



IEEE Custom Integrated Circuits Conference

An In-Memory-Computing Charge-Domain Ternary CNN Classifier

Xiangxing Yang¹, Keren Zhu¹, Xiyuan Tang¹, Meizhi Wang¹, Mingtao Zhan², Nanshu Lu¹, Jaydeep P. Kulkarni¹, David Z. Pan¹, Yongpan Liu², Nan Sun^{1,2}

¹University of Texas at Austin, Austin, TX

²Tsinghua University, Beijing, China

04/26/2021



IEEE
SOLID-STATE
CIRCUITS SOCIETY™

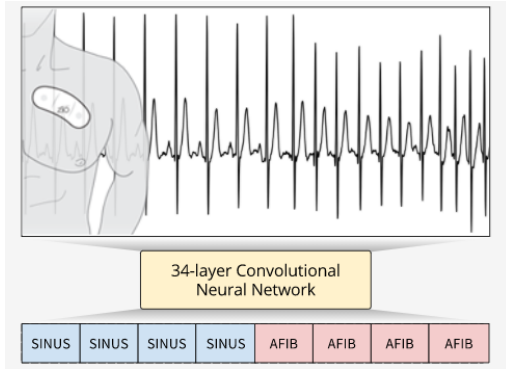


Outline

- **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

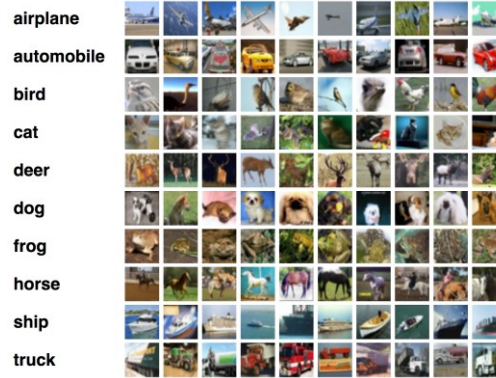
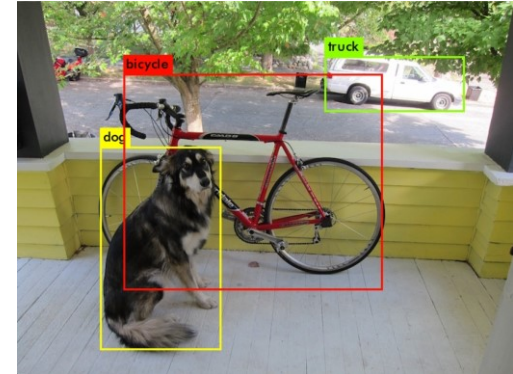


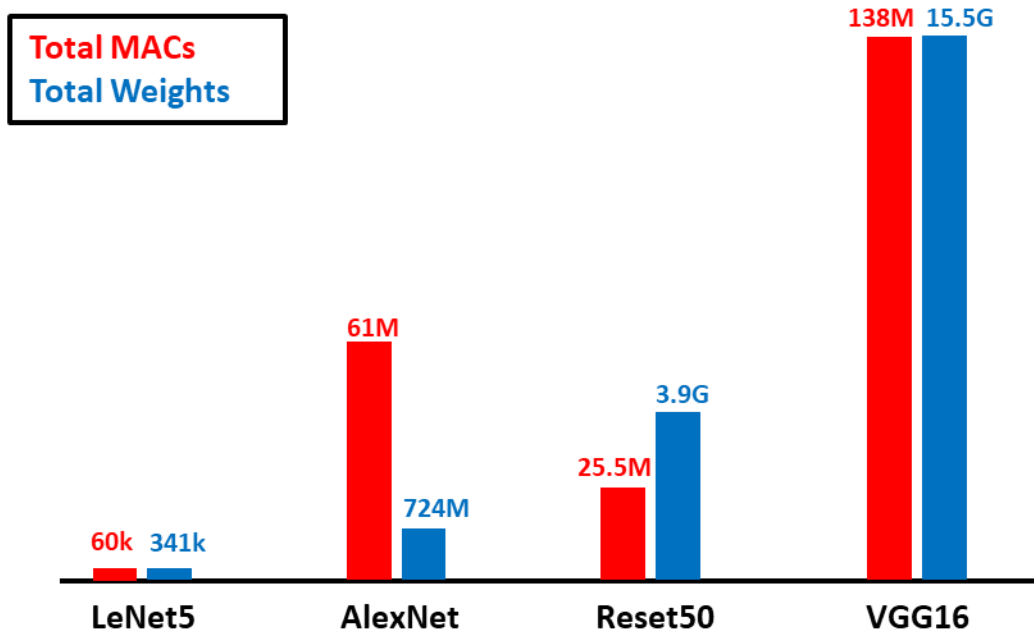
Image Classification



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

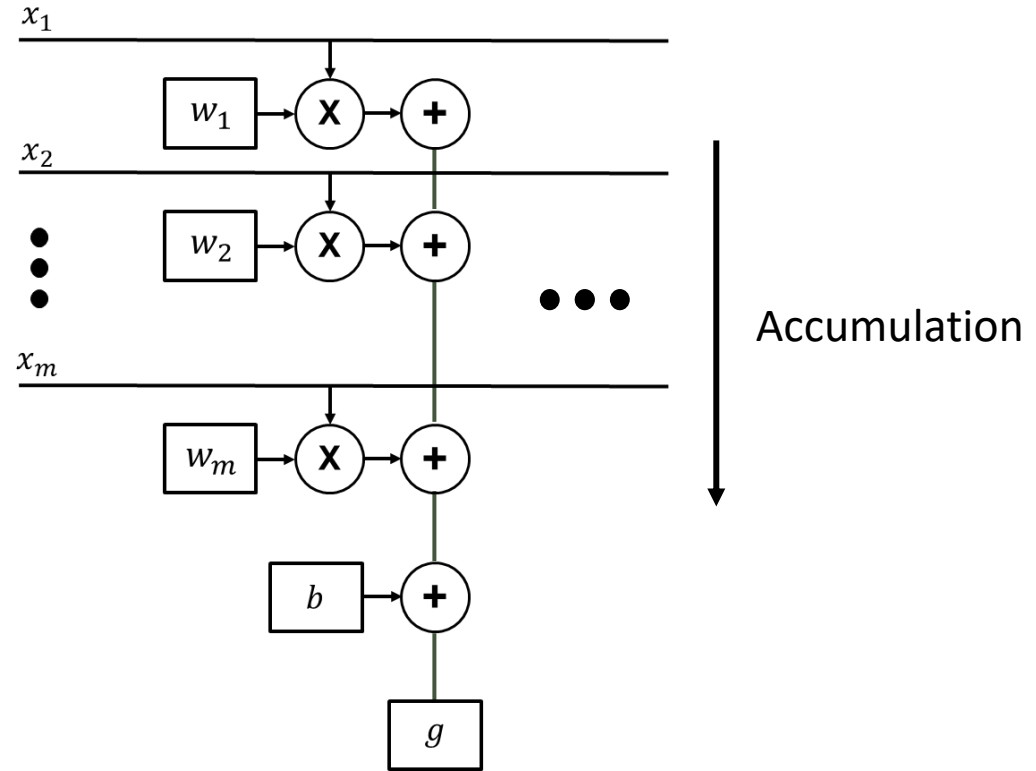
$$h = g \left[\left(\sum_{i=1}^m w_i * x_i + b \right) \right]$$

x_i : Input activation

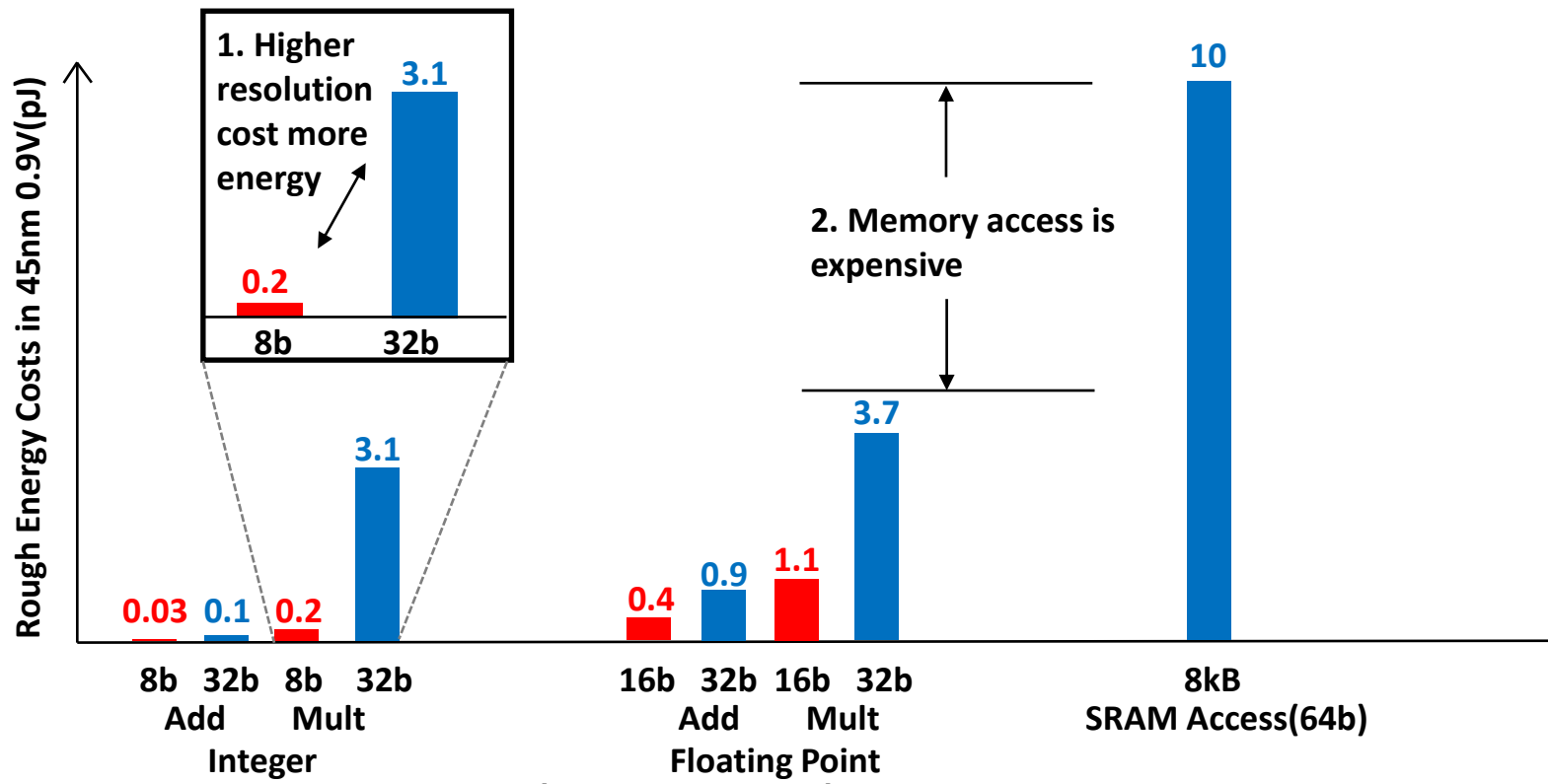
w_i : Weight b: Bias

h : Output to next layer

g : Activation function



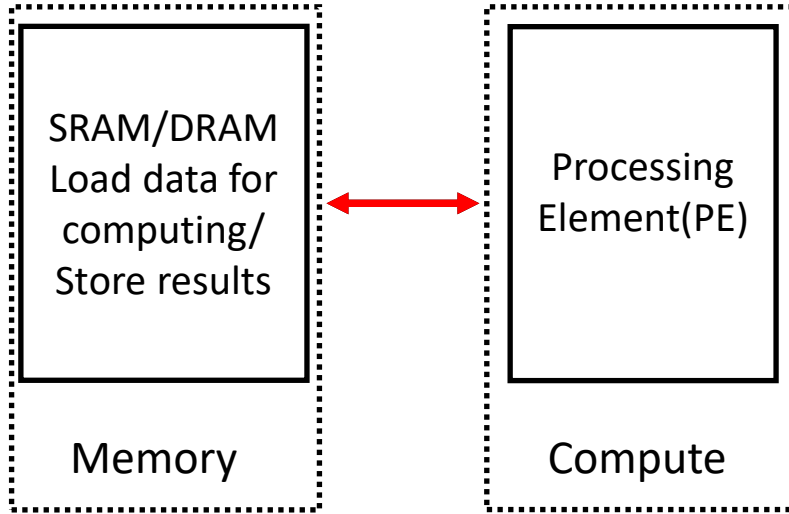
Challenges on Energy Efficient NN Inference



