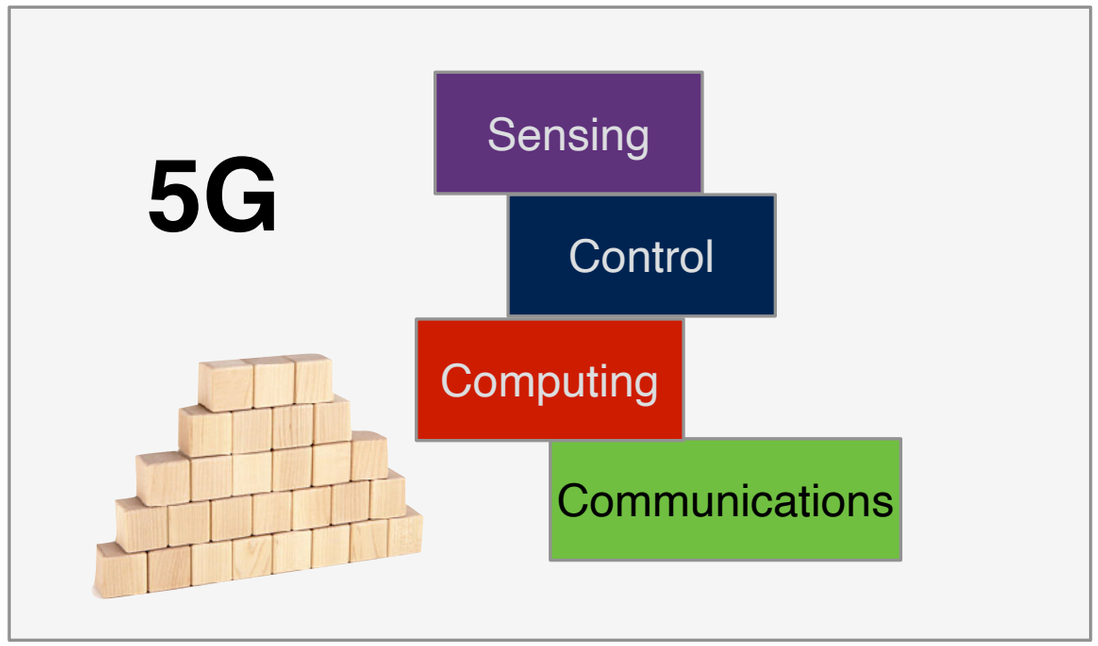
Memristor Empowered Ultra-fast Baseband Processing

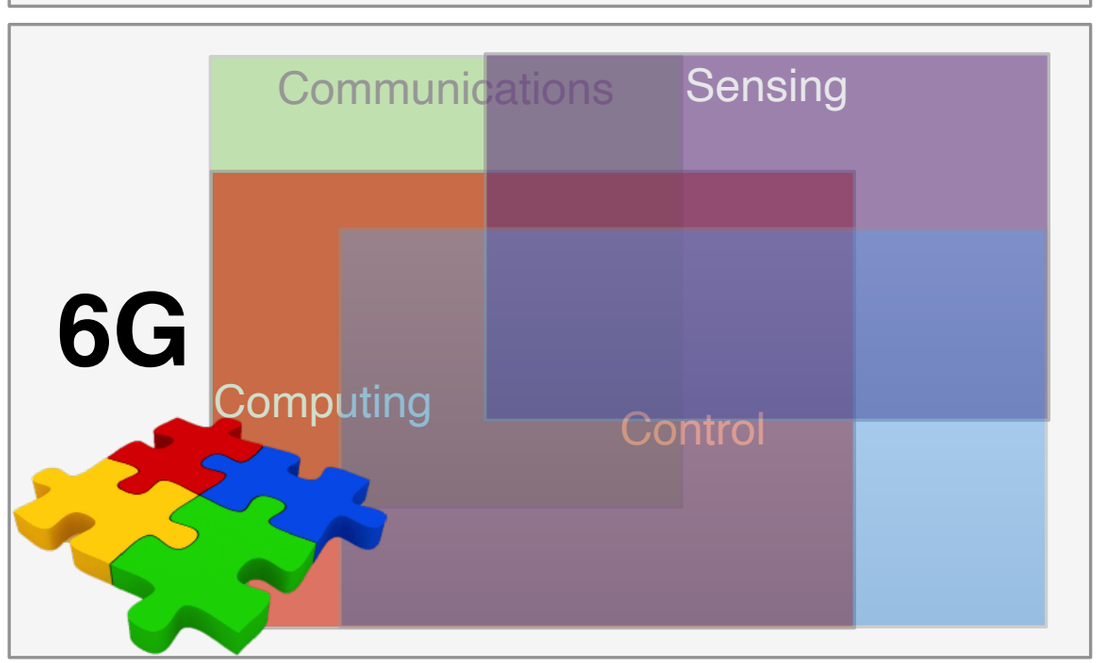
Kaibin Huang

Dept. of Electrical & Electronic Engineering
The University of Hong Kong
Hong Kong

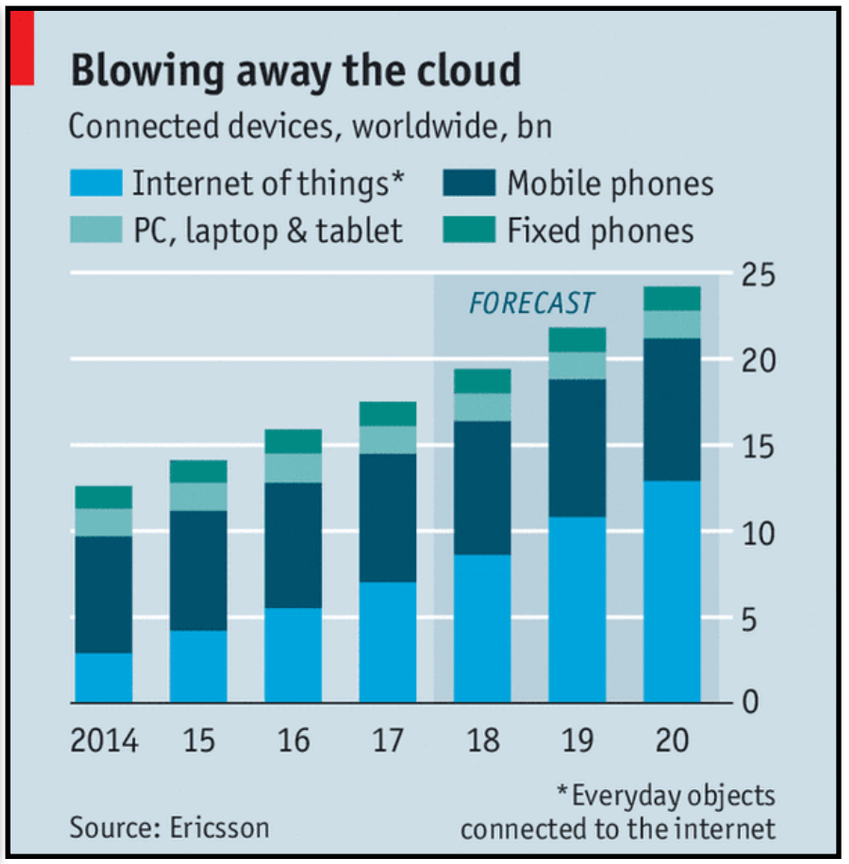
Acknowledgement: Parts of presentation were created by Zhongrui Wang and Qunsong Zeng

6G — Fusion of Communication and Computing

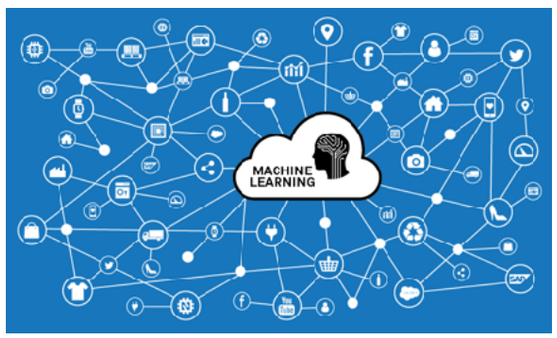




Revolution in Computing — “Living on the Edge”



About 150 trillion gigabytes of data will need analysis by 2025 (Forbes)



Machine Learning

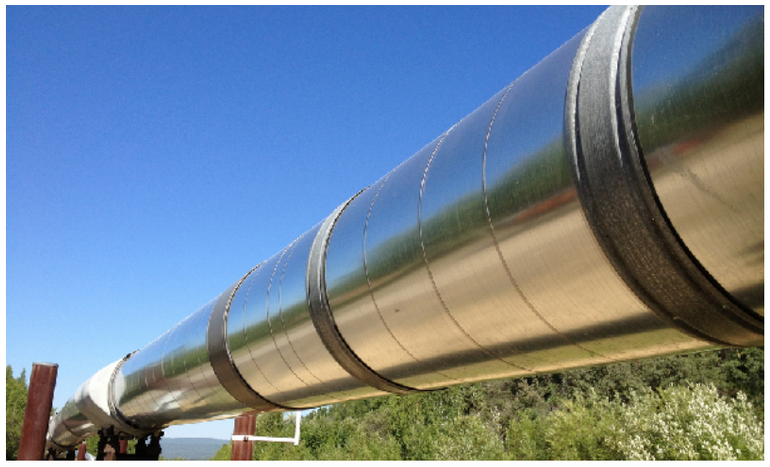
Artificial Intelligence

From Shannon 1.0 to Edge AI

Shannon 1.0 — Rate Maximization

"*Given a constraint on distortion,*

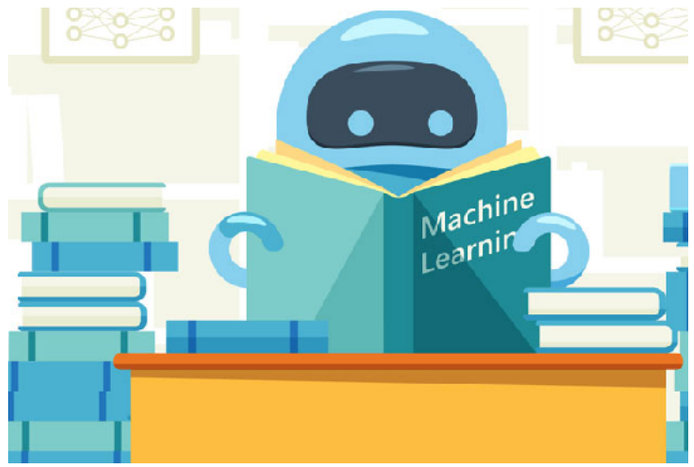
transmit as much data as possible"



Shannon 2.0 — Fast Edge Intelligence

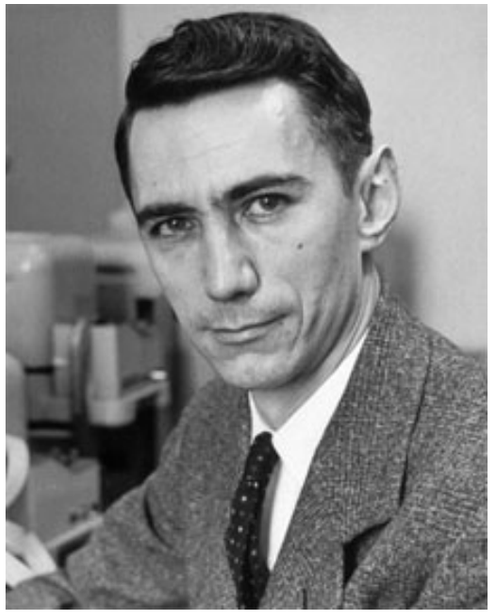
"*Given a constraint on learning/decision accuracy,*

distill or use intelligence as fast as possible"

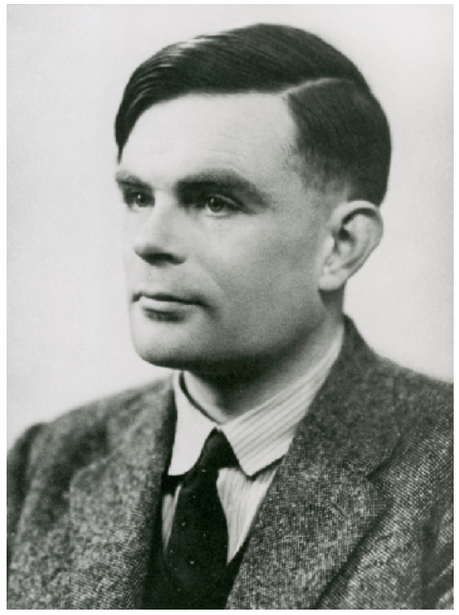


G. Zhu, D. Liu, Y, Du, C. You, J. Zhang, and K. Huang, "Towards an Intelligent Edge: Wireless Communication Meets Machine Learning", IEEE Commun. Magazine, 2020.

