IEEE Xplore® Digital Library

ISCAS 2020

2020 IEEE International Symposium on Circuits and Systems
Virtual, October 10-21, 2020

CAS

MINT: *M*ixed-precision RRAM-based *IN*-memory *T*raining Architecture

Hongwu Jiang, Shanshi Huang, Xiaochen Peng and Shimeng Yu
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA

2020 IEEE International Symposium on Circuits and Systems
Virtual, October 10-21, 2020

Georgia Tech School of Electrical and Computer Engineering

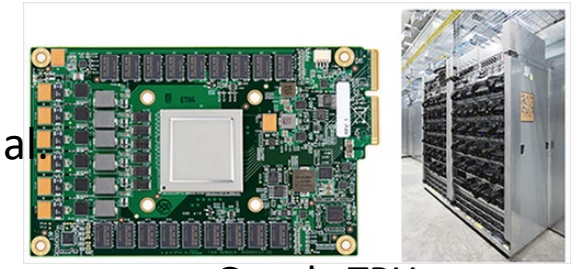
Email: shimeng.yu@ece.gatech.edu
Website: http://shimeng.ece.gatech.edu/

Outline

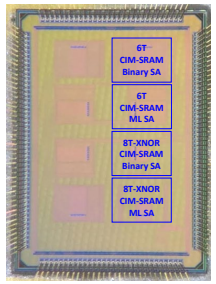
- Background and Motivation
- Weight Mapping Strategy for Training
  - Transpose Weight Matrix
  - Mapping Dataflow for Training in CIM
- Proposed MINT Architecture
  - RRAM Subarray Design
  - Overall Architecture
- Evaluation
  - Impact of ADC quantization and RRAM Non-idealities
  - Hardware Performance Benchmarking
- Conclusion

Background and Motivation

[1] N. P. Jouppi, et al. ISCA 2017



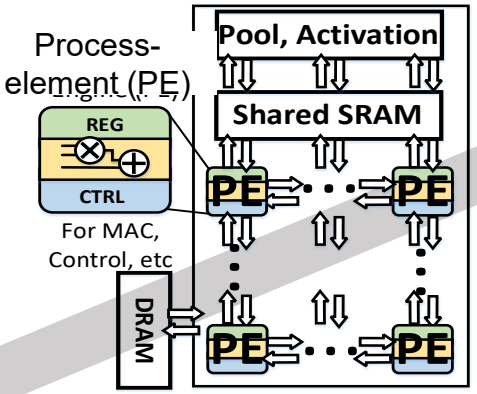
Google TPU

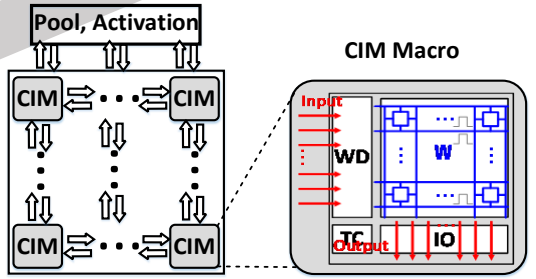


[2] R. Liu, et al. *DAC*, 2018

CHIP SUMMARY	
Technology	65nm CMOS
Unit-Macro Size	4Kb (64x64b)
6T SRAM bit-cell size	0.5μm X 1.05μm
8T XNOR-SRAM bit-cell size	0.51μm X 1.91μm
6T SRAM energy-efficiency	>100 TOPS/W
8T XNOR-SRAM energy-efficiency	>30 TOPS/W

NVIDIA GPU





Memory Wall

inference

training

CPU

GPU

Near memory

ASIC design accelerator

In memory computing

Compute-in-memory (CIM): the weight are stored in memory array, while the activations are loaded in as input to WLs: Parallel access, eliminate MAC units and weights movement

