HỌC VIỆN KỸ THUẬT QUÂN SỰ

Hội nghị nghiên cứu trẻ 2025

Thực thi hiệu quả phần cứng tăng tốc tính toán AI tại biên kết hợp giữa mạng nơ ron hạng nhẹ và tính toán trong bộ nhớ

(Efficient Edge-AI Hardware Acceleration Enabled by the Synergy of Lightweight Neural Networks and In-Memory Computing)

Trịnh Quang Kiên, Đào Thị Ngà, Đinh Văn Ngọc,

Phạm Thị Nhẫn, Nguyễn Văn Tình, Dương Quang Mạnh

OUTLINE

- **Introduction**
- **System and Architectural Approach: BNN & BSNN**
- **Hardware Architecture – In-Memory Computing**
- **Our work: BNN and BSNN on STT-MRAM and HDC**
- **Conclusions & Outlook**

INTRODUCTION: EDGE AI

Edge-AI enables digital transformation

https://www.cdotrends.com/story/16179/ai-edge-enabling-digital-transformation

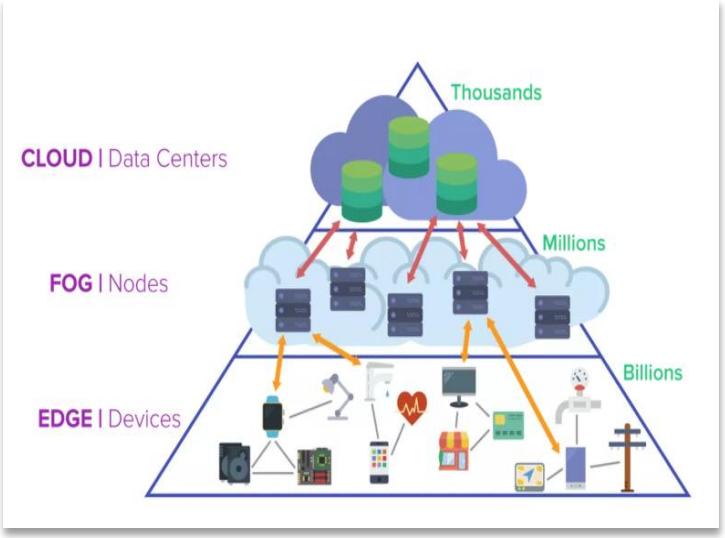
Key benefits:

- Real-time processing/Scalability
- Diminished latency
- Reduced bandwidth
- Data privacy

From IBM: "*Edge artificial intelligence refers to the deployment of AI algorithms and AI models directly on local edge devices such as sensors or Internet of Things (IoT) devices, which enables real-time data processing and analysis without constant reliance on cloud infrastructure*". (ibm.com)

Growing market for:

- Healthcare
- Smart farm
- Transportation
- Manufacturing
- Power grid
- Specialized (military, security)



FACT

In 2022, the global edge AI market was valued at USD 14,787.5 million and is expected to grow to USD 66.47M by 2023 Grand View Research, Inc

INTRODUCTION: WHERE IS DATA CENTER?

- Datacenter energy consumption account for only 1% of total electricity energy consumption worldwide

The Microsoft Aruze DC (used for ChatGPT now!)

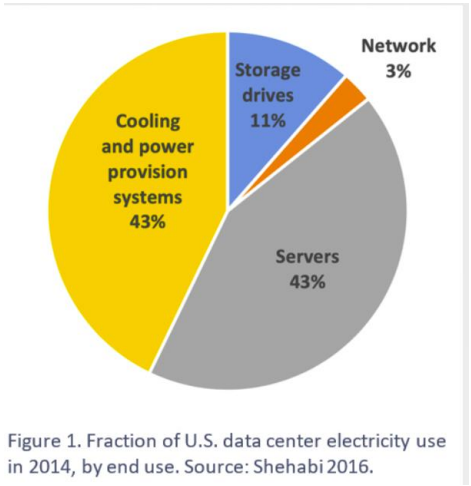
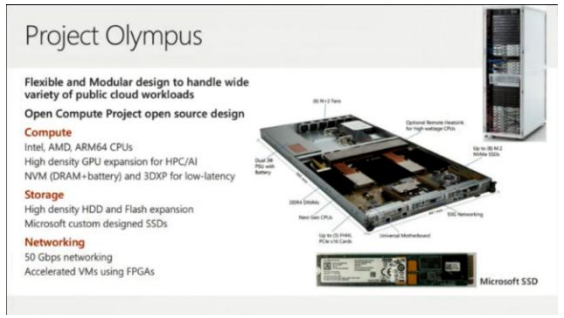
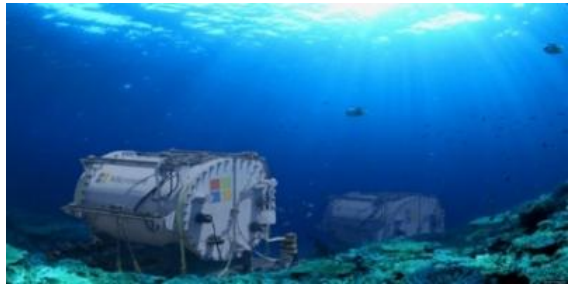


Figure 1. Fraction of U.S. data center electricity use in 2014, by end use. Source: Shehabi 2016.





ChatGPT is hosted on a Microsoft Azure data center in San Antonio, Texas. The service runs on Kubernetes on over 7,500 virtual machines to handle the prompts and API calls. ChatGPT draws on nearly 570 gigabytes of data to answer the prompt and provide helpful information for users. Hardware accelerator based on FPGA
https://agio.com/where-is-chatgpt-hosted/#gref May 1, 2024

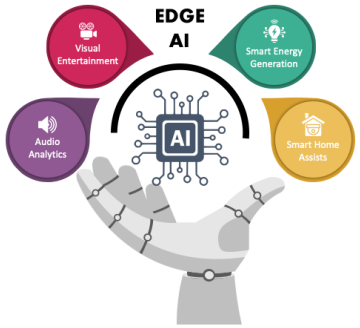
China's subsea datacenter
https://www.scmp.com/news/china/science/article/3299313/chinas-subsea-data-centre-could-power-7000-deepseek-conversations-second-report?

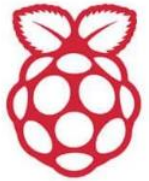
The China's underwater DC is equivalent of 30,000 high-end gaming computers operating simultaneously and could support 7,000 conversations per second with the Chinese AI chatbot DeepSeek, CCTV reported (19 Feb 2025)

INTRODUCTION: EDGE-AI CHALLENGES

Edge devices constraints:

limited in resources, form factor, computing capability, power budget, etc.





Hardware platforms:

- Raspberry Pi (3, 4, 5): Cortex-A+ generic GPU
- Jetson (Orin) nano: Cortex-A + CUDA cores GPU
- Google Coral (Mini/Micro): Cortex-A + dedicated GPU
- Intel Neural stick: Intel VPU (Vision)

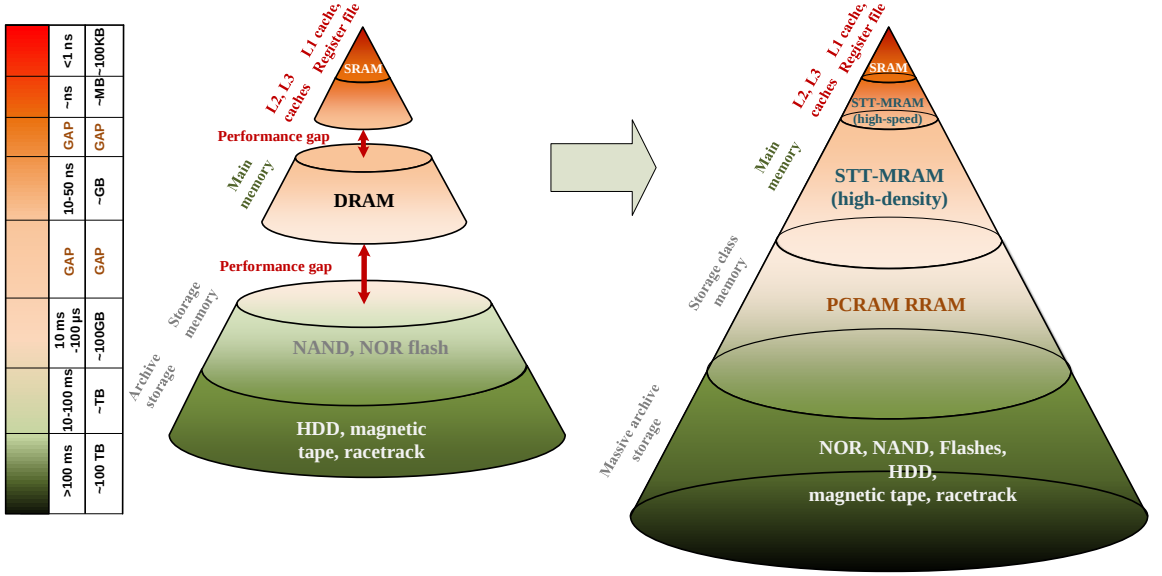
FACT

These devices are designed for inference, with limited network models

https://qengineering.eu/deep-learning-with-raspberry-pi-and-alternatives.html

Current computer-centric architecture limitation:

- Moving data is more expensive than processing it
- AI processing is both computing and memory-intensive



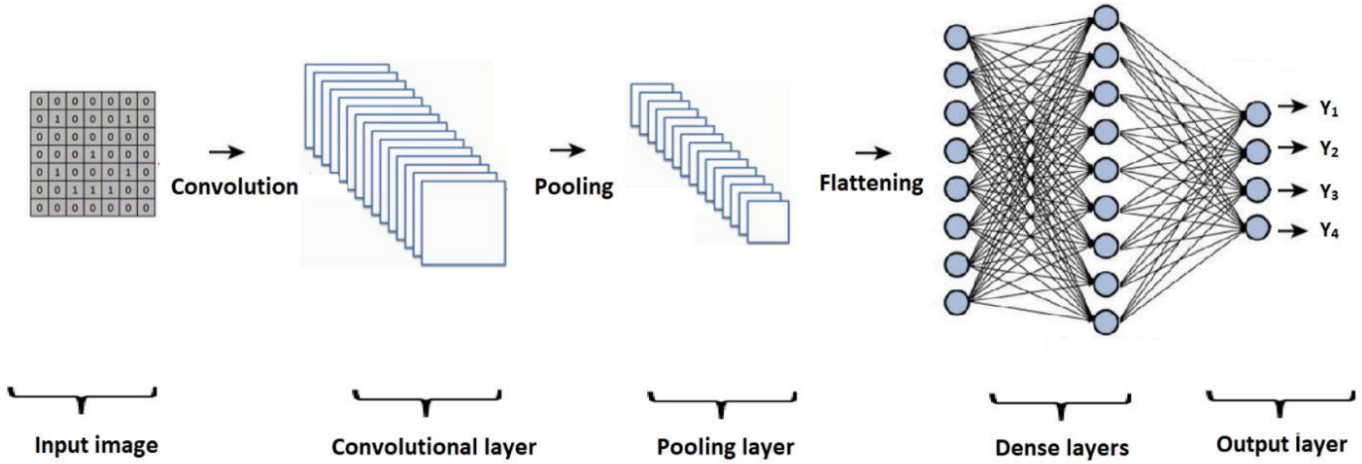
NUS scholar bank: Trinh Quang Kien Ph.D thesis

Ubiquitous Edge AI would need a more revolutionized approach

- Lightweight neural network to fit resources-constrained edge-devices
- Innovative approaches to solve the memory bottleneck

ARCHITECTURE APPROACH: DNN FOR EDGE AI?

Architecture	Model	Accuracy	# Params
AlexNet	AlexNet	89.67%	27.31M
VGG	VGG11	91.60%	14.50M
	VGG13	93.66%	14.68M
	VGG16	93.42%	20M
	VGG19	92.87%	25.31M
ResNet	ResNet18	92.36%	11.19M
	ResNet34	92.39%	21.31M
	ResNet50	92.04%	23.59M
	ResNet101	91.52%	42.66M
	ResNet152	91.30%	58.38M
DenseNet	DenseNet121	91.86%	3.27M
	DenseNet161	92.69%	12.30M
	DenseNet169	91.31%	5.99M
	DenseNet201	91.61%	8.5M
Inception	GoogleNet	86.91%	6.07M
	Inception V3	94.25%	19.33M



FACT

*(Conceptual) VGG16 occupied **500MB** (**140M** 32-bit FP parameters) and performs **16M** FP operations for an inference*

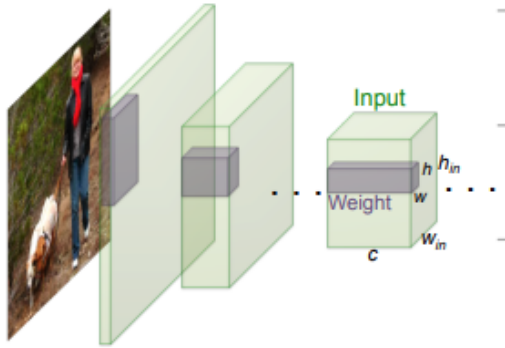
Optimizing/Simplifying the DNN

- **Network pruning (remove redundancy)**
- **Weight quantization (our work)**
- Low-rank parameter factorization (optimizing tensor/matrix operation)
- Compact convolutional filter (filter optimization)
- Knowledge distillation (convert to more compact networks with the similar performance)

- The DNN requires large memory allocation for the weight, on-chip memory is too small for such a configuration
- Current compute-centric architecture is not suited for such heavy floating point tensor operations

Haotong Qin, et all, Binary neural networks: A survey, Elsevier Pattern Recognition, Volume 105, 2020

