CICC

IEEE Custom Integrated Circuits Conference

An In-Memory-Computing Charge-Domain Ternary CNN Classifier （Best Student Paper Candidate）

Xiangxing Yang[1], Keren Zhu[1], Xiyuan Tang[1], Meizhi Wang[1], Mingtao Zhan[2], Nanshu Lu[1], Jaydeep P. Kulkarni[1], David Z. Pan[1], Yongpan Liu[2], Nan Sun[1,2]

[1]University of Texas at Austin, Austin, TX

[2]Tsinghua University, Beijing, China

04/26/2021

IEEE SOLID-STATE CIRCUITS SOCIETY

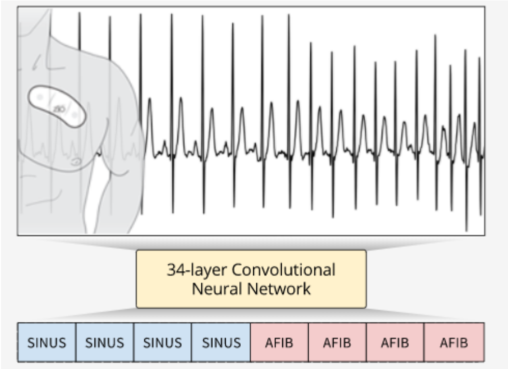
IEEE

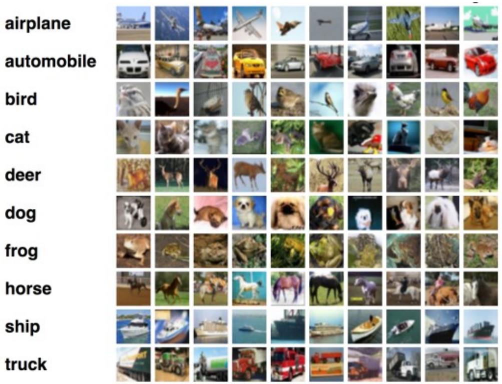
Outline

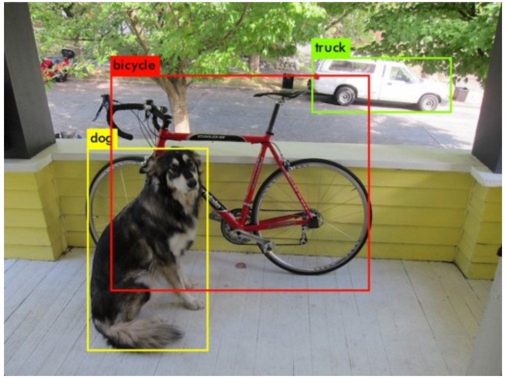
- **Motivations**
- Existing Works
- Theoretical Concept of the Proposed Work
- Circuit Implementation
- Measurement results
- Summary

Quest for Energy Efficient Edge Computing System

Increasing need from various applications:





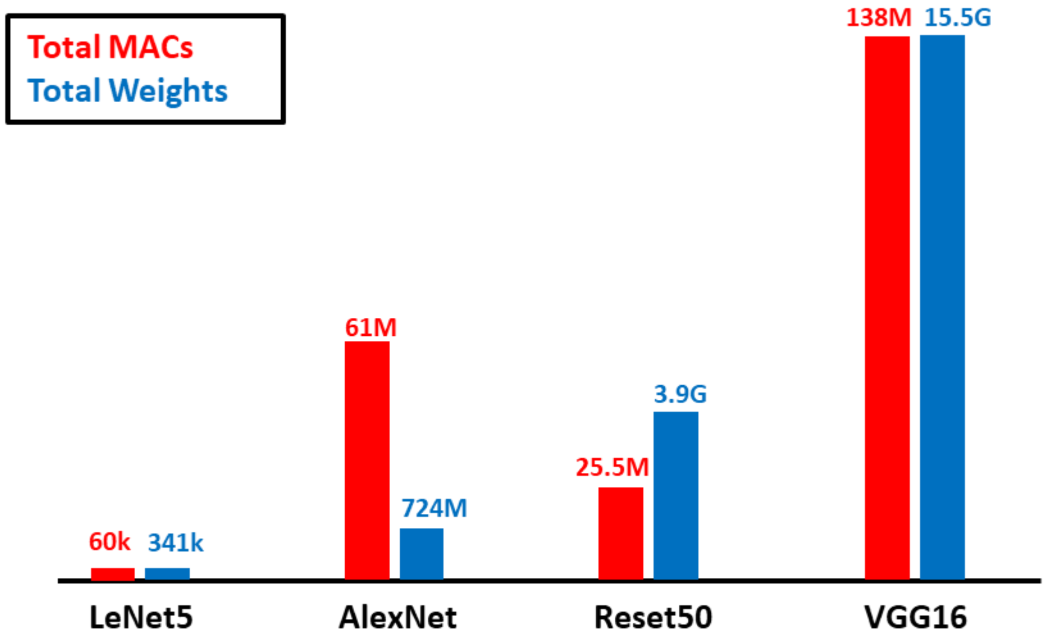


Pattern Recognition Image Classification Object recognition

Challenges on Energy Efficient NN Inference

- High computation energy

- High memory access energy



[V. Sze, Proceedings of the IEEE 105.12 2017]

Challenges on Energy Efficient NN Inference

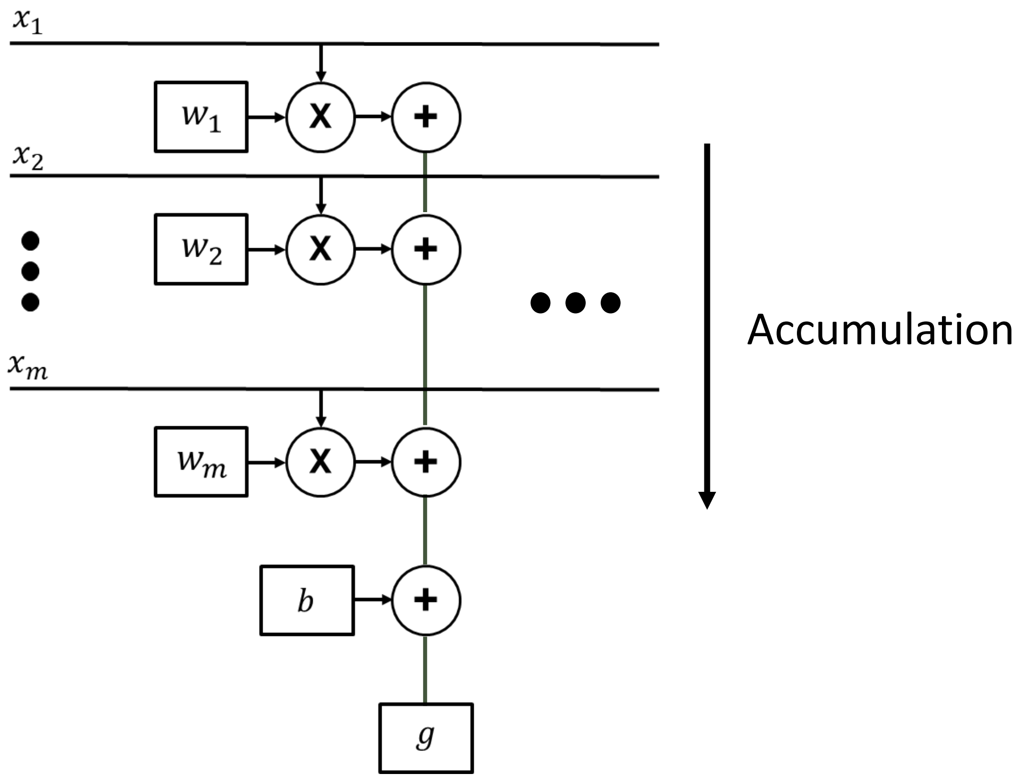
$$h = g\left[\left(\sum_{i=1}^{m} w_i * x_i \quad + b\right)\right]$$

x_i : Input activation

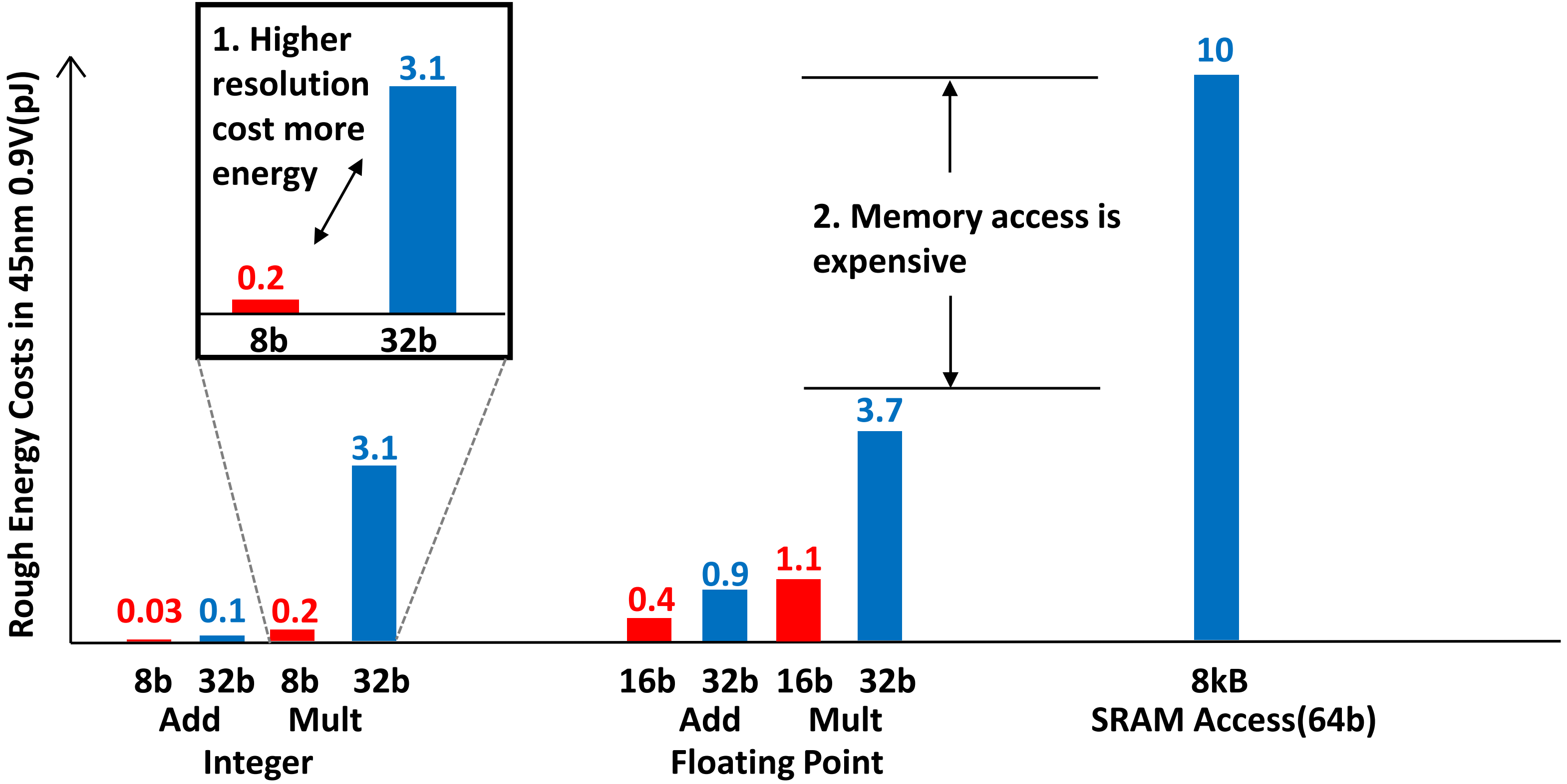
w_i : Weight b: Bias

h: Output to next layer

g: Activation function



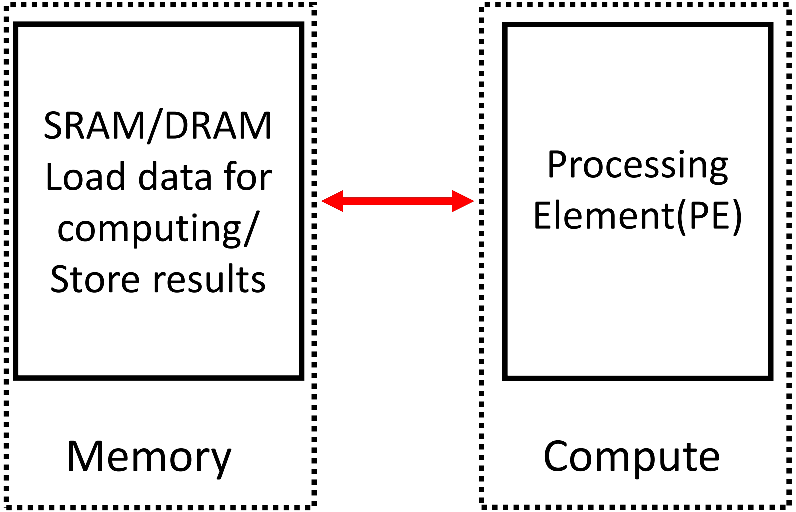
Challenges on Energy Efficient NN Inference