[M. Horowitz, ISSCC 2014]

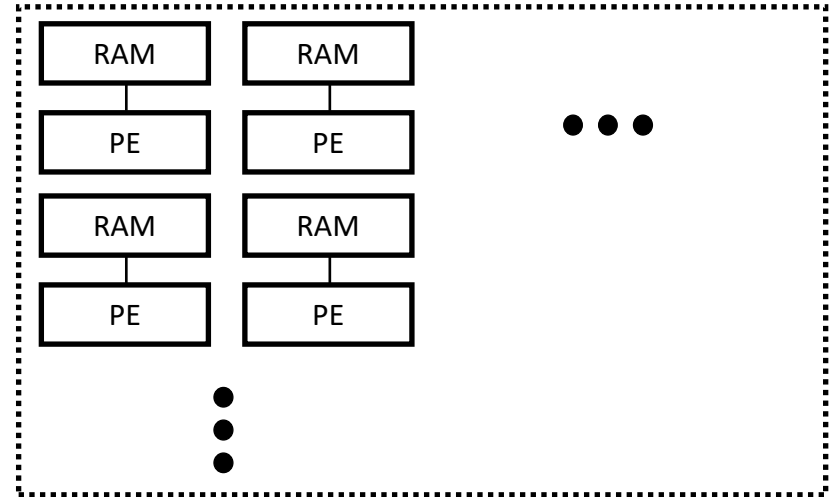
Solutions to Energy Efficient NN Inference

- Conventional computing:



Memory access can easily dominate energy/throughput

- In-memory-computing:

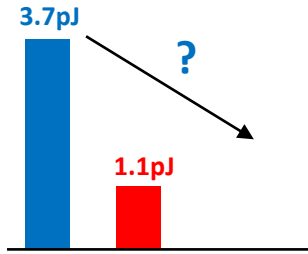


Minimized data movement from distributed memory

Solutions to Energy Efficient NN Inference

- Reduced Resolution Network:

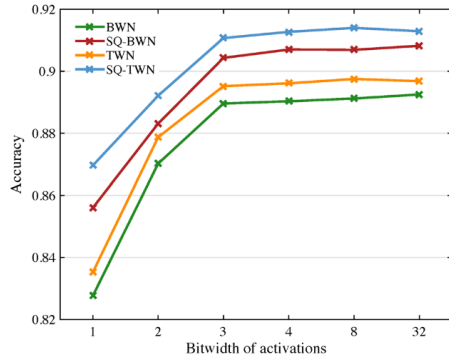
32b Floating point → ?



Multiplying energy cost

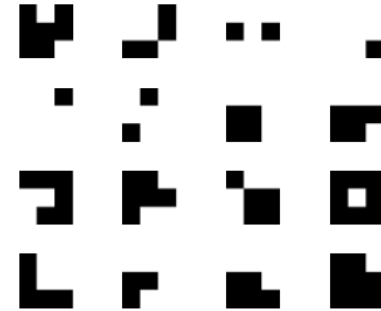
Solutions to Energy Efficient NN Inference

- Reduced Resolution Network:



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

[Y. Dong, IJCV 2019]



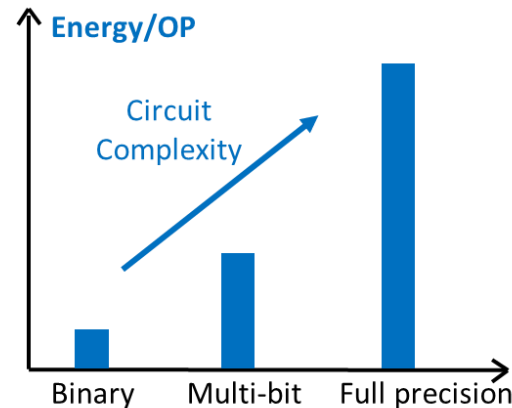
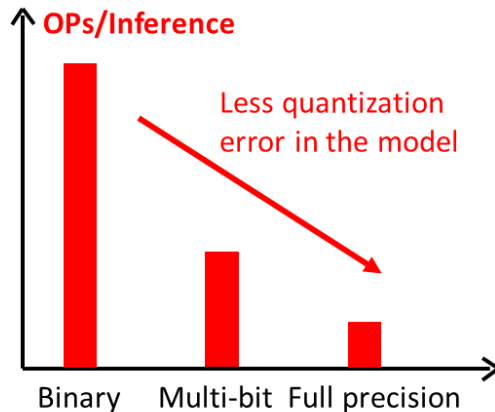
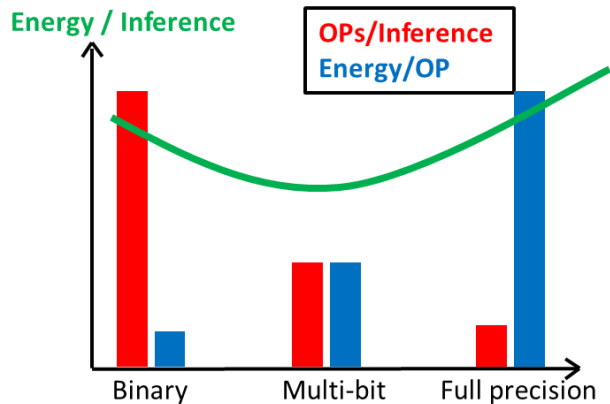
Visualization of filters from binary
neural network

[M. Courbariaux, arXiv 2016]

Energy Cost of NN Inference

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

[B. Murmann, ISSCC 19 Tutorial]



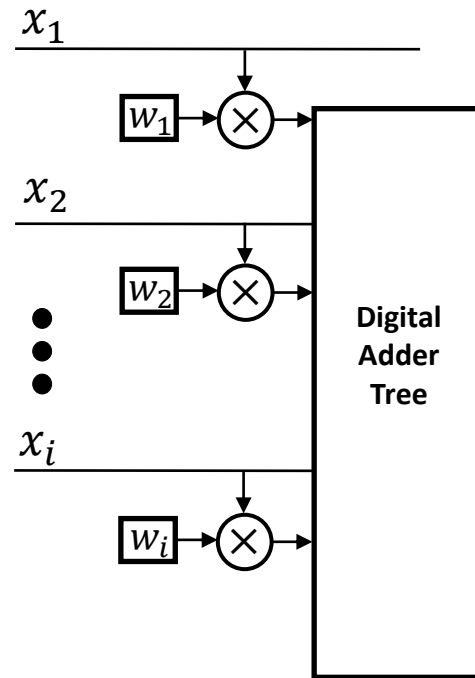
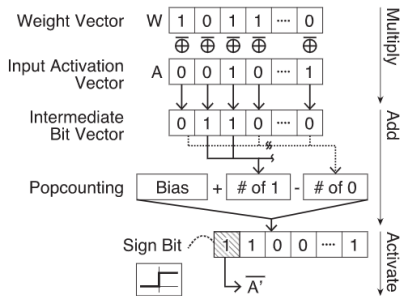
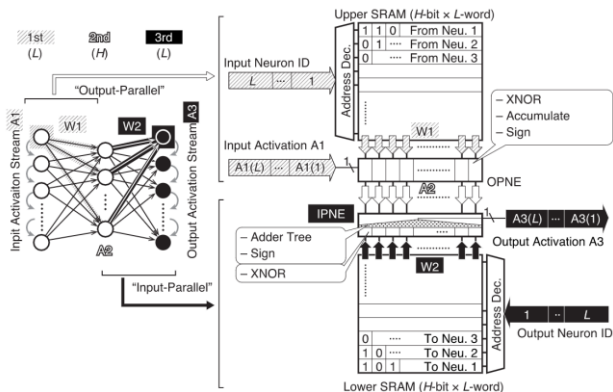
Outline

- 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 😞

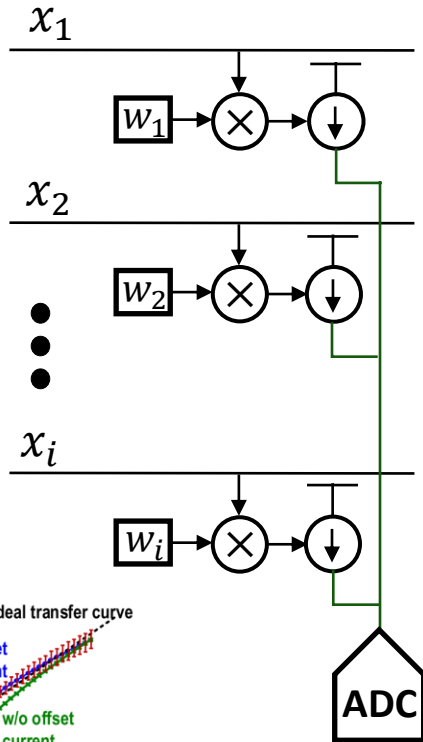
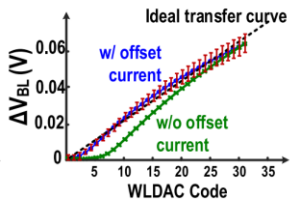
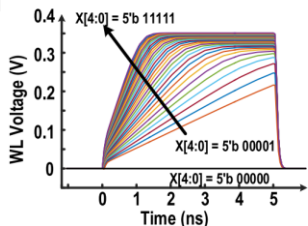
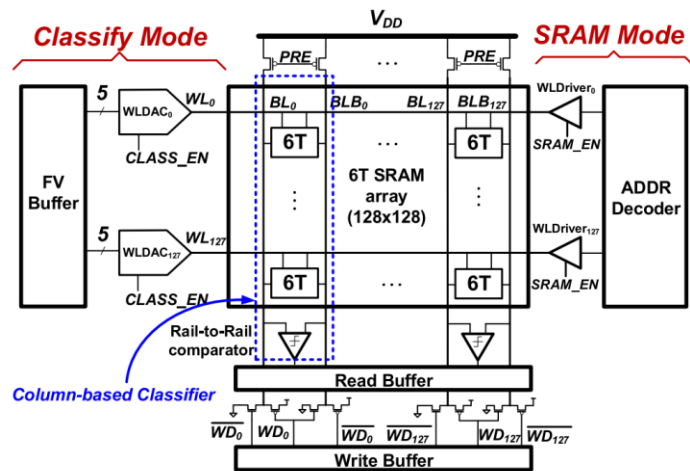


[K. Ando, JSSC 18]