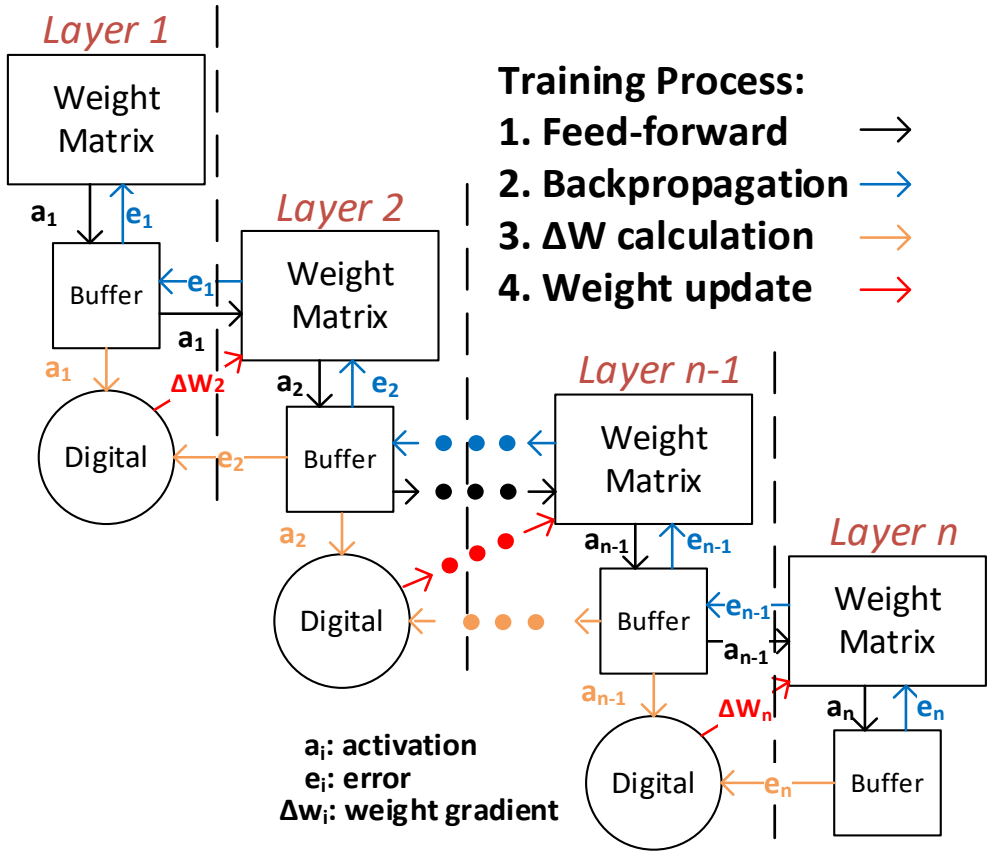
Background and Motivation

- Low-precision DNN Training
  - WAGE is proposed as a low precision training method in [1].
  - Friendly for hardware since it uses fixed range quantization for activations and weights

Method	k_W	k_A	k_G	k_E	Opt	BN	MNIST	SVHN	CIFAR10	ImageNet
BC	1	32	32	32	Adam	✓	1.29	2.30	9.90	-
BNN	1	1	32	32	Adam	✓	0.96	2.53	10.15	-
BWN[1]	1	32	32	32	withM	✓	-	-	-	43.2/20.6
XNOR	1	1	32	32	Adam	✓	-	-	-	55.8/30.8
TWN	2	32	32	32	withM	✓	0.65	-	7.44	34.7/13.8
TTQ	2	32	32	32	Adam	✓	-	-	6.44	42.5/20.3
DoReFa[2]	8	8	32	8	Adam	✓	-	2.30	-	47.0/ -
TernGrad[3]	32	32	2	32	Adam	✓	-	-	14.36	42.4/19.5
WAGE	2	8	8	8	SGD	✗	**0.40**	**1.92**	6.78	51.6/27.8