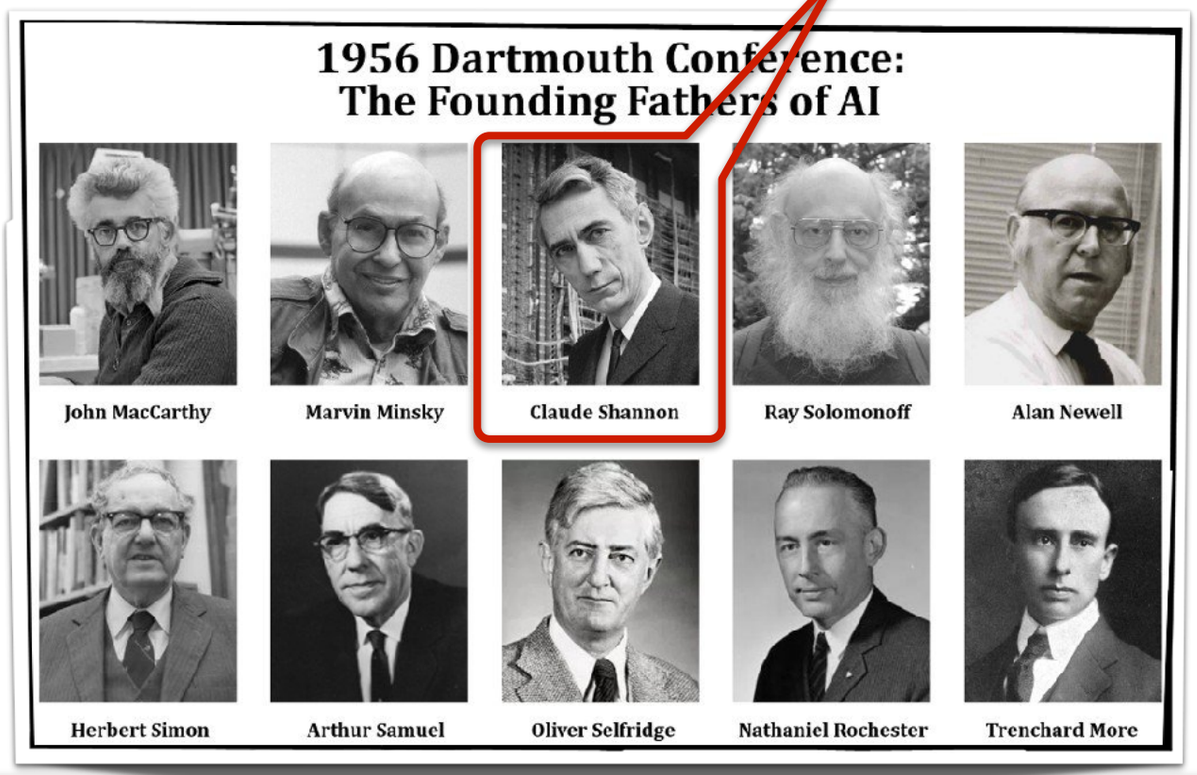
6G — Shannon Meets Turing



Claude Shannon
(Father of information theory)

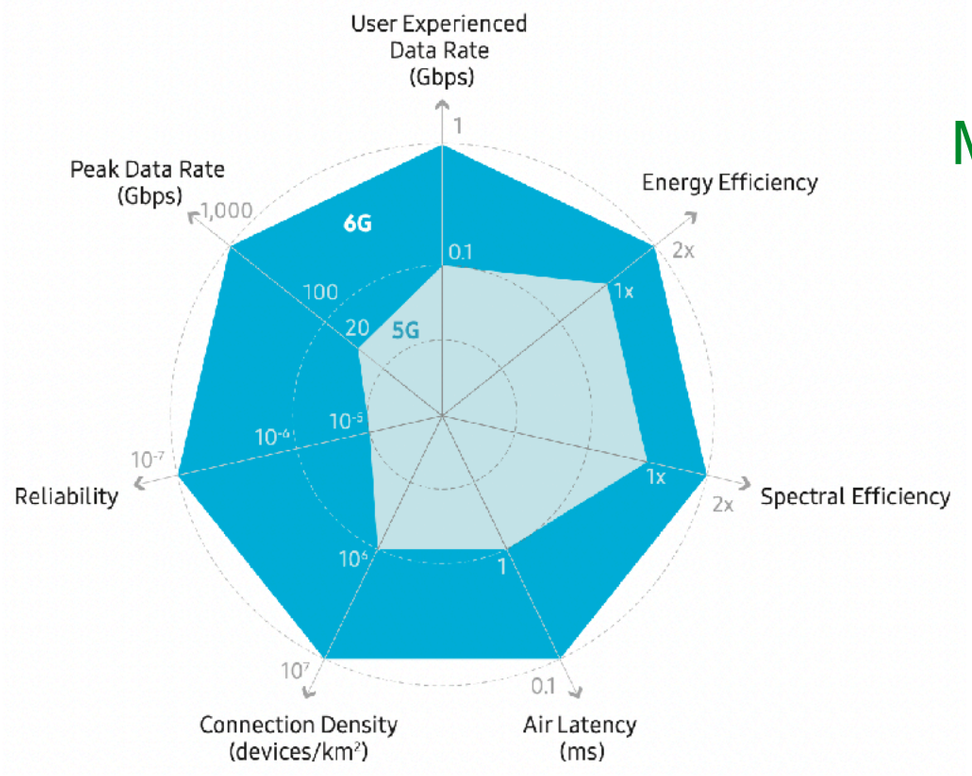


Alan Turing
(Father of AI)

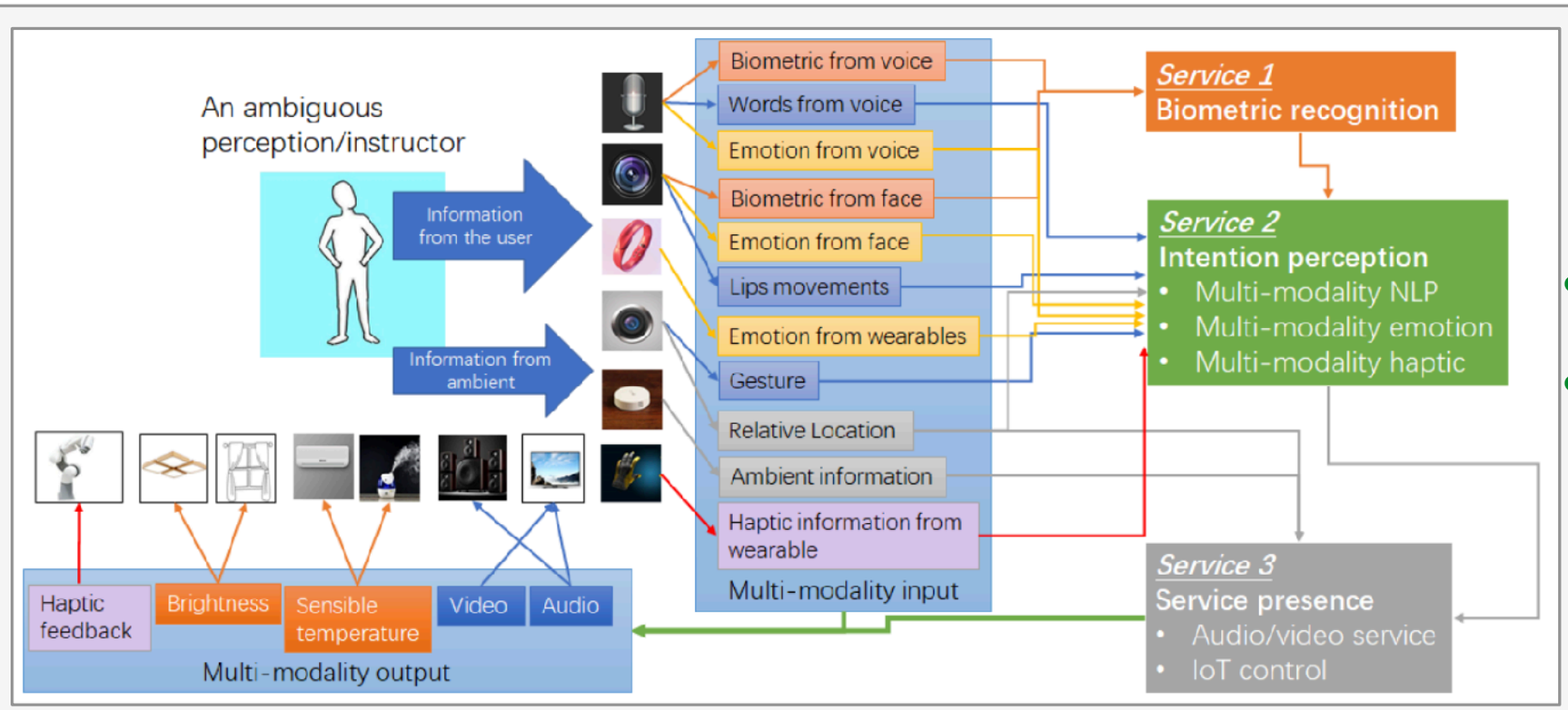


6G Sub-millisecond Latency

Peak Data Rate
= 1000 Gb/s!



Minimum Air Latency
= 0.1 ms!