ARCHITECTURE APPROACH: BINARIZING NETWORK



Network Variations	Operations used in Convolution	Memory Saving (Inference)	Computation Saving (Inference)	Accuracy on ImageNet (AlexNet)
Standard Convolution — Real-Value Inputs: 0.11 -0.21 … -0.34 / -0.25 0.61 … 0.52; Real-Value Weights: 0.12 -1.2 … 0.41 / -0.2 0.5 … 0.68	+ , − , ×	1x	1x	%56.7
Binary Weight — Real-Value Inputs: 0.11 -0.21 … -0.34 / -0.25 0.61 … 0.52; Binary Weights: 1 -1 … 1 / -1 1 … 1	+ , −	~32x	~2x	%56.8
BinaryWeight Binary Input (XNOR-Net) — Binary Inputs: 1 -1 … -1 / -1 1 … 1; Binary Weights: 1 -1 … 1 / -1 1 … 1	XNOR , bitcount	~32x	~58x	%44.2

BNN problems

- Accuracy degradation (inevitable)
- Unconventional training method required
  - Conversion method
  - Direct training using approximate gradients

BNN optimization

- Naive (original) BNN [BC, BNN]
- Minimizing quantization error [BWN, ABC-Net]
- Improve Network Loss function [DL-BNN, MD-Net]
- Reduce the gradient error [DSQ, IR-Net]

[BC] M. Courbariaux, et. al., Binaryconnect: Training deep neural networks with binary weights during propagations, NeurIPS, 2015
[BNN] I. Hubara et. al., Binarized neural networksNeurIPS, 2016
[XOR-net, BWN] M. Rastegari et. al., Xnor-net: Imagenet classification using binary convolutional neural networks, ECCV, 2016
[ABC-Net]X. Lin et. al., Towards accurate binary convolutional neural network, NeurIPS, 2017
[BNN-DL] R. Ding et. al., Regularizing activation distribution for training binarized deep networks, IEEE CVPR, 2019
[MS-Net] Y. Xu, et. al., A main/subsidiary network framework for simplifying binary neural networks, IEEE CVPR, 2019.
[DSQ] R. Gong, et. al., Differentiable soft quantization: Bridging full-precision and low-bit neural networks, IEEE ICCV, 2019.
[IR-Net] H. Qin et. al., Forward and backward information retention for accurate binary neural networks.

ARCHITECTURE APPROACH: BNN PERFORMANCE

Method	Bit-with (W/A)	Topology	Acc. (%)
Full precision	32/32	ResNet-18 [61]	69.6
	32/32	ResNet-34 [61]	73.3
	32/32	ResNet-50 [61]	76.0
	32/32	VGG-Variant [61]	72.0
	32/32	AlexNet [61]	57.1
BinaryConnect [59]	1/32	AlexNet	35.4
BNN [57]	1/1	AlexNet	27.9
BWN [58]	1/32	AlexNet	56.8
XNOR-Net [58]	1/1	AlexNet	44.2
DoReFa-Net [60]	1/1	AlexNet	43.6
ABC-Net [71]	1/32	ResNet-18	62.8
INQ [84]	2/32	ResNet-18	66.0
BNN-DL [85]	1/1	AlexNet	41.3
Main/Subsidiary Network [87]	1/1	ResNet-18	50.1
DSQ [63]	1/32	ResNet-18	63.7
IR-Net [96]	1/32	ResNet-18	62.9

Image Net benchmark

Method	Bit-with (W/A)	Topology	Acc. (%)
Full precision	32/32	ResNet-20 [61]	92.1
	32/32	ResNet-32 [92]	92.8
	32/32	ResNet-44 [92]	93.0
	32/32	VGG-11 [87]	83.8
	32/32	NIN [87]	84.2
BinaryConnect [59]	1/32	VGG-Small	91.7
BNN [57]	1/1	VGG-Small	89.9
BWN [58]	1/32	VGG-Small	90.1
DoReFa-Net [60]	1/32	ResNet-20	90.0
Main/Subsidiary Network [87]	1/1	VGG-11	82.0
	1/1	ResNet-18	86.4
BNN-DL [85]	1/1	VGG-Small	90.0
DSQ [63]	1/1	VGG-Small	91.7
IR-Net [63]	1/32	ResNet-20	90.2

Cifar-10 benchmark

Haotong Qin, et all, Binary neural networks: A survey, Elsevier Pattern Recognition,2020

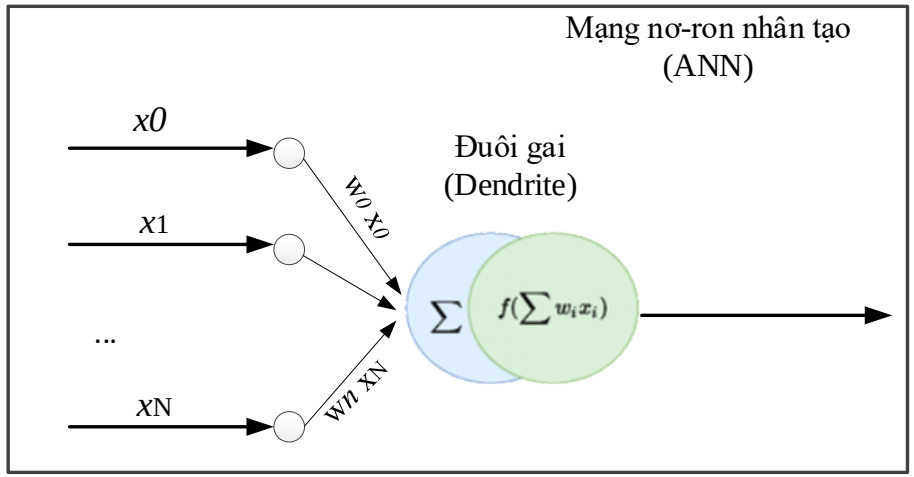
ARCHITECTURE APPROACH: BINARY SPIKING NETWORK

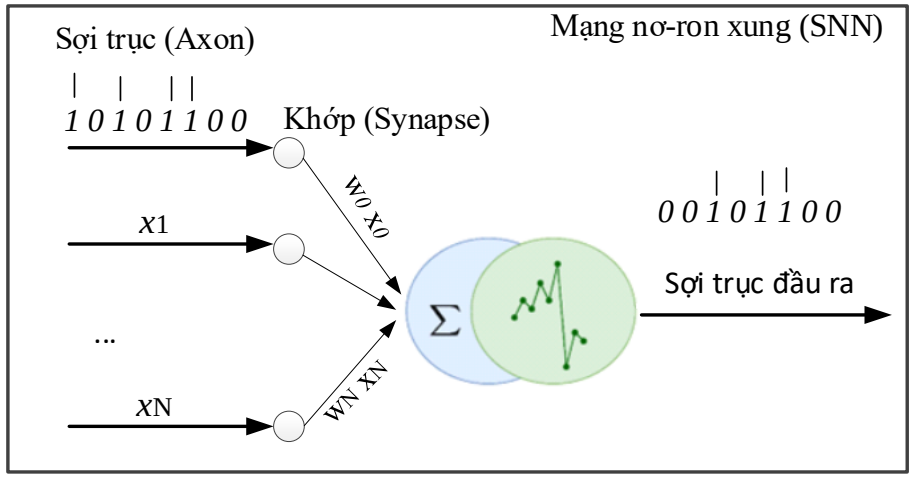
Spiking neural network (SNN)

- Third generation
- Data traversal in the format of spike
- Mimic better the brain function
- Simplify the data structure

Binary SNN

- The SNN with the weights are binary
- Inherit the advantage of BNN and the simplicity of data representation in SNN
- Fault tolerant (allow spike error)





FACT

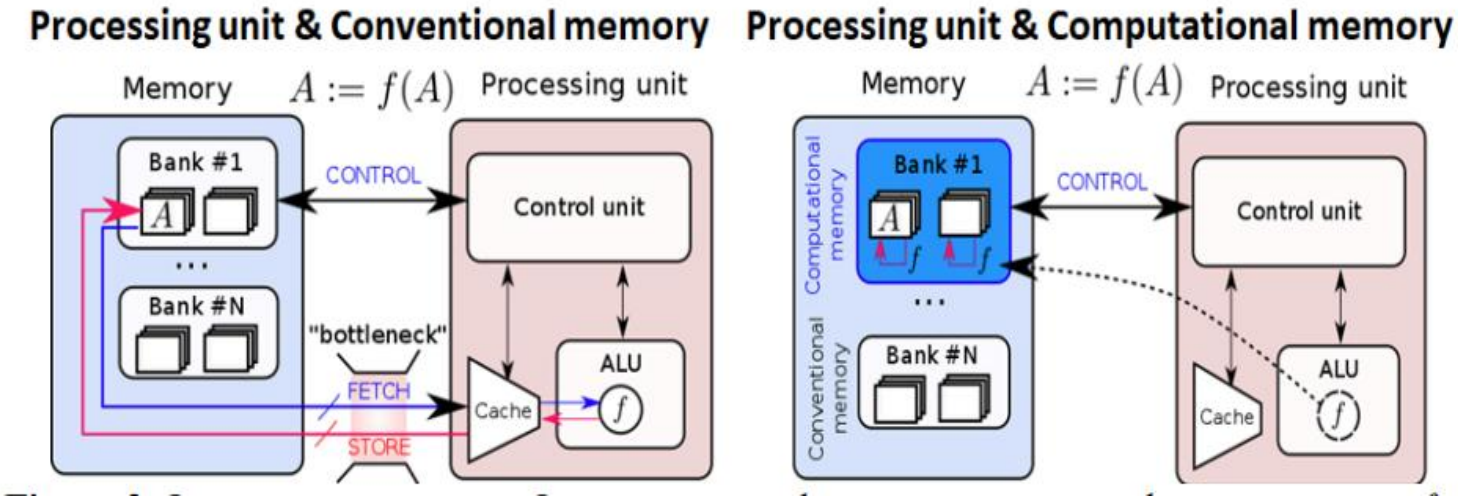
Brain (our) takes 12-25W only
(For the processing scale of 100M GB bytes of memory and a computing speed of 10^18 FLOPS. This is in the EXAFLOP range)

W. Maass. "Networks of spiking neurons: the third generation of neural network models." *Neural networks* 1997

HARDWARE ARCHITECTURE: IN-MEMORY COMPUTING

In-memory computing

- Eliminate the expensive data movement
- Solve the exponential growth in of data in modern applications



In-memory computing techniques

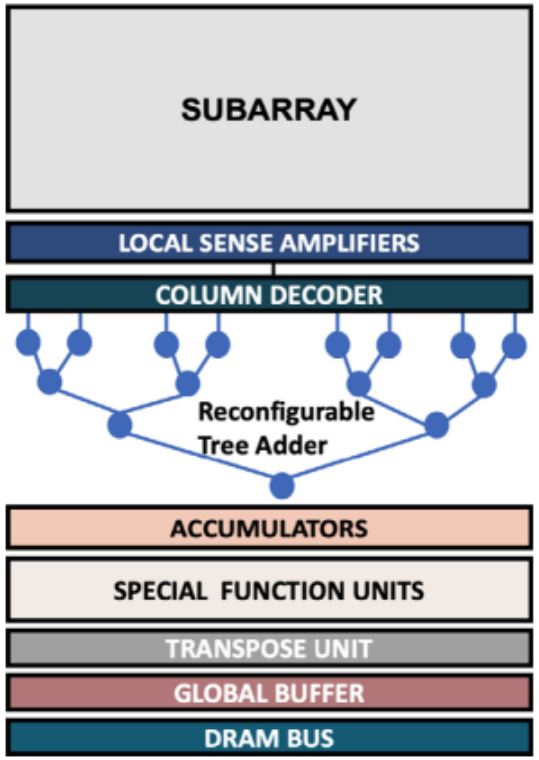
- **Processing at the data array (our works)**
- Processing at Row-buffer
- Processing at a unit near the memory banks (near-memory computing)

Memory Technologies

- DRAM/SRAM
- HBM/HMC
- Emerging resistive memories: STT-MRAM, PCM-RAM, ReRAM

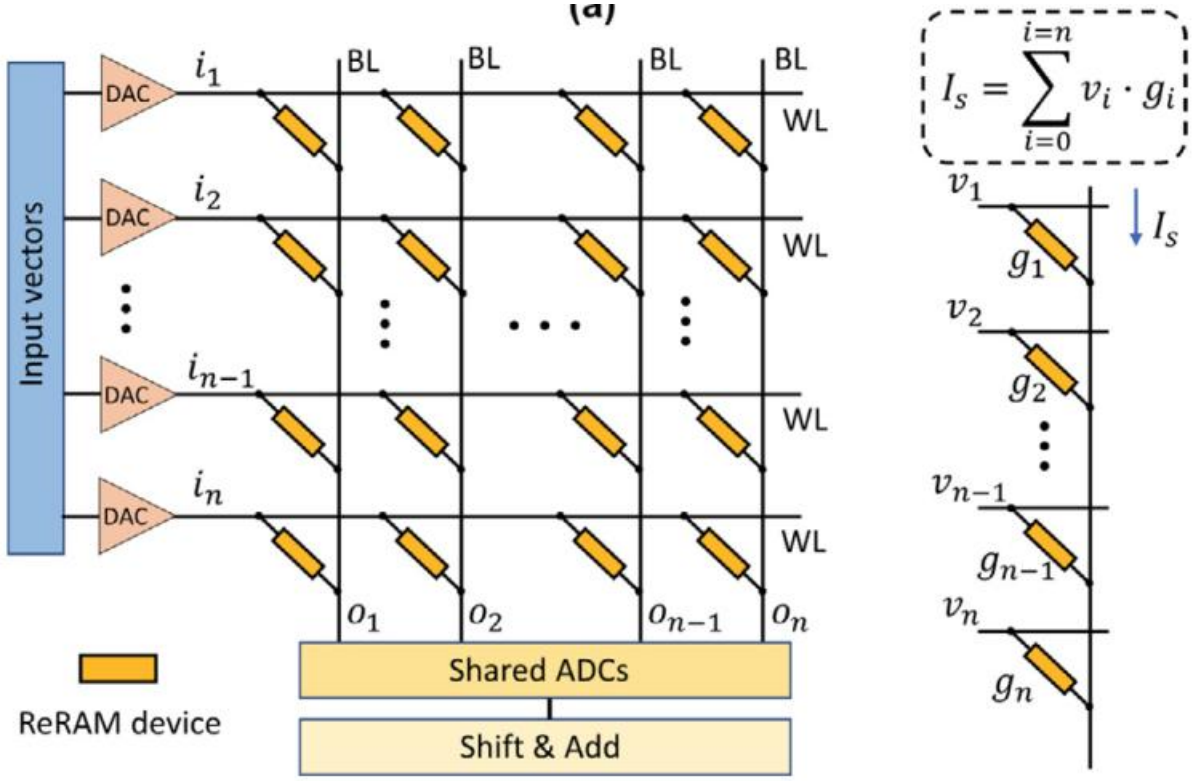
Mehonic, Adnan et. al. (2020). Memristors -- from In-memory computing, Deep Learning Acceleration, Spiking Neural Networks, to the Future of Neuromorphic and Bio-inspired Computing