Existing works

- **Current Domain:**

- High throughput 😊
- PVT-robustness 😞
- Consumes static current 😞

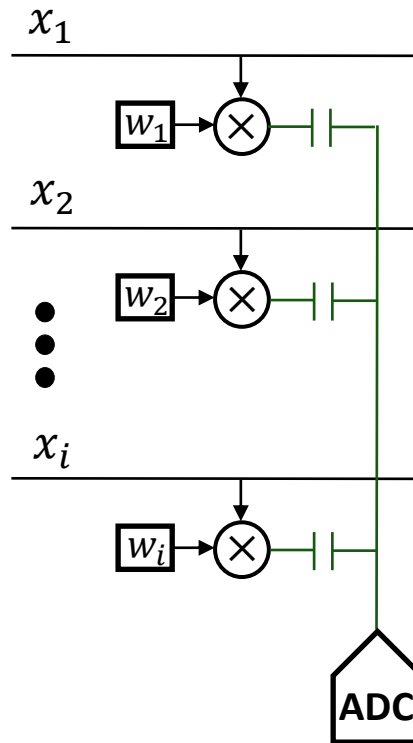
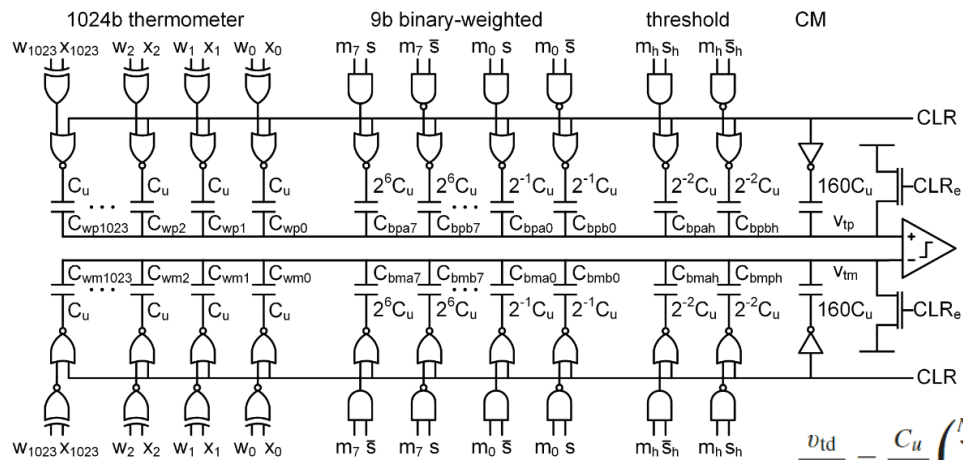


[J. Zhang, JSSC 17]

Existing works

• Charge Domain:

- High throughput 😊
- No static current 😊
- Large operations/inference 😞



[D. Bankman, JSSC 18]

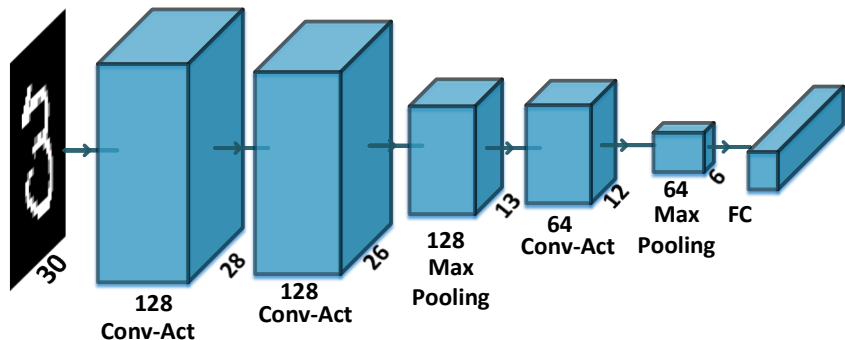
$$\frac{v_{td}}{V_{DD}} = \frac{C_u}{C_{tot}} \left(\sum_{i=0}^{N-1} w_i x_i + b \right).$$

Outline

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

Comparison of Model Size

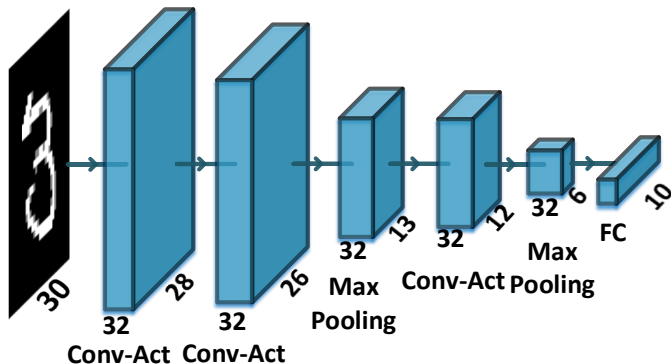
Baseline test: 98% Accuracy on MNIST



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

1b Resolution
 1.38×10^8 OPs

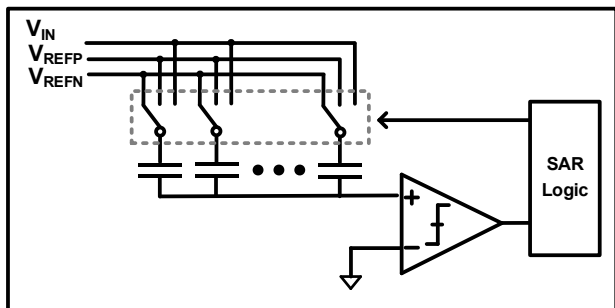
~4x Bigger model size



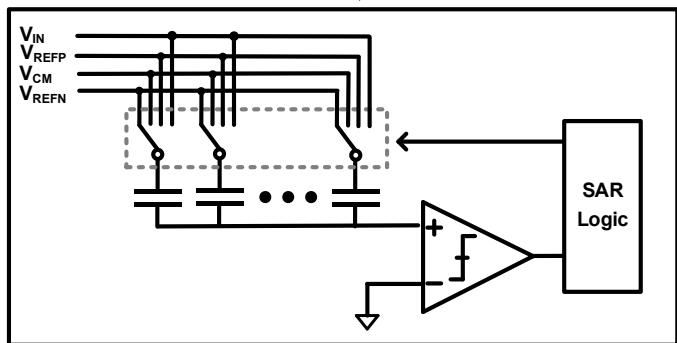
Layer	Type	Size	Channel	Filter Size
1	CONV-TN	30x30	1(input)	2x2
2	CONV-TN	28x28	32	
2p	MAX POOL	26x26		
3	CONV-TN	13x13		
3p	MAX POOL	12x12		
4	FC	(Flatten 6x6x32) 1152 - 10		

1.5b Resolution
 3.57×10^7 OPs
 {w,x from -1,0,1}

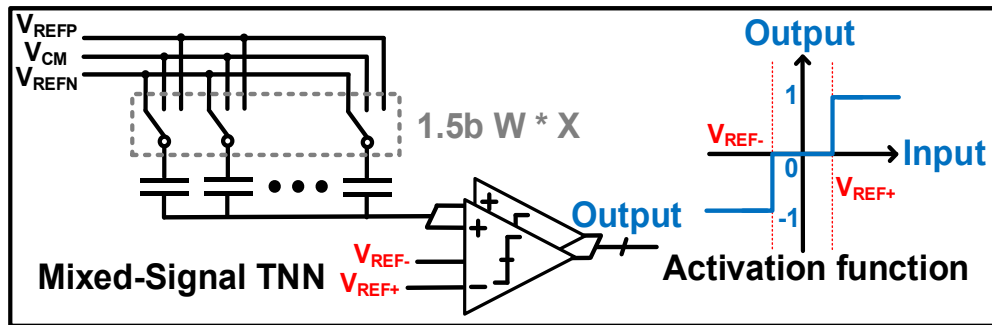
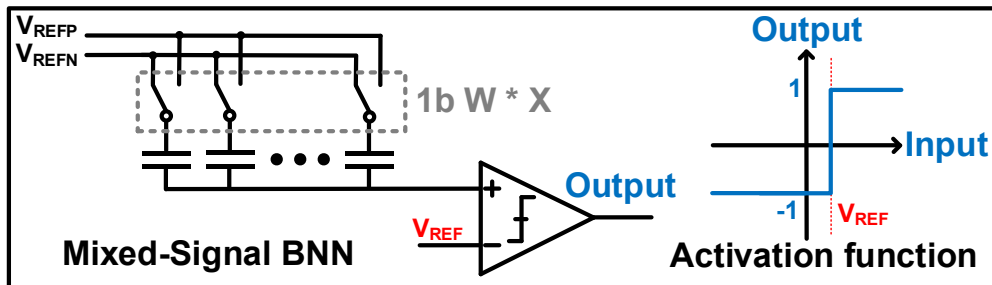
Mixed Signal BNN vs TNN



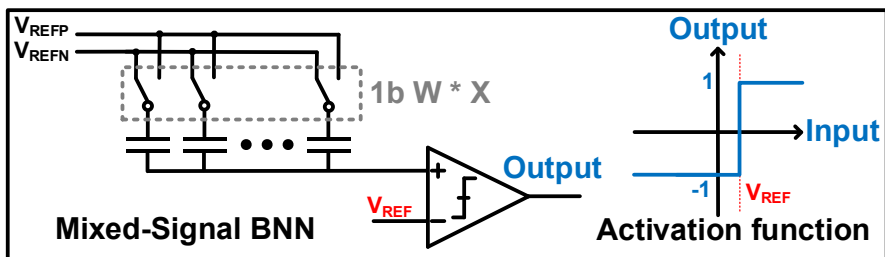
SAR ADC



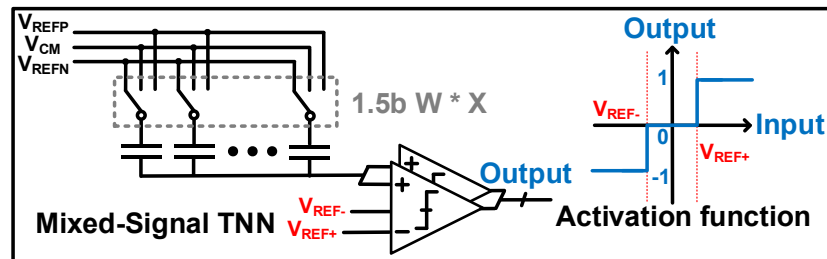
SAR ADC with V_{CM} based switching



Mixed Signal BNN vs TNN



VS



	Hardware Complexity	Operations Inference (@same accuracy)	\times Energy Operation (CDAC signal swing)	= Energy Inference
BNN	😊😊	😞	😞	😞
TNN	😊	😊	😊	😊