[1] S. Wu, et al. *ICLR*, 2018

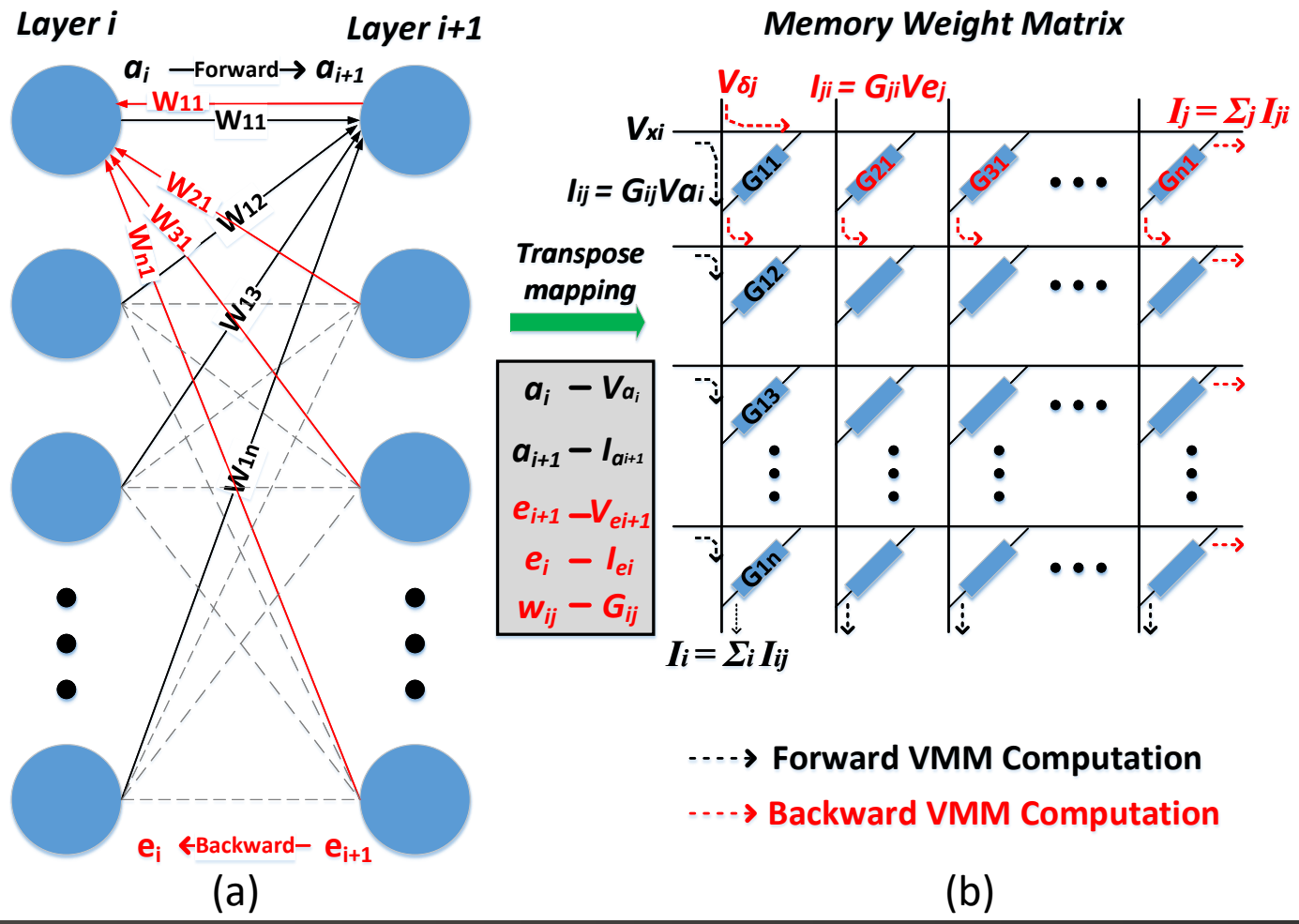
- Workflow for DNN training in CIM



Outline

- Background and Motivation
- Weight Mapping Strategy for Training
  - Transpose Weight Matrix
  - Mapping Dataflow for Training in CIM
- Proposed MINT Architecture
  - RRAM Subarray Design
  - Overall Architecture
- Evaluation
  - Impact of ADC quantization and RRAM Non-idealities
  - Hardware Performance Benchmarking
- Conclusion