HARDWARE ARCHITECTURE: IN-MEMORY COMPUTING



PIM-DRAM

- Processing at the row-buffer, not modify much the sub-array circuit

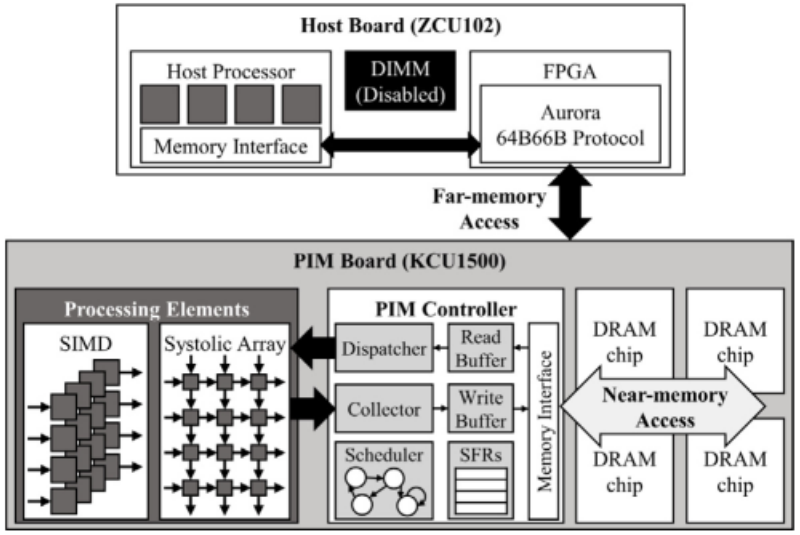


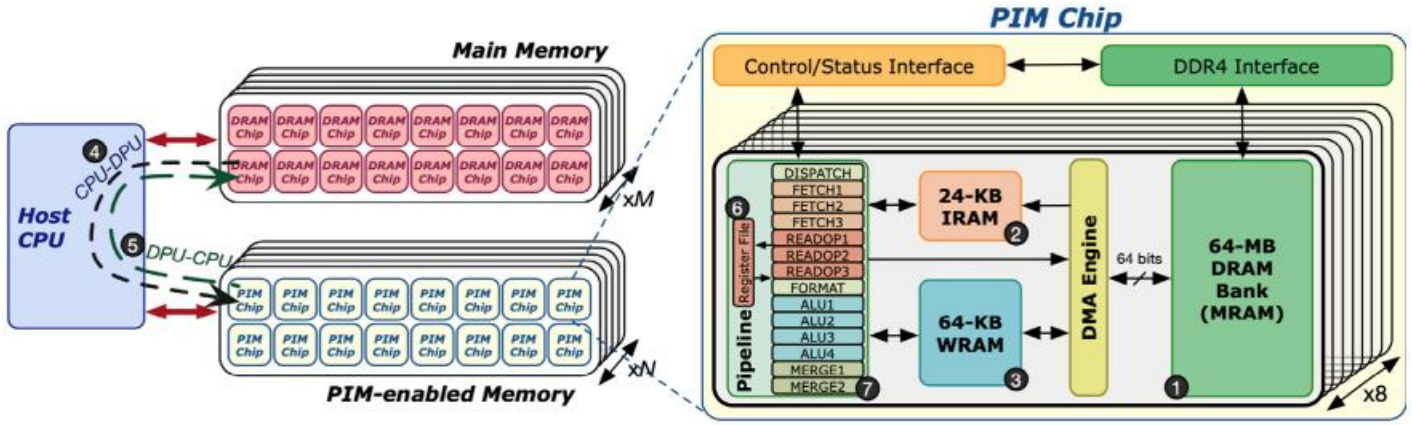
RERAM -crossbar

- Using simple DAC, ADC to convert between analog & digital domain
- The resulted current at each row represent the MAC

[RRAM] Y. Long, T. Na, S. Mukhopadhyay, ReRAM-based processing-in-memory architecture for recurrent neural network acceleration, IEEE Trans. Very Large Scale Integr. (VLSI) Syst. 26 (12) (2018) 2781–2794
[PIM-DRAM]

HARDWARE ARCHITECTURE: IN-MEMORY COMPUTING





PIMCaffe

PIM-emulating FPGA platform with SIMD and systolic array computing engines that can perform vector and matrix multiplication on the PIM device

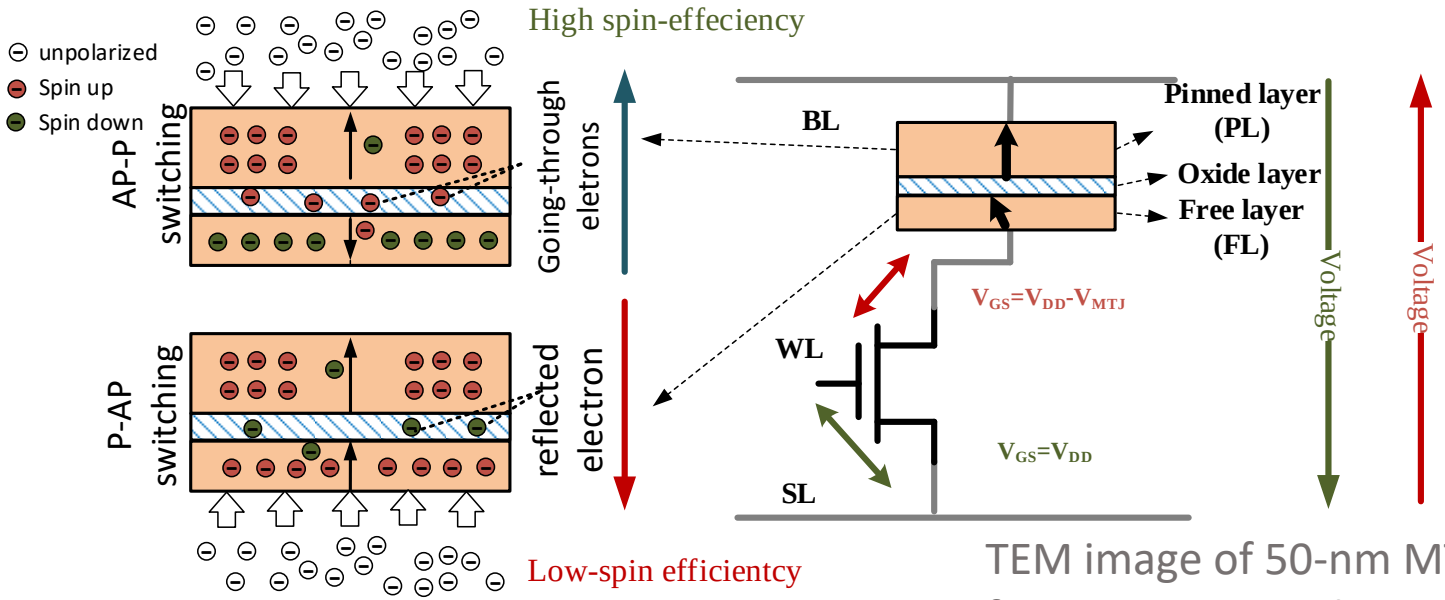
UPMEM

- PIM system with a host CPU, standard DRAM main memory, DDR4 DIMM with several PIM chips. Each PIM chip consists of 8 DRAM processing units (DPUs), and each DPU has access to a 64 MB DRAM bank

[PIMCaffe] W. Jeon, et. al.,, PIMCaffe: Functional evaluation of a machine learning framework for in-memory neural processing unit, IEEE Access, 2021

[UPMEM J. Gómez-Luna, et. Al., Benchmarking a new paradigm: An experimental analysis of a real processing-in-memory architecture, 2021

OUR WORK: PROPOSED STT-MRAM IMC CIRCUITS



TEM image of 50-nm MTJ from TDK Jan *et al.*, 2014

- STT-MRAMs: Spin-Transfer Torque Magnetoresistive Memories
  - employs Magnetic Tunnel Junction as memory element: high resistance state (1), low-resistance state (0)
  - the simplest bitcell: 1 transistor + 1 MTJ
  - Write (read) bitcell by injecting direct large (small) current

OUR WORK: BITCELL CIRCUIT CONFIGURATION

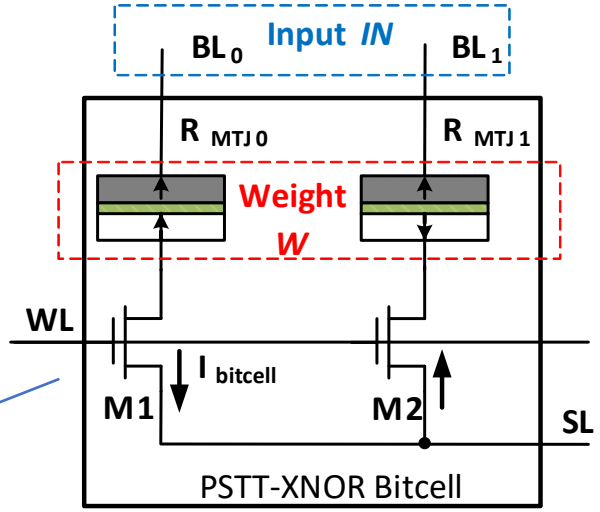
- **A dual complementary 2T2J bit-cell ⇔ a (synapse) weight**
  - (+1): $(R_0,R_1)=(R_P, R_{AP})$
  - (-1): $(R_0,R_1)=(R_{AP}, R_P)$
- **A dual complementary bit-lines (BL_0,BL_1) ⇔ an input neuron**
  - (+1): $(BL_0, BL_1)=(V_{BL}, 0)$
  - (-1): $(BL_0, BL_1)=(0, V_{BL})$
- **SL voltage ⇔ Output (XNOR product)**
  - (+1): high voltage
  - (-1): low voltage

Binary multiplication

$$I_{bitcell} = \frac{V_{BL}}{R_{MTJ,0} + R_{access,0} + R_{MTJ,1} + R_{access,1}} = \frac{V_{BL}}{\sum R}$$

$$V_{SL} = V_{BL} - I_{bitcell} * \left(R_{MTJ,0} + R_{access,0}\right) = \frac{V_{BL}}{\sum R}\left(R_{MTJ,1} + R_{access,1}\right)$$

$$SL_{margin} \sim \frac{V_{BL}}{\sum R}(R_{AP} - R_P) \sim TMR\,\frac{V_{BL}}{\sum R}R_P$$



Voltage generated a floating SL is (*IN* XNOR *W*)

MTJ_0, MTJ_1 / BL_0, BL_1	AP , P (-1)	P , AP (+1)
V_{BL} , 0 (+1)	Low (-1)	High (+1)
0 , V_{BL} (-1)	High (+1)	Low (-1)

OUR WORK: BINARIZED MAC COMPUTATION

- **Considered a row of M bitcells with merge SL terminal**

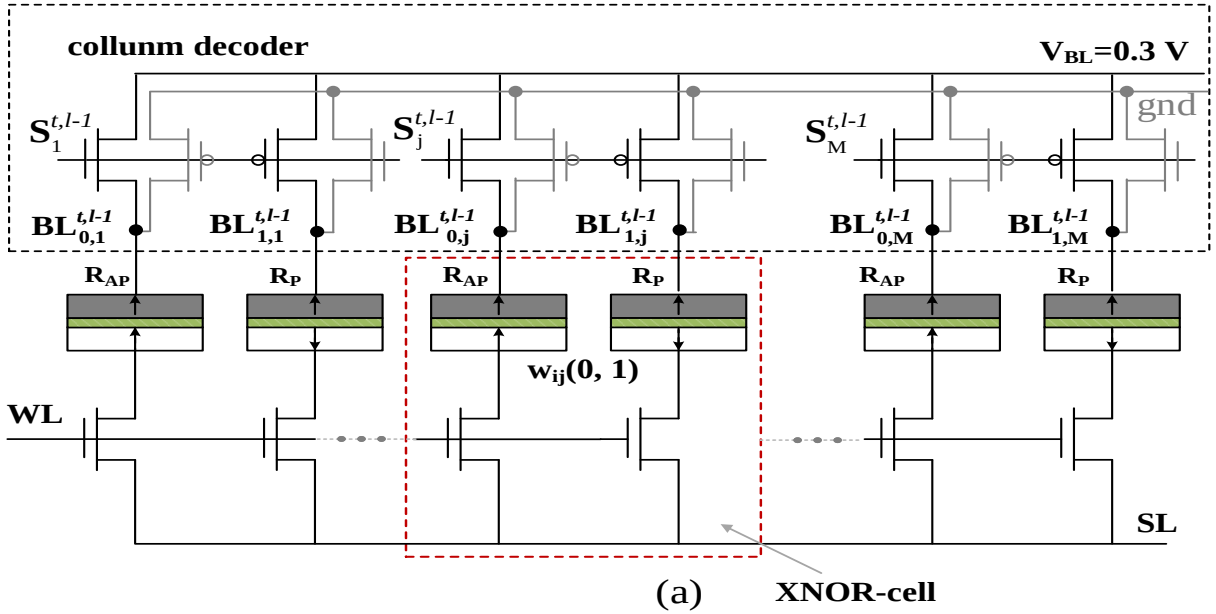
$$V_{SL} = V_{BL}\left(\frac{M-K}{M}\cdot\frac{R_{bitcell}}{R_{AP}+R_{access}} + \frac{K}{M}\cdot\frac{R_{bitcell}}{R_P+R_{access}}\right)$$
$$\sim constant * K$$