OPs/Inference ↓ 75%

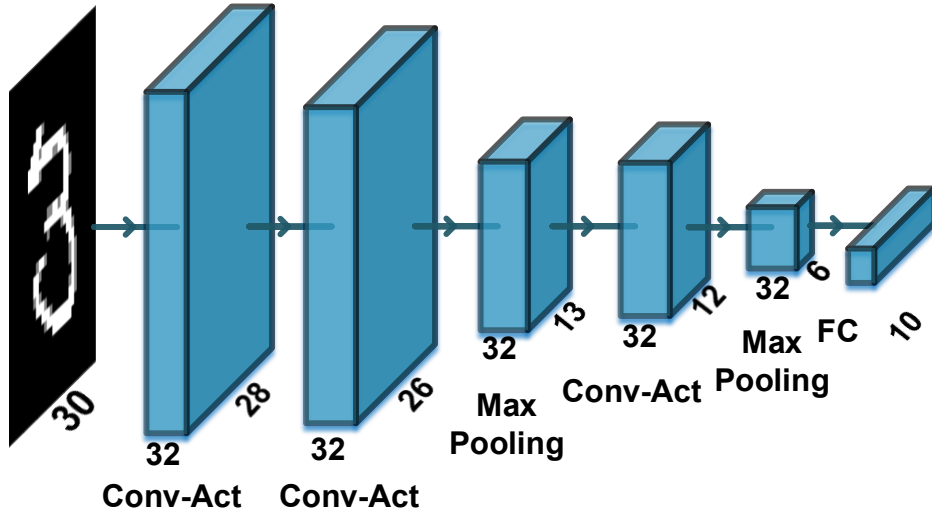
Energy/Operation ↓ 31%

Energy/Inference ↓ 82%

Outline

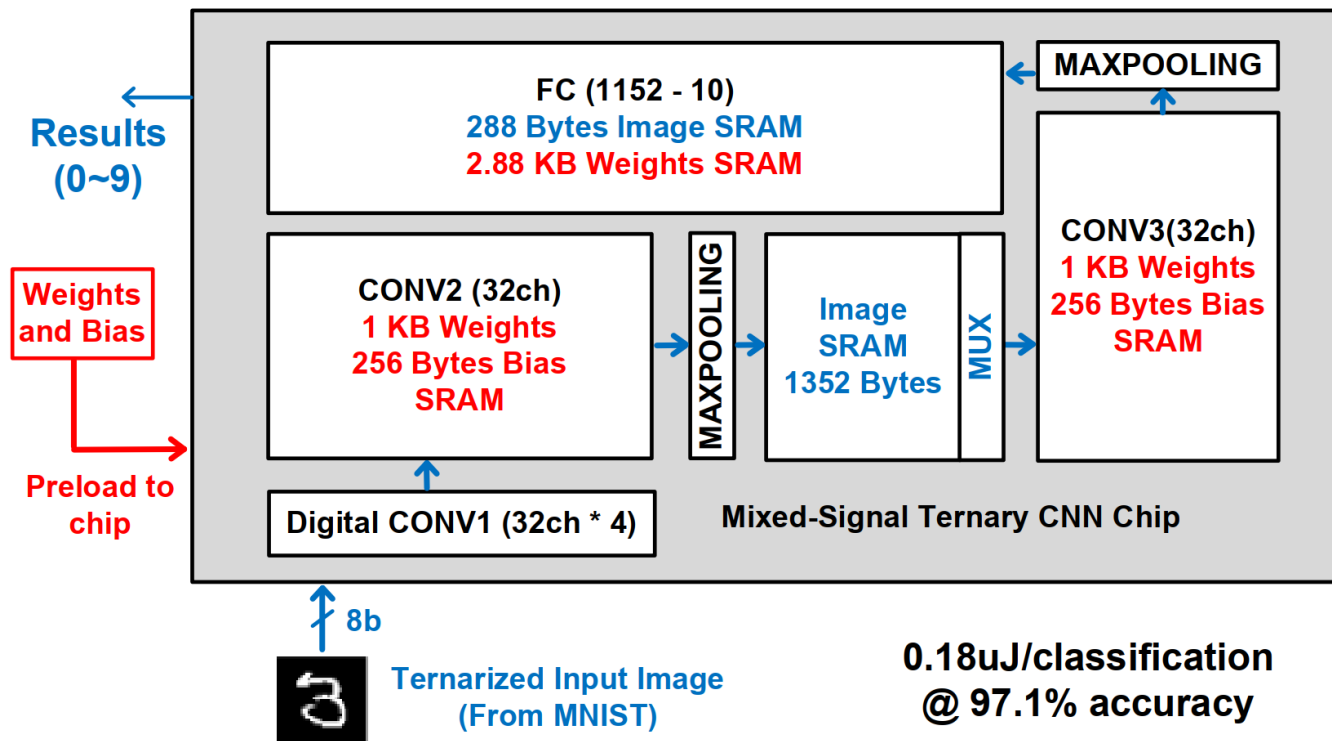
- Motivations
- Existing Works
- Theoretical Concept of the Proposed Work
- **Circuit Implementation**
- Measurement results
- Summary

On-chip Neural Network Model

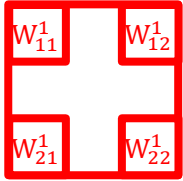


Layer	Type	Size	Channel	Filter Size	Dilated
1	CONV-TN	30x30	1(input)	2x2	2
2	CONV-TN	28x28	32		2
2p	MAX POOL	26x26			1
3	CONV-TN	13x13			1
3p	MAX POOL	12x12			1
4	FC	(Flatten 6x6x32) 1152 - 10			

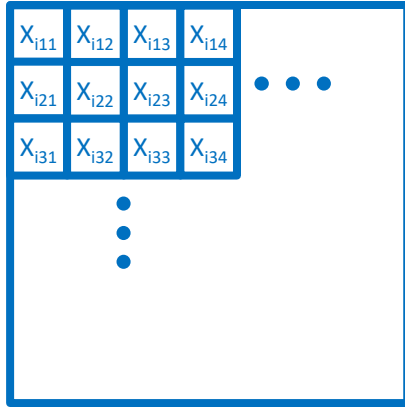
Chip Architecture



CONV1 – Example of One-Channel Convolution



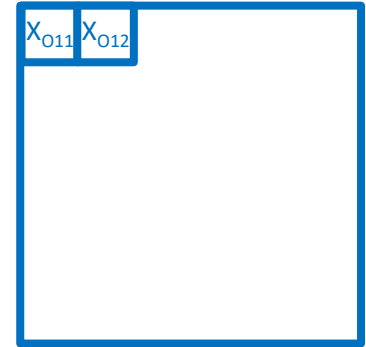
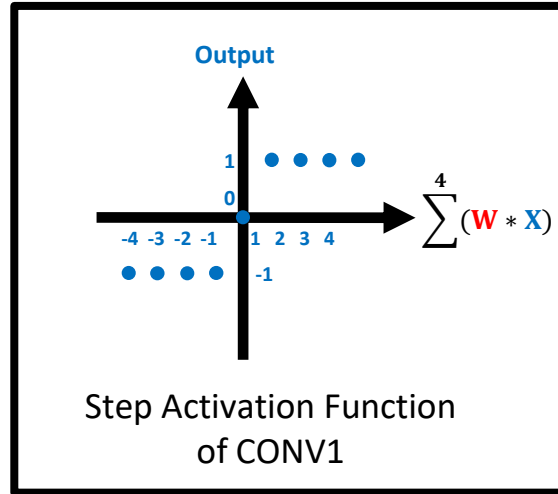
Filter0 2x2
Dilated L = 2



Ternarized
Input Image
1ch

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

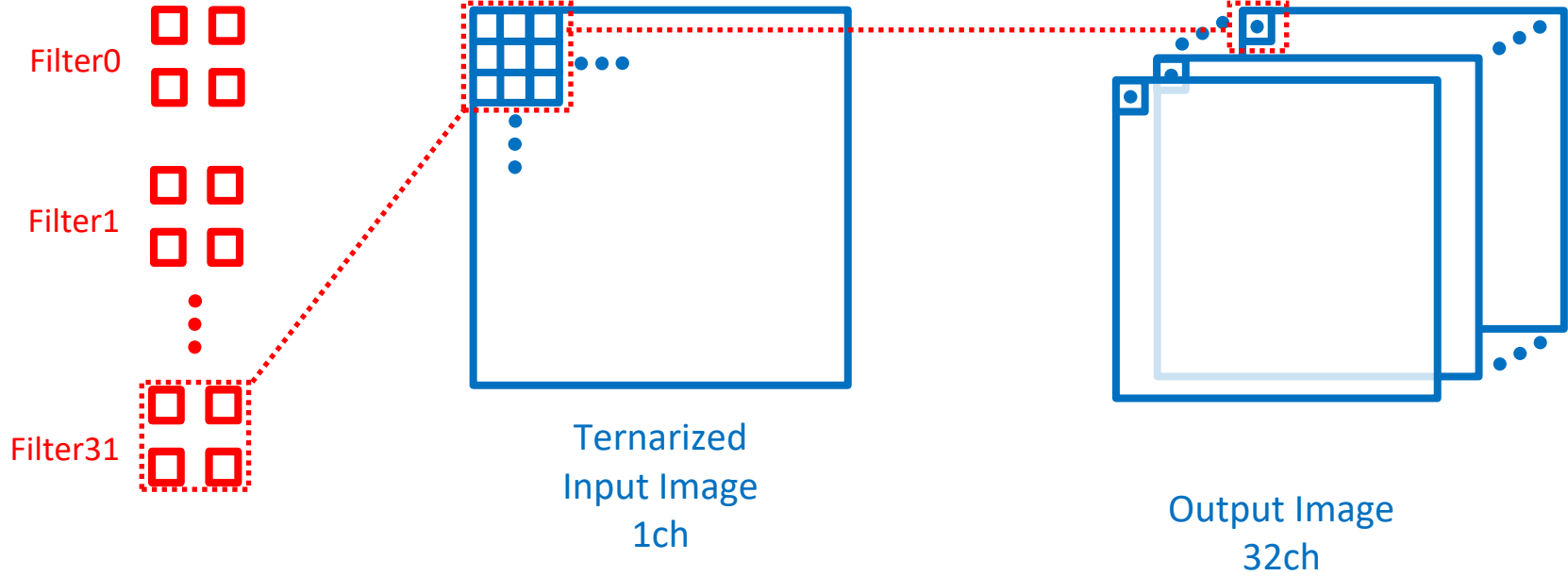
$$X_{O12} = \text{STEP}(W_{11} * X_{i12} + W_{12} * X_{i14} + W_{21} * X_{i32} + W_{22} * X_{i34})$$



Output Image
1ch

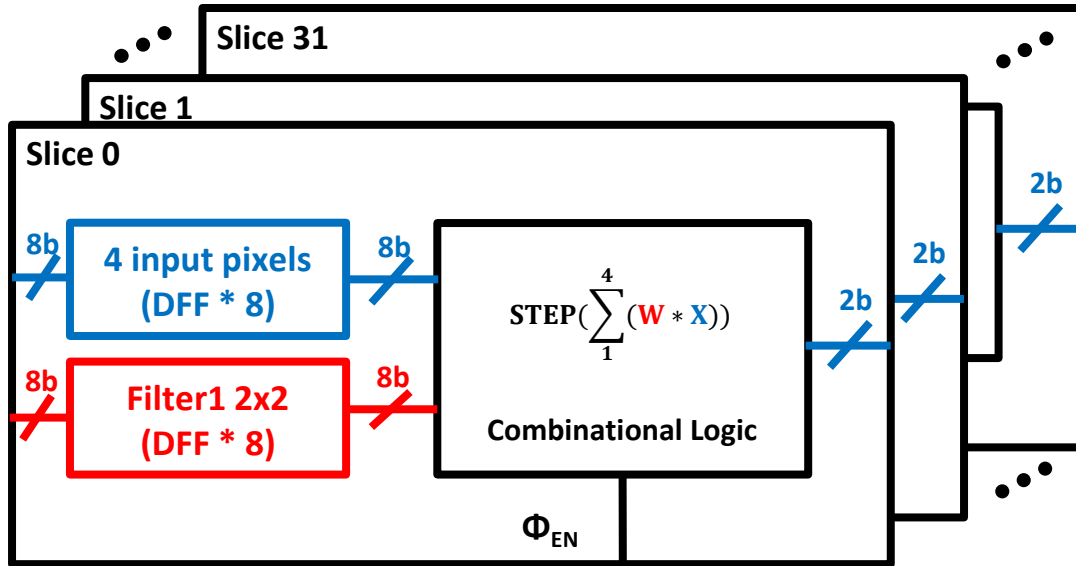
$$W, X \in \{-1, 0, 1\}$$

CONV1 – Example of 32-Channel Convolution

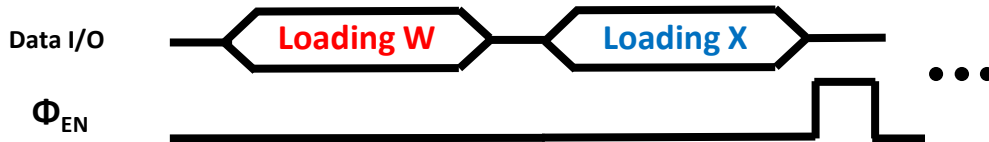
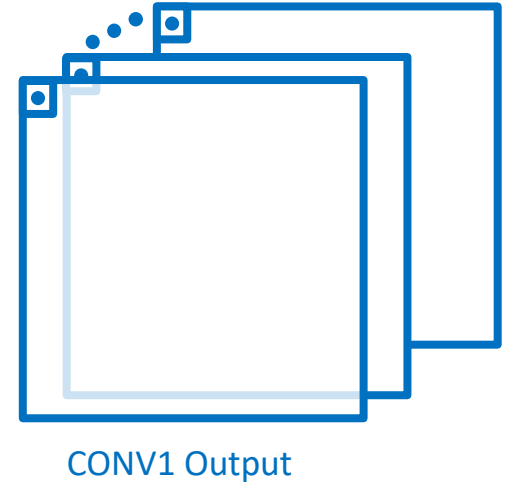