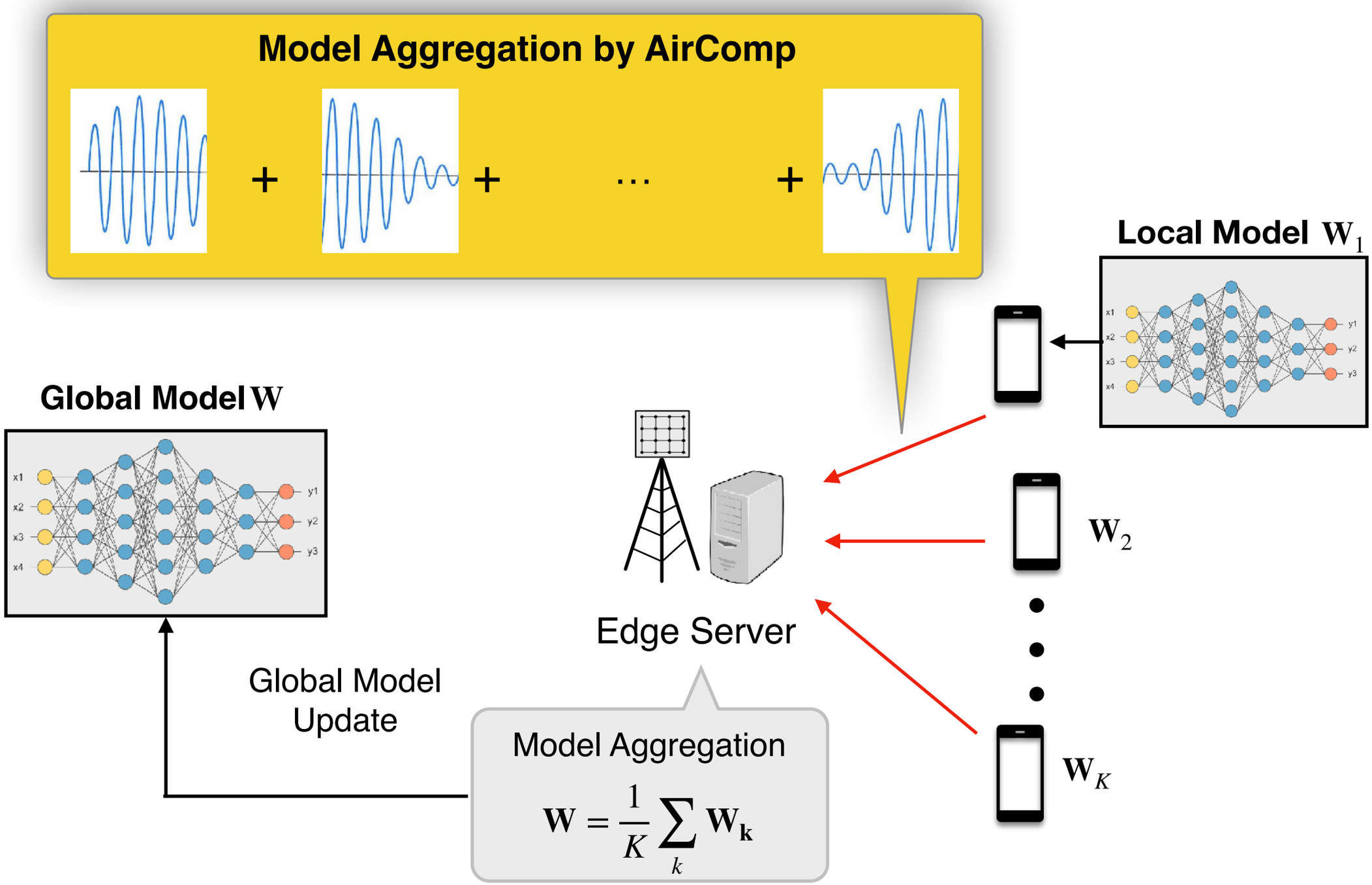
VR/AR Applications

- Tactile delay < 50 ms
- Visual delay < 15 ms

Martin Cooper with 1G Phone

Question A:
Is analog communication dead?

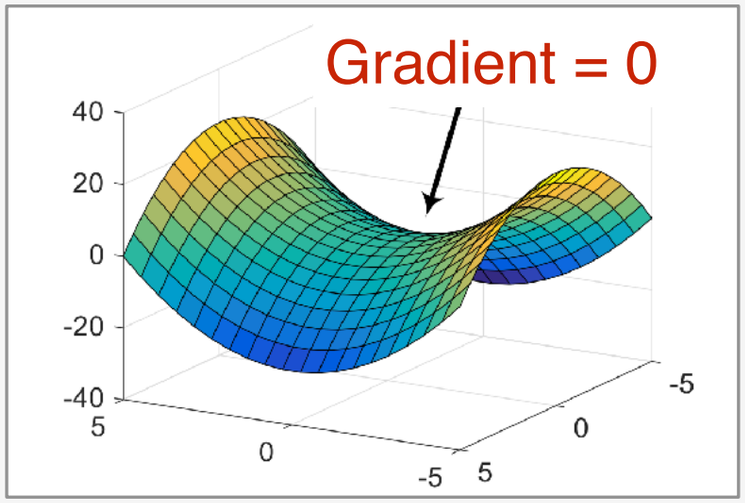
Over-the-Air Computing

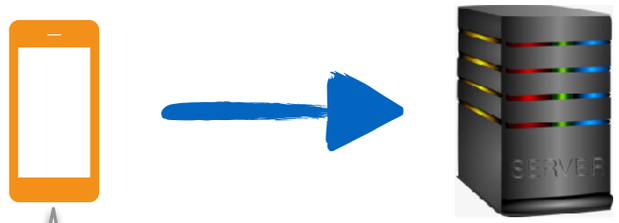


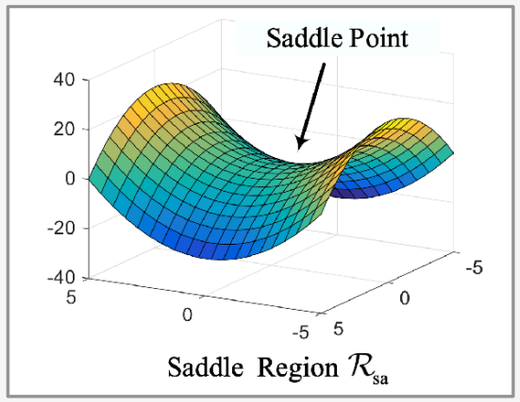
G. Zhu, Y. Wang, and K. Huang, "Broadband Analog Aggregation for Low-Latency Federated Edge Learning," IEEE TWC, 2020.

Turning Channel Noise into Accelerator

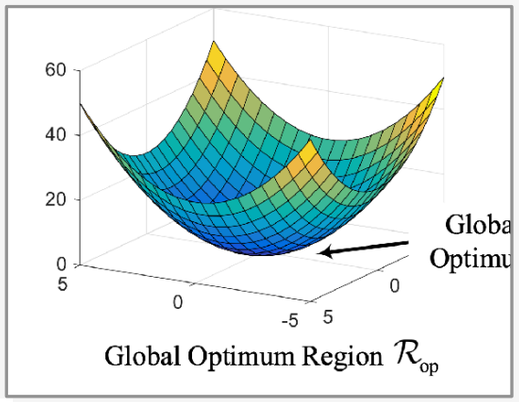
What is the problem?



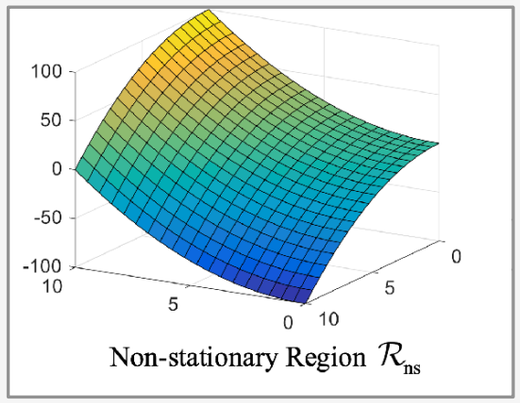




Low Power



High Power

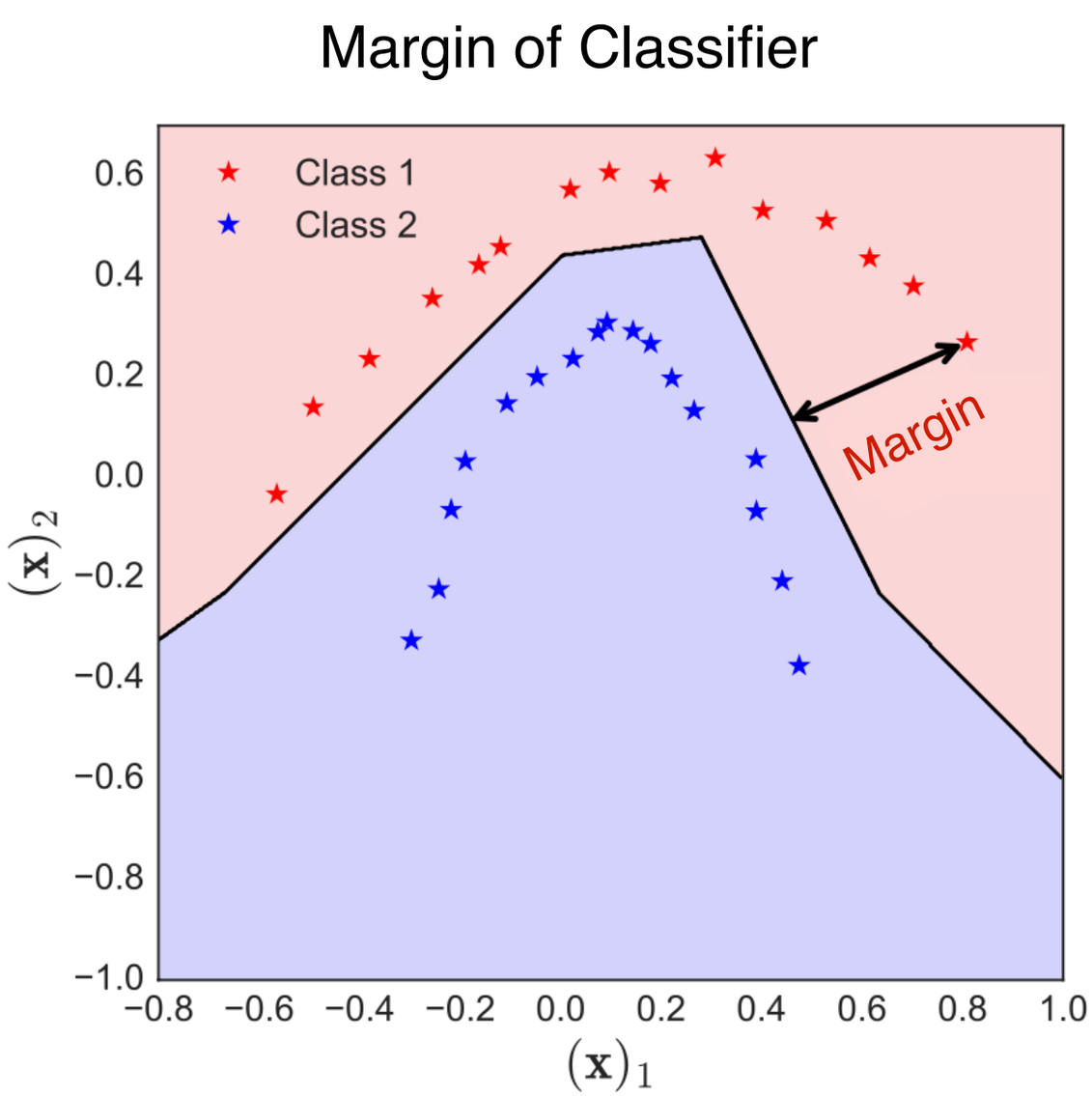


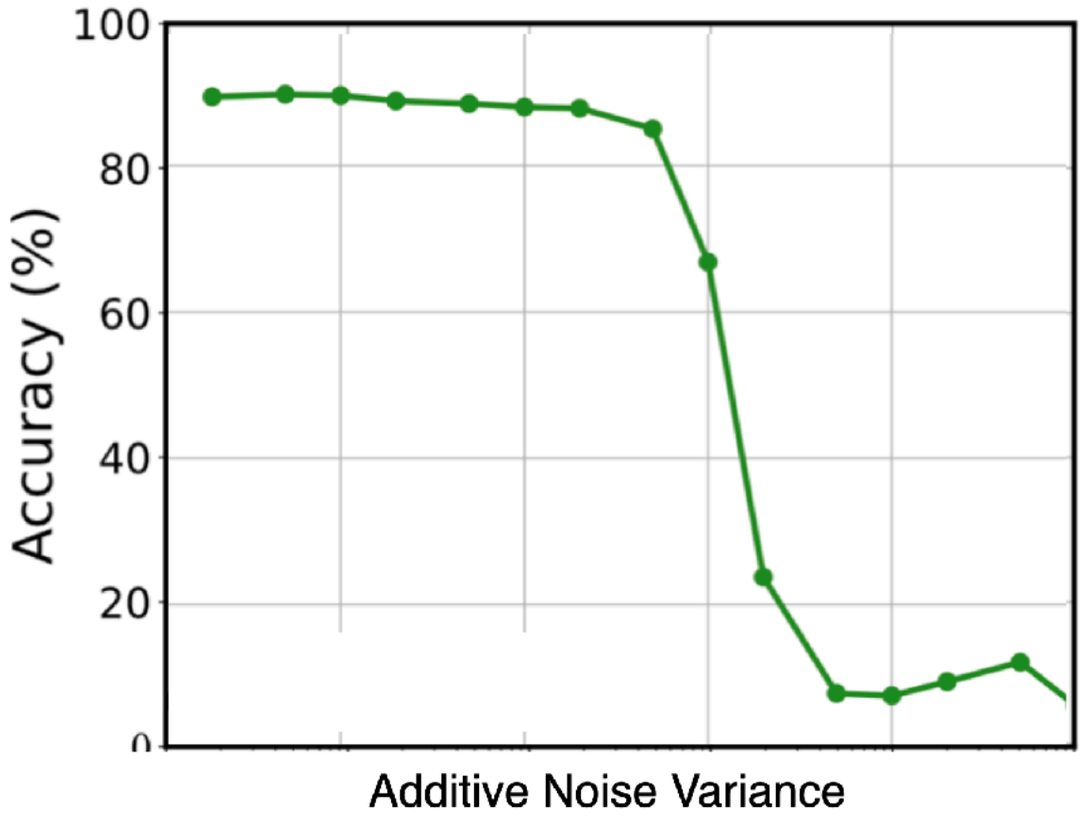
Medium Power

Solution - Region-aware Power Control

Z. Zhang, G. Zhu, R. Wang, V. K. N. Lau, and K. Huang, "Turning Channel Noise into an Accelerator for Over-the-Air Principal Component Analysis," TWC 2022