M number of cells in a row

$R_{bitcell}$ is the overall resistance seen from the SL,

K : number of XNOR outputs (i.e., $s_j^{t,l-1} \oplus w_{ij}$) equal to +1 across the entire row of M bitcells

- **V_{SL} linearly dependence on K, i.e., binarized MAC operation**

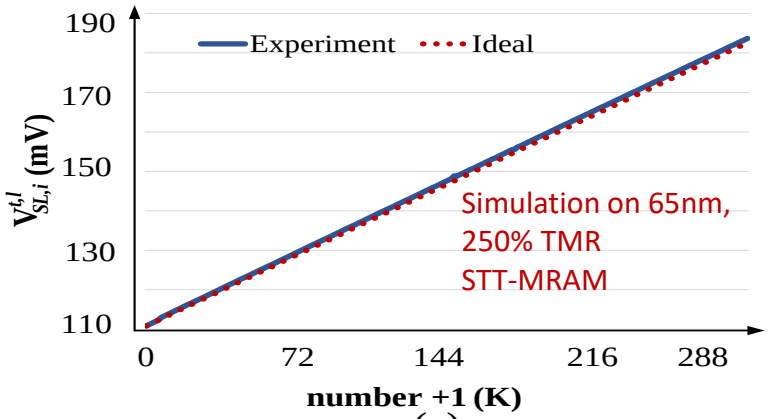


(a)

Equivalent to XNOR operation for a 2T-2R bitcell

MTJ_0, MTJ_1 / BL_0, BL_1	AP, P (0)	P, AP (1)
V_{BL}, 0 (1)	Low (0)	High (1)
0, V_{BL} (0)	High (1)	Low (0)

(b)



(c)

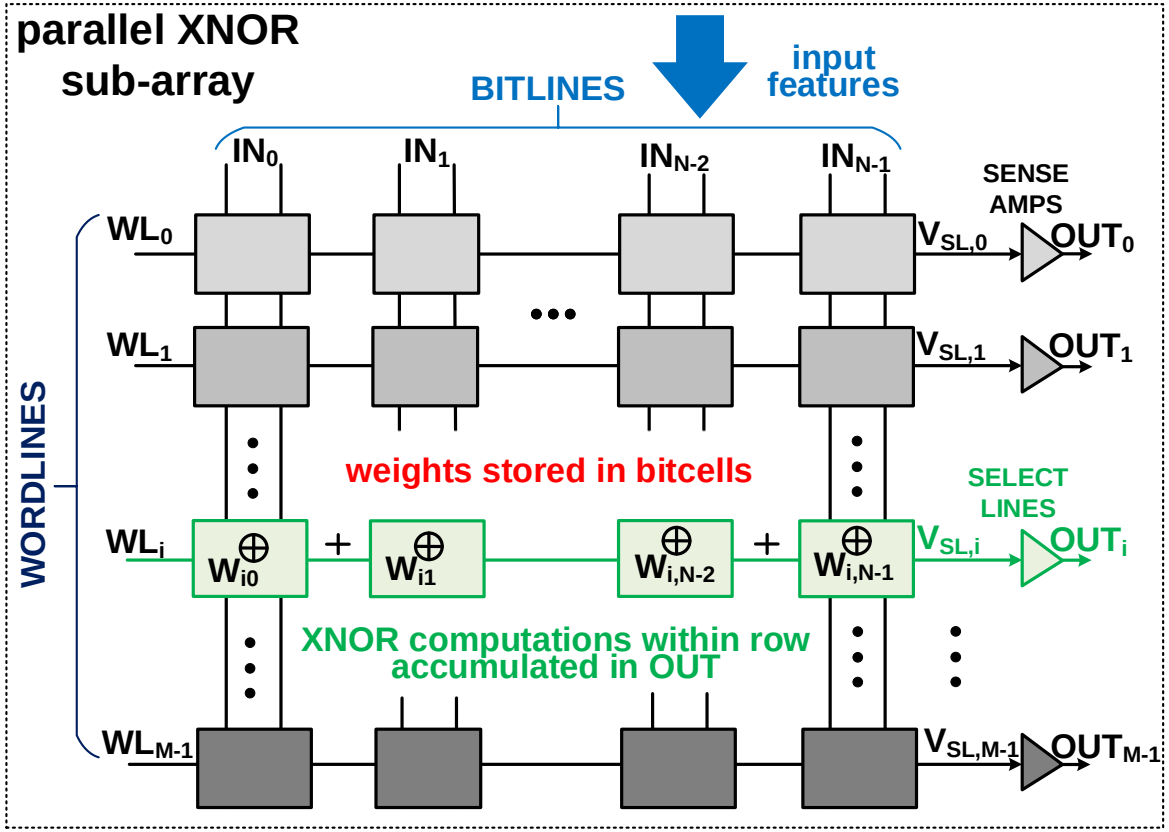
OUR WORK: ARRAY ORGANIZATION AND COMPUTATION

- **The logic output equals to the accumulation of the XNORs**

$$OUT_i = \sum_{j=0}^{N-1} \overline{W_{ij} \oplus IN_j}$$

- **The logic output is data-dependent**
  - (+1): $N_1 \geq \frac{N}{2}$
  - (-1): $N_1 < \frac{N}{2}$
- **Activating all WLs simultaneously => independently readout across all rows**

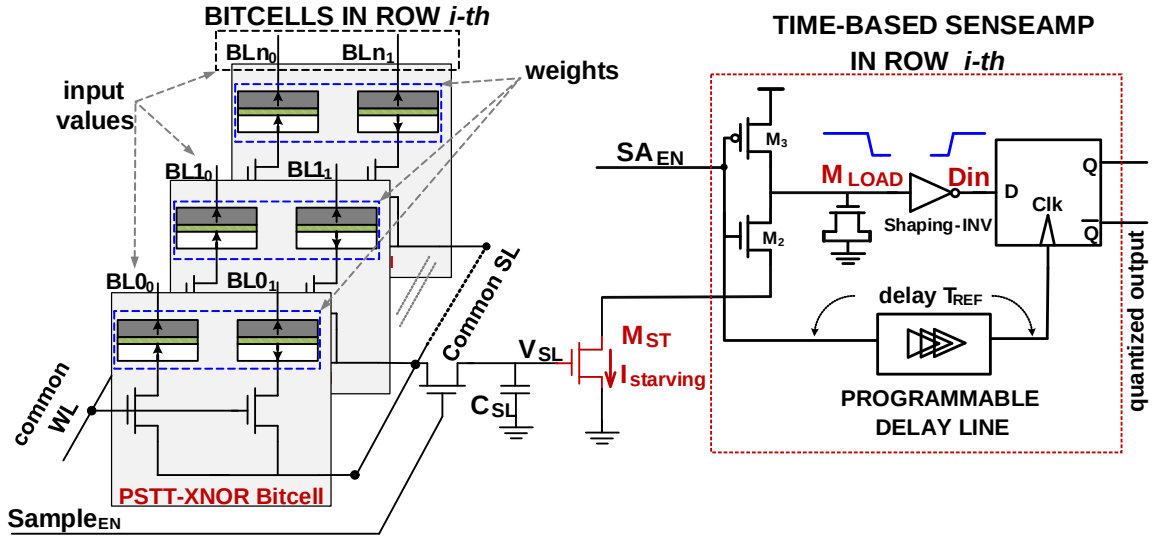
- **The maximum degree of accumulation-level parallelism of M => full utilization of the array**

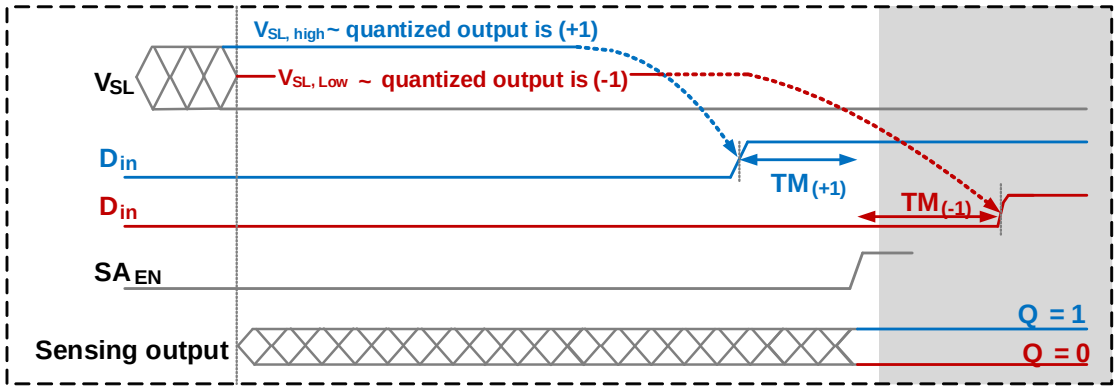


N is the XNOR vector size
N_1 is the number of XNOR resutls as (+1)

OUR WORK: CIRCUIT DESIGN FOR TIME-BASED SENSE AMPLIFIER

- **The V_{SL} voltage is converted to time domain**
  - TBS senseamp converts V_{SL} into $I_{starving}$
  - $I_{starving}$ is then converted to time
  - The discharge time at the LOAD node is inversely proportional to the discharge current $I_{starving}$, hence V_{SL}
- **Time-based comparator is a latch**
  => No analog reference needed
- **Key-design transistors**
  - M_{ST}: modulates the starving current $I_{starving}$
  - M_{LOAD}: adjusting output C_{LOAD}
- **The delay line has negligible variation**
- **Latency-efficient: sensing time ~10 ns**





OUR WORK: BNN-SYSTEM-LEVEL COMPARISON WITH PRIOR ARTS

TABLE IV. COMPARISON WITH PRIOR ART

	VLSI'21 [34]	DATE'18 [8]	IEDM'18 [15]	IEEE Trans. Magn.'18 [17]	TCAD'20 [18]	ISCAS'19 [35]	This work
memory type	SRAM	RRAM	STT-MRAM	STT-MRAM	STT-MRAM	STT-MRAM	STT-MRAM
technology	28 nm	65 nm	45 nm	45 nm	45 nm	22 nm	65 nm
single-memory access MAC operations	YES	YES	YES	NO	NO	YES	YES
neural network supported	high-precision DNN	BNN	BNN	XNOR-Net	binary-weight CNNs	high-precision DNN	BNN
activations / weights	5b /1b	1b/1b	1b/1b	1b/1b	1b/1b	4b/5b	1b/1b
sub-array size	1152x256	128x128	128x256	NA	512x256	64x576	128x128
DAC required	YES	YES	NO	NO	NO	YES	NO
max throughput (TOPS)	6.144	NA	NA	NA	NA	NA	3.28
max energy efficiency (TOPS/W)	5796	141	NA	0.0169	0.455	NA	311 [a] (319 [b])
accuracy in MLP-MNIST / CNN-CIFAR-10	NA / 91.1%	98.43% / 86.08%	< 95% / N/A	NA/ NA	NA/ NA	98%/ 91%	98.42% / 80.01%

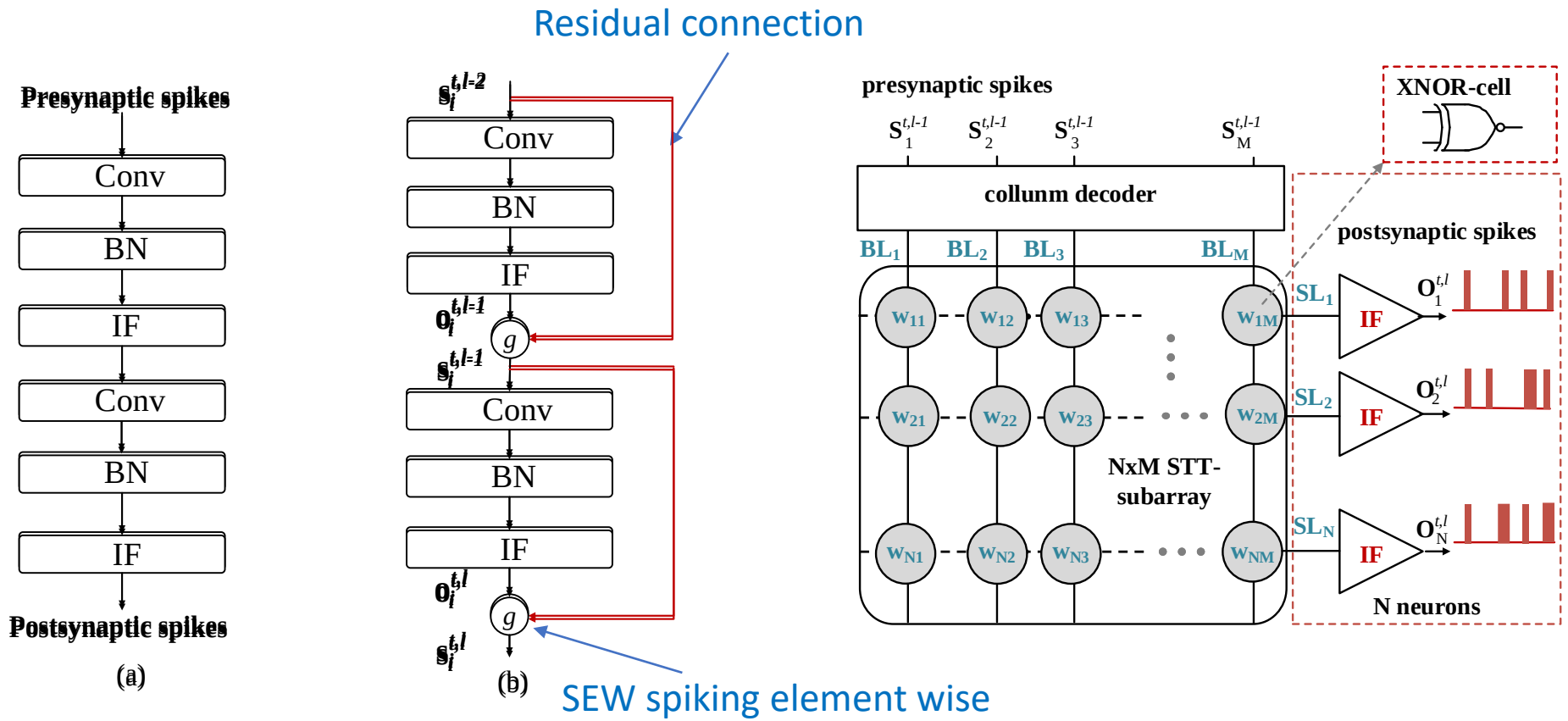
[a] This work with second-order SL boosting

[b] This work without SL boosting

[JETCAS-BNN] T. N Pham, Q. K. Trinh, Ik-Joon, and Massimo Alioto, "STT-BNN: A Novel STT-MRAM In-Memory Computing Macro for Binary Neural Networks", in IEEE Journal on Emerging and Selected Topics in Circuits and Systems,, June 2022.