$$W, X \in \{-1, 0, 1\}$$

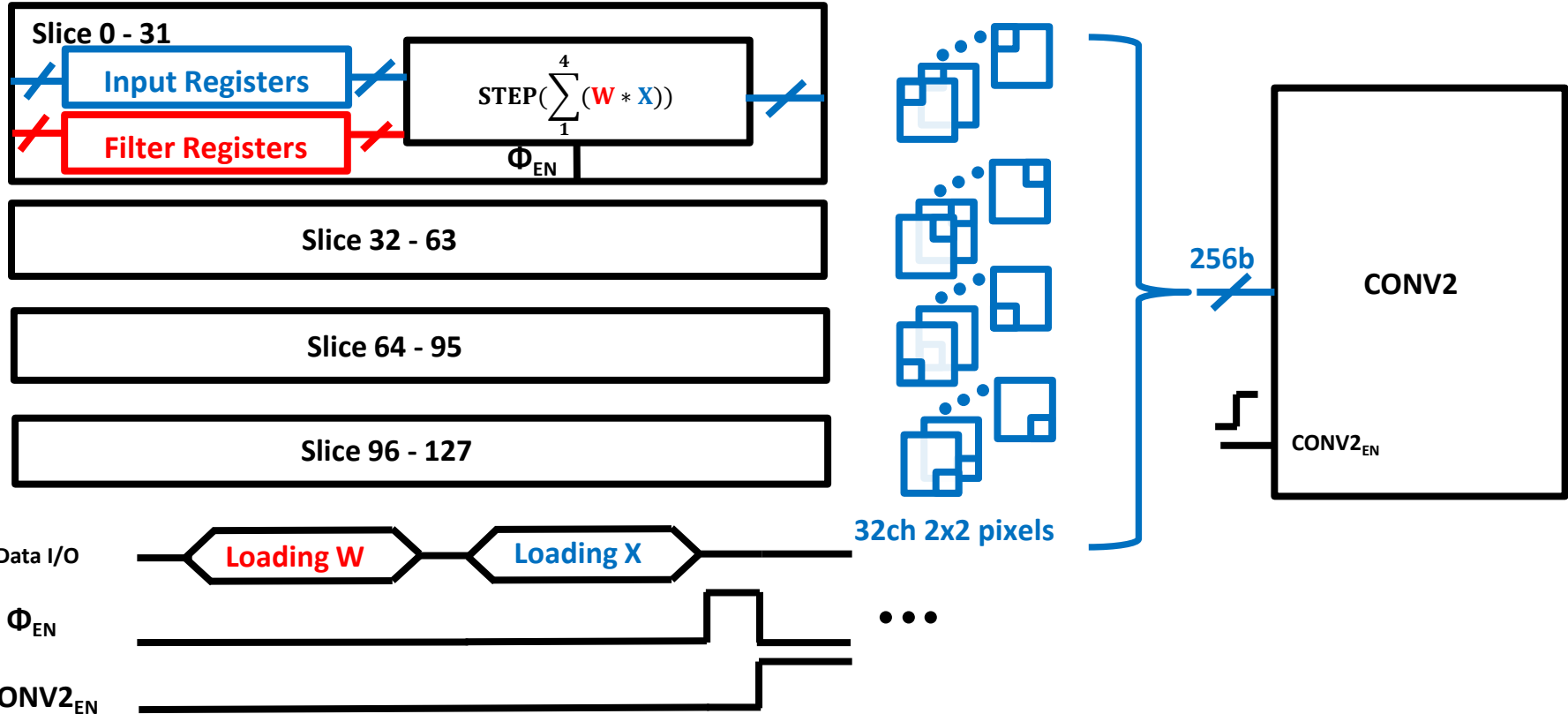
CONV1 – Digital Implementation



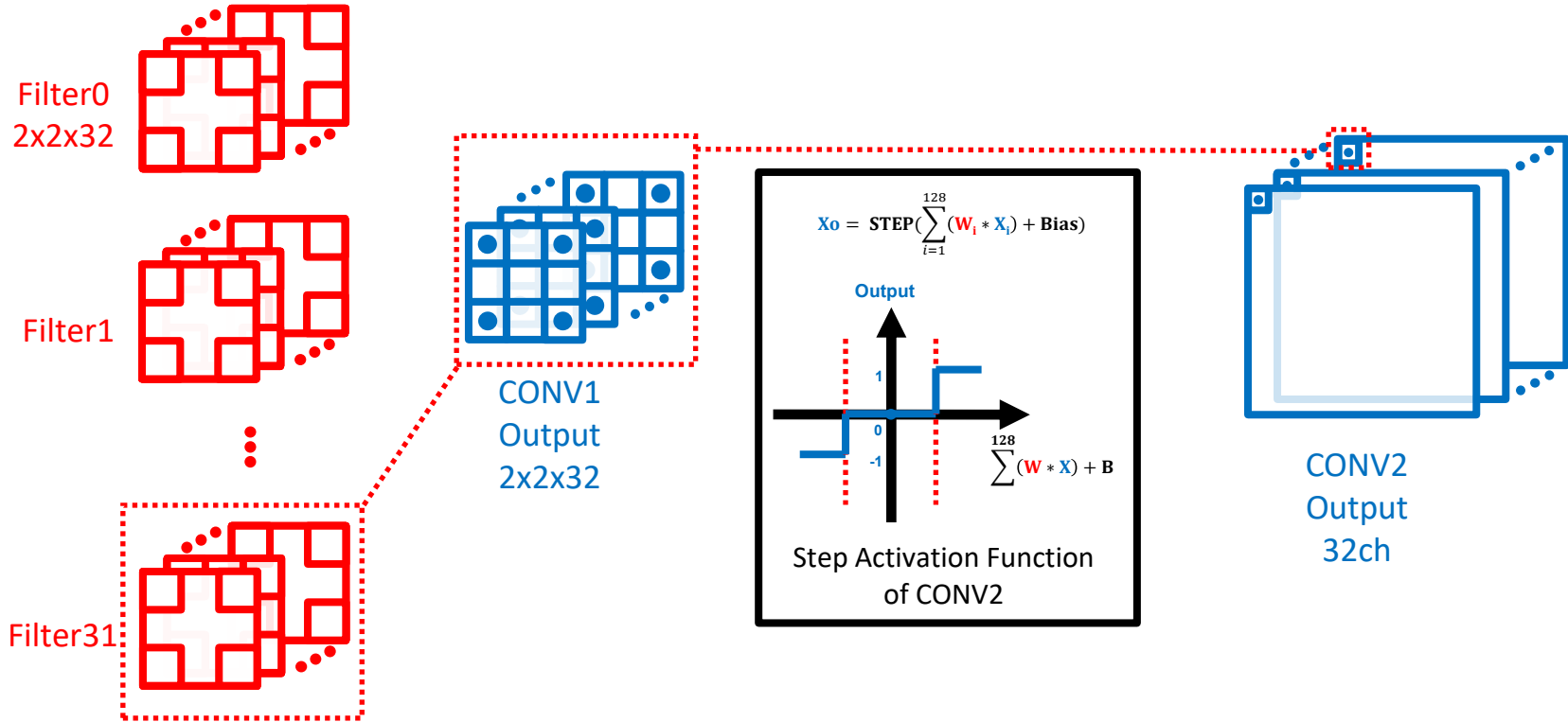
One Pixel Output (32 ch)