Transpose Weight Matrix

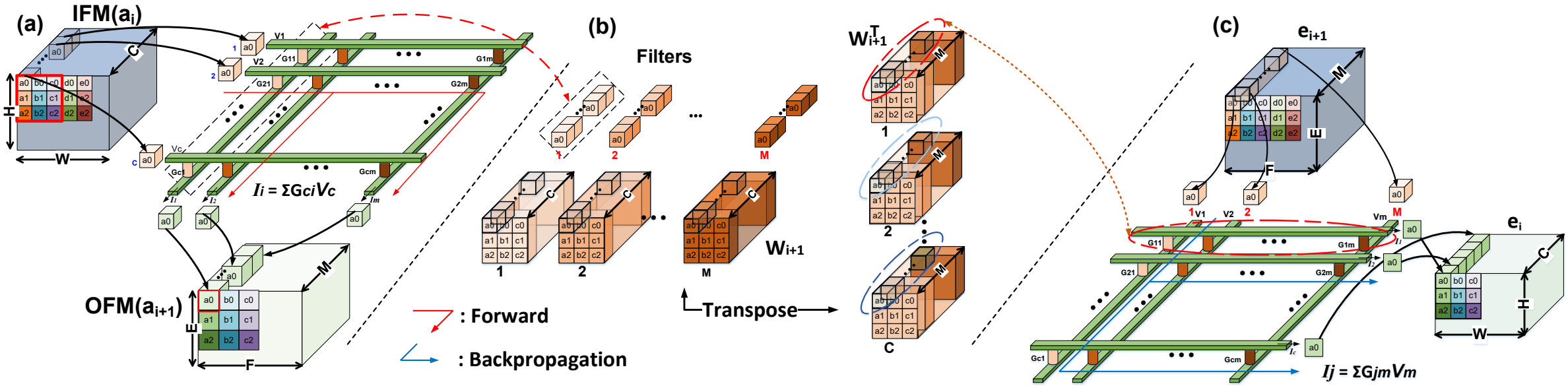


- (a) shows a schematic of two hidden layers of neurons with weighted interconnection.
- (b) shows the mapping methods between weight matrix and resistor crossbar array.
- Device conductance values *Gij* represent the weights *Wij*.
- The crossbar array performs the weighted summations during the **forward** and **backward** propagations.

Outline

- Background and Motivation
- Weight Mapping Strategy for Training
  - Transpose Weight Matrix
  - Mapping Dataflow for Training in CIM
- Proposed MINT Architecture
  - RRAM Subarray Design
  - Overall Architecture
- Evaluation
  - Impact of ADC quantization and RRAM Non-idealities
  - Hardware Performance Benchmarking
- Conclusion

Mapping Dataflow for Training in CIM

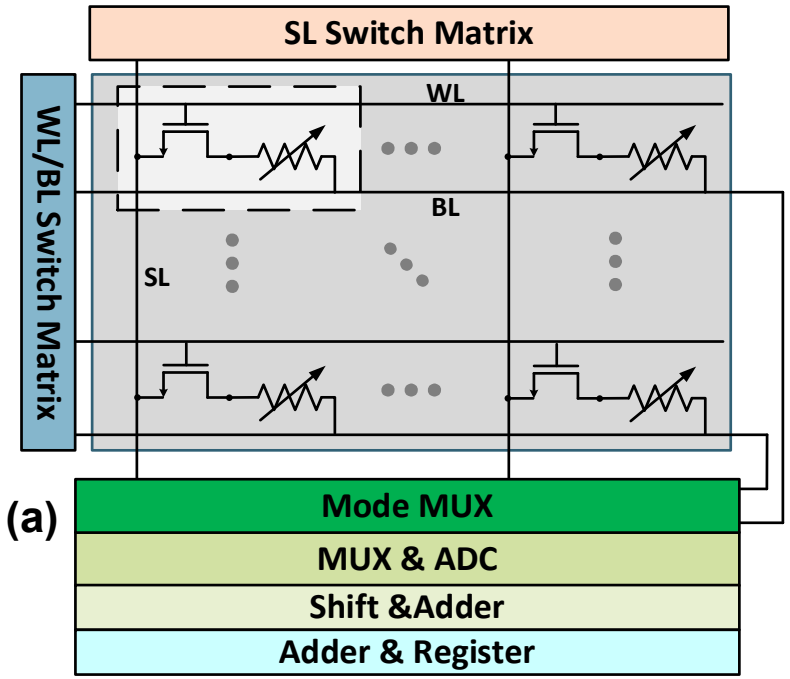


- The feed-forward process of the CIM array is shown in Fig. (a) while the details of error calculation is shown in Fig. (c). Transpose weight mapping scheme is shown in (b)
- With such transpose array and weight mapping strategy, FF and BP can be performed within the same array.