[ISCAS-BNN] T. -N. Pham, Q. -K. Trinh, I. -J. Chang and M. Alioto, "STT-MRAM Architecture with Parallel Accumulator for In-Memory Binary Neural Networks," *2021 IEEE ISCAS*, Daegu, Korea, 2021.

OUR WORK#2: BSNN AND IMC USING STT-MRAM



STT-BSNN: An In-Memory Deep Binary Spiking Neural Network Based on STT-MRAM VT Nguyen, QK Trinh, R Zhang, Y Nakashima - IEEE Access, 2021

Q-M. Duong, Q-K. Trinh, V-T. Nguyen, Đ-H. Đao, D-M. Luong, V-P. Hoang, J. Deepu, L. Lin, "A Low-Power Charge-Based Integrate-and-Fire Circuit for Binarized-Spiking Neural Network", on International Journal of Circuit Theory and Applications

OUR WORK#2: BSNN SOLUTION: BSNN BASED ON STT-MRAM

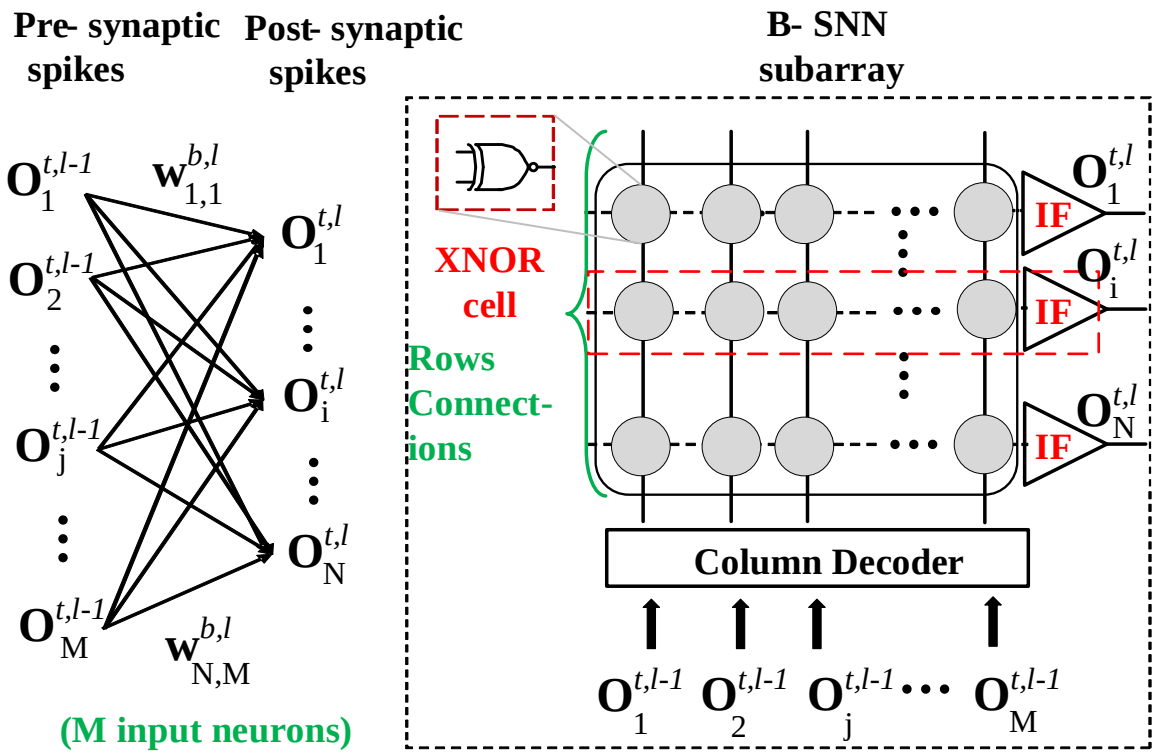
- **Integrated and fire model**

$$Integration: \begin{cases} \hat{u}_i^{t,l} = \hat{u}_i^{t-1,l} + \sum_{j=1}^{M} \overline{w_{ij}^{u,l} \oplus o_j^{t,l-1}} - \rho \\ \hat{\theta}_i^l = \frac{\sigma}{\alpha} \cdot \theta_i^l \end{cases}$$

$$Firing: o_i^{t,l} = \begin{cases} 1, & if\ \hat{u}_i^{t,l} > \hat{\theta}_i^l \\ 0, & otherwise \end{cases}$$

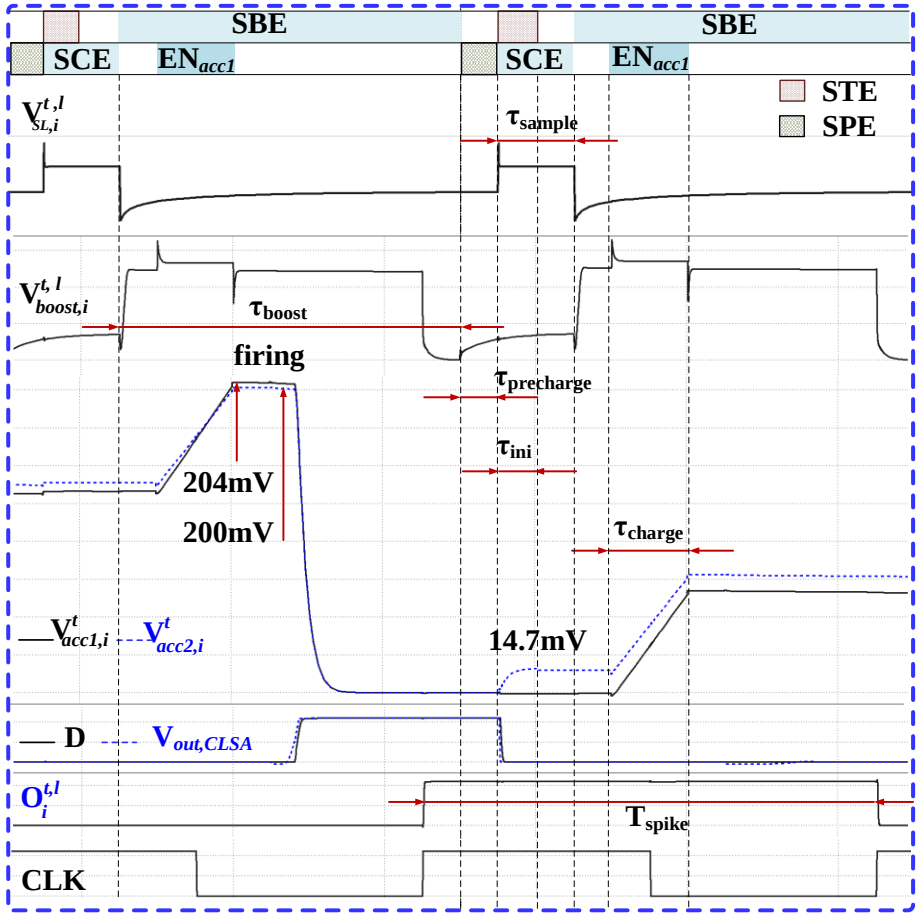
$$Resetting: \hat{u}_i^{t,l} = 0$$

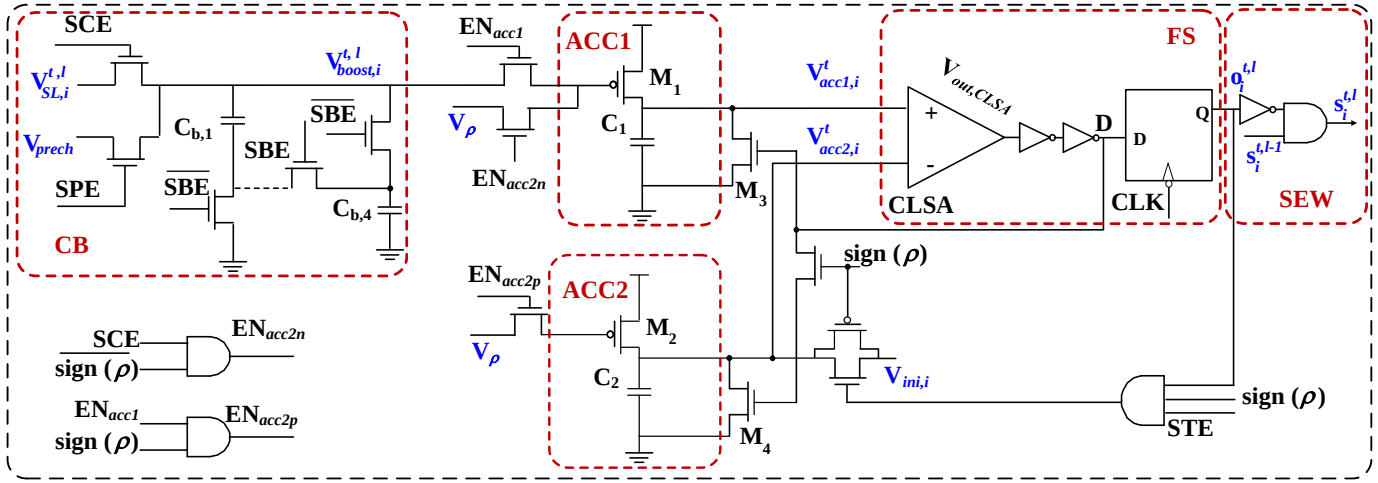
- **Required binarized MAC operation (based on XNOR cell) and constant subtraction**



STT-BSNN: An In-Memory Deep Binary Spiking Neural Network Based on STT-MRAM VT Nguyen, QK Trinh, R Zhang, Y Nakashima - IEEE Access, 2021

OUR WORK#2: DYNAMIC THRESHOLD IF CIRCUIT





Integration:

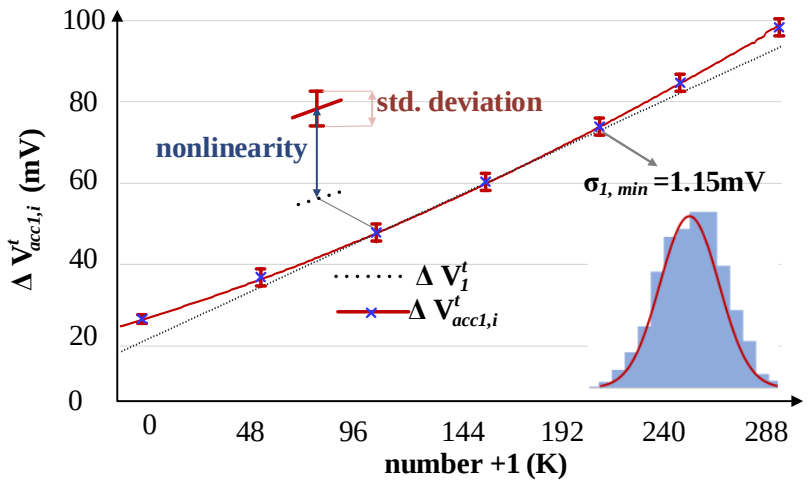
$$\begin{cases} \hat{u}^{t,l}_{xnor,i} = \hat{u}^{t-1,l}_{xnor,i} + \sum_{j=1}^{M} \overline{w^{u,l}_{ij} \oplus s^{t,l-1}_{j}} \\ \hat{\theta}^{t,l}_{dyn,i} = \hat{\theta}^{t-1,l}_{dyn,i} + \rho, \qquad \hat{\theta}^{t=0,l}_{dyn,i} = \hat{\theta}^{l}_{i} \end{cases}$$