Noise Tolerance of Edge Inference

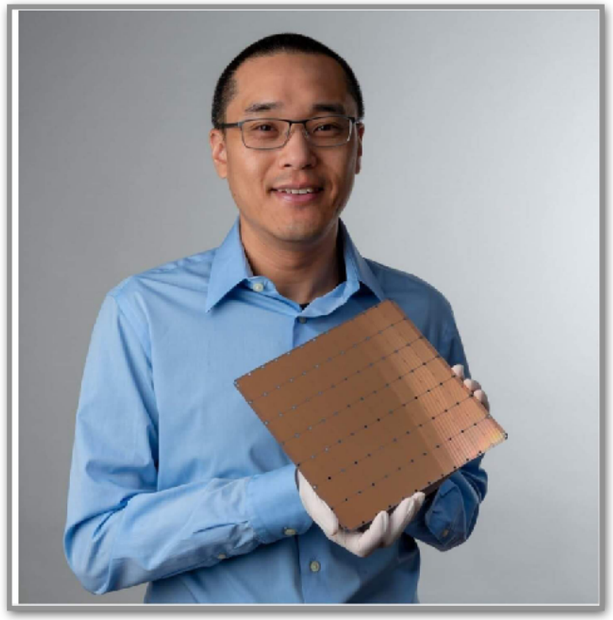




Question B - Is analog computing dead?

Analog Computer	Digital Computer
Specialized for one problem	Flexible due to Boolean algebra
Small errors can accumulate	Resilient to noise
Cannot get same answer twice	Reproducible results
?	Advent of solid-state electronics allows VLSI

Alan Turing used analog computer to crack Nazi's enigma code in WWII



Cerebras Wafer-size chip - 1.2 trillion transistors and 400,000 AI cores.

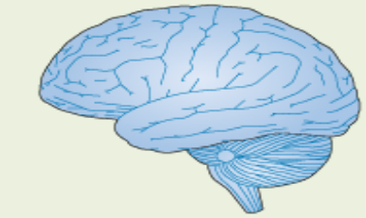
Outline

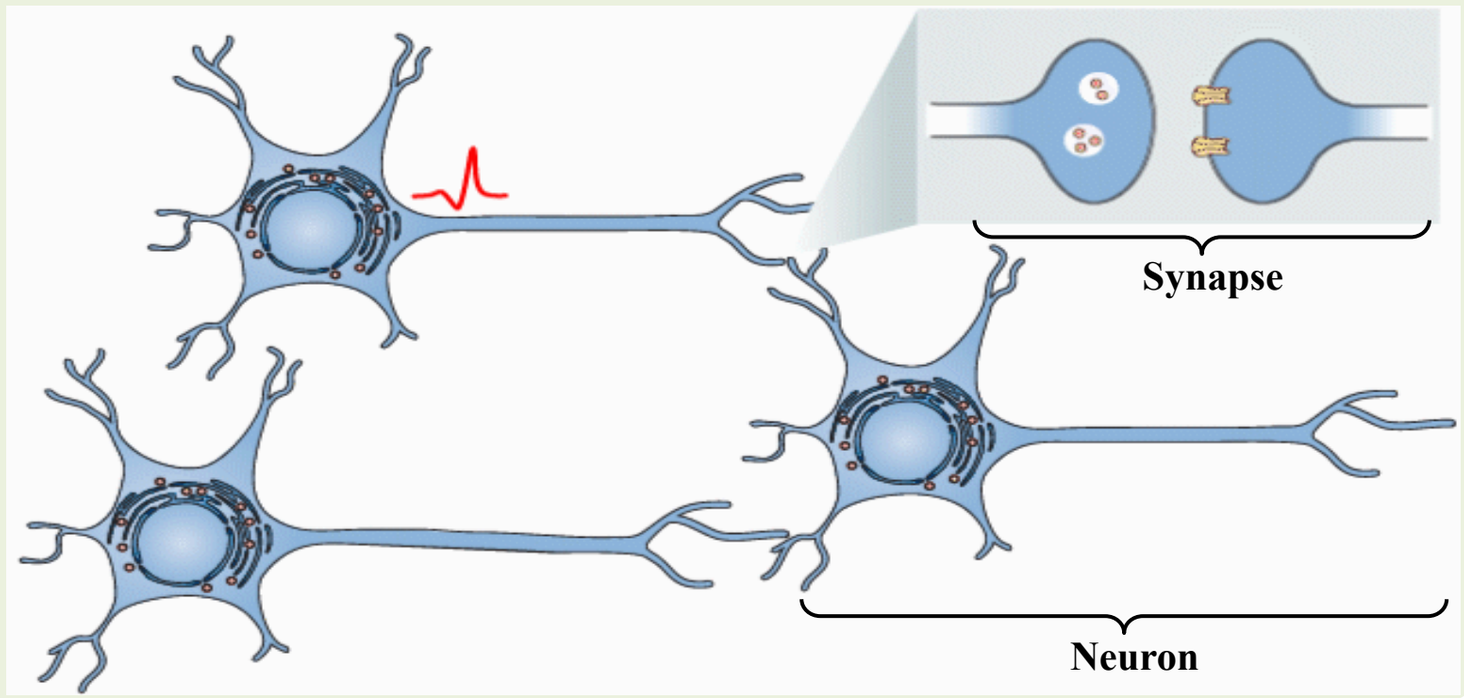
I. Analog Neuromorphic Computing

II. Memristor Empowered Ultra-fast Baseband

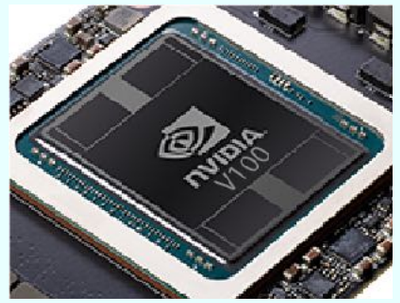
The digital “brain”

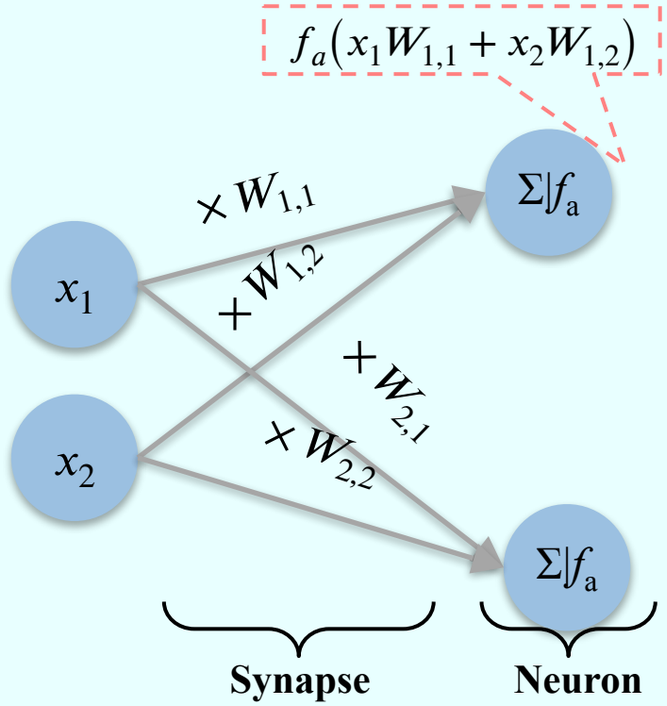
Synapses/neurons:
Hardware
Neuron outputs:
Temporal binary spikes
Synaptic weight modulation:
Local plasticity





Synapses/neurons:
Software
Neuron outputs:
Real numbers
Synaptic weight modulation:
Global optimization

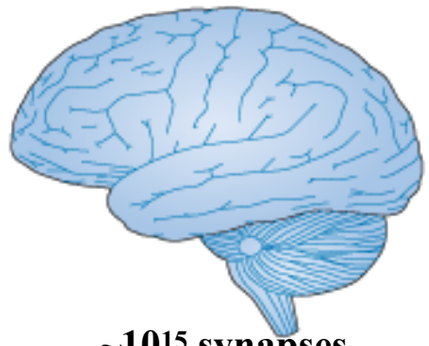




Are digital chips approaching the performance of the brain?

~10^{10} transistors (~10^{8} software/hardware synapses)
~10pJ per operation [1-2]

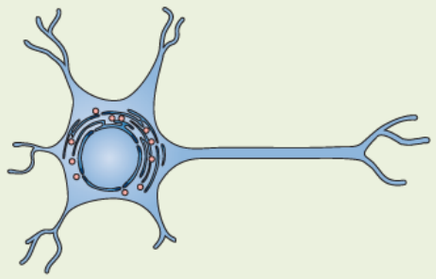
VS



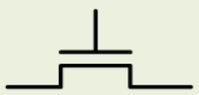
~10^{15} synapses
~10fJ per operation [2]

Expensive computing unit

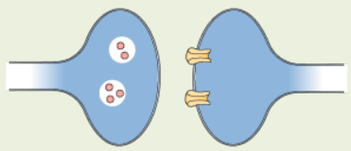
1×



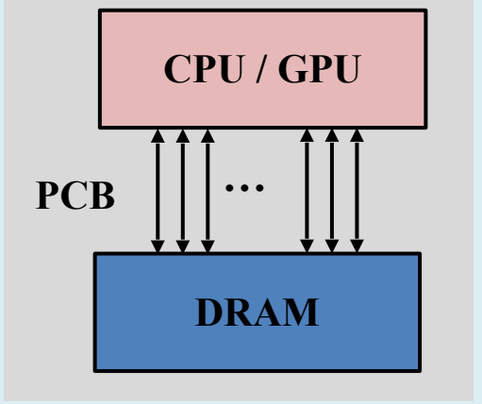
~16000×



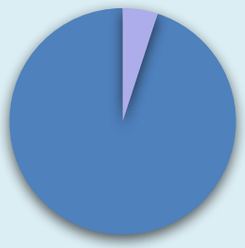
1000×



Separated memory / processing units



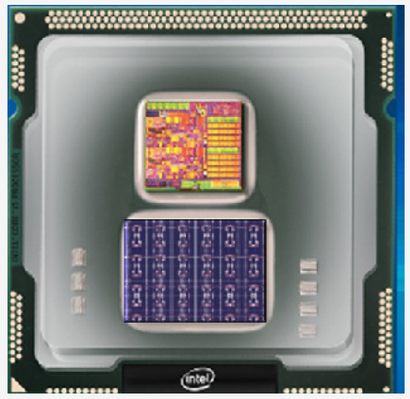
Computation energy



Register file, cache,
DRAM energy [3]