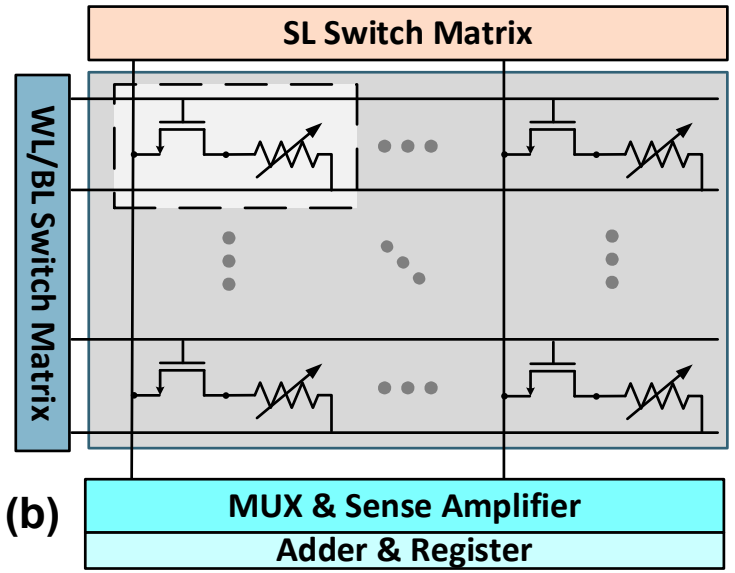
Outline

- Background and Motivation
- Weight Mapping Strategy for Training
  - Transpose Weight Matrix
  - Mapping Dataflow for Training in CIM
- Proposed MINT Architecture
  - RRAM Subarray Design
  - Overall Architecture
- Evaluation
  - Impact of ADC quantization and RRAM Non-idealities
  - Hardware Performance Benchmarking
- Conclusion

RRAM Subarray Design



(a). Computing subarray with expensive ADC



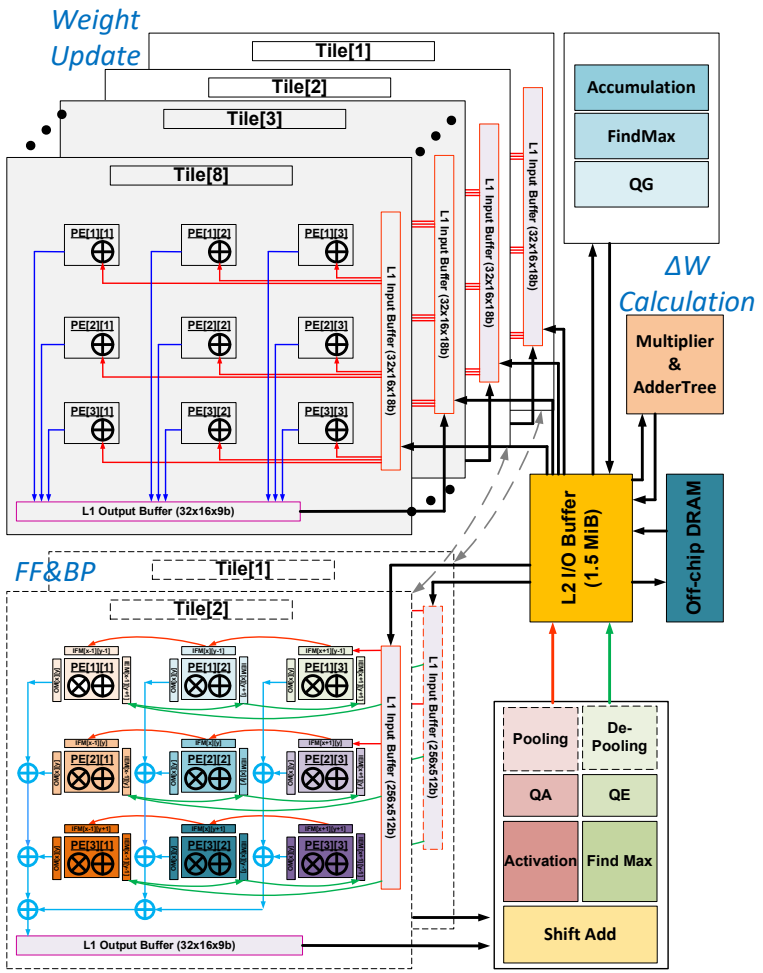
(b). Storage subarray with simple SA

- One-transistor-one-resistor (1T1R) pseudo-crossbar structure
- Each cell only has binary on-state or off-state
- Two types of subarrays are proposed:
  - computing subarrays
  - storage subarrays