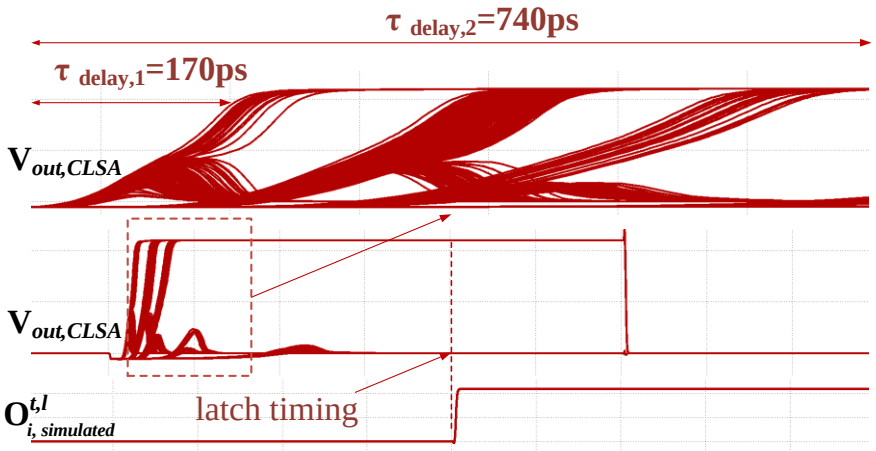
was here

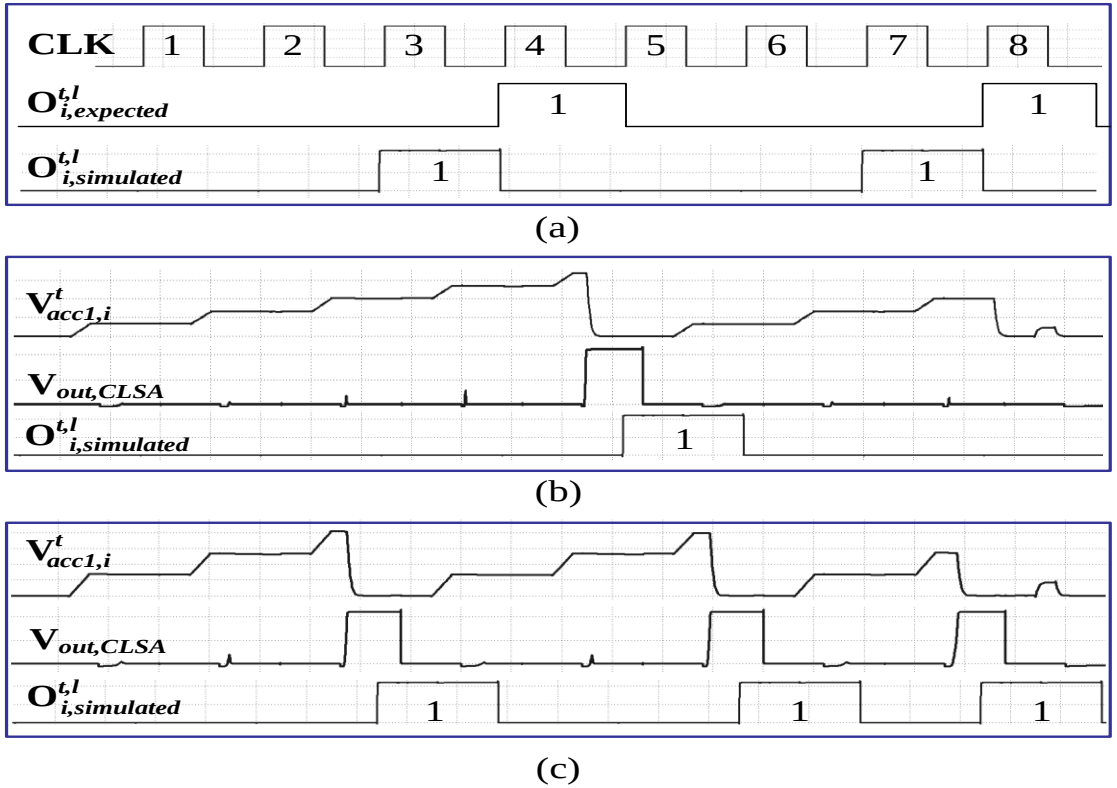
$$Firing: o^{t,l}_i = \begin{cases} 1, & if\ \hat{u}^{t,l}_{xnor,i} > \hat{\theta}^{t,l}_{dyn,i} \\ 0, & otherwise \end{cases}$$

$$Resetting: \hat{u}^{t,l}_{xnor,i} = 0\ and\ \hat{\theta}^{t,l}_{dyn,i} = \hat{\theta}^{l}_{i}$$

OUR WORK#2: DYNAMIC THRESHOLD IF: NONIDEALITY IMPACT



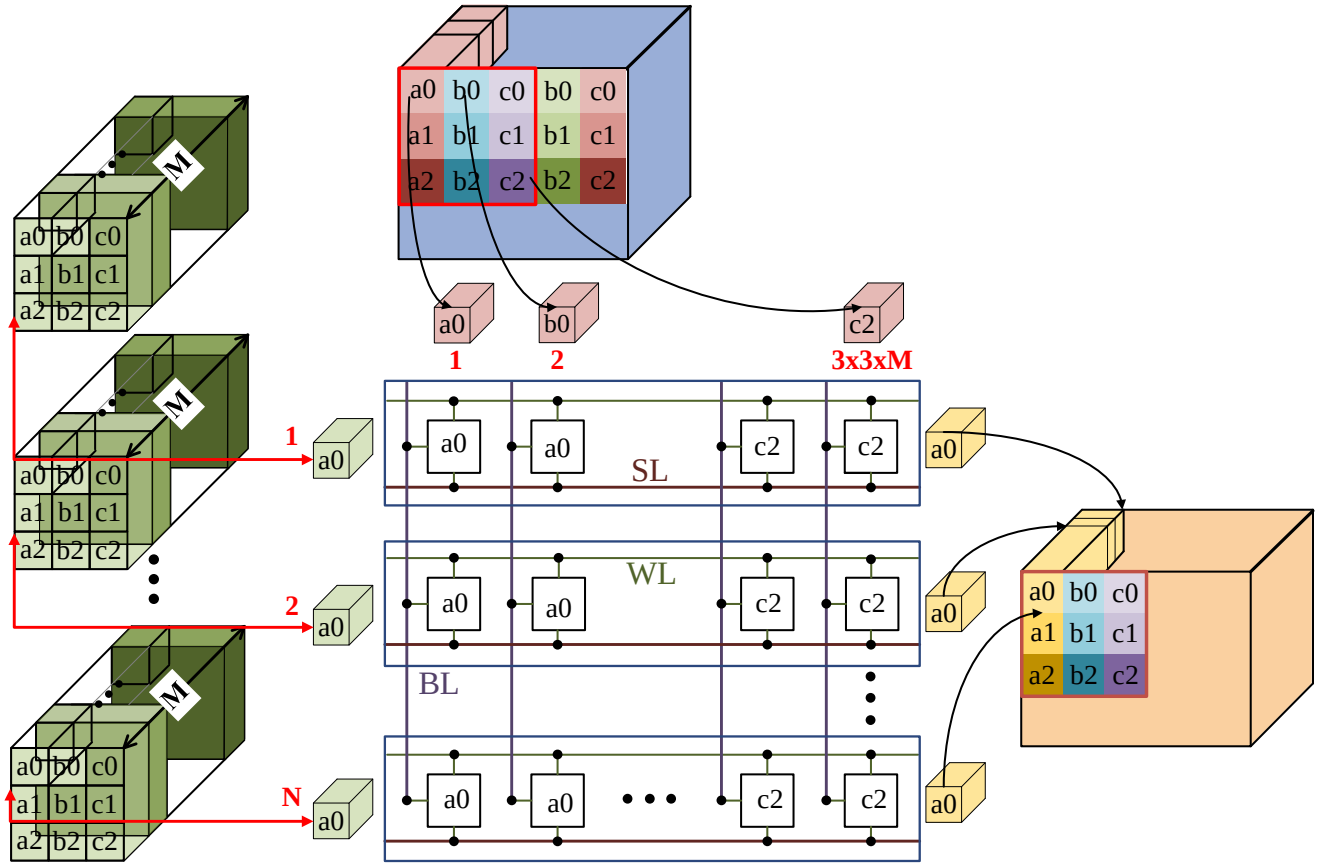




- **The V_{SL} voltage is accumulated every time step and is stored in C_1**
- **The constant subtraction is done via C2 by a self termination discharging circuit (controled by M_{s5})**

OUR WORK#2: MAPPING IMC MACRO TO BSNN MODEL

- **Row size equal to kernel size multiplied by number of input feature map**
- **Spike travels across the network continuously**
- **First and last layer may require full precision (rather than binarized) for adequate classification accuracy**



Example of mapping BSNN circuit macro to a convolution layer with M feature map

MAPPING IN-MEMORY MACRO TO BSNN MODEL

	IJCNN'19 [16]	TNNSL'20 [17]	VLSI'20 [19]	Our wwork BSNN ACCESS'21*	Our work BSNN* (Wiley ITC)
synapse	MTJ	MTJ	MTJ	MTJ	MTJ
neuron	MTJ	Digital	MTJ	Analog	Analog
technology	45 nm	28 nm	N/A	65 nm	65 nm
network type	BSNN	SNN	BSNN	BSNN	BSNN
structure	3 Conv	FC layers	2 Conv	2/7 Conv	7 Conv
neuron	sigmoid	IF	Possion	IF	IF
weights	+1/-1	+1/0[c]	+1/-1	+1/0	
spiking rate (MHz)	N/A	83	0.1	166	**285**
energy/ synapse (fJ)	36[b]	8.87	N/A	5.48	**5.10**
area/neuron (F^2)[a]	N/A	~15×10^5	6×10^3	32×10^3	19×10^3
accuracy MNIST/ CIFAR-10	N/A/ 70.3%	91.5%/ N/A	~97.4%/ N/A	97.92%/ 83.85%	N/A/82.01% (σ_{BSNN} = 0.29%)

[a]The area is normalized to F^2, with F is the technology feature size.

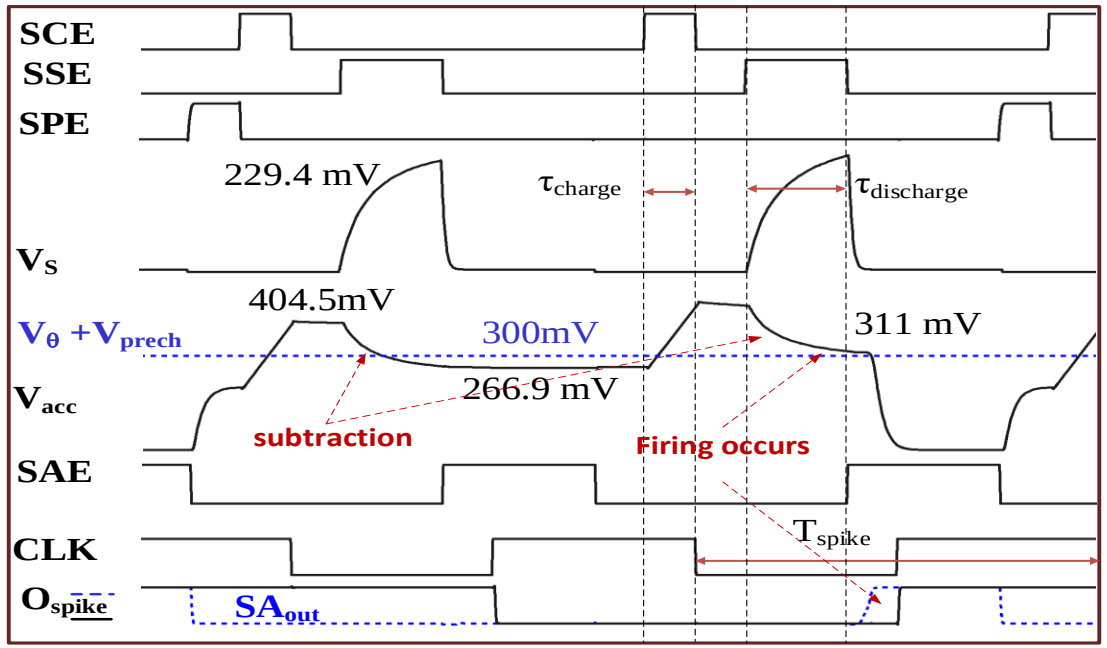
[b]The maximum energy consumption per spiking event for a synapse, as reported in [39], [40].

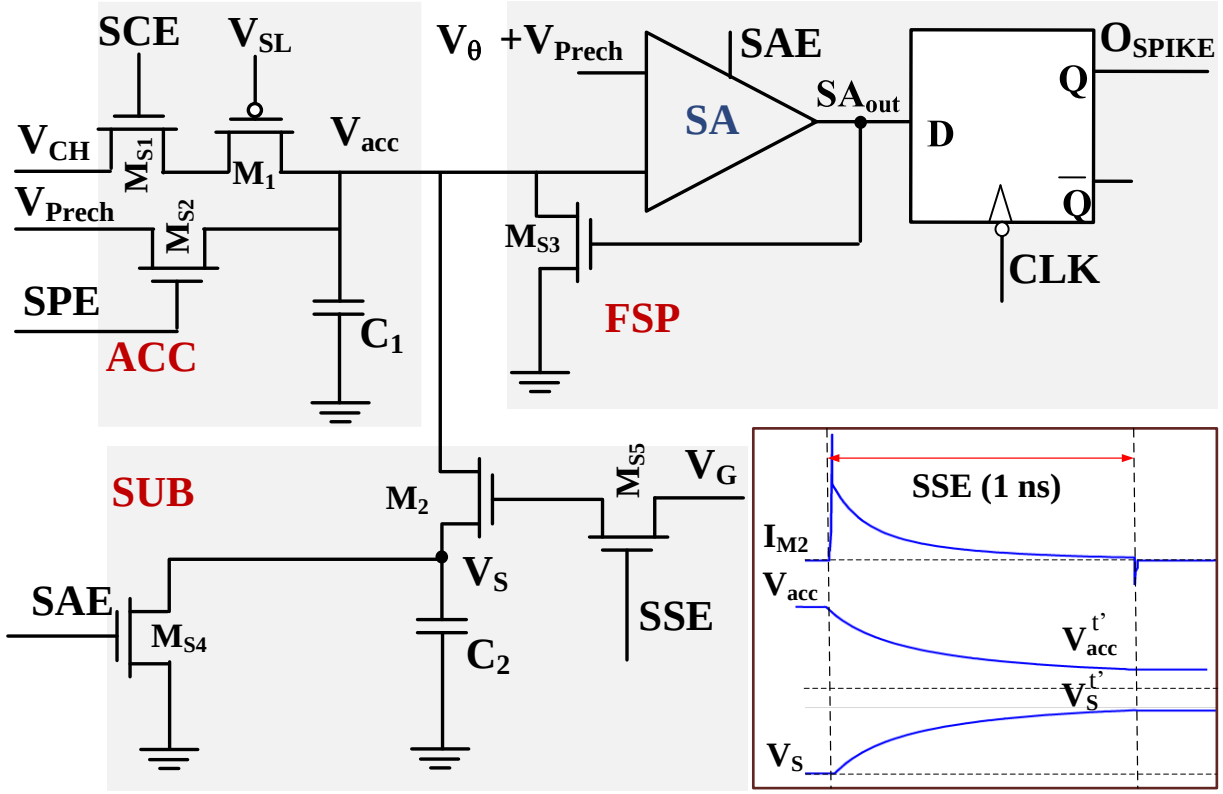
[c]The full-precision weights are converted into stochastic bits (+1/0) in each time step.