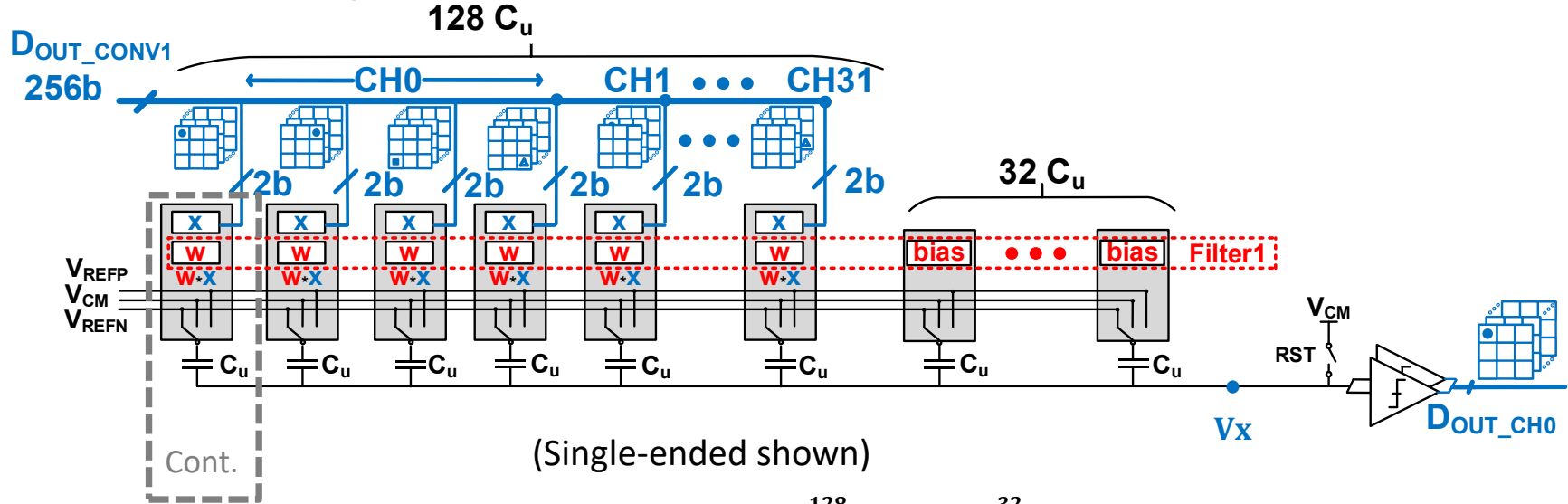
CONV1 – Digital Implementation



CONV2 – Example of 32-Channel Convolution



CONV2 – Implementation of One-Channel SC Neuron

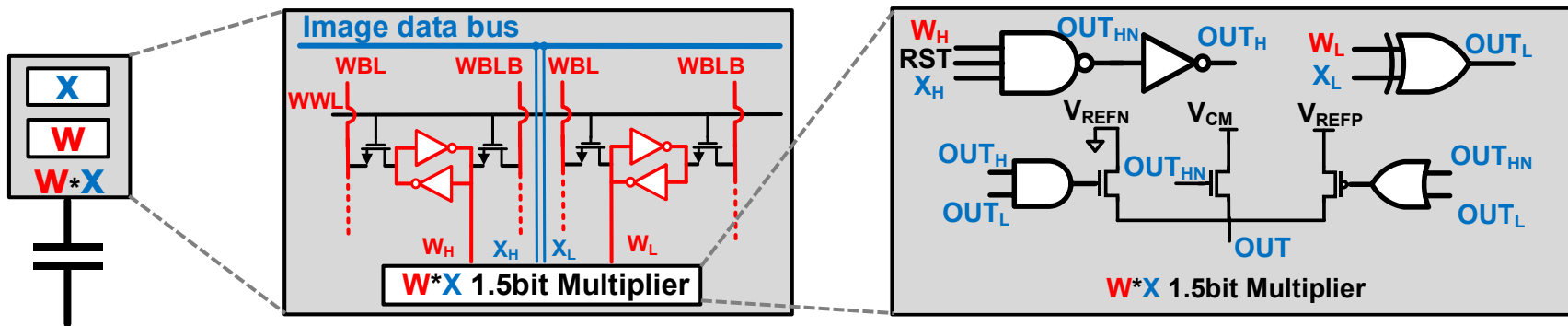


$$V_x = \frac{C_u}{C_{Total}} * (V_{REFP} - V_{REFN}) * \left(\sum_{i=1}^{128} (W_i * X_i) + \sum_{i=1}^{32} Bias_i \right)$$

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

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

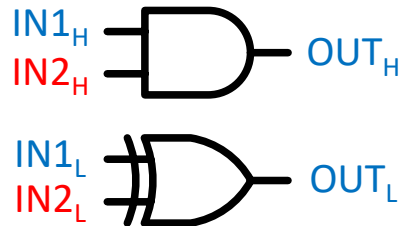
CONV2 – Synapse Design



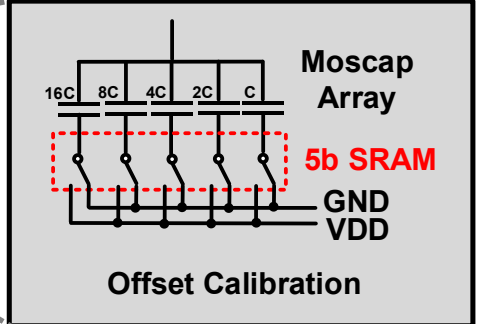
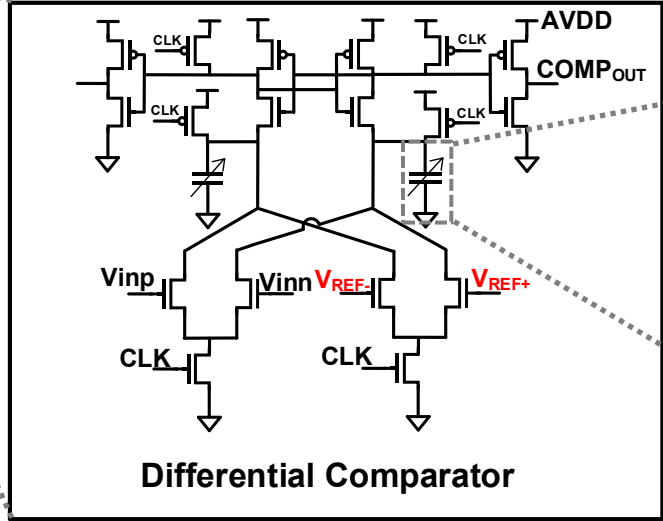
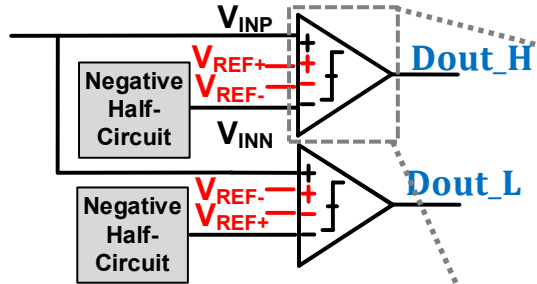
DEC	BIN	Voltage
1	10	V_{REFP}
-1	11	V_{REFN}
0	0X	V_{CM}

Encoding for simplicity:

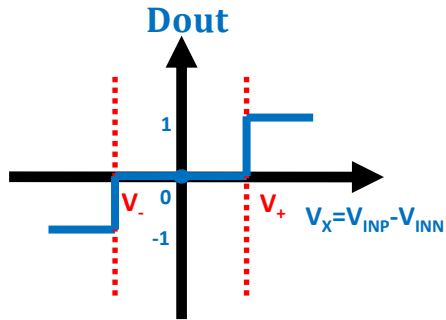
1.5b Multiplier \rightarrow



CONV2 – Comparator Design



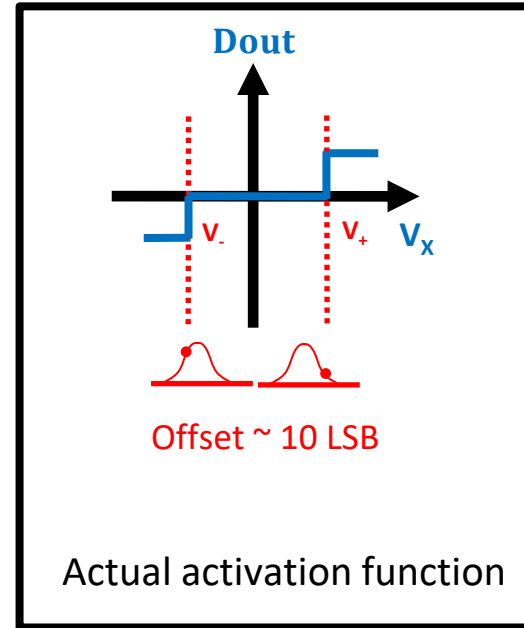
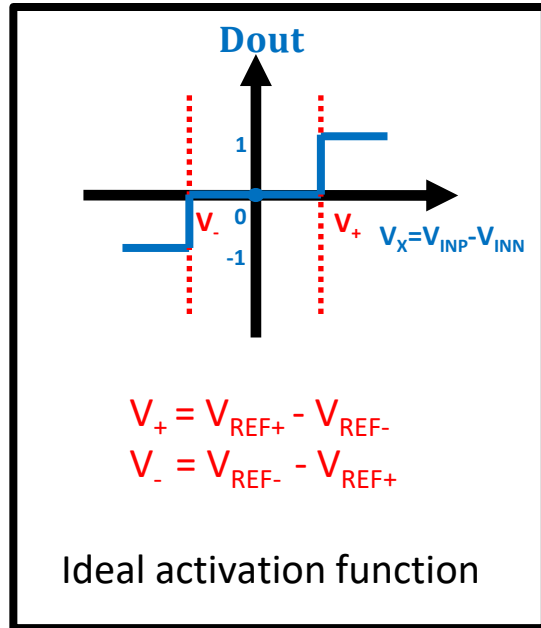
$Dout = STEP (V_x)$



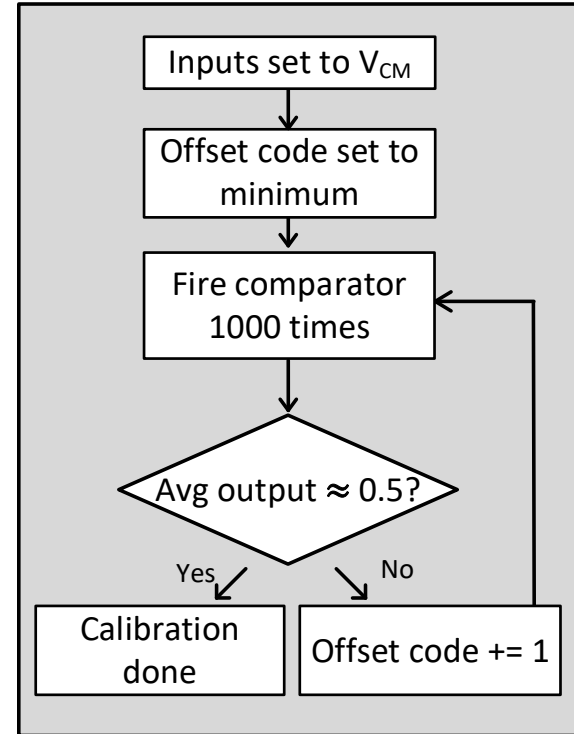
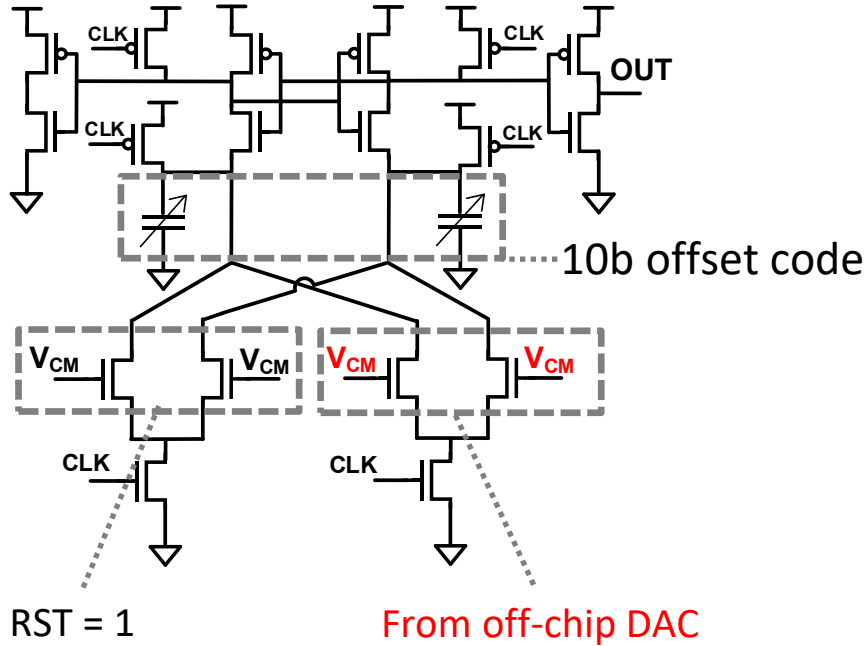
$V_+ = V_{REF+} - V_{REF-}$
 $V_- = V_{REF-} - V_{REF+}$