[M. Horowitz, ISSCC 2014]

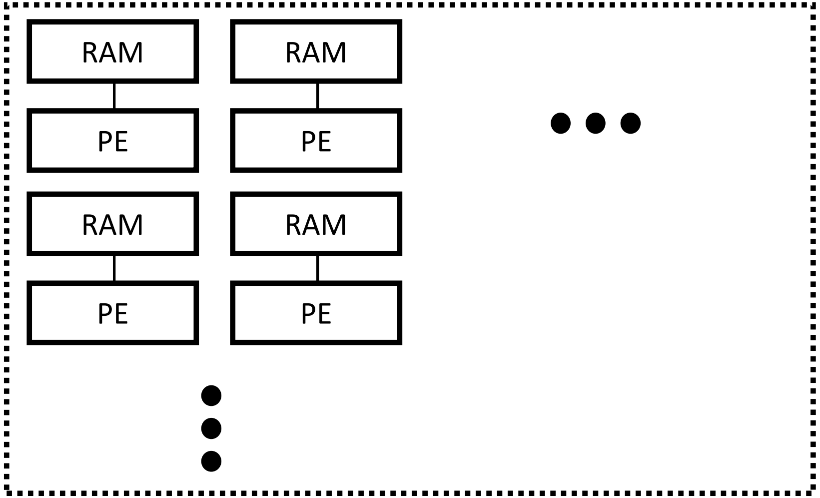
Solutions to Energy Efficient NN Inference

- Conventional computing:



Memory access can easily dominate energy/throughput

- In-memory-computing:

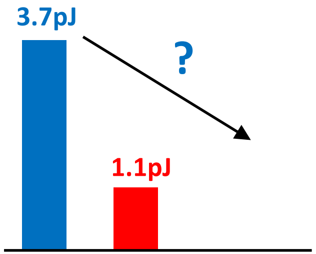


Minimized data movement from distributed memory

Solutions to Energy Efficient NN Inference

- Reduced Resolution Network:

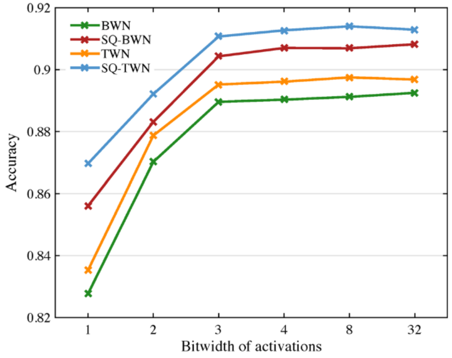
32b Floating point → ?



Multiplying energy cost

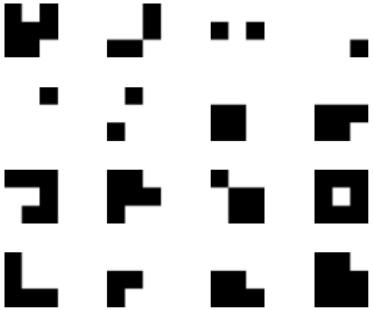
Solutions to Energy Efficient NN Inference

- Reduced Resolution Network:



CIFAR-10, ResNet-56
Activations are quantized to
1/2/3/4/8/32b

[Y. Dong, IJCV 2019]



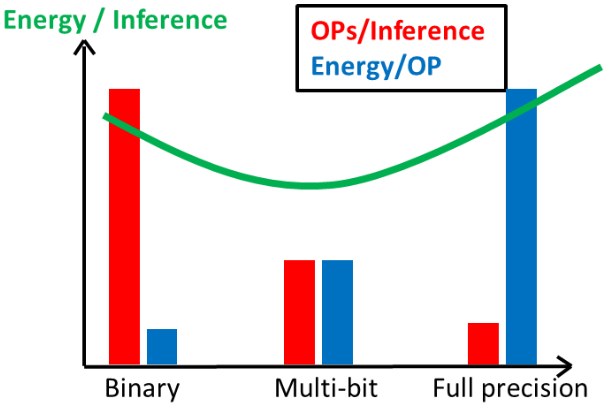
Visualization of filters from binary neural network

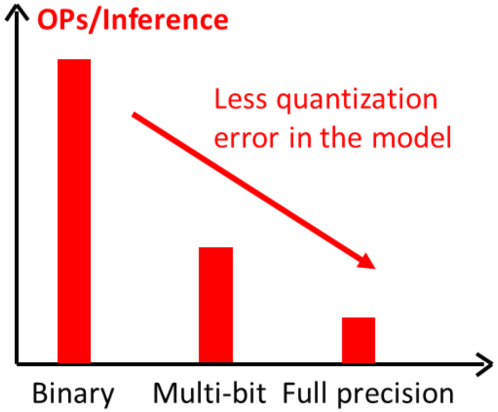
[M. Courbariaux, arXiv 2016]

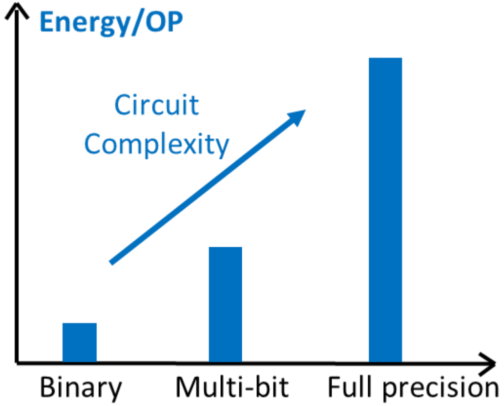
Energy Cost of NN Inference

$$Power = Rate \times \frac{Energy}{Inference} = Rate \times \frac{Operations}{Inference} \times \frac{Energy}{Operation}$$

[B. Murmann, ISSCC 19 Tutorial]





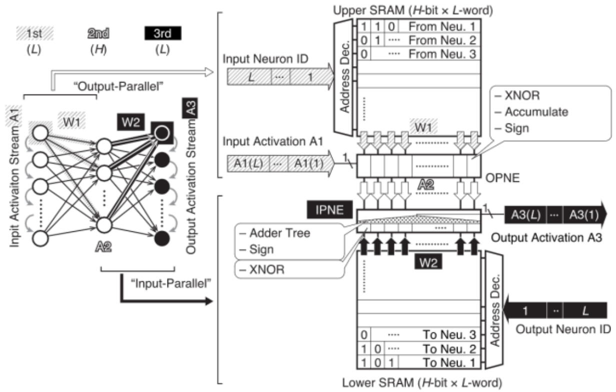


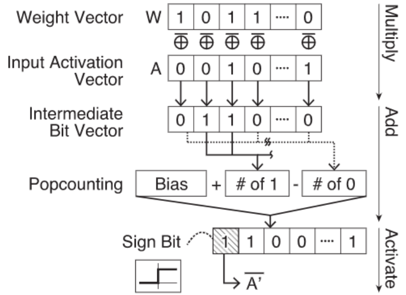
Outline

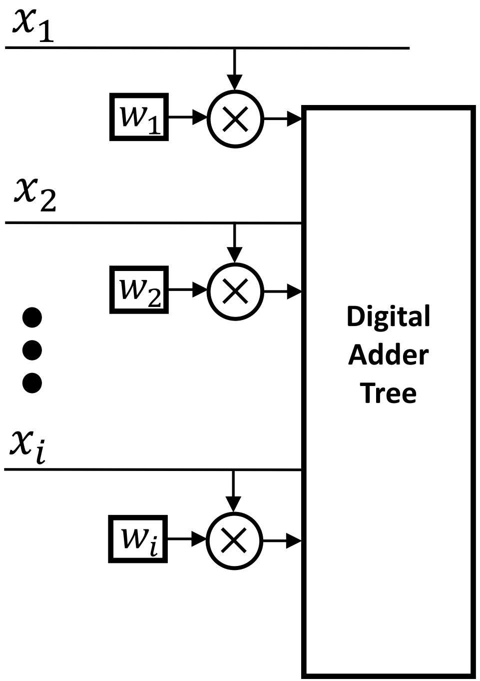
- Motivations
- **Existing Works**
- Theoretical Concept of the Proposed Work
- Circuit Implementation
- Measurement results
- Summary

Existing works

- **Digital Domain:**
  - **Bit error free ☺**
  - **High power from digital adder tree ☹**
  - **Low throughput ☹**



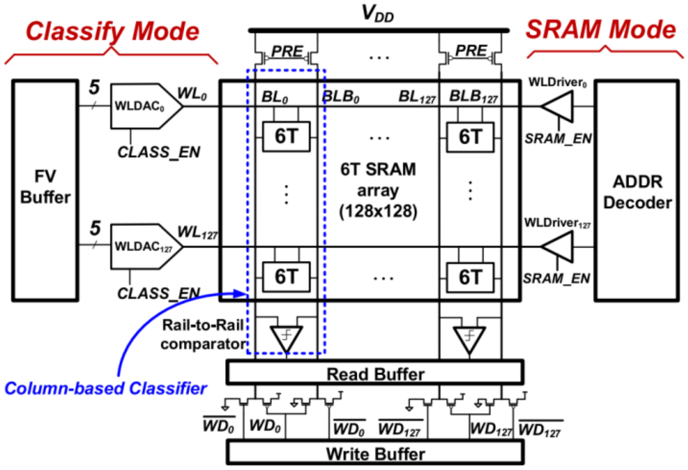


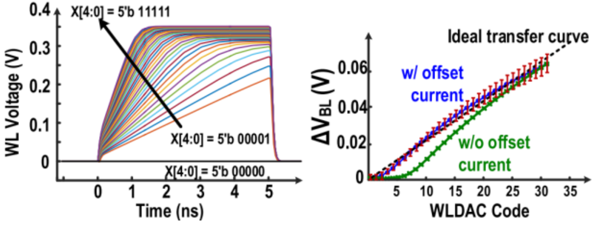


[K. Ando, JSSC 18]

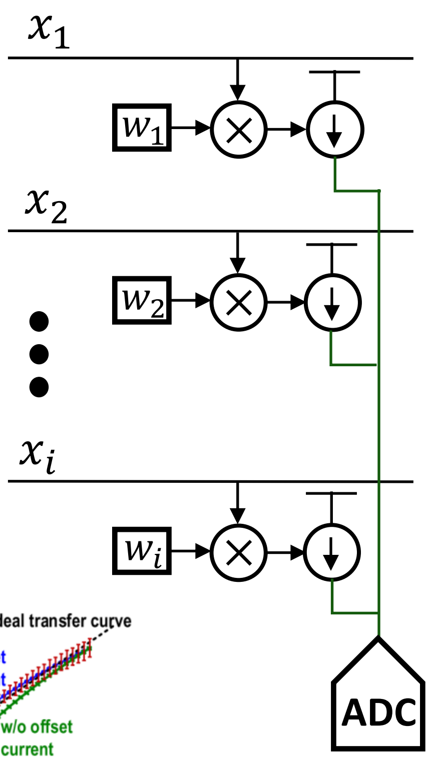
Existing works

- **Current Domain:**
  - High throughput ☺
  - PVT-robustness ☹
  - Consumes static current ☹