STT-BSNN: An In-Memory Deep Binary Spiking Neural Network Based on STT-MRAM VT Nguyen, QK Trinh, R Zhang, Y Nakashima - IEEE Access, 2021

OUR WORK#2: A LOW-POWER CHARGE-BASED IF CIRCUIT FOR BSNN

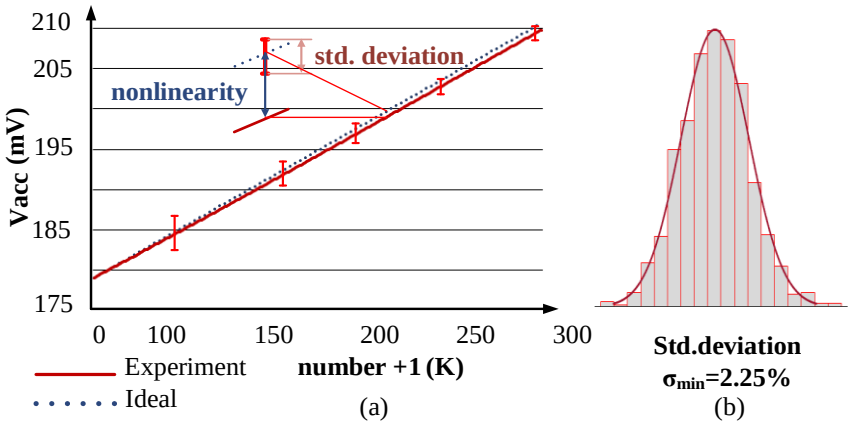
- The V_{SL} voltage is accumulated every time step and is stored in C_1
- The constant subtraction is done via C2 by a self termination discharging circuit (controlled by M_{s5})

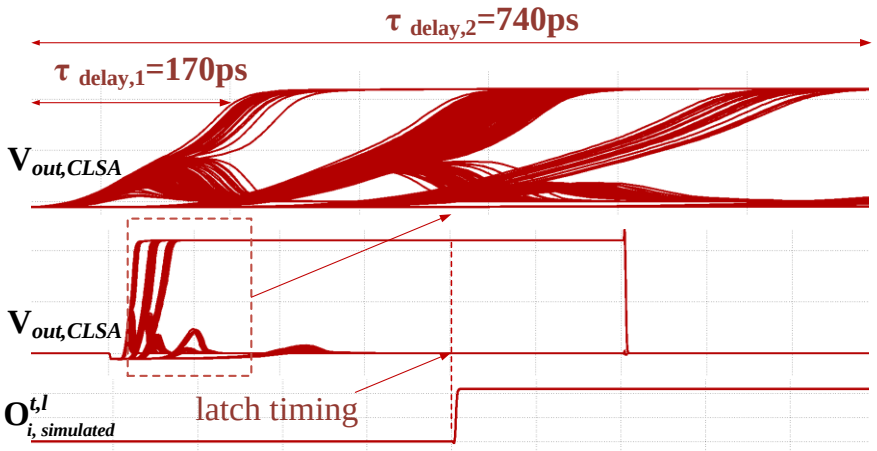


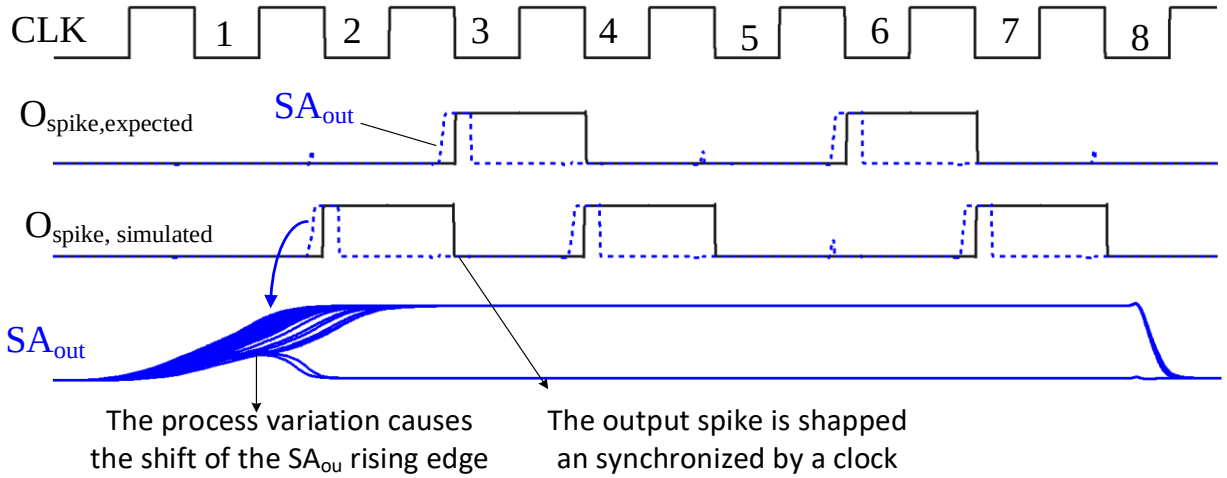


IF circuit and example simulation waveform (65 nm CMOS with STT-MRAM (250% TMR)

OUR WORK#2: SUBTRACTION IF CIRCUIT: NON-IDEALITY AND ERROR TOLERANT



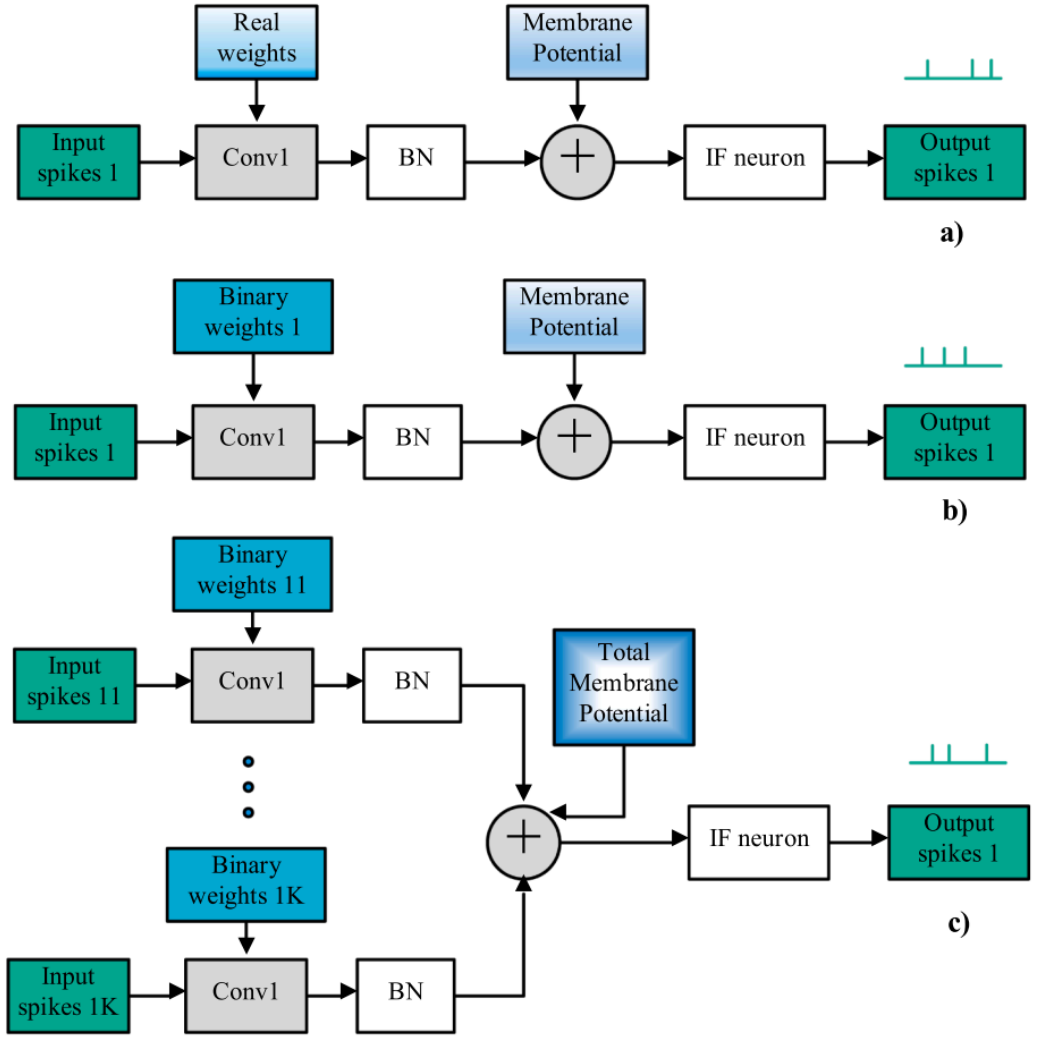




- **Similar to the Dynamic IF circuit, this IF is prone to process variations and other device non-ideality effect.**

OUR WORK#3: NUTS-BSNN - FULLY BINARIZED SNN

- ✓ **SNN** with real weights
- ✓ **FBW-BSNN** with binary weights but the first layer using the real weights
- ✓ **NUTS-BSNN:** we successful binarize all layer of the BSNN without accuracy degradation



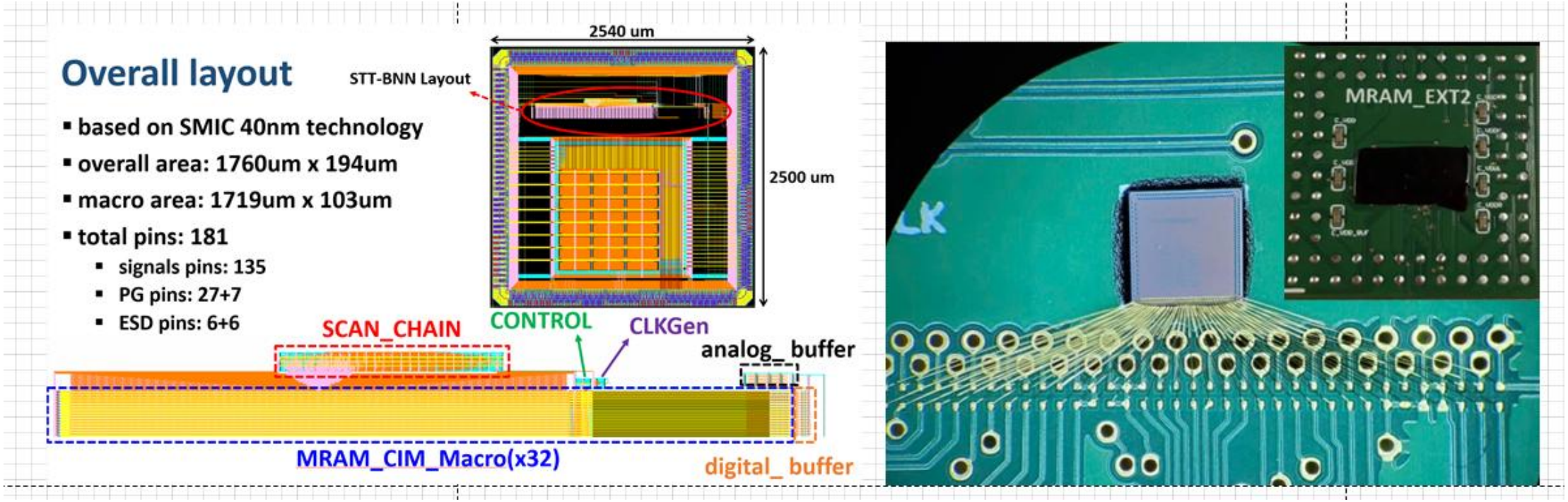
[FWB-SNN] V. -N. Dinh, et. al., "FBW-SNN: A Fully Binarized Weights-Spiking Neural Networks for Edge-AI Applications". ICICDT, 2022

[NUTS-BSNN]. V. -N. Dinh, , et. al., “NUTS-BSNN: A Non-uniform Time-step Binarized Spiking Neural Networks with Energy-Efficient In-memory Computing Macro”. Neurocomputing,, 2023