Outline

- Background and Motivation
- Weight Mapping Strategy for Training
  - Transpose Weight Matrix
  - Mapping Dataflow for Training in CIM
- Proposed MINT Architecture
  - RRAM Subarray Design
  - Overall Architecture
- Evaluation
  - Impact of ADC quantization and RRAM Non-idealities
  - Hardware Performance Benchmarking
- Conclusion

Overall Architecture



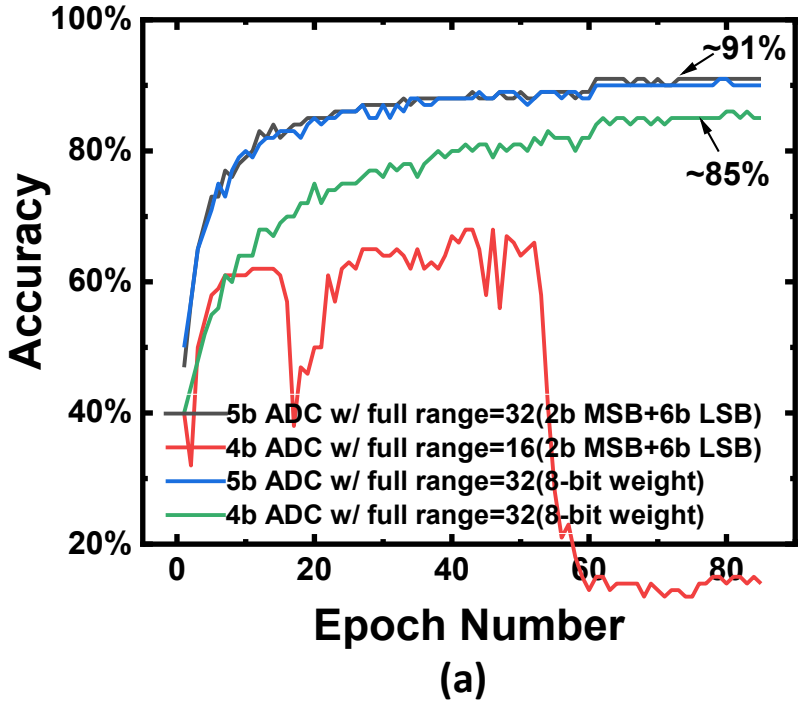
Top-level architecture for one convolution layer

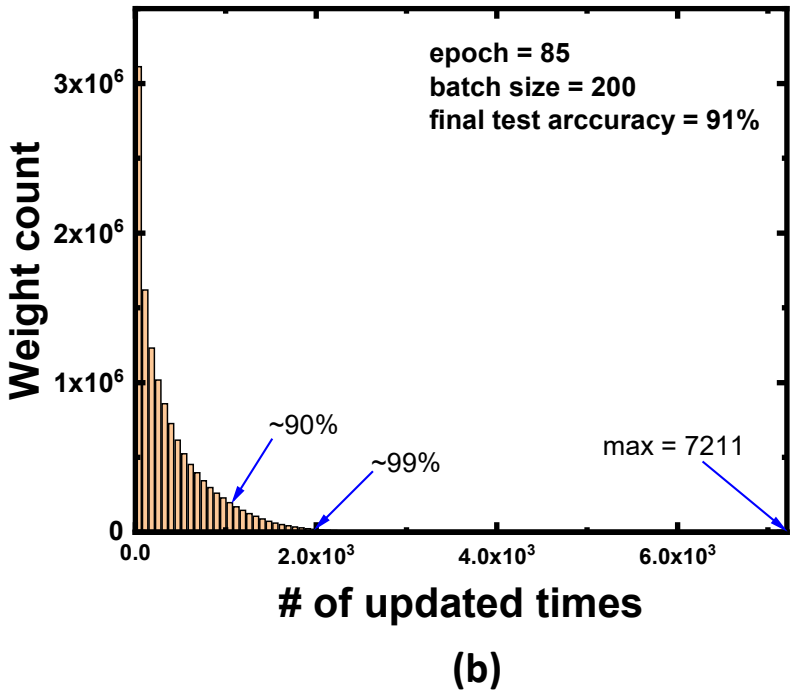
- Weight bits with different significance are stored on different tiles
- First 2 MSBs of the weight for convolution in FF and BP, corresponding to computing subarrays in Tile[1] and Tile[2]. Tile[3-8] consist of regular storage arrays for the other 6 LSBs of the weight.
- Digital MAC computation for gradient calculation
- Accumulation & FindMax block for weight update