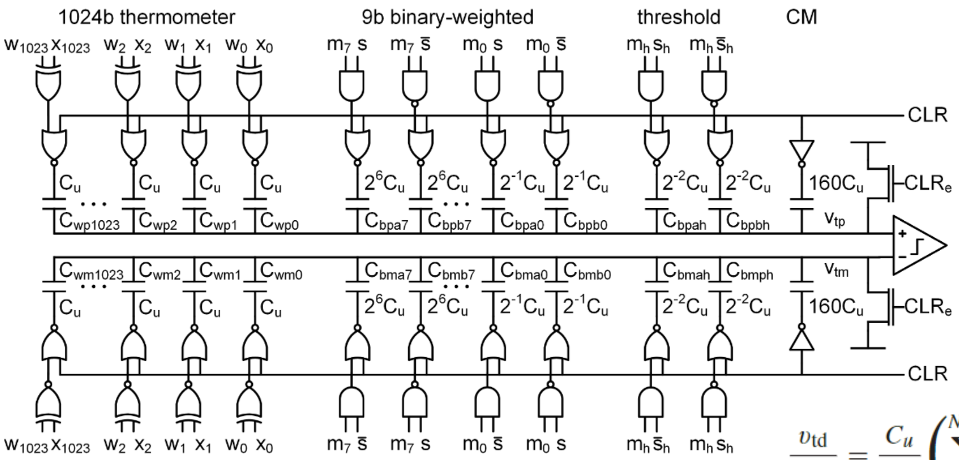


[J. Zhang, JSSC 17]



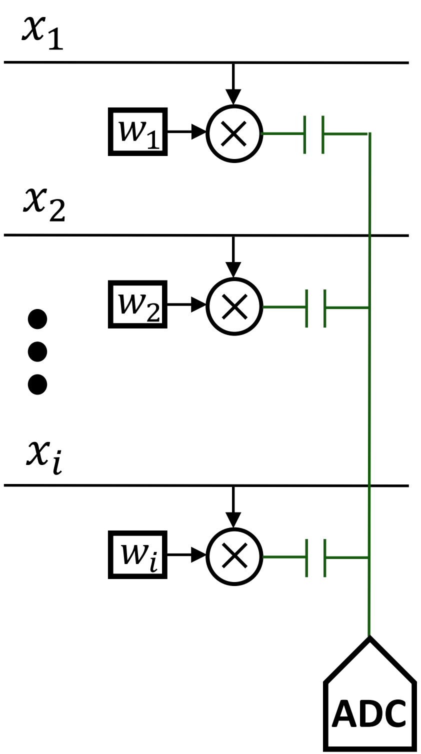
Existing works

- **Charge Domain:**
  - **High throughput ☺**
  - **No static current ☺**
  - **Large operations/inference ☹**



$$\frac{v_{\mathrm{td}}}{V_{\mathrm{DD}}} = \frac{C_u}{C_{\mathrm{tot}}}\left(\sum_{i=0}^{N-1} w_i x_i + b\right).$$

[D. Bankman, JSSC 18]

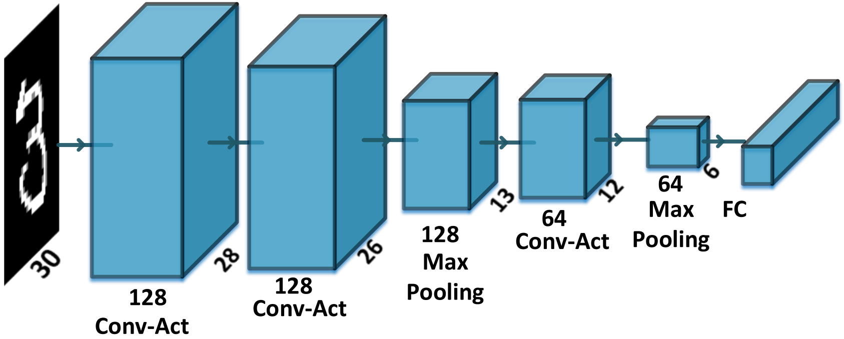


Outline

- Motivations
- Existing Works
- **Theoretical Concept of the Proposed Work**
- Circuit Implementation
- Measurement results
- Summary

Comparison of Model Size

Baseline test: 98% Accuracy on MNIST

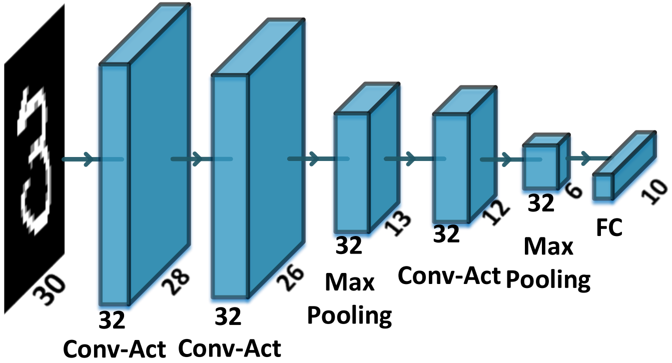


Layer	Type	Size	Channel	Filter Size
1	CONV-TN	30x30	1(input)	2x2
2	CONV-TN	28x28	128	
2p	MAX POOL	26x26		
3	CONV-TN	13x13	64	
3p	MAX POOL	12x12		
4	FC	(Flatten 6x6x64) 2304 - 10		

1b Resolution

1.38×10^8 OPs

~4x Bigger model size



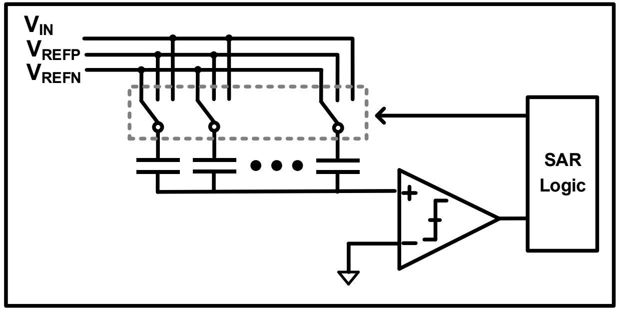
Layer	Type	Size	Channel	Filter Size
1	CONV-TN	30x30	1(input)	2x2
2	CONV-TN	28x28	32	
2p	MAX POOL	26x26		
3	CONV-TN	13x13		
3p	MAX POOL	12x12		
4	FC	(Flatten 6x6x32) 1152 - 10		

1.5b Resolution

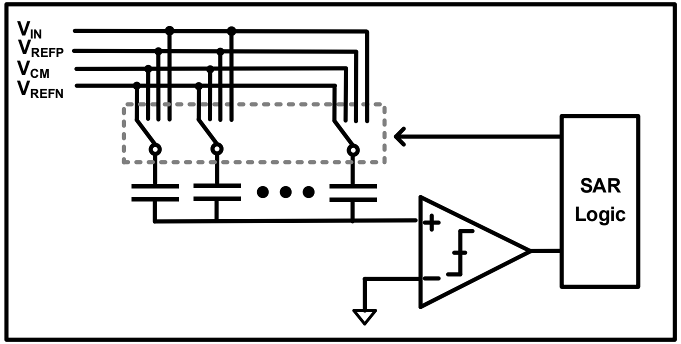
3.57×10^7 OPs

{w,x from -1,0,1}

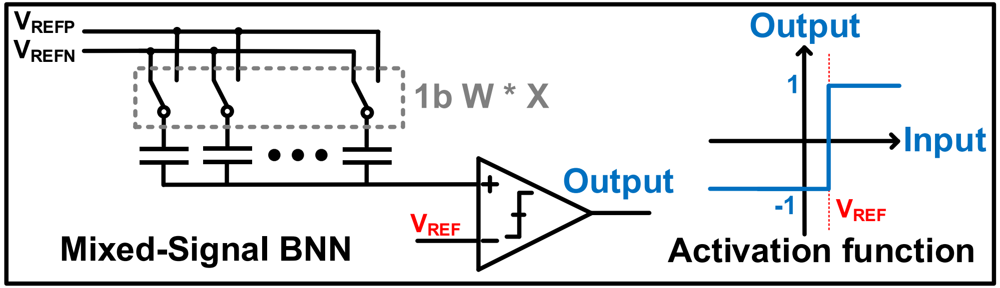
Mixed Signal BNN vs TNN

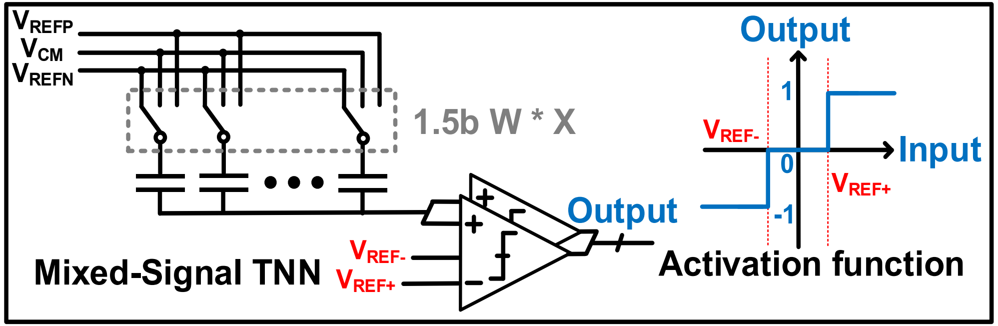


SAR ADC

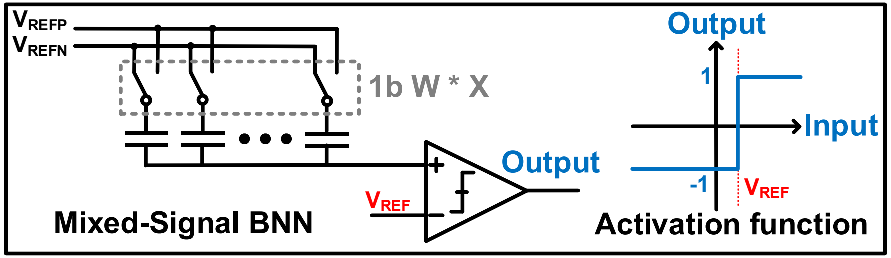


SAR ADC with V_{CM} based switching

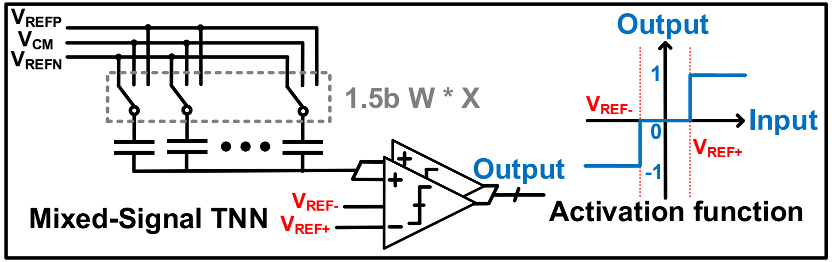




Mixed Signal BNN vs TNN



VS



	Hardware Complexity	Operations / Inference (@same accuracy)	X	Energy / Operation (CDAC signal swing)	=	Energy / Inference
BNN	😊 😊	😐		😐		😐
TNN	😊	😊		😊		😊