OUR WORK#2: NUTS-BSNN COMPARISON

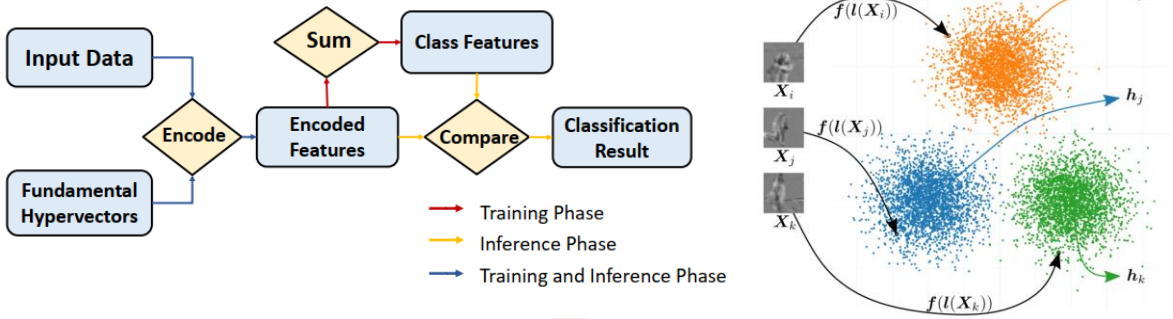
Mạng nơ-ron	Trọng số lớp đầu và đầu ra	Phương pháp huấn luyện	Time steps	Độ chính xác (%)
Fashion-MNIST				
BS4NN [73]' NPL 22	Giá trị nhị phân	Huấn luyện trực tiếp	256	87,30
FBW-SNN [CT2]	Giá trị nhị phân	Huấn luyện trực tiếp	14	91,49
BSNN [This work]	Giá trị thực	Huấn luyện trực tiếp	14	92,58
NUTS-BSNN (K = 5) [CT4]	**Giá trị nhị phân**	**Huấn luyện trực tiếp**	**14**	**93,25**
CIFAR-10				
BSNN [69]' TCDS 20	Giá trị thực	Chuyển đổi	100	90,19
BSNN [70]' ICSICT 20	Giá trị thực	Chuyển đổi	150	80,52
ReStoCNet [72]' FN 19	Giá trị nhị phân	Hybrid-STDP	500	66,23
STT-BSNN [40]' IA 21	Giá trị thực	Huấn luyện trực tiếp	8	83,85
FBW-SNN [CT2]' ICICDT 22	Giá trị nhị phân	Huấn luyện trực tiếp	14	82,86
BSNN [this work]	Giá trị thực	Huấn luyện trực tiếp	14	84,37
NUTS-BSNN (K = 6) [CT4]	**Giá trị nhị phân**	**Huấn luyện trực tiếp**	**14**	**88,71**
CIFAR-100				
B-SNN [71]' FN 20	Giá trị nhị phân	Chuyển đổi	148	62,71
BSNN [69]' TCDS 20	Giá trị thực	Chuyển đổi	300	62,02
B-SNN [91]' HPBD&IS 21	Giá trị thực	Huấn luyện trực tiếp	8	59,11
BSNN [this work]	Giá trị thực	Huấn luyện trực tiếp	14	61,49
NUTS-BSNN (K = 6) [CT4]	**Giá trị nhị phân**	**Huấn luyện trực tiếp**	**14**	**70,31**

OUR WORK#3: EDGE-AI STT-MRAM IMPLEMENTATION



OUR WORK#4: HYPERDIMENSIONAL COMPUTING

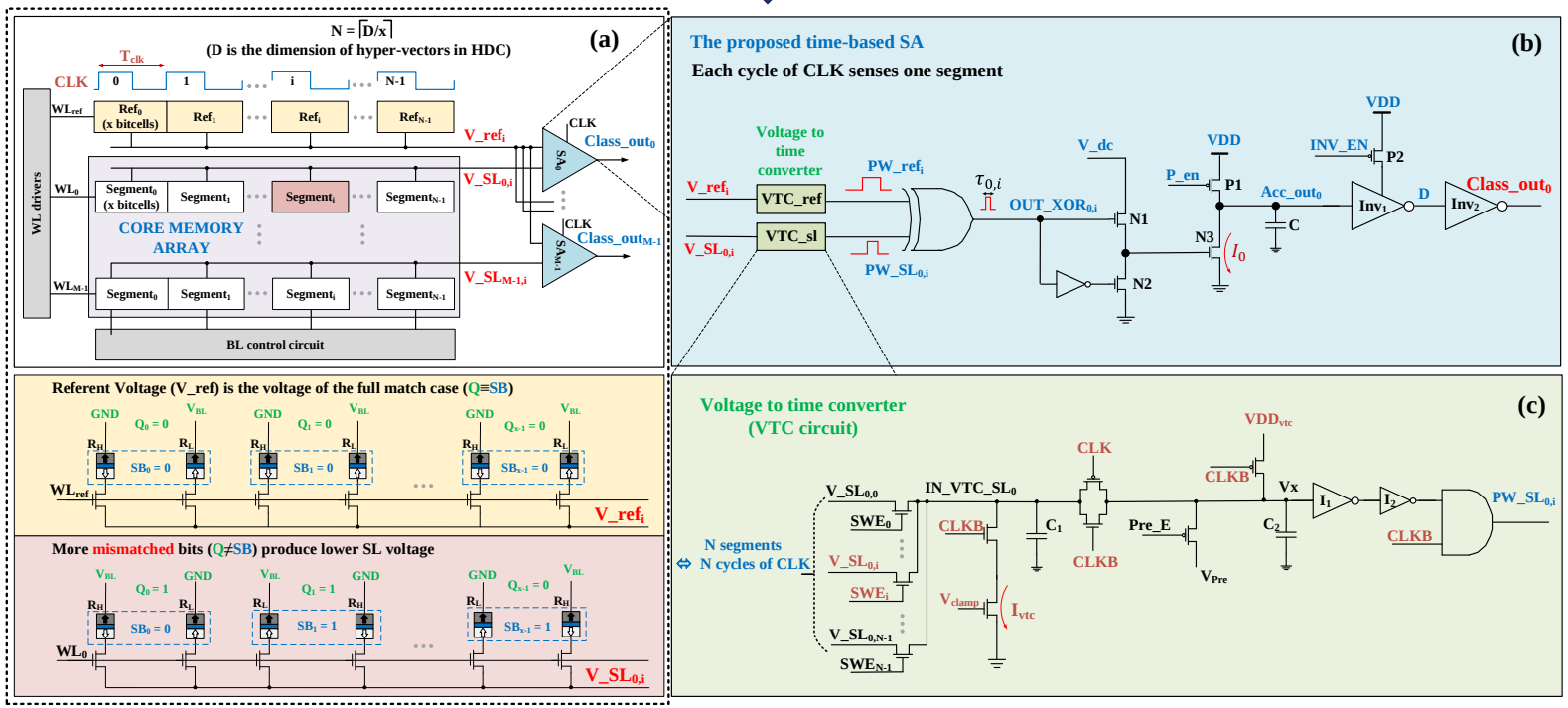
HDC: an alternative to neural network with straightforward and fast training



Segmentation:

To handle the long vector calculation

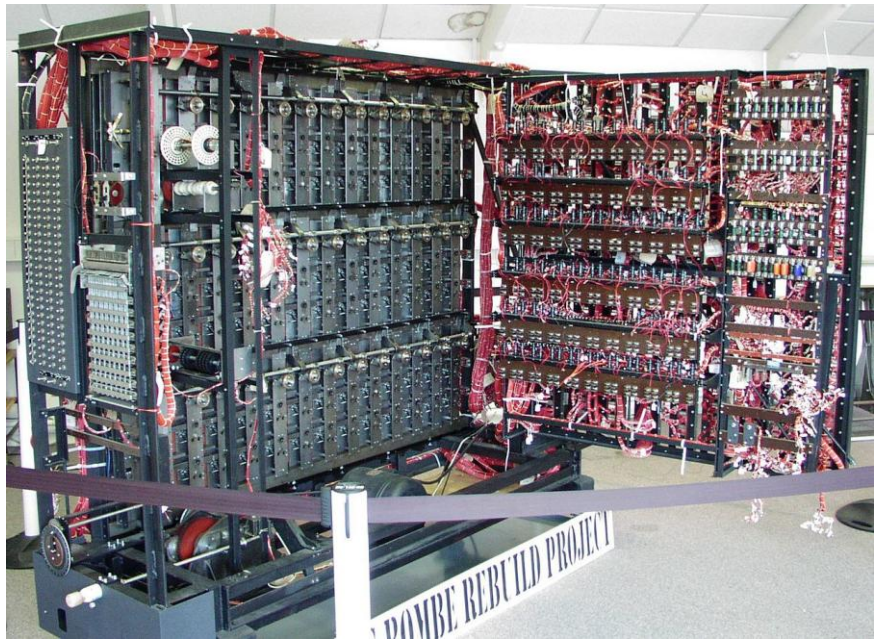
In-memory computing architecture: massive parallel computational macro



Analog-based computational circuit for ultra-low energy consumption

[FWB-SNN] T.N Pham, Quang-Kien Trinh et. Al, "STT-HDC: An Efficient Time-domain In-memory Hyper-dimensional Computing Design Based on STT-MRAM", IEEE ACCESS, 2025

The first computers



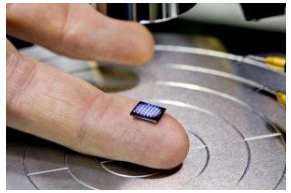
BOMBE (1940, Alan Turing): the first electro-mechanical computer for decrypting the Germany Enigma message. £100,000, 2.1x1.98x0.61 m^3,1 ton. Each bombe 08 mounted drums, which were in three groups of 12 triplets.

https://en.wikipedia.org/wiki/Bombe
https://vi.wikipedia.org/wiki/ENIAC

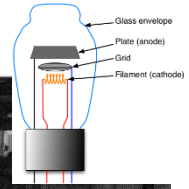
How to make computer smaller, faster, more reliable and more energy efficient?

Like this?

or this?



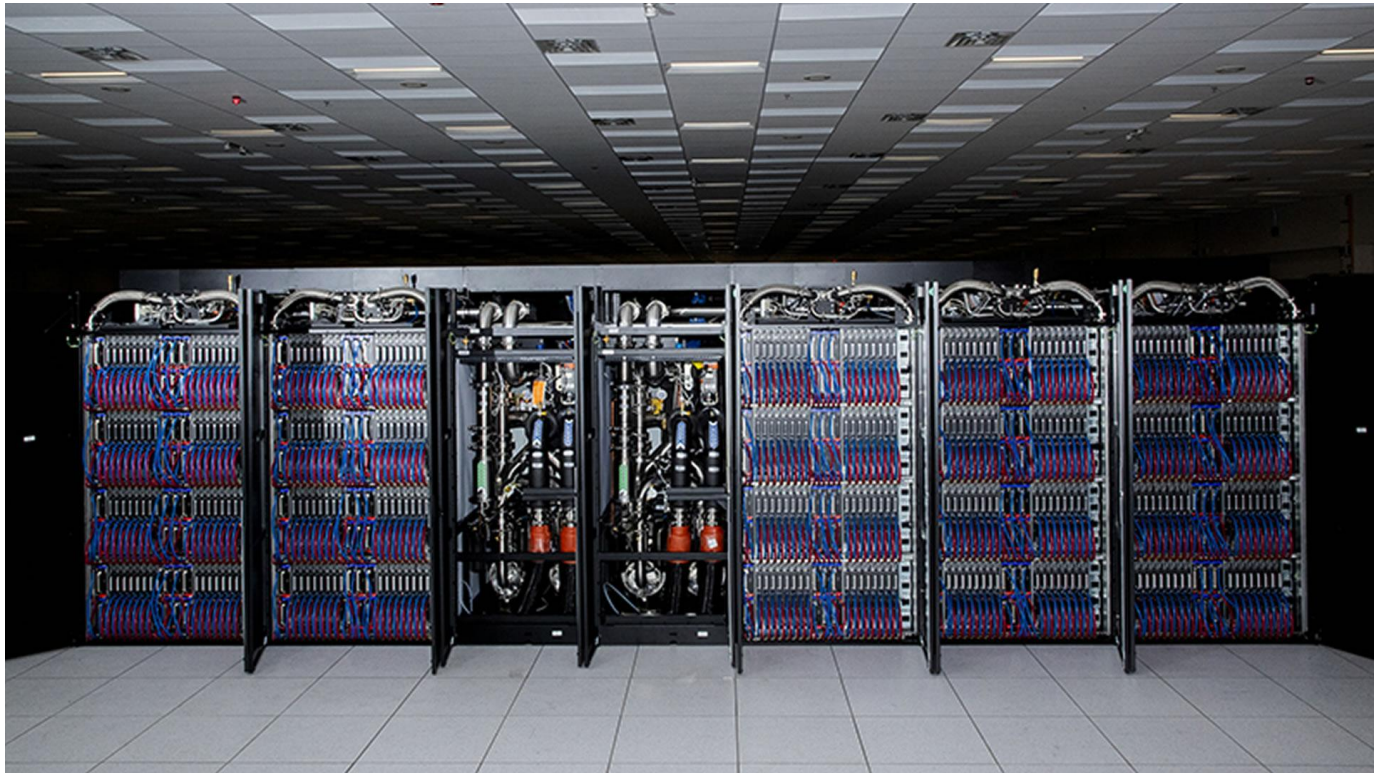
ENIAC (1945): the first electronic computer designed by Pennsylvania scientists using **vacuum tubes** for calculating Artillery firing tables.
18,000 Vacuum tubes (24m long x 2x6m high), too complex, poor reliable, and supper power hungry



The current AI monster

This AI computer has the size similar to the first electronic computer (1948)

Any possibility and when the AI computer can be miniaturized to be a USB stick?



2024 Venado Peak : 10 Exa-FLOPS An AI-enable supercomputer by Los Alamos National Laboratory

CONCLUSIONS

- **Edge-AI is growing part of the future computing system, including the military applications**
- **Lightweight network such as BNN and BSNN is well suited for Edge-devices**
- **The successful of IMC based on emerging technology with approximate computing could be the key for breaking the computing limit, i.e. EXAFLOPS (10^18) range**

THANK YOU