CONV2 – Effect of Comparator Offset

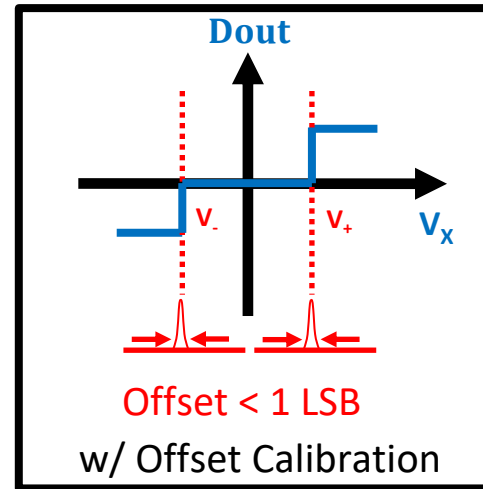
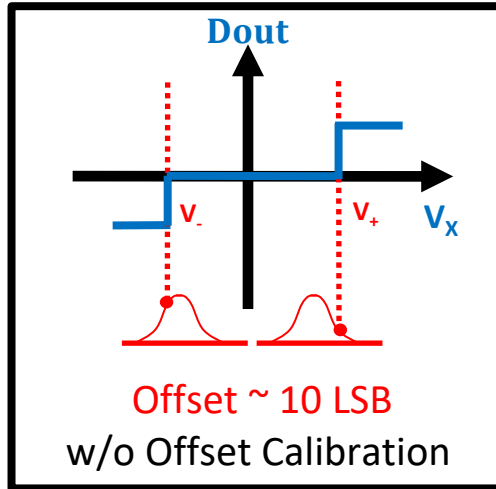
$$Dout = \text{STEP}(V_x)$$



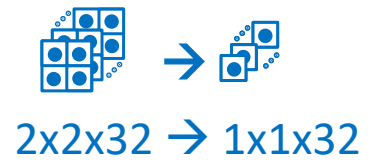
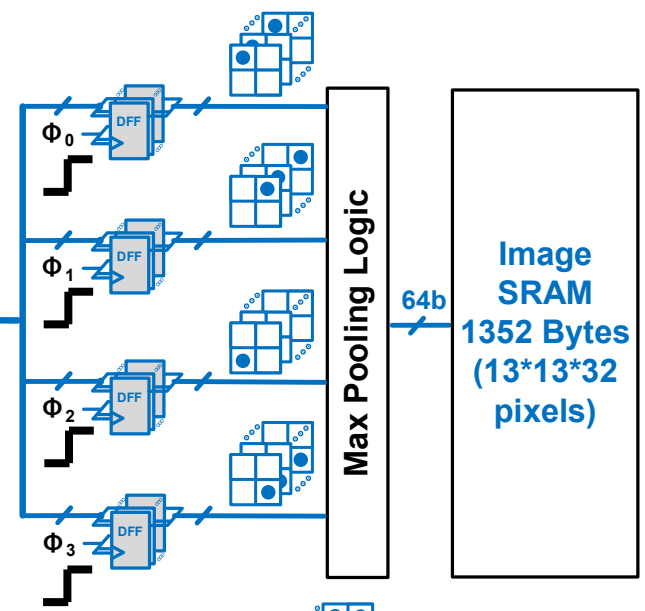
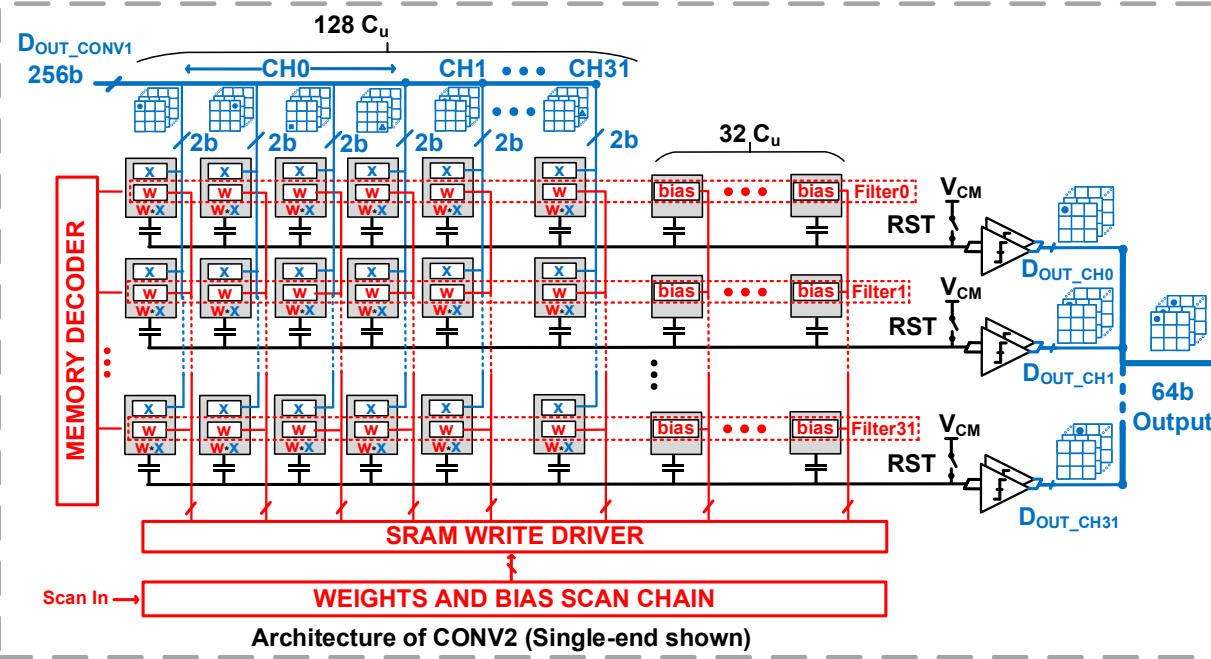
CONV2 – Foreground Comparator Offset Calibration



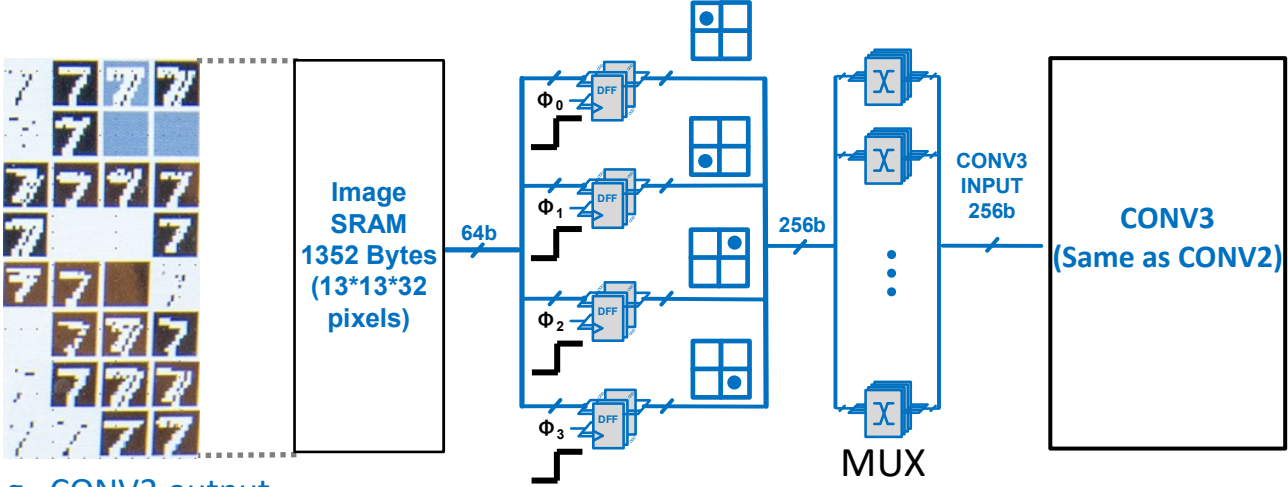
CONV2 – Foreground Comparator Offset Calibration



CONV2 – Maxpooling

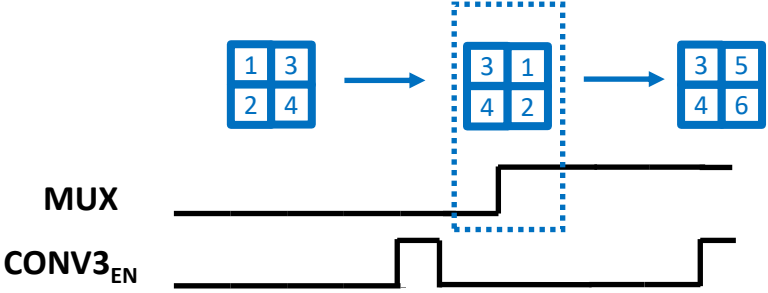
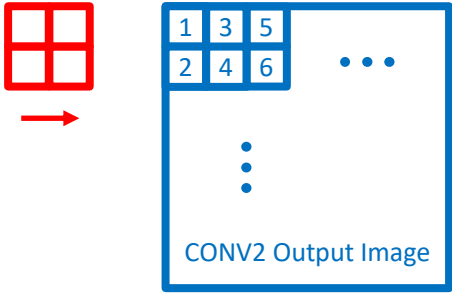


Datapath from CONV2 to CONV3



E.g. CONV2 output

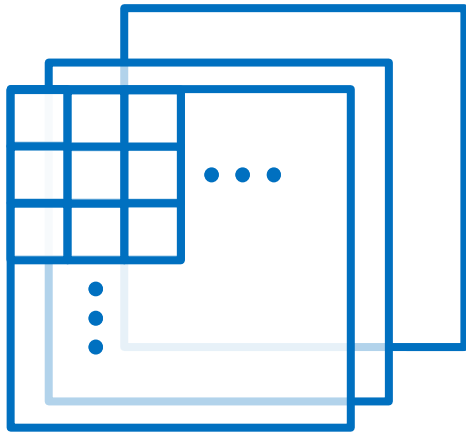
CONV3 Filter Window



Only need to read 2 pixels from SRAM

x144

FC Layer Operation



CONV3 Output Image
6x6x32



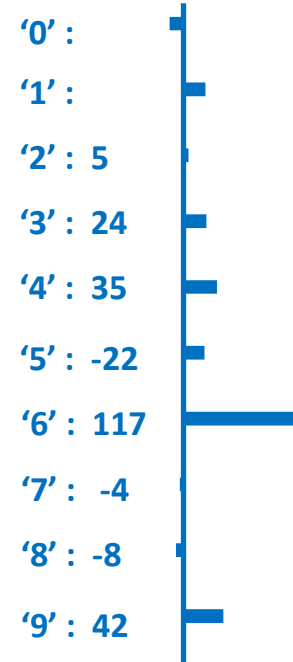
Flattened
1152 x 1

*



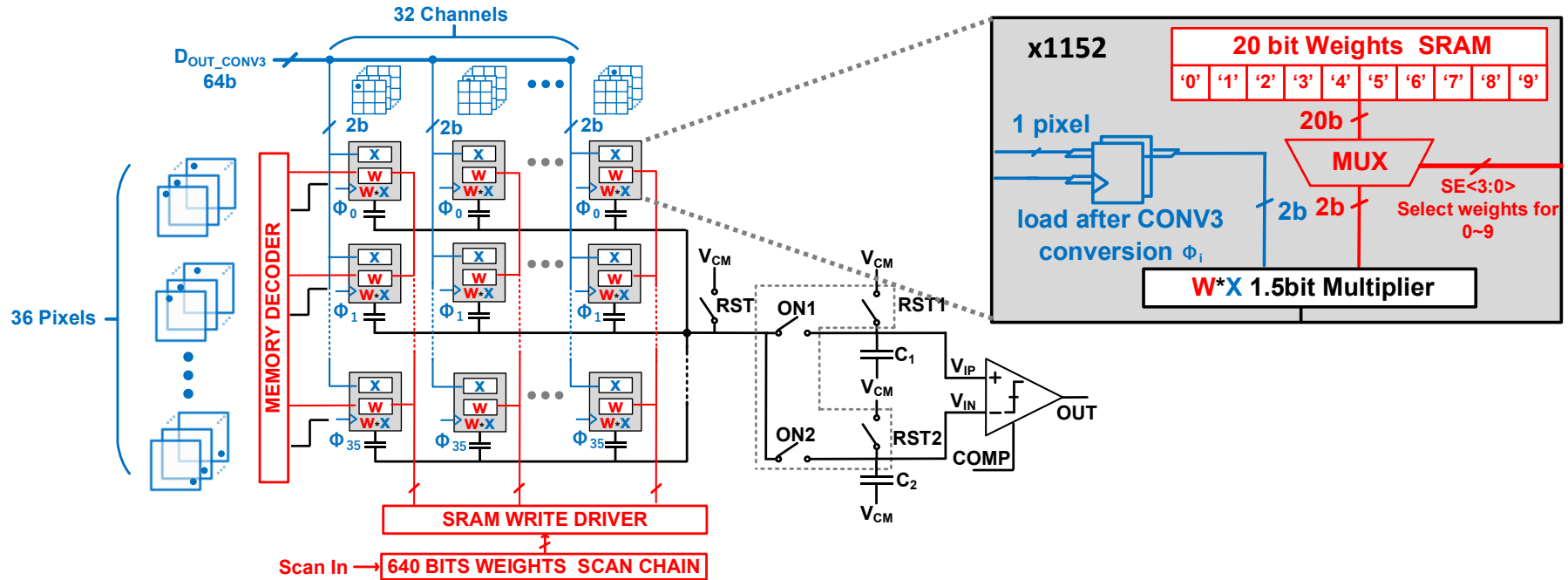
Weights
for '0-9'

$$\text{Logit} = \left(\sum_{i=1}^{1152} (W_i * X_i) \right)$$
$$W_i, X_i \in \{-1, 0, 1\}$$