OPs/Inference ↓ 75%

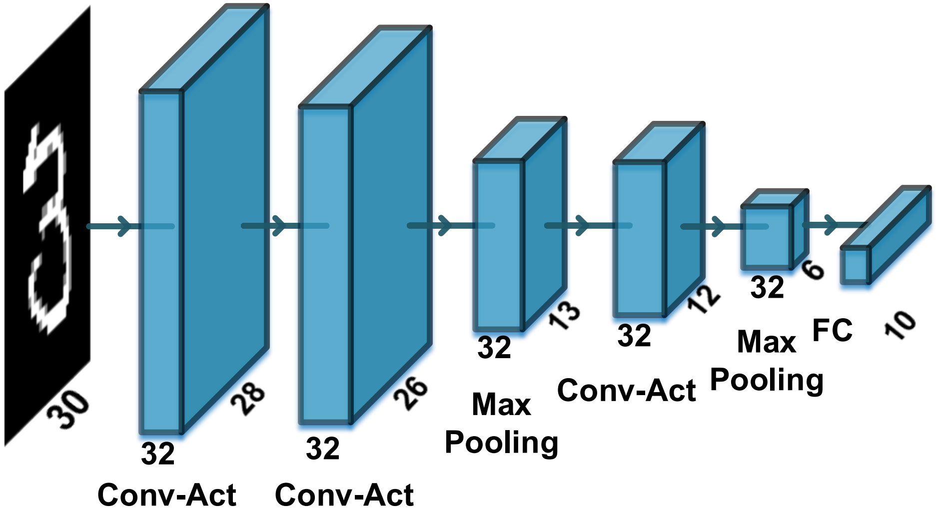
Energy/Operation ↓ 31%

Energy/Inference ↓ 82%

Outline

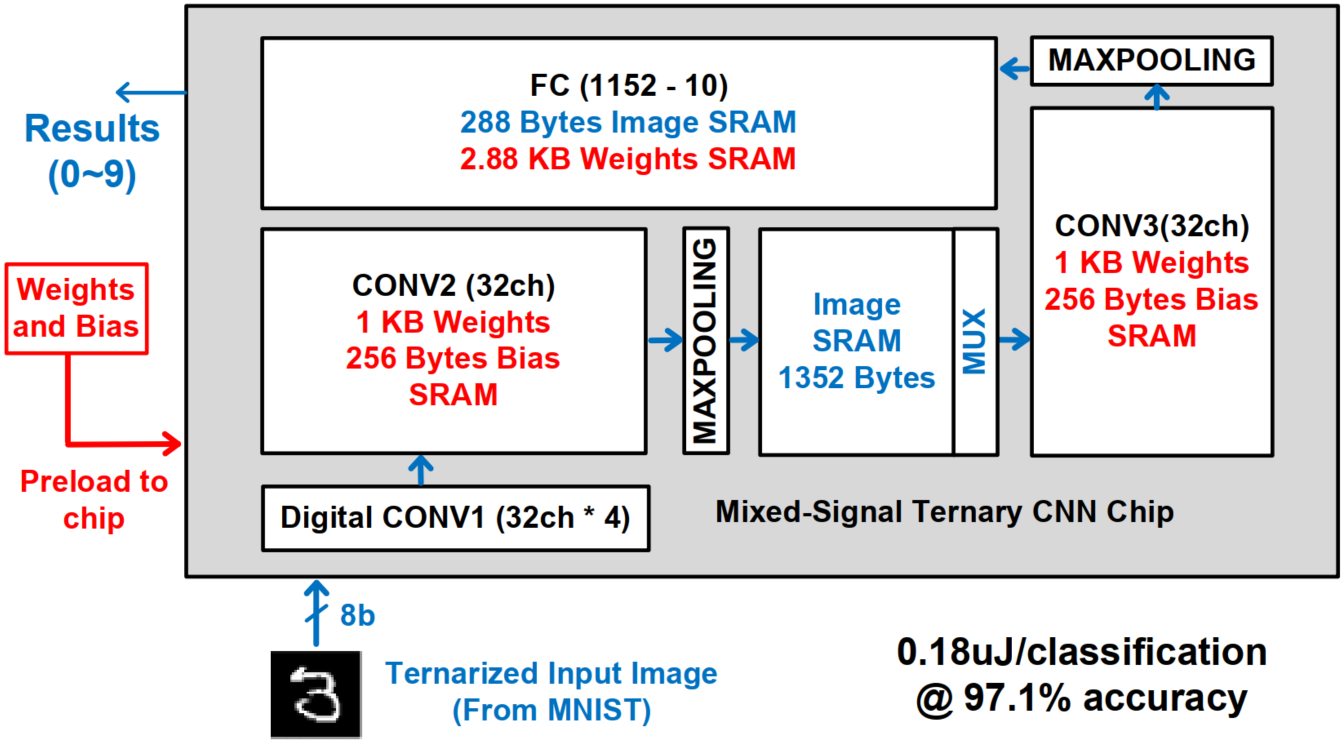
- Motivations
- Existing Works
- Theoretical Concept of the Proposed Work
- **Circuit Implementation**
- Measurement results
- Summary

On-chip Neural Network Model



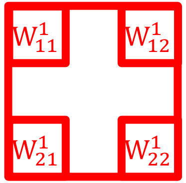
Layer	Type	Size	Channel	Filter Size	Dilated
1	CONV-TN	30x30	1(input)	2x2	2
2	CONV-TN	28x28	32		2
2p	MAX POOL	26x26			1
3	CONV-TN	13x13			1
3p	MAX POOL	12x12			1
4	FC	(Flatten 6x6x32) 1152 - 10			

Chip Architecture

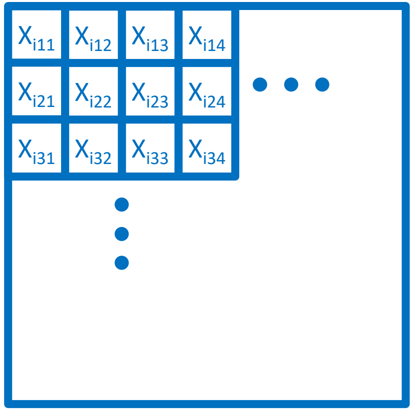


0.18uJ/classification
@ 97.1% accuracy

CONV1 – Example of One-Channel Convolution



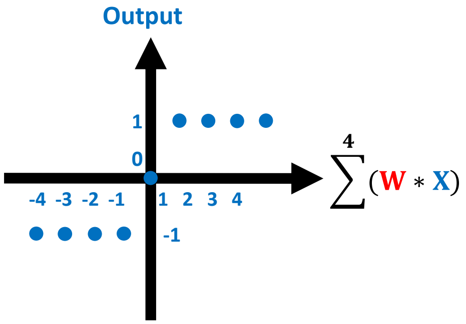
Filter0 2x2
Dilated L = 2



Ternarized
Input Image
1ch

$$X_{O11} = \text{STEP}(W_{11} * X_{i11} + W_{12} * X_{i13} + W_{21} * X_{i31} + W_{22} * X_{i33})$$

$$X_{O12} = \text{STEP}(W_{11} * X_{i12} + W_{12} * X_{i14} + W_{21} * X_{i32} + W_{22} * X_{i34})$$