Outline

- Background and Motivation
- Weight Mapping Strategy for Training
  - Transpose Weight Matrix
  - Mapping Dataflow for Training in CIM
- Proposed MINT Architecture
  - RRAM Subarray Design
  - Overall Architecture
- Evaluation
  - Impact of ADC quantization and RRAM Non-idealities
  - Hardware Performance Benchmarking
- Conclusion

Impact of ADC quantization and RRAM Non-idealities





(a). Training performance for different ADC quantization range and resolution. (b). Statistics of weight update frequency

- VGG-8 on CIFAR10
- 128*128 array size
- No accuracy loss with 5-bit ADC

- 50K Image per epoch
- 99% RRAM flips < 2000 times
- Today's RRAM endurance > 10^6

Outline

- Background and Motivation
- Weight Mapping Strategy for Training
  - Transpose Weight Matrix
  - Mapping Dataflow for Training in CIM
- Proposed MINT Architecture
  - RRAM Subarray Design
  - Overall Architecture
- Evaluation
  - Impact of ADC quantization and RRAM Non-idealities
  - Hardware Performance Benchmarking
- Conclusion

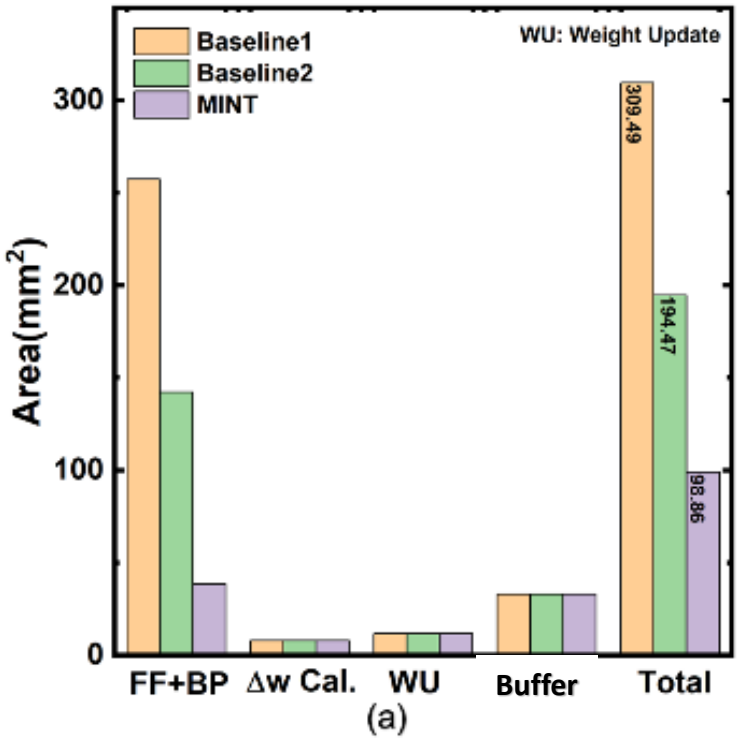
Hardware Performance Benchmarking

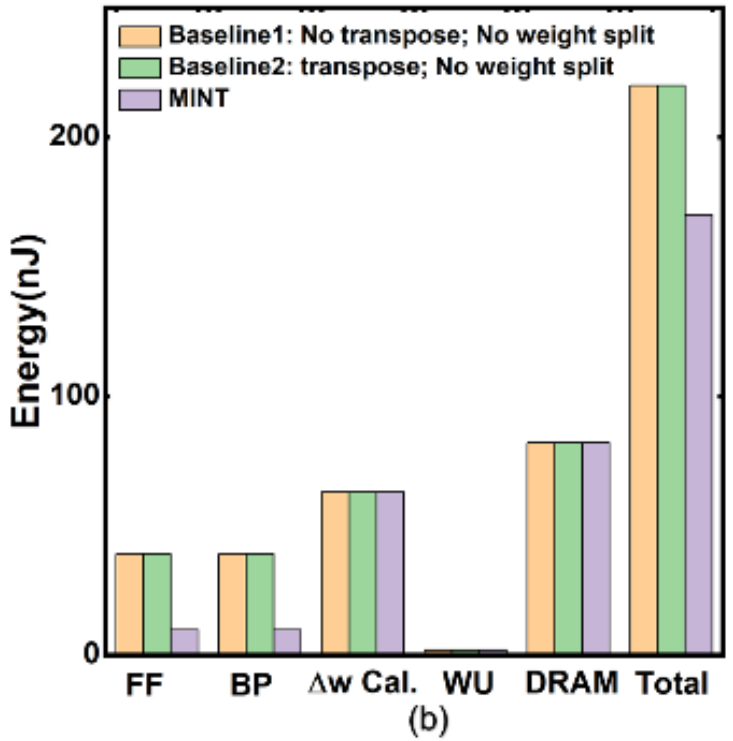
- 8-layer VGG-like network on CIFAR-10 dataset; split the *8-bit* weight into *2-bit* MSBs and *6-bit* LSBs
- Built with a modified *NeuroSim* [1] framework. (32nm technode)
- Inter-layer parallelism scheme
- Table shows the chip-level parameters including the hardware configuration, precision, area and energy for key circuit modules.

[1] P.Y. Chen, et al. *IEDM*, 2017

Main block	Spec.	Energy(pJ/op)	Area(mm^2)
Subarray Level			
RRAM array	Size: 128x128 Precision:1-bit	99.85	0.0003
MUX & Decoder		3.68	0.0013
ADC	Number: 32 Precision:5-bit	327.92	0.0019
Shift Add	Number: 32 Precision:12-bit	71.17	0.0009
WL/SL SwitchMatrix and other		15.74	0.0006
Subarray Total		518.36	0.0050
PE Level			
Subarray	Size: 4x4	8293.77	0.08
Adder Tree	Number: 64 Precision: 12-bit	54.77	0.05
