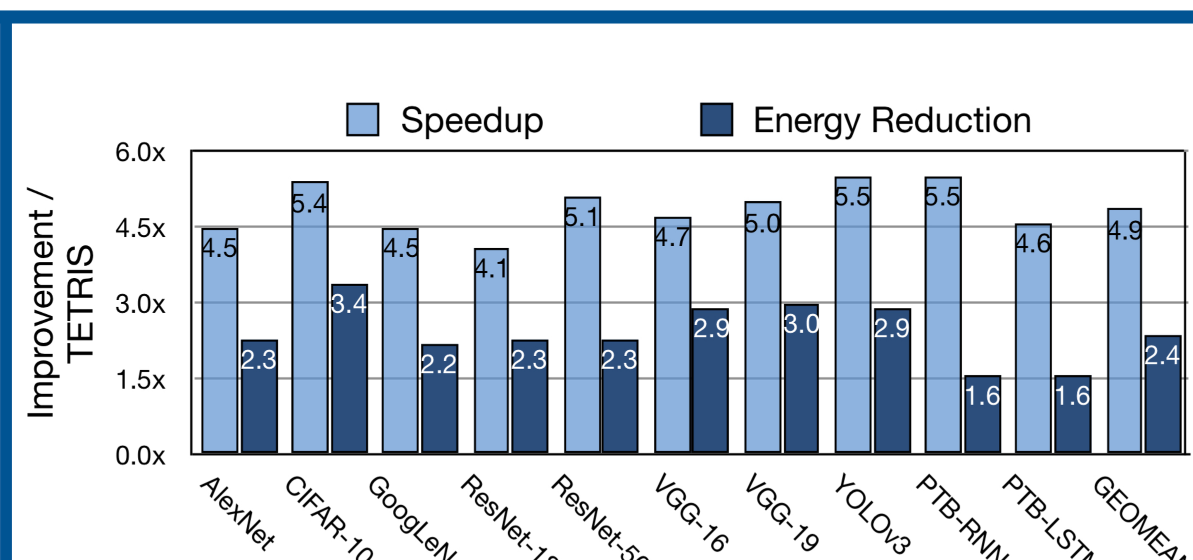


Switched-Capacitor design enables storage of the intermediate results over multiple cycles and reduces A/D conversion rate.

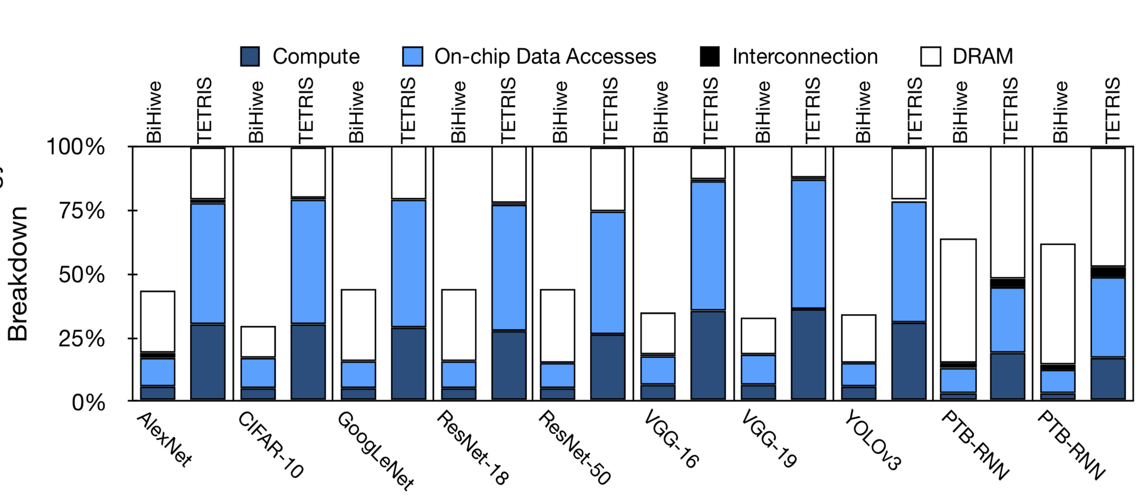
BiHiWe Compilation Stack



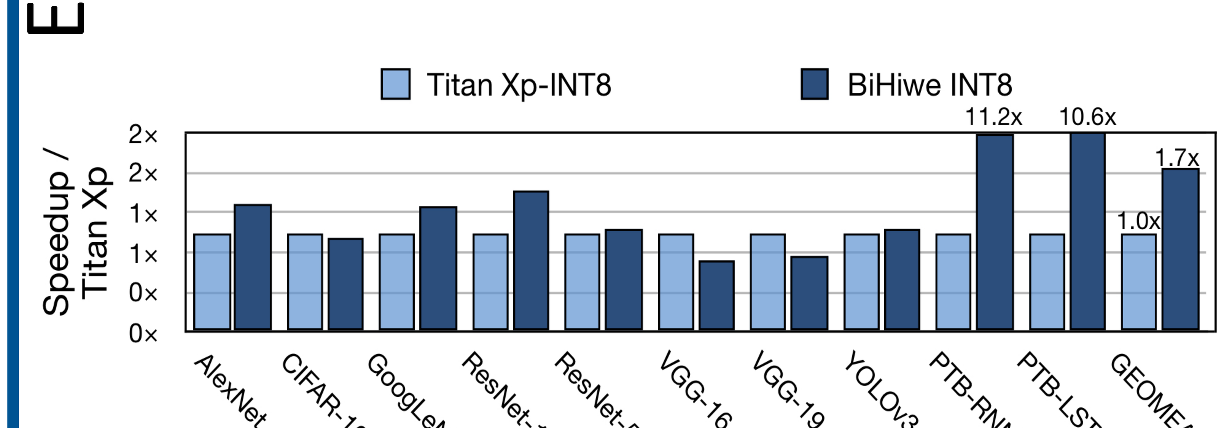
Evaluation



4.9x speedup and 2.4x energy reduction over Tetris an optimized 3D-stacked fully-digital accelerator for DNNs.



Energy reduction breakdown of BiHiWe compared to Tetris.



BiHiWe delivers 66.1x Performance-per-Watt compared to Nvidia Titan Xp while running 1.7x faster.

DNN Model	Dataset	Top-1 Accuracy (With non-idealities)	Top-1 Accuracy (After fine-tuning)	Top-1 Accuracy (Ideal)	Accuracy Loss
AlexNet	Imagenet	53.12%	56.64%	57.11%	0.47%
YOLOv3	Imagenet	75.92%	77.1%	77.22%	0.21%
CIFAR-10	CIFAR-10	90.82%	91.01%	91.03%	0.02%
VGG-16	Imagenet	70.31%	71.28%	71.46%	0.18%
VGG-19	Imagenet	73.24%	74.20%	74.52%	0.32%
ResNet-18	Imagenet	66.91%	68.96%	68.98%	0.02%
ResNet-50	Imagenet	74.5%	75.21%	75.25%	0.04%
GoogLeNet	Imagenet	67.15%	68.39%	68.72%	0.33%
PTB-RNN	Penn TreeBank	1.1 BPC	1.6 BPC	1.1 BPC	0.0 BPC
PTB-LSTM	Penn TreeBank	97 PPW	170 PPW	97 PPW	0.0 PPW

BiHiWe has no virtual impact on the accuracy of the DNN models.