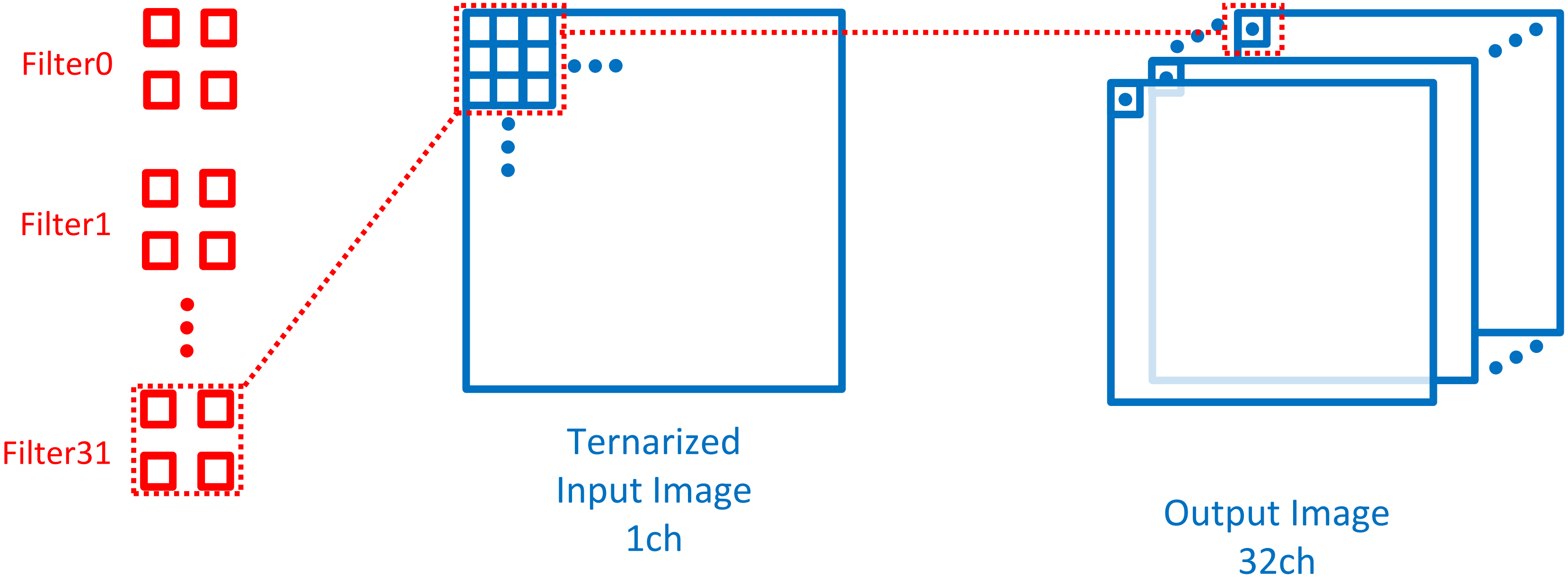


Step Activation Function
of CONV1

Output Image
1ch

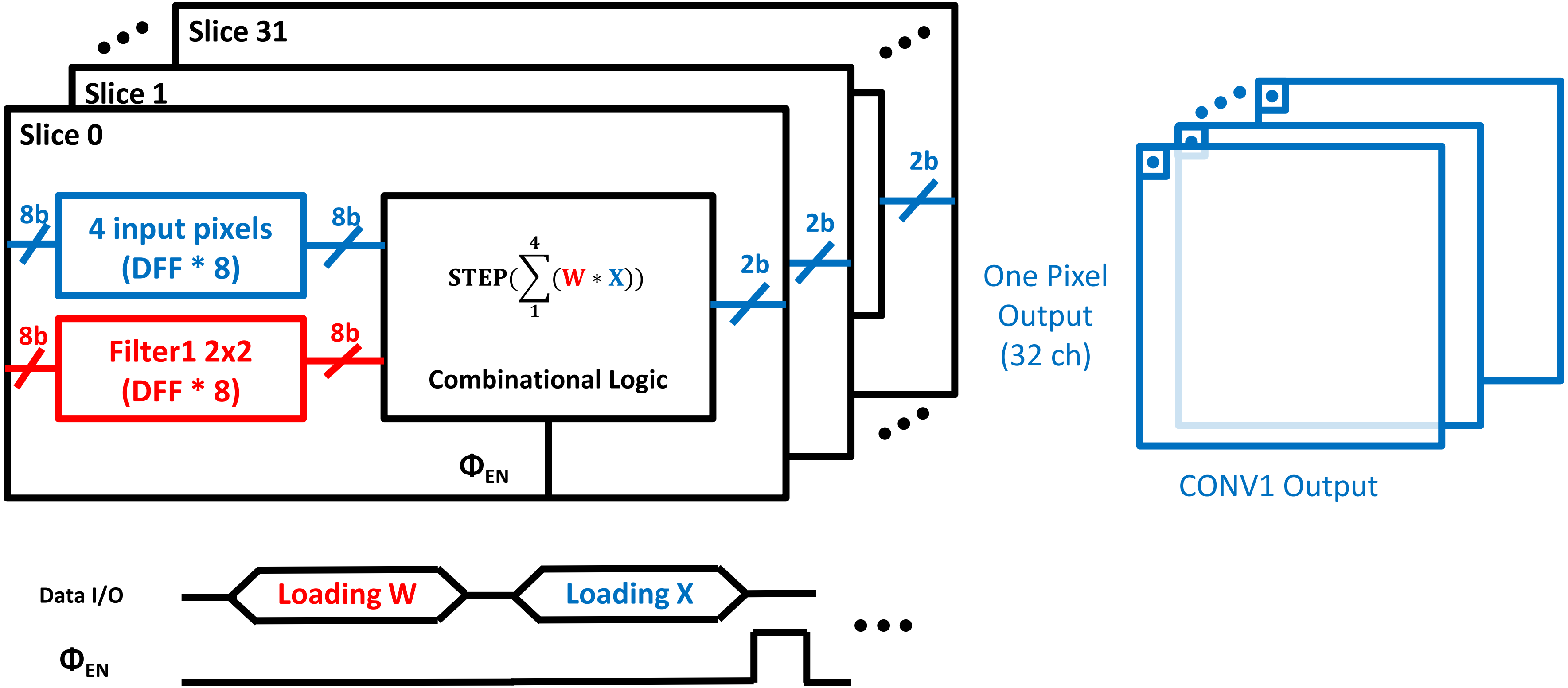
$W, X \in \{-1, 0, 1\}$

CONV1 – Example of 32-Channel Convolution

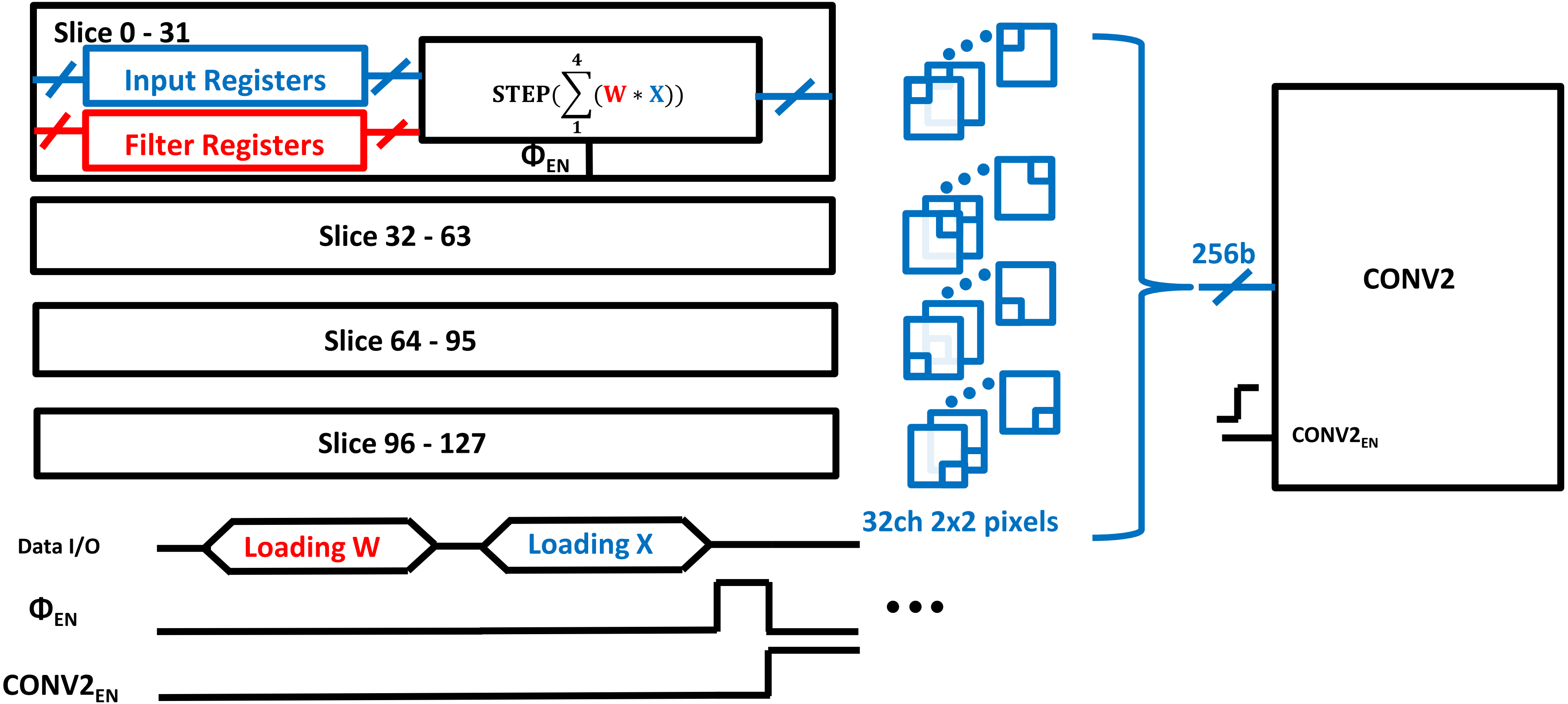


W,X $\in$ **{-1,0,1}**

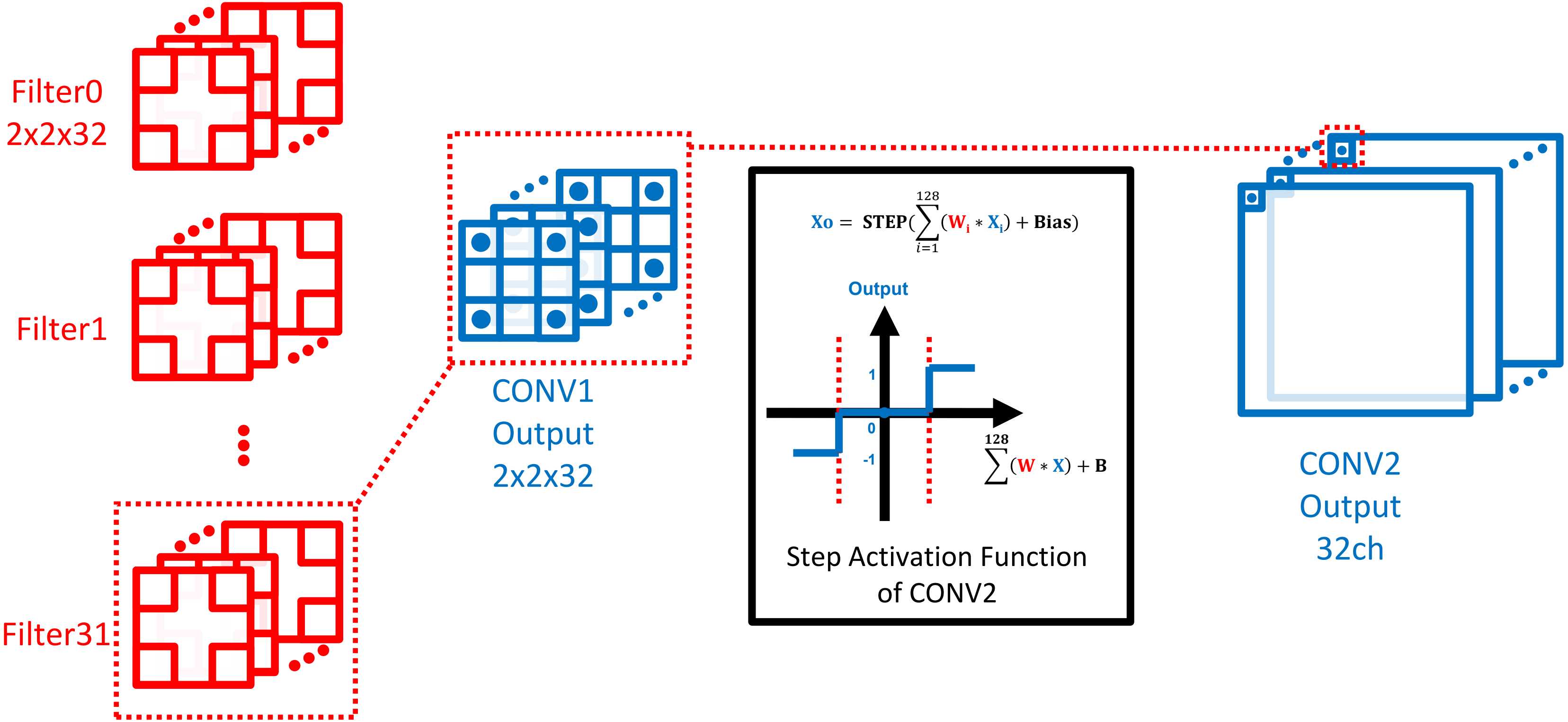
CONV1 – Digital Implementation



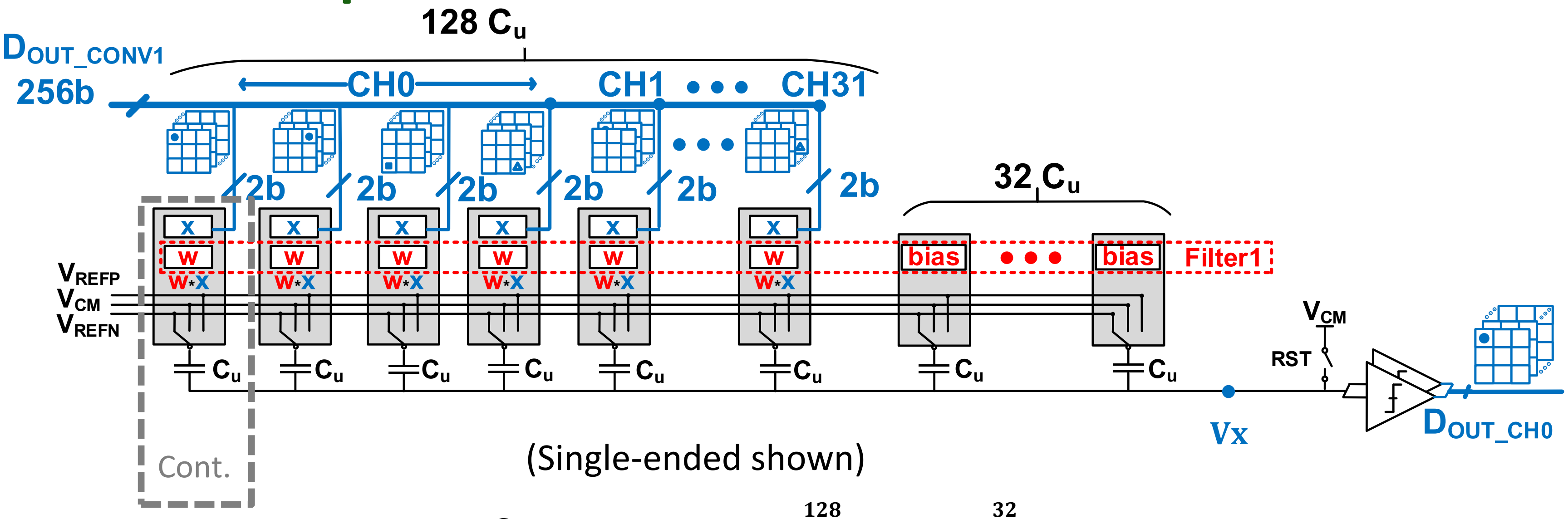
CONV1 – Digital Implementation



CONV2 – Example of 32-Channel Convolution



CONV2 – Implementation of One-Channel SC Neuron



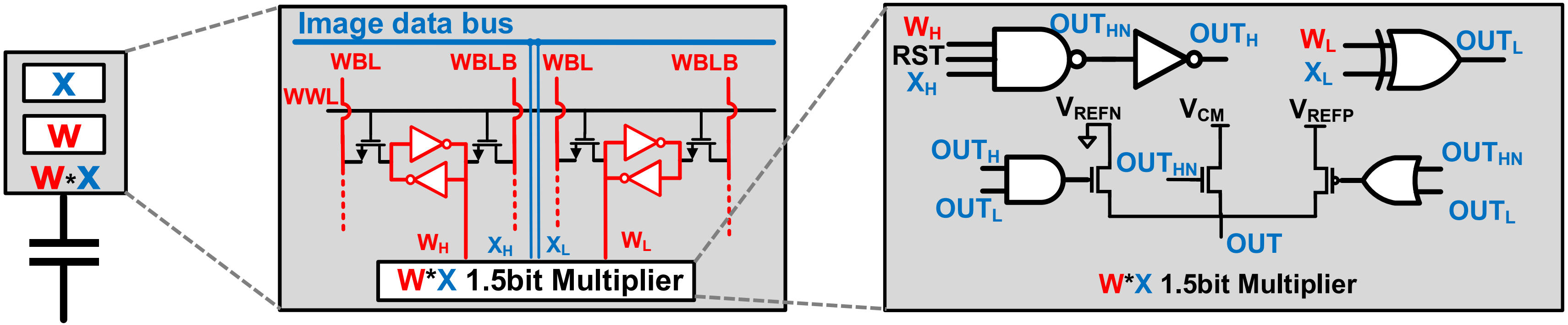
(Single-ended shown)

$$Vx = \frac{C_u}{C_{Total}} * (V_{REFP} - V_{REFN}) * (\sum_{i=1}^{128} (W_i * X_i) + \sum_{i=1}^{32} Bias_i)$$