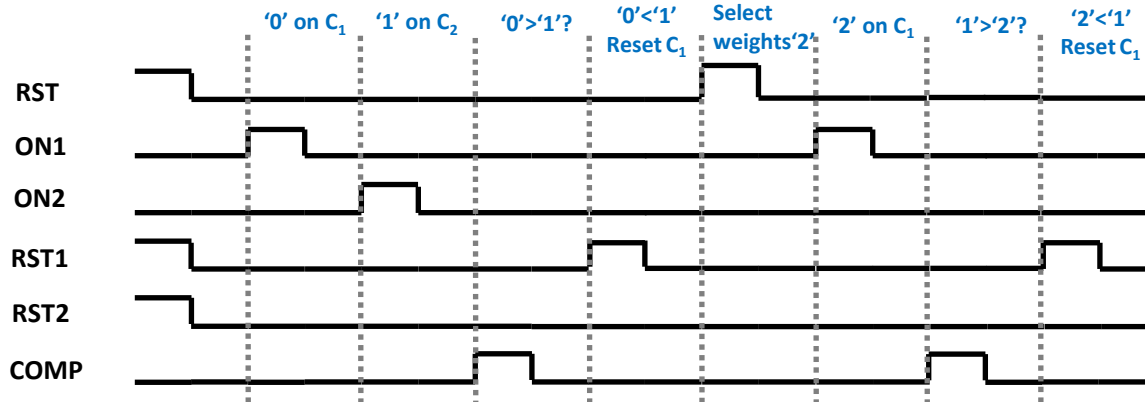
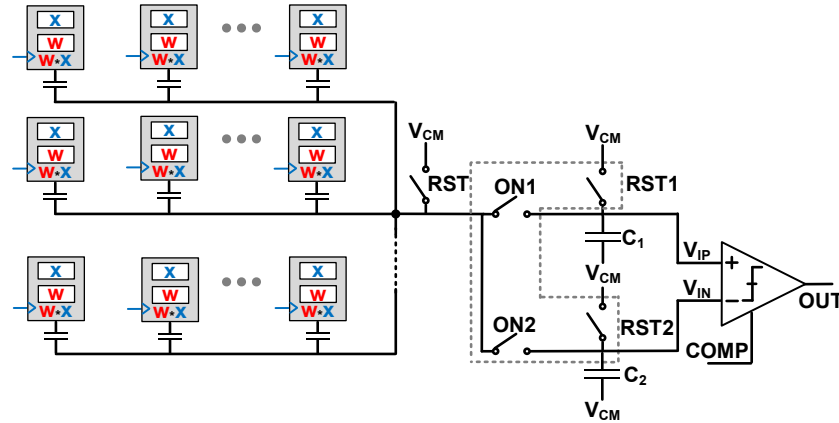
Classification Result : 6

FC Layer Implementation



(Single-ended shown)

FC Layer Implementation

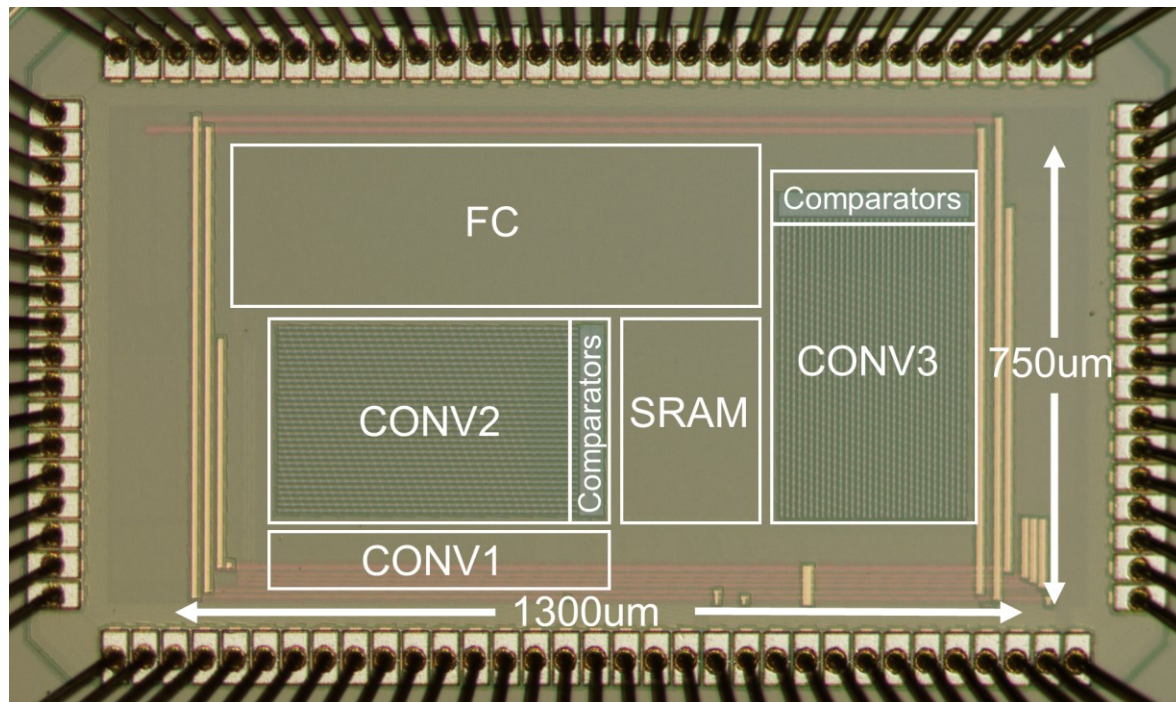


Outline

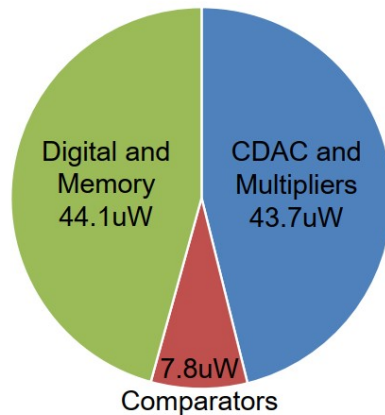
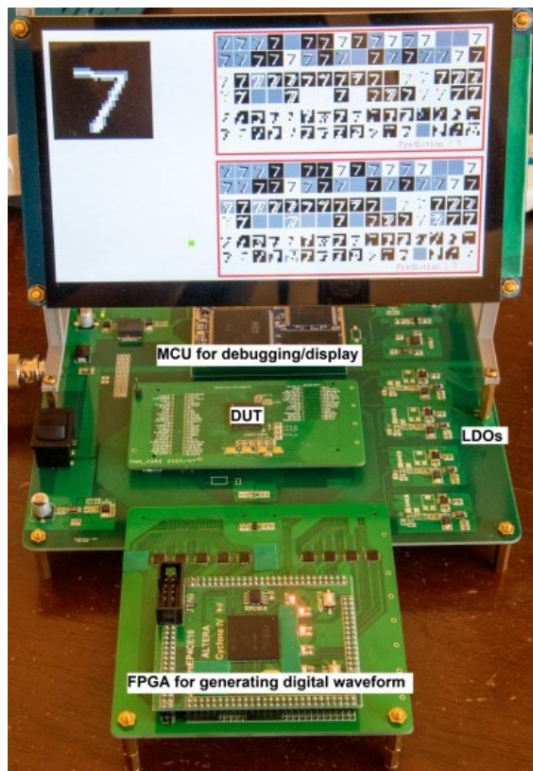
- Motivations
- Existing Works
- Theoretical Concept of the Proposed Work
- Circuit Implementation
- **Measurement results**
- Summary

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 _{REFF}	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(mm ²)	0.98		3.9	4.6	5.85	0.055	12.6
Area Eff.(GOPs/mm ²)	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

¹Based on SC neuron

²Based on MACs energy efficiency

³10 runs average on 10,000 test set images.

Outline

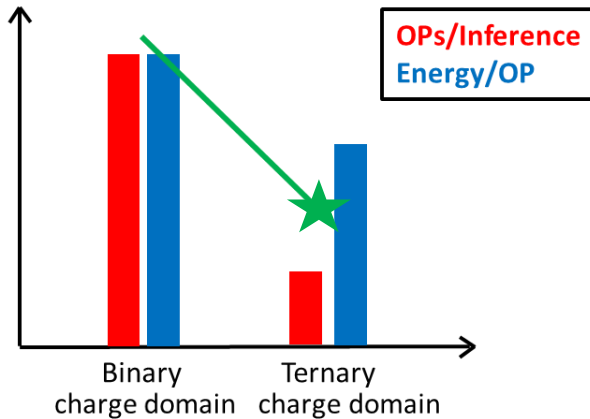
- Motivations
- Existing Works
- Theoretical Concept of the Proposed Work
- Circuit Implementation
- Measurement results
- **Summary**

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% ↓ $\frac{\text{Operations}}{\text{Inference}}$
 - 31% ↓ $\frac{\text{Energy}}{\text{Operation}}$

82% ↓ $\frac{\text{Energy}}{\text{Inference}}$

Energy / Inference



OBRIGADO
 gracias
 どうも
 ARIGATO
 grazas
 GRAZZI
 THANKS
 qujan
 PALDIES
 DANKU
 danke
 DANKU
 takk
 MERSI
 merci
 謝謝
 謝謝
 danke
 DANKU
 takk
 MERSI
 merci
 謝謝
 謝謝
 danke
 OBRIGADO
 danke schön
 KÖSZI
 سپاس
 PALDIES

 ありがとう
 TEŞEKKÜR EDERİM
 MOLTE GRAZIE
 GO RAIBH MAITH AGAT
 THANK YOU
 благодаря
 TAK
 どうも
 asante
 muchas gracias
 vielen dank
 DANKU
 OBRIGADO
 mesī
 감사합니다
 TEŞEKKÜR EDERİM
 NA GODE
 muchas gracias
 DZLEKI
 Gràcies
 MULTUMESC
 DANKU
 OBRIGADO
 danke
 KÖSZI
 благодаря
 СПАСИБО
 多謝
 شكراً