L1 Buffer	Size: 128*128	0.05/bit	0.043
Output Buffer	Size: 32*54	0.01/bit	0.003
PE Total		8348.60	0.13
Tile Level			
PE	Size: 3x3	74643.93	1.17
Adder Tree	Number: 64 Precision: 16-bit	485.14	0.08
L2 Buffer	Size:256*512	0.10/bit	0.3
Output Buffer	Size: 16*32*9	0.014/bit	0.007
Tile Total		75129.10	1.60
Layer Level			
Shift Add	Number:64 Precision:17-bit	208.57	3.20
Chip Level			
ReLU+Find Max		45.40	0.006
Digital Gradient Calculation and Process		0.022	18.8
Global Buffer	Size:12*1024*1024	0.4/bit	32.90

Hardware Performance Benchmarking





(a). Breakdown of chip area. (b). Breakdown of energy consumption.

- Two baselines(8-bit):
  - (1) without transpose subarray design and MSB/LSB splitting;
  - (2) with transpose subarray but without MSB/LSB splitting;
- Total area is reduced to only 31.9% of the Baseline 1
- The energy saving of MINT in FF & BP is 4× compared with Baselines

Outline

- Background and Motivation
- Weight Mapping Strategy for Training
  - Transpose Weight Matrix
  - Mapping Dataflow for Training in CIM
- Proposed MINT Architecture
  - RRAM Subarray Design
  - Overall Architecture
- Evaluation
  - Impact of ADC quantization and RRAM Non-idealities
  - Hardware Performance Benchmarking
- Conclusion

Conclusion

- Proposed mixed-precision RRAM-based in-memory training architecture, namely *MINT*, supporting DNN training
  - Transpose RRAM crossbar design
  - Splitting MSB/LSB to reduce hardware overhead (in particular ADCs)
- Evaluate the impact of ADC quantization and RRAM non-idealities
- Architecture-level performance
  - Achieving 4.46 TOPs/W, which shows great advantage compared to GPU and digital ASIC designs
  - The area of MINT is only ~30% of the prior CIM designs.
  - On-chip buffer capacity is the limitation

Acknowledgment

- This work is supported by NSF, and ASCENT, one of the SRC/DARPA JUMP Centers.

Questions?