$$C_{Total} \approx 160\ C_u$$

$$(W_i * X_i),\ Bias \in \{ V_{REFP}, V_{CM}, V_{REFN} \}$$

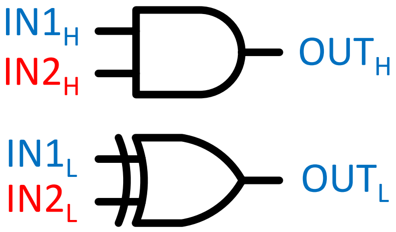
CONV2 – Synapse Design



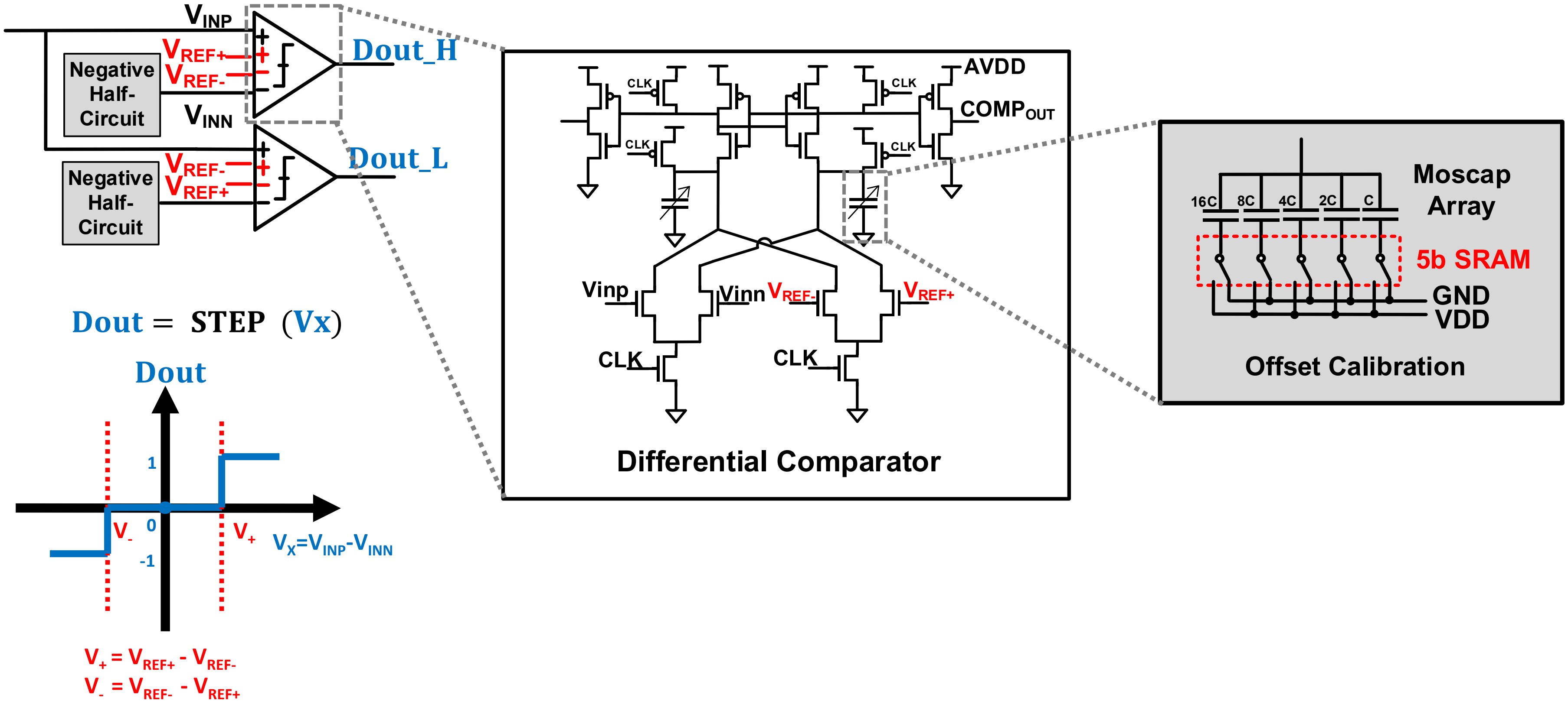
DEC	BIN	Voltage
1	10	V_{REFP}
-1	11	V_{REFN}
0	0X	V_{CM}

Encoding for simplicity:

1.5b Multiplier →

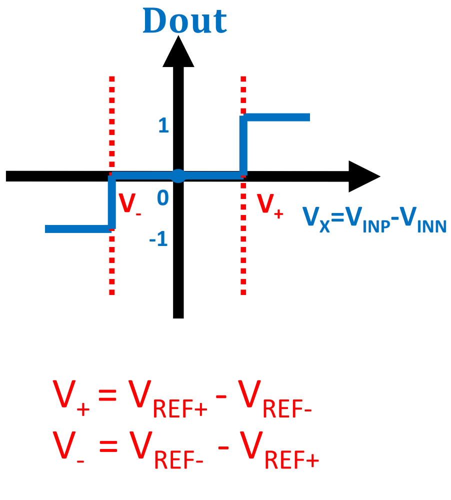


CONV2 – Comparator Design

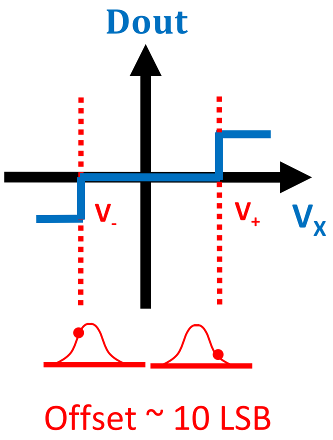


CONV2 – Effect of Comparator Offset

$\text{Dout} = \text{STEP}\ (\text{Vx})$

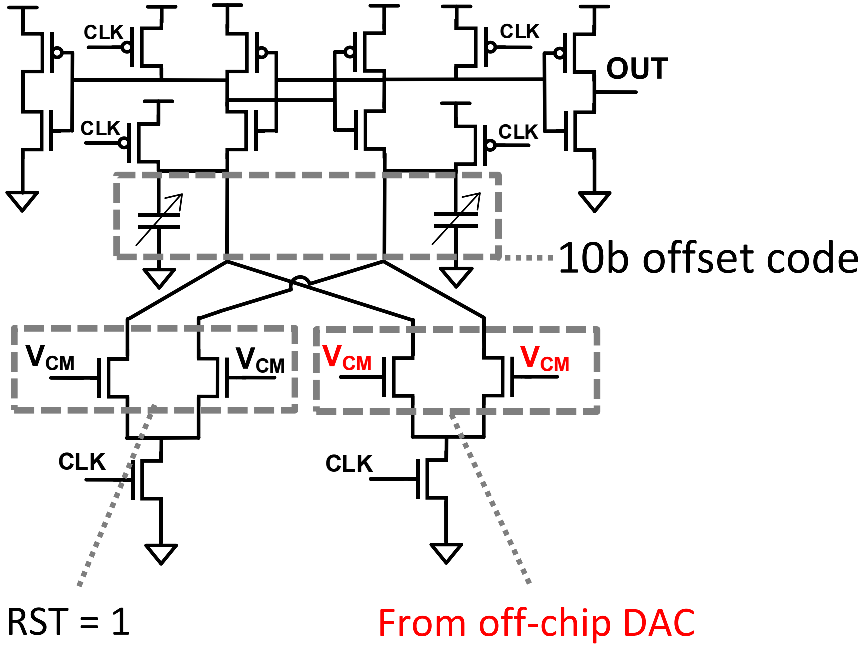


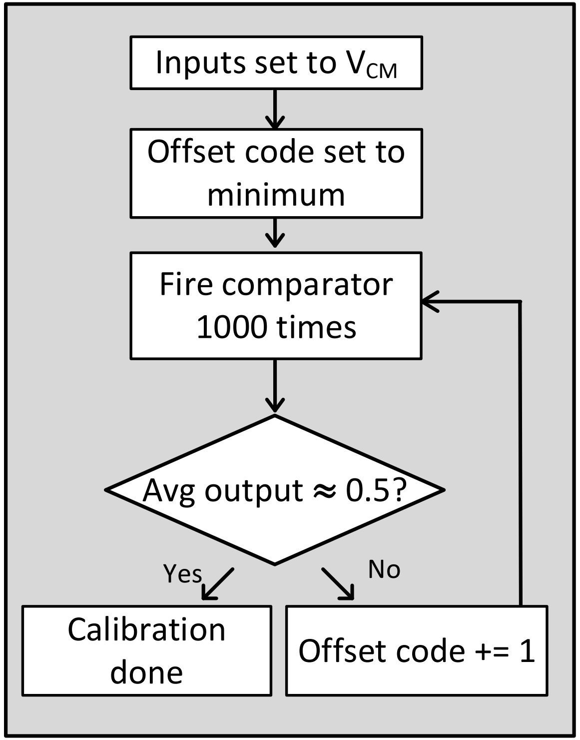
Ideal activation function



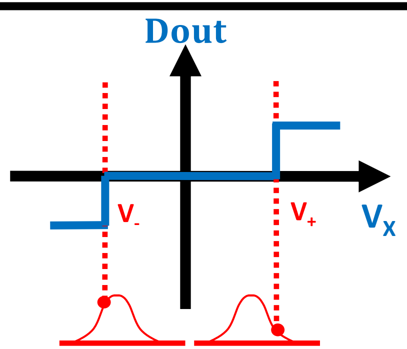
Actual activation function

CONV2 – Foreground Comparator Offset Calibration

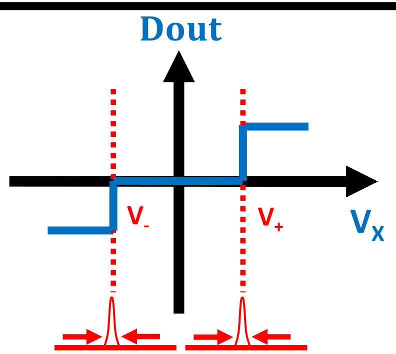




CONV2 – Foreground Comparator Offset Calibration

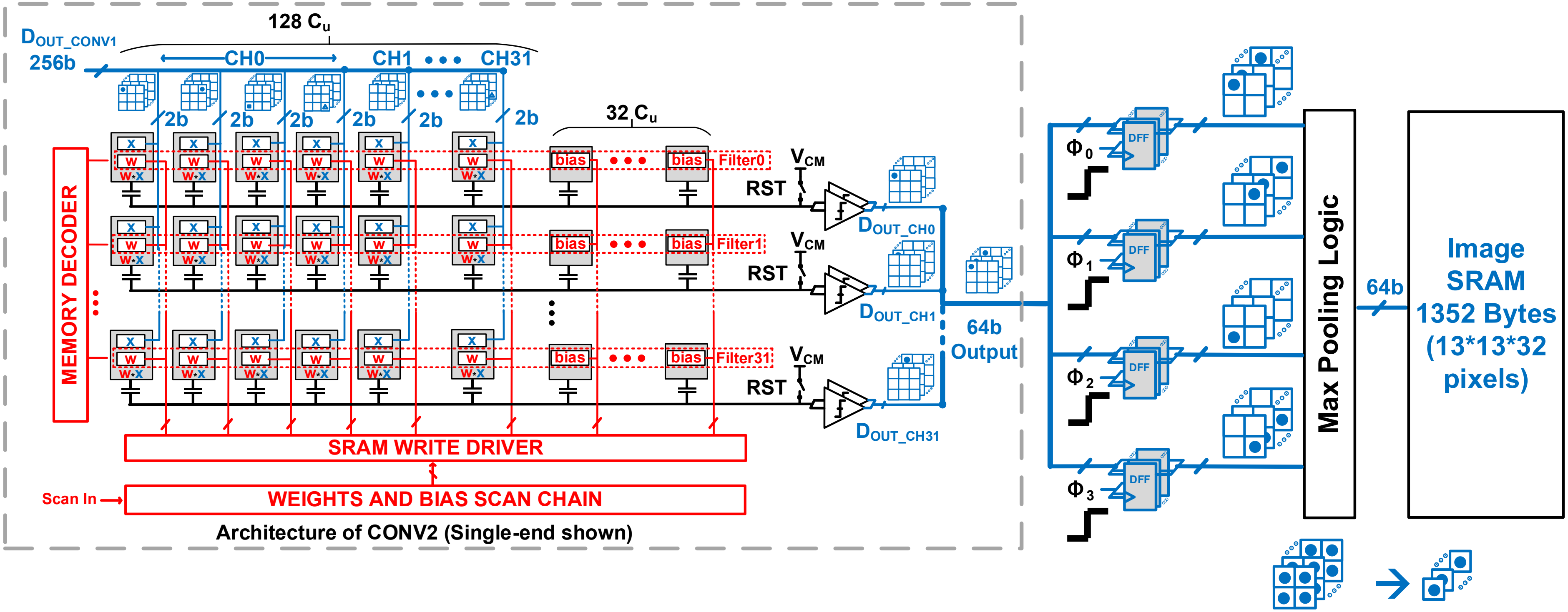


Offset ~ 10 LSB
w/o Offset Calibration



Offset < 1 LSB
w/ Offset Calibration

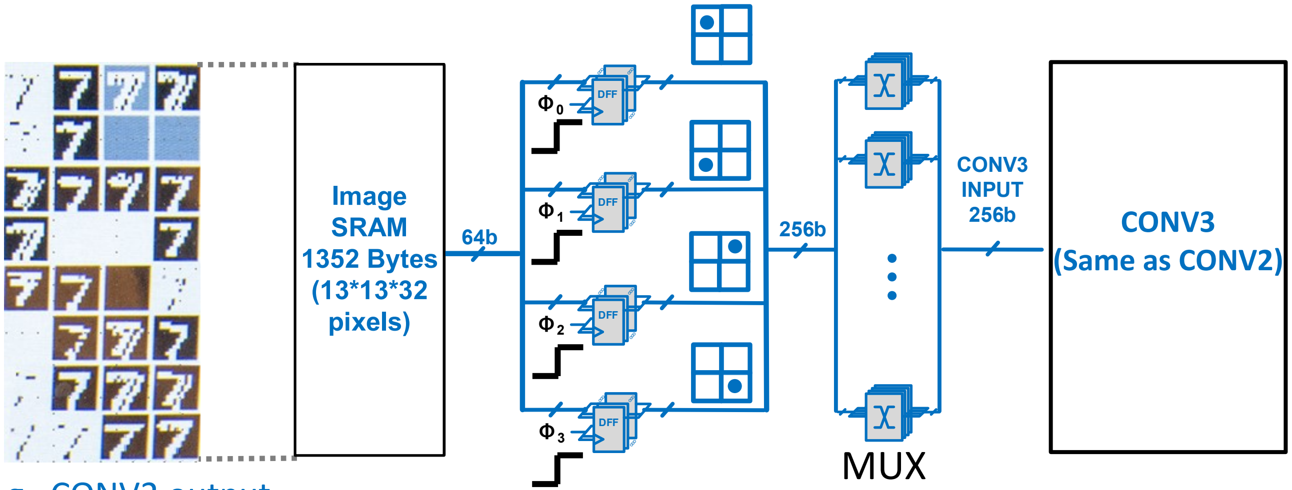
CONV2 – Maxpooling



Architecture of CONV2 (Single-end shown)