[1] NVidia [2] S. Furber, J. Neural Eng. **13**, 051001 (2016) [3] Horowitz, ISSCC 2014

Why Neuromorphic Computing

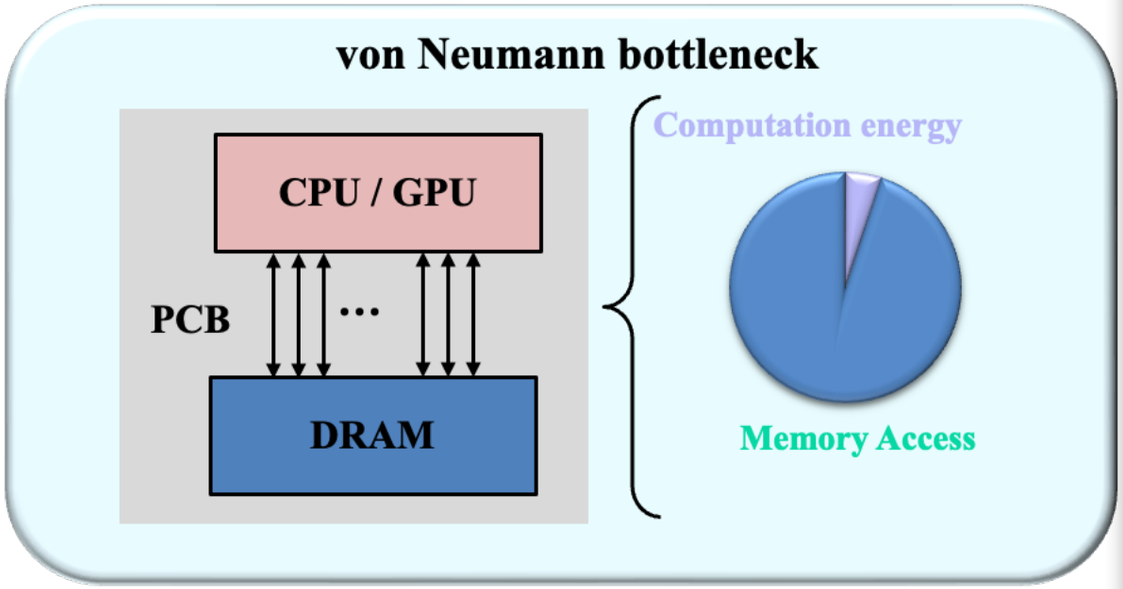


vs

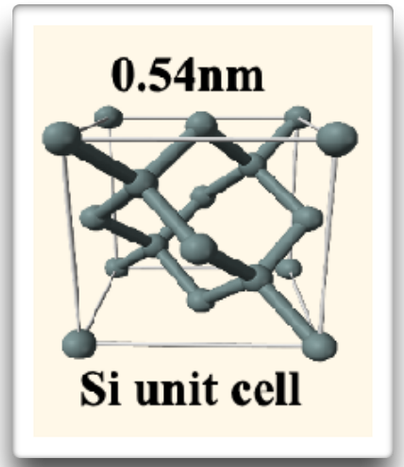
10^{10} transistors vs. 10^{15}synapse

10^{-12} vs. 10^{-15} J per operation

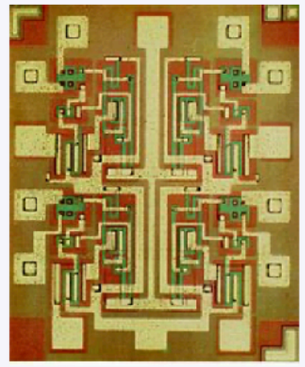
Von Neumann Bottleneck



Approaching Transistor Scaling Limit



Semiconductor device fabrication



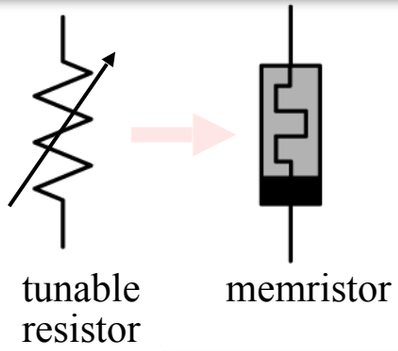
MOSFET scaling (process nodes)

10 μm – 1971
6 μm – 1974
3 μm – 1977
1.5 μm – 1981
1 μm – 1984
800 nm – 1987
600 nm – 1990
350 nm – 1993
250 nm – 1996
180 nm – 1999
130 nm – 2001
90 nm – 2003
65 nm – 2005
45 nm – 2007
32 nm – 2009
22 nm – 2012
14 nm – 2014
10 nm – 2016
7 nm – 2018
5 nm – 2020
3 nm – 2022

Future

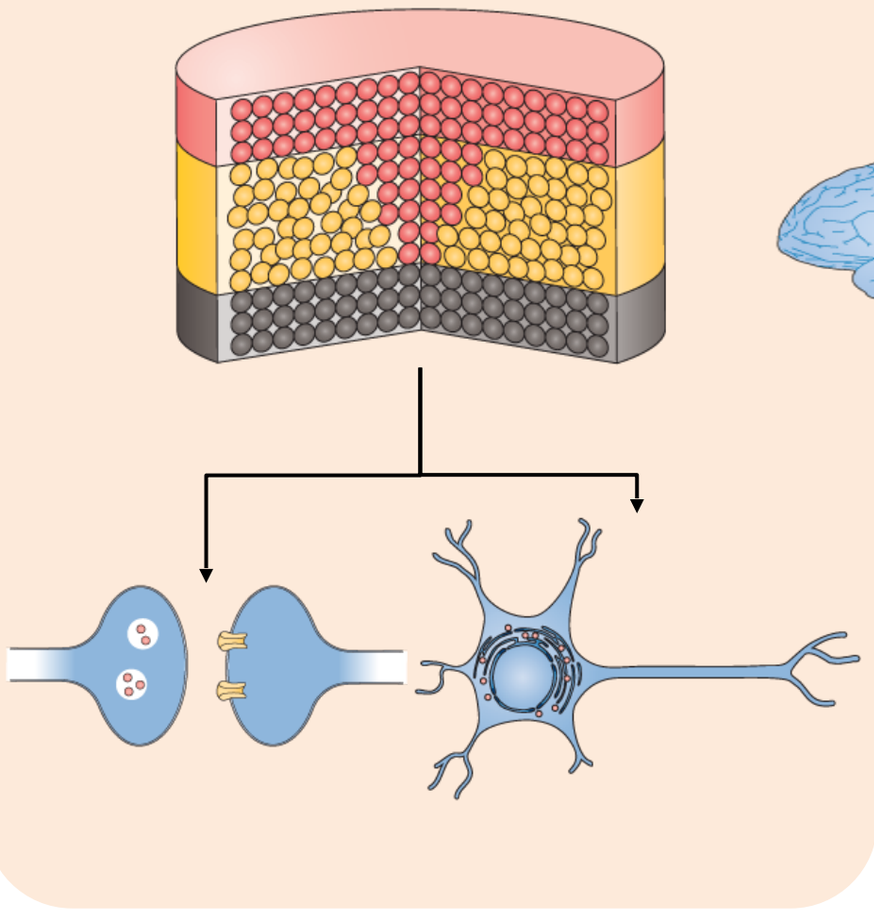
2 nm ~ 2024

Possible Solution — In-Memory Computing

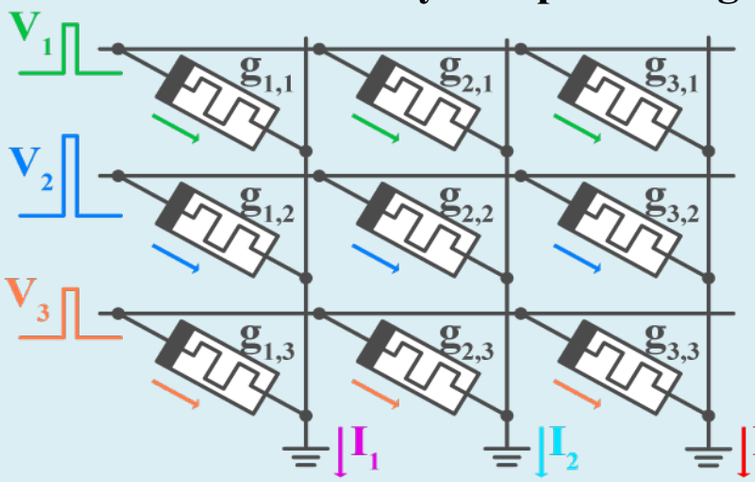


- **"Mem" + "resistor" :** Resistance with memory of electrical signal history.
- Conceptualised and named by Prof. Leon Chua in 1971
- Demonstrated by Hewlett Packard laboratory in 2008

Electrochemical reactions to simulate synaptic behavior



Colocation of memory and processing



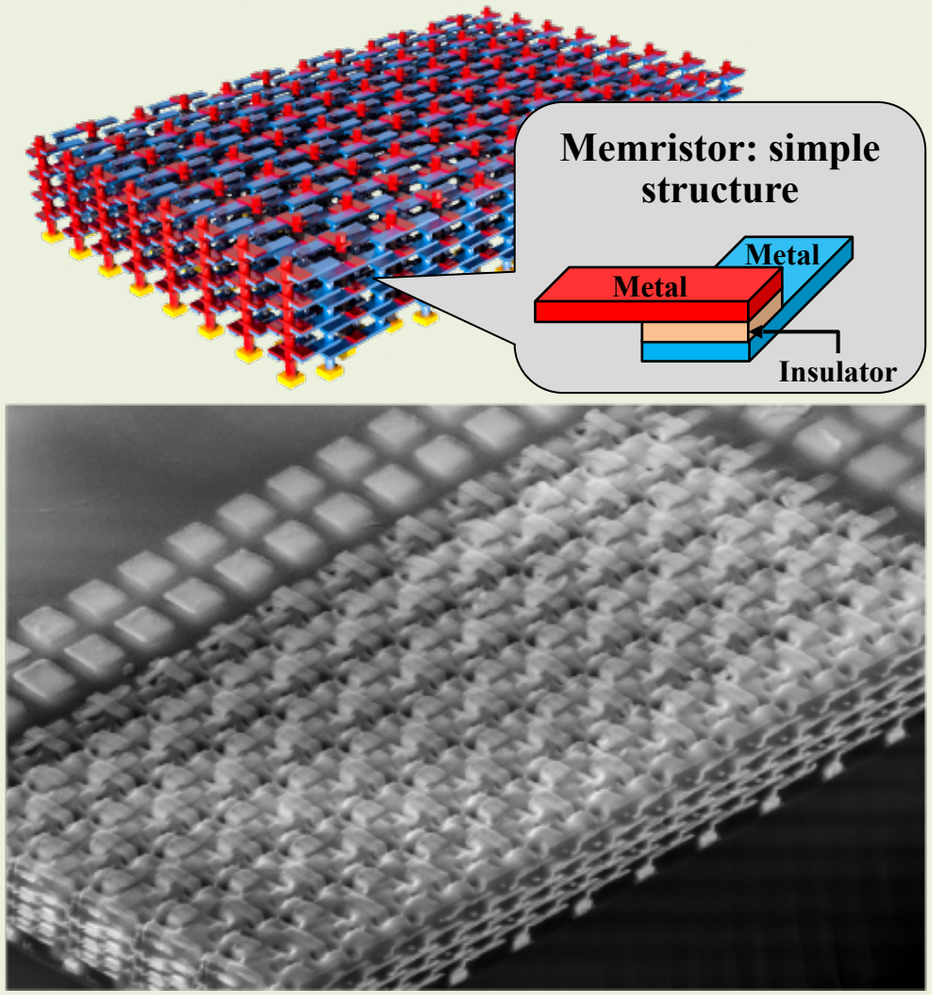
$I_1 = V_1 g_{1,1} + V_2 g_{1,2} + V_3 g_{1,3}$ Ohm's law + Kirchhoff's law

$$\begin{bmatrix} g_{1,1} & g_{1,2} & g_{1,3} \\ g_{2,1} & g_{2,2} & g_{2,3} \\ g_{3,1} & g_{3,2} & g_{3,3} \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix} = \begin{bmatrix} I_1 \\ I_2 \\ I_3 \end{bmatrix}$$

Z. Wang *et al.*, Nat. Rev. Mater. doi:10.1038/s41578-019-0159-3

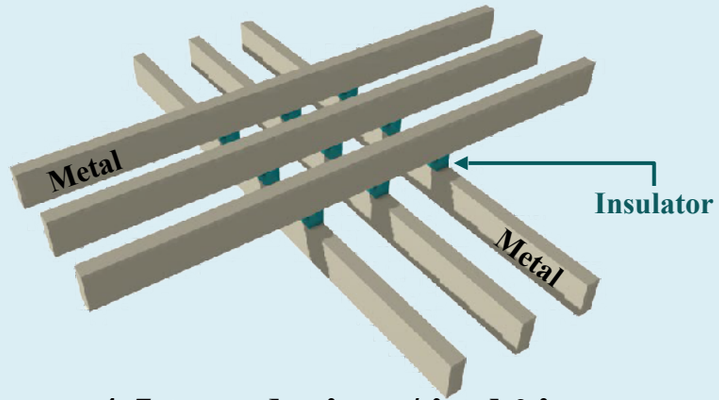
Why memristors? Stack-ability and Scalability

8-layer stacking of memristors

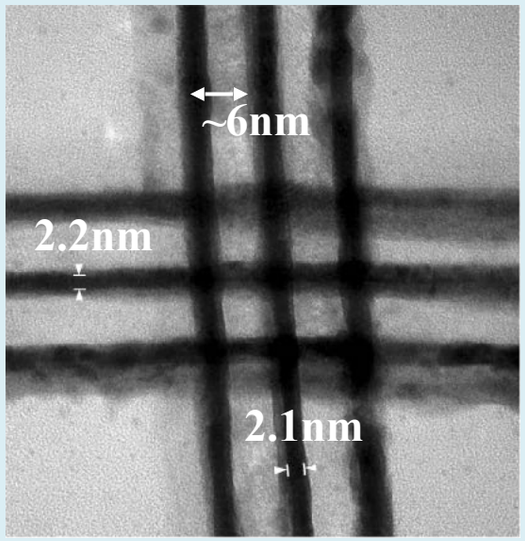


P. Lin *et al.*, Nat. Nanotechnol. **14**, 35-39

2nm × 2nm memristors



~4.5 tera-devices / inch² !



P. Lin *et al.*, Nat. Electron., in press

Outline

I. Analog Neuromorphic Computing

II. Memristor Empowered Ultra-fast Baseband

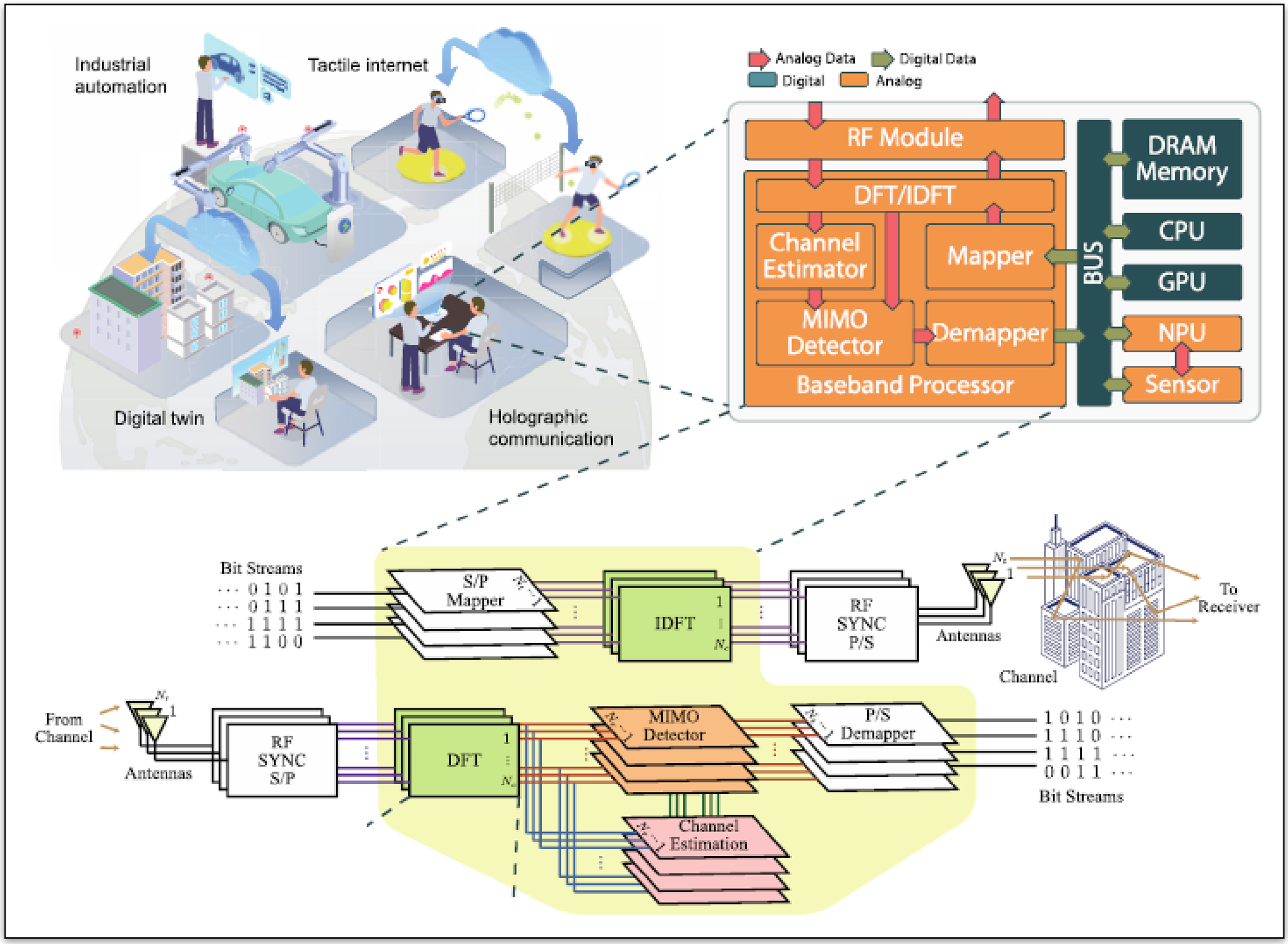
- Kaibin Huang
- Zhongrui Wang
- Qunsong Zeng
- Jiawei Liu

- Yida Li
- Jun Lan
- Yi Gong

SUSTech
Southern University of Science and Technology

Q. Zeng, J. Liu, J. Lan, Y. Gong, Z. Wang, Y. Li, and K. Huang, “Realizing Ultra-Fast and Energy-Efficient Baseband Processing Using Analogue Switching Memory”, [Online] http://arxiv.org/abs/2205.03561.

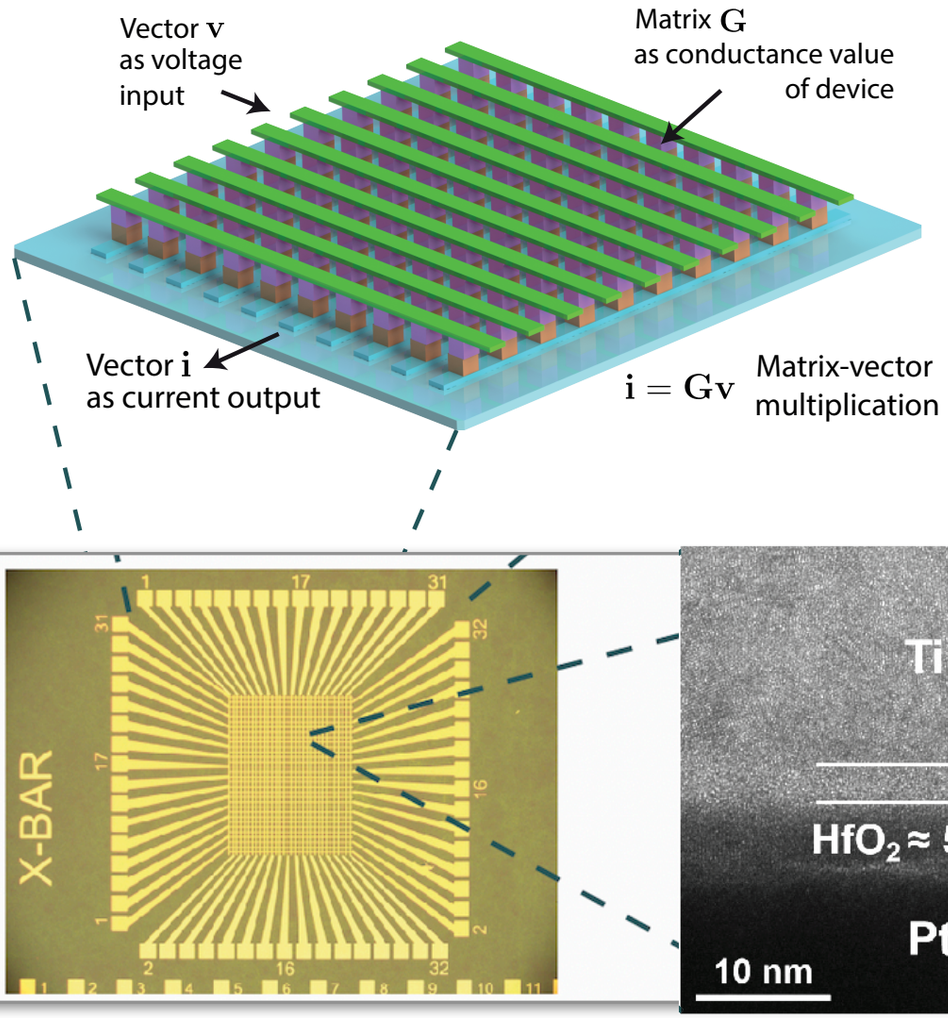
In-Memory Empowered Ultra-Fast 6G Communication

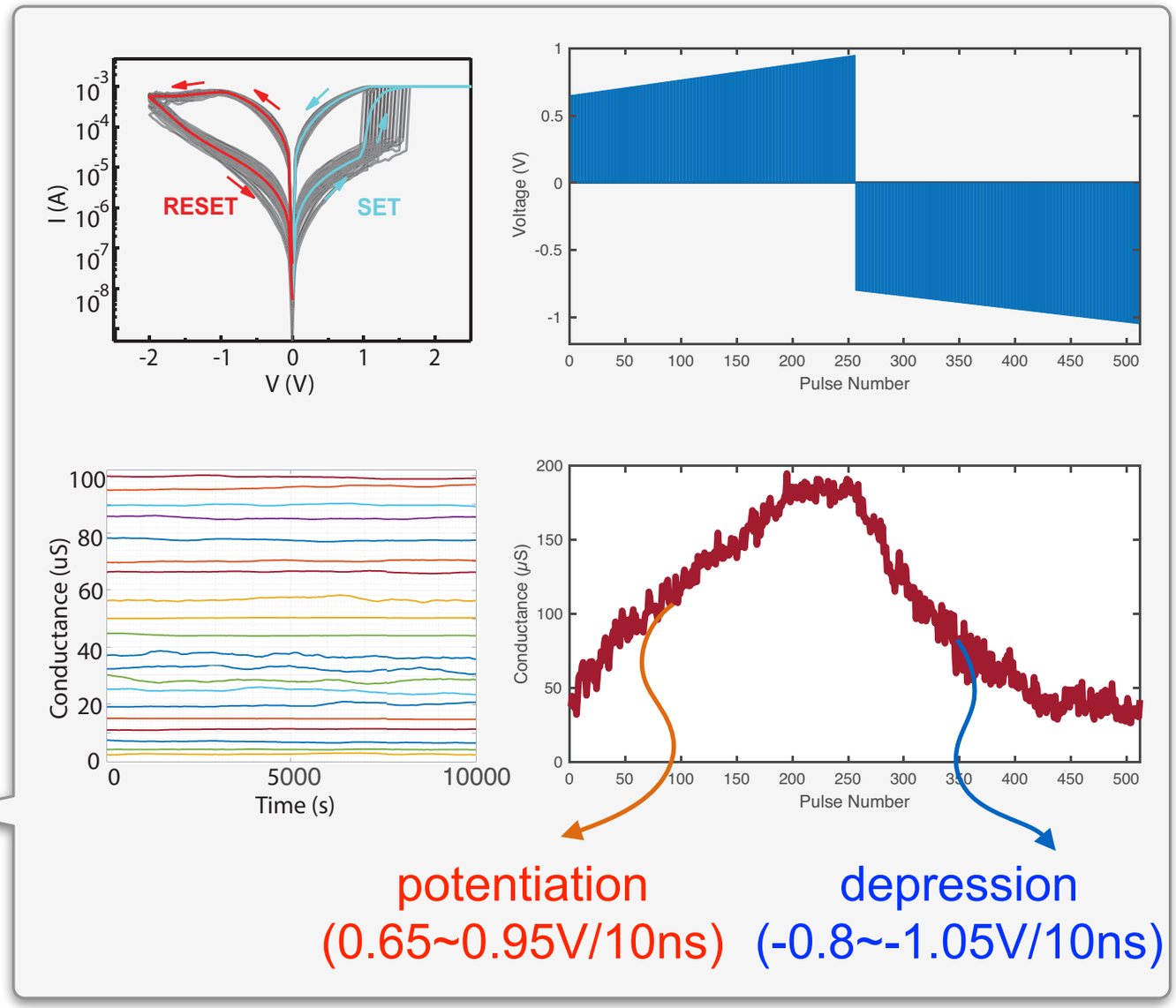


Analog (ICM)

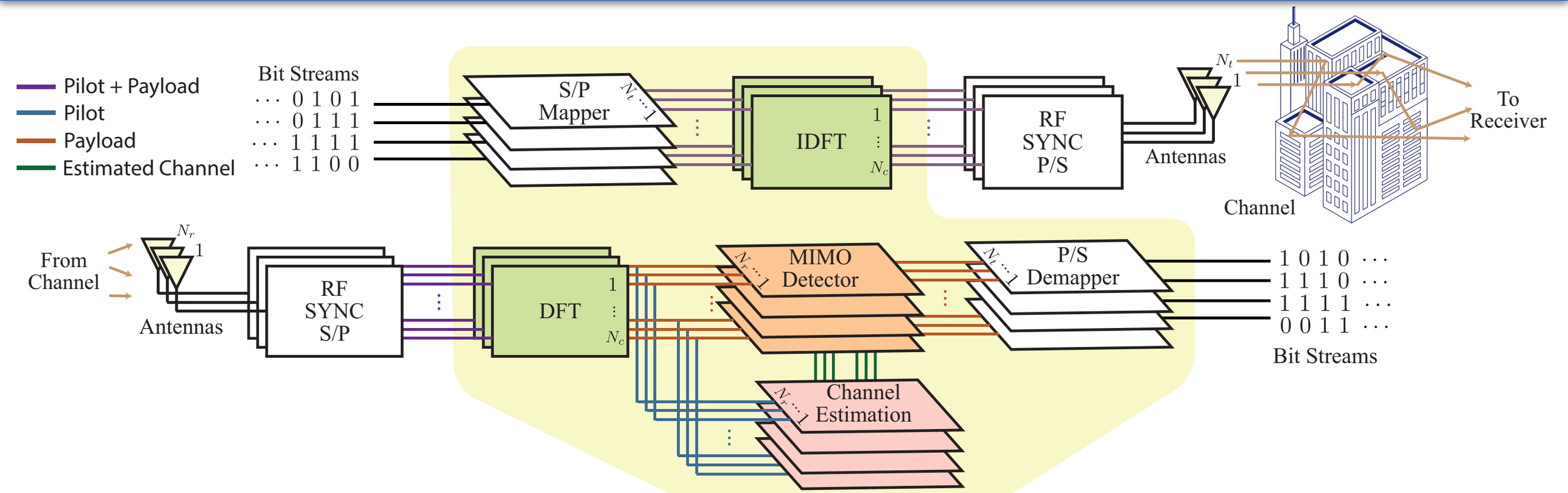
Digital

Fabrication: Resistive Random-Access Memory





MIMO-OFDM Transceiver



- Key modules of MIMO-OFDM transceiver:
  - **OFDM: Orthogonal frequency-division multiplexing**
    - **IDFT:** inverse discrete Fourier transform (Tx)
    - **DFT:** discrete Fourier transform (Rx)
  - **MIMO: Multiple-input multiple-output**
    - **MIMO detection:** recover signal by channel inversion (Rx)
    - **Channel estimation:** obtain channel state information (Rx)

Design: DFT Module

Highlight:

- DFT operation in **one-step** (i.e., $O(1)$ complexity).
- Traditional FFT algorithms complexity $O(N_c \log N_c)$.

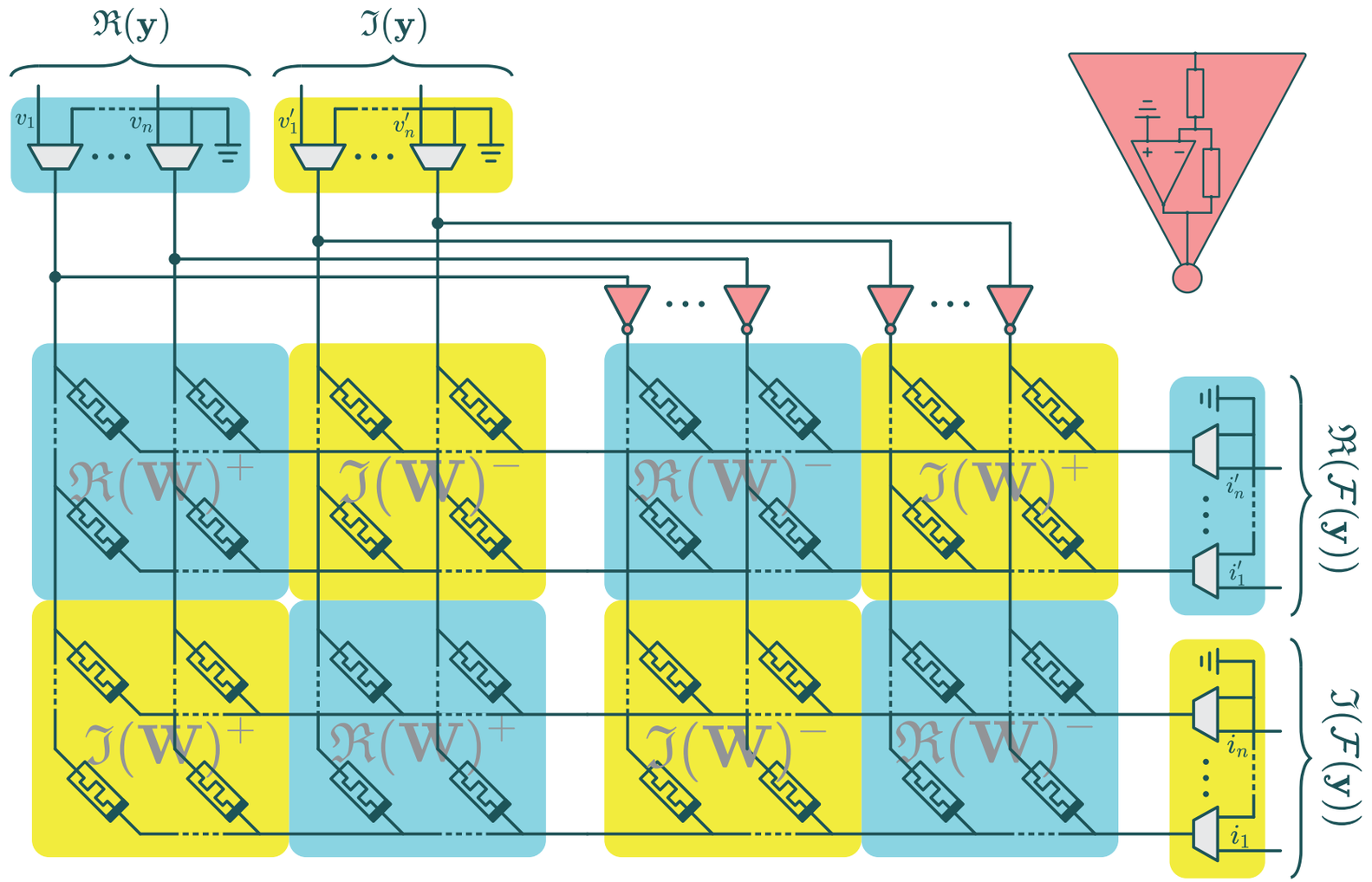
Real mappings:

matrix

$$\mathscr{R}(\mathbf{W}) = \begin{bmatrix} \Re(\mathbf{W}) & -\Im(\mathbf{W}) \\ \Im(\mathbf{W}) & \Re(\mathbf{W}) \end{bmatrix}$$

vector

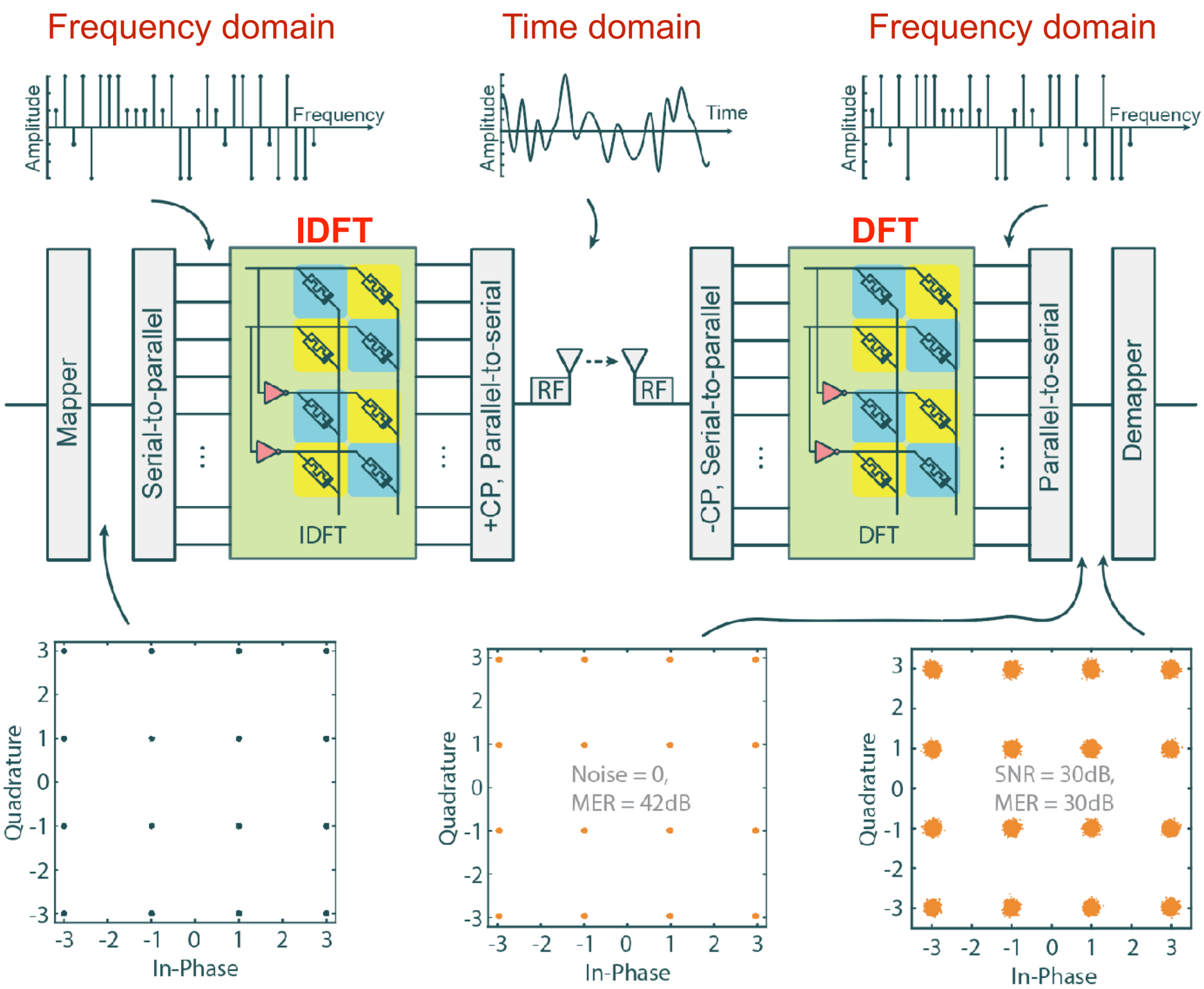
$$\mathscr{T}(\mathbf{x}) = \begin{bmatrix} \Re(\mathbf{x}) \\ \Im(\mathbf{x}) \end{bmatrix}$$



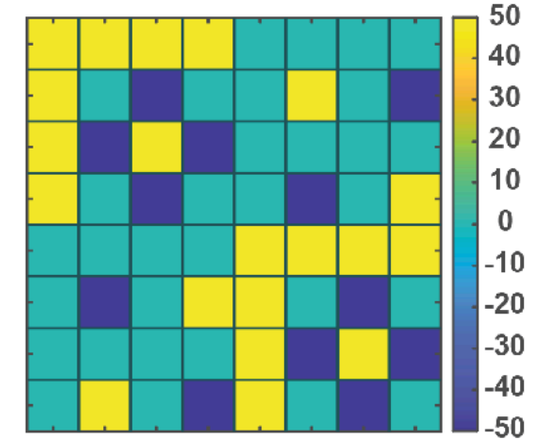
DFT operation: $\mathscr{F}(\mathbf{y}) = \mathbf{W}\mathbf{y}$ $\Rightarrow$ $\mathscr{R}(\mathbf{W}\mathbf{y}) = \mathscr{R}(\mathbf{W})\mathscr{T}(\mathbf{y})$

Validation: OFDM System

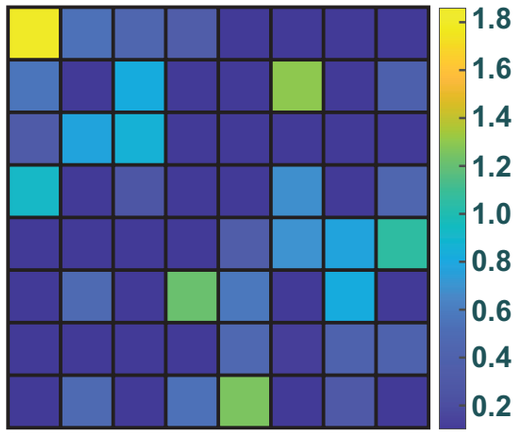
- Hardware implementation of OFDM system:



DFT matrix written into differential RRAM arrays

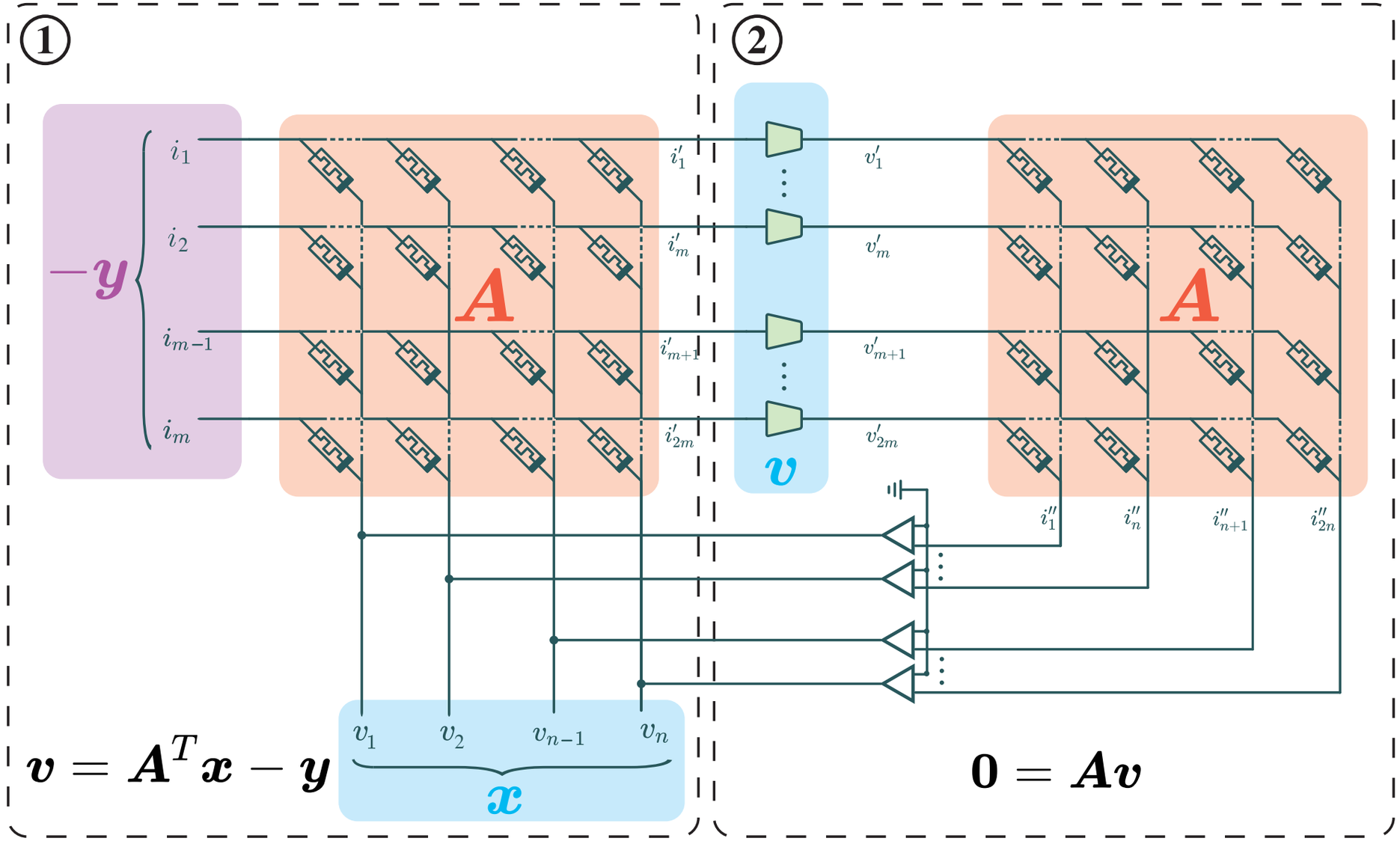


Real mapped DFT matrix (μS)



Error matrix (%)

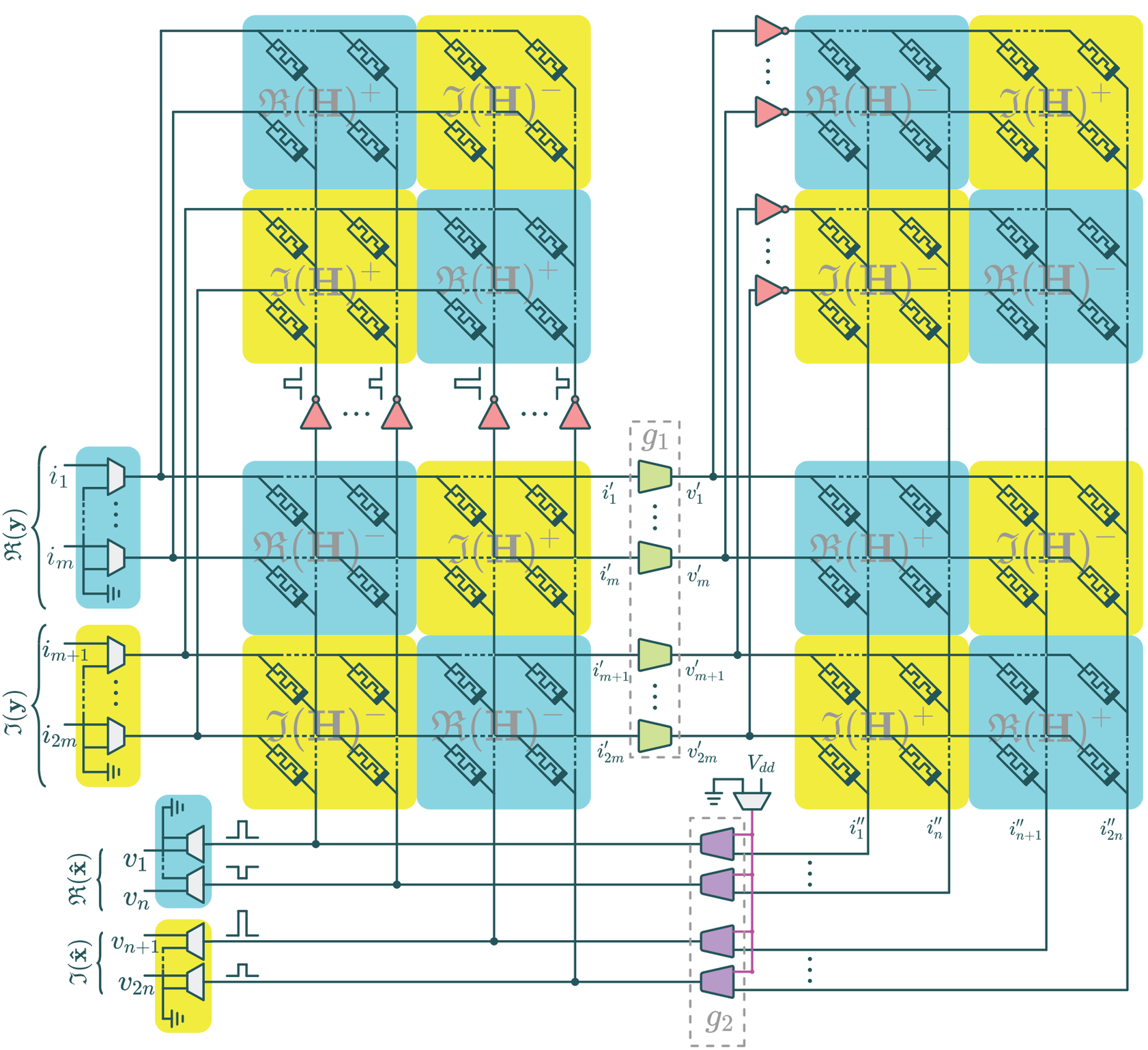
Matrix Inversion Using RRAM Crossbar



① and ② $\Rightarrow$ $AA^T x = Ay$ $\Rightarrow$ $x = \left(AA^T\right)^{-1} Ay$