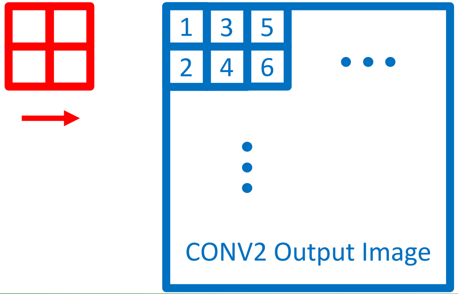
2x2x32 → 1x1x32

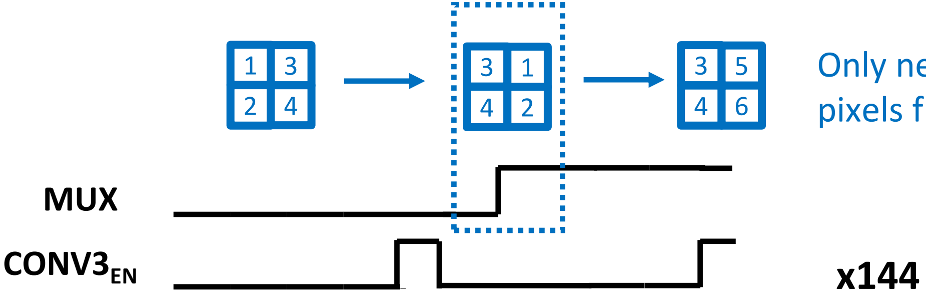
Datapath from CONV2 to CONV3



E.g. CONV2 output

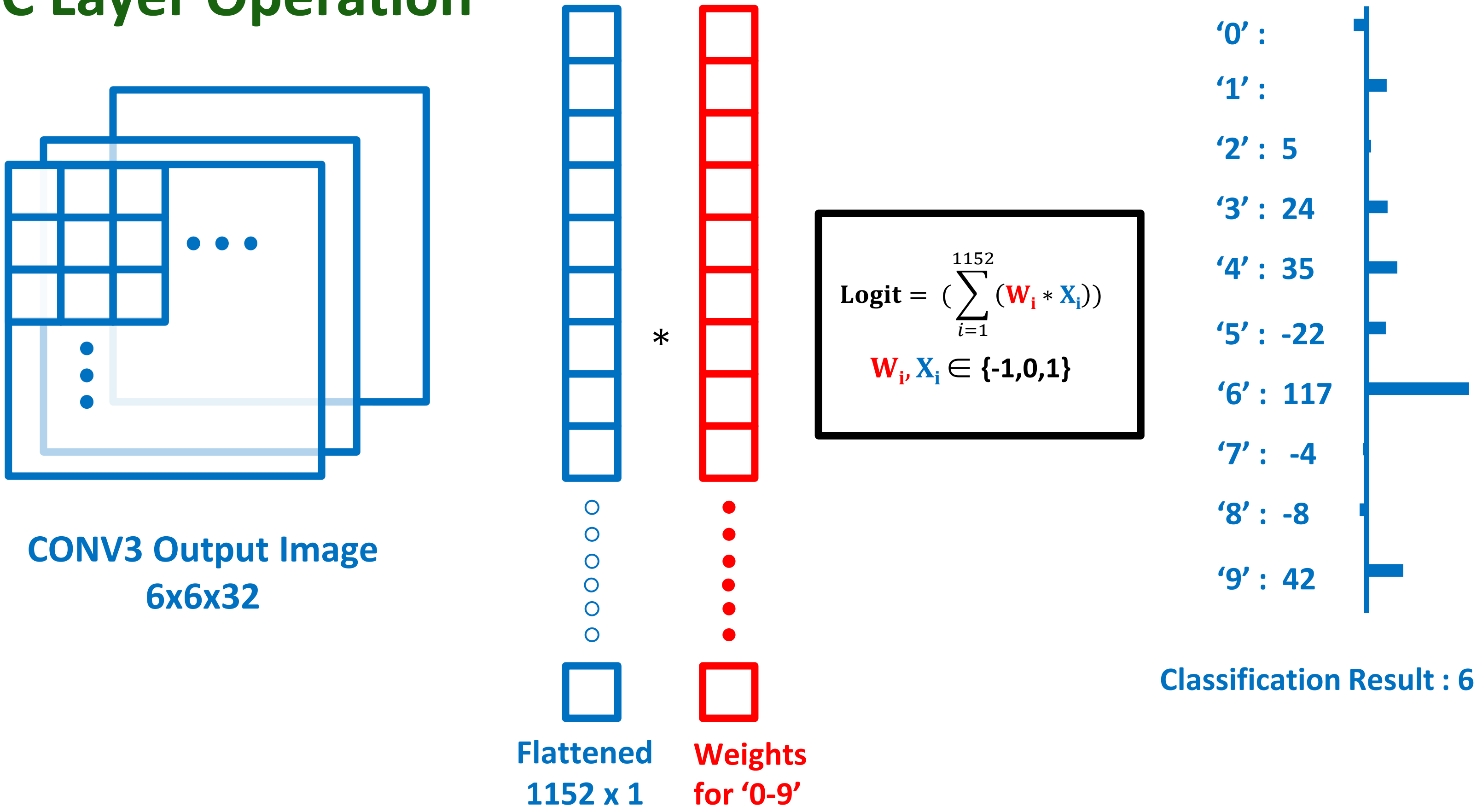
CONV3 Filter Window





Only need to read 2 pixels from SRAM

FC Layer Operation

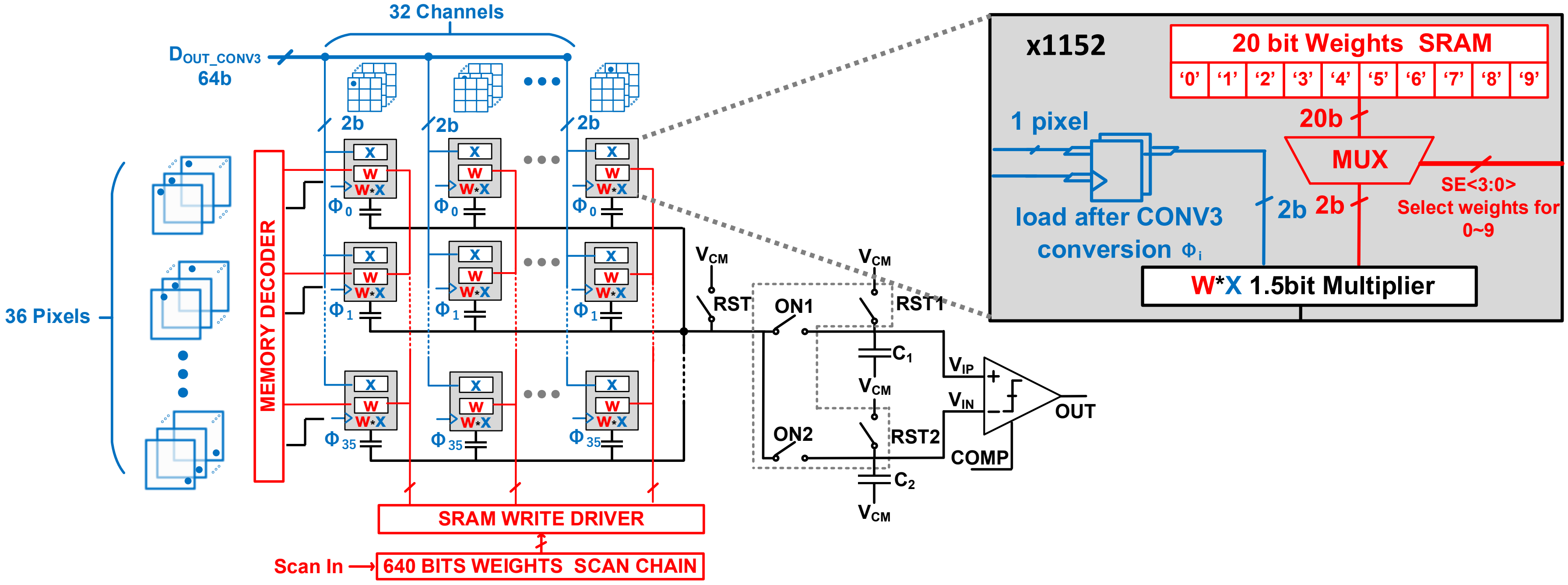


$$\text{Logit} = \left(\sum_{i=1}^{1152} (W_i * X_i)\right)$$

$$W_i, X_i \in \{-1,0,1\}$$

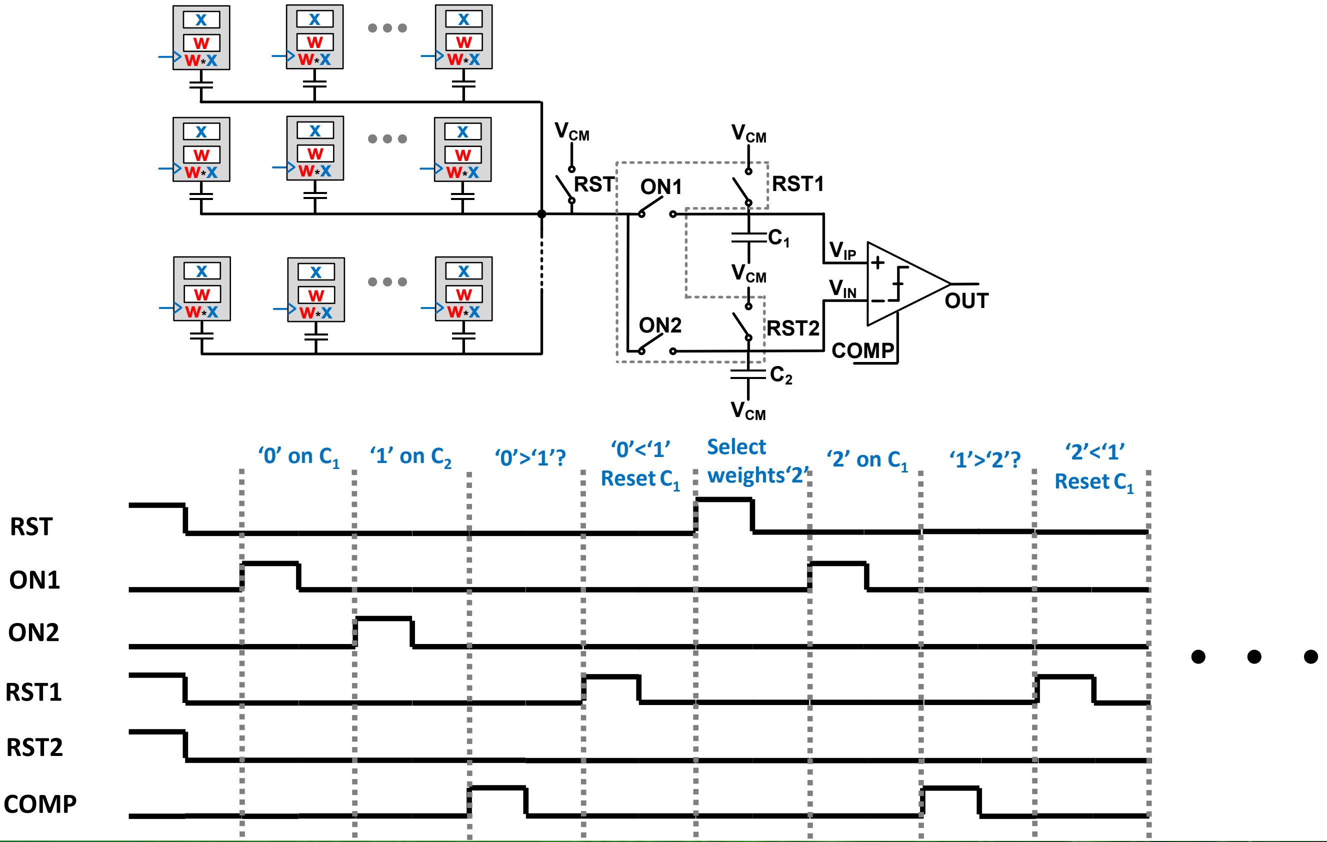
Classification Result : 6

FC Layer Implementation



(Single-ended shown)

FC Layer Implementation

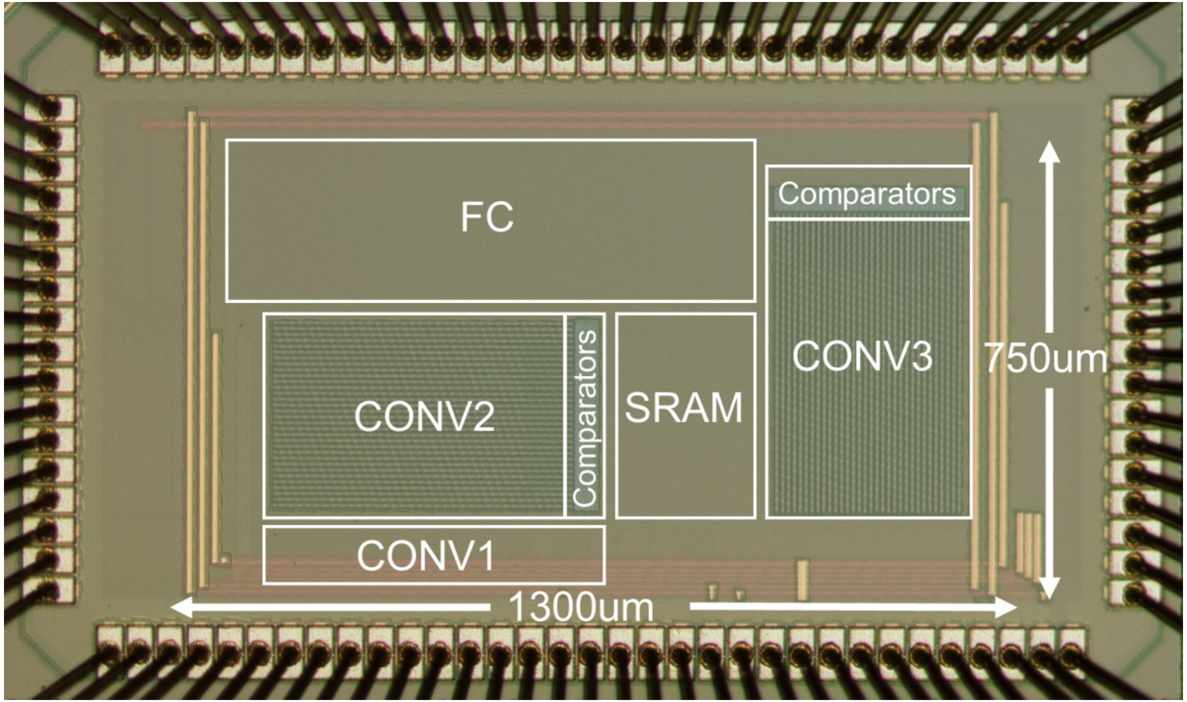


Outline

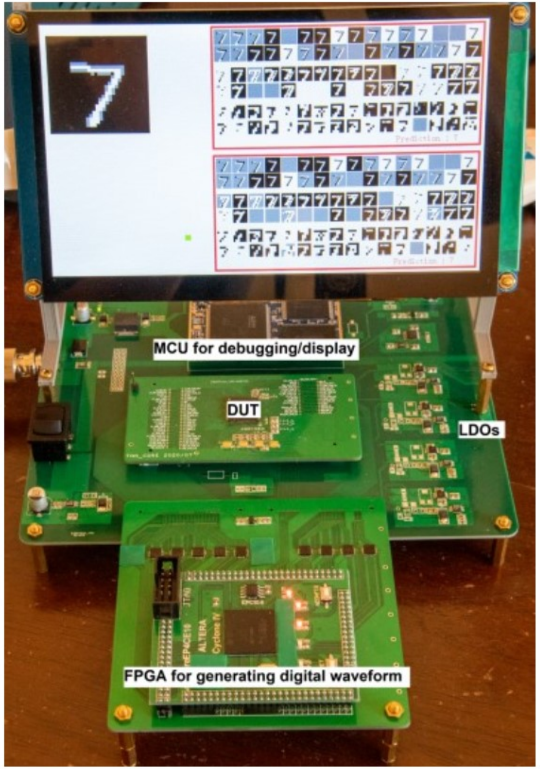
- Motivations
- Existing Works
- Theoretical Concept of the Proposed Work
- Circuit Implementation
- **Measurement results**
- Summary

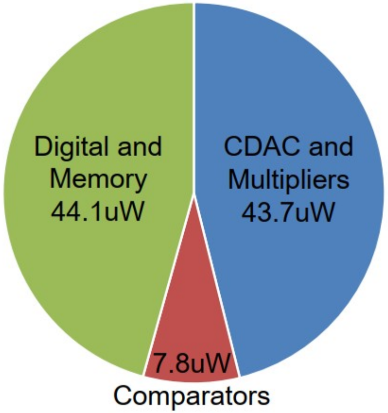
Die Photo

- **40nm LP CMOS**
- **Active Area: 0.98mm²**
- **Supply: 0.8V/0.7V/0.9V**



Measurement Results





97.1% accuracy @ 549 FPS

Power domain	Voltage	Energy
DVDD	0.7 V	44.1 uW
AVDD	0.8 V	7.8 uW
V_{REFP}	0.9 V	37.8 uW
V_{CM}	0.45 V	5.9 uW

Comparison table

	This work		JSSC'18 K. Ando [1]	ISSCC'18 D. Bankman [2]	JSSC'20 Y. Cheng [3]	CICC'20 C. Yu [4]	JSSC'19 H. Valavi [5]
Technology	**40nm**		65nm	28nm	55nm	65nm	65nm
Circuit Type	**Mixed-Signal Charge-domain**		Digital	Mixed-Signal Charge-domain	Mixed-Signal Current-domain	Mixed-Signal Current-domain	Mixed-Signal Charge-domain
Bit Precision	**1.5b**		1/1.5b	1b	1-8b	1-5b	1b
Area(mm2)	**0.98**		3.9	4.6	5.85	0.055	12.6
Area Eff.(GOPS/mm2)	**469[1]**		105	67	N/A	N/A	1498
Operating VDD(V)	**0.8/0.7/0.9**		0.55-1.0	0.8/0.8	0.9	0.8/0.45	0.94/0.68/1.2
Energy Eff.(TOPS/W)	**556[2]**		2.3-6.0	532	40.2	490-15.8	866
Dataset	**MNIST**		MNIST	CIFAR-10	MNIST	MNIST	MNIST
Accuracy	**97.1%[3]**		90.1%	86.05%	98.56%	96.2%	98.6%
FPS	**549**		N/A	237	N/A	N/A	651
Power(mW)	**0.096**		N/A	0.899	N/A	N/A	N/A
Operations / Inference	**TNN**	**BNN (simu)**	N/A	N/A	N/A	N/A	$5.3x10^8$
	$3.57x10^7$	**$1.38x10^8$**					
MACs Energy / Inference	**0.09uJ**	**0.52uJ**	N/A	N/A	N/A	N/A	0.8uJ
Total Energy / Inference	**0.18uJ**	**0.7uJ**	N/A	3.8uJ	N/A	N/A	N/A
All operations on chip	**Yes**		No	Yes	No	No	No

[1]Based on SC neuron
[2]Based on MACs energy efficiency
[3]10 runs average on 10,000 test set images.

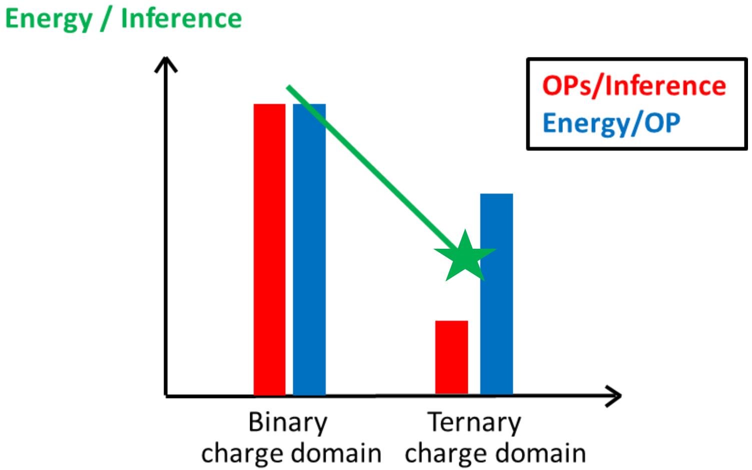
Outline

- Motivations
- Existing Works
- Theoretical Concept of the Proposed Work
- Circuit Implementation
- Measurement results
- **Summary**

Summary

- **A 1.5b charge domain ternary CNN classifier is proposed:**
  - Fully on-chip NN with lowest energy/inference reported for >97% MNIST accuracy
  - Compared to BNN with same accuracy:
    - 75% ↓ $\frac{Operations}{Inference}$
    - 31% ↓ $\frac{Energy}{Operation}$

82% ↓ $\frac{Energy}{Inference}$




THANK YOU
hvala
danke
asante
OBRIGADO
gracias
どうも
ARIGATO
DANK U
takk
MERSI
merci
謝謝
obrigado
danke schön
köszi
سپاس
PALDIES
muchas gracias
ありがとう
TEŞEKKÜR EDERIM
MOLTE GRAZIE
GO RAIBH MAITH AGAT
grazas
GRAZZI
THANKS
qujan
TAK
благодаря
TAK
どうも
vielen dank
grazie
DZIEKI
спасибо
MULTUMESC
PALDIES
OBRIGADO
mesi
TACK
Gràcies
多謝
شكراً
danke
DANK U
감사합니다
köszi
TEŞEKKÜR EDERIM
NA GODE
muchas gracias
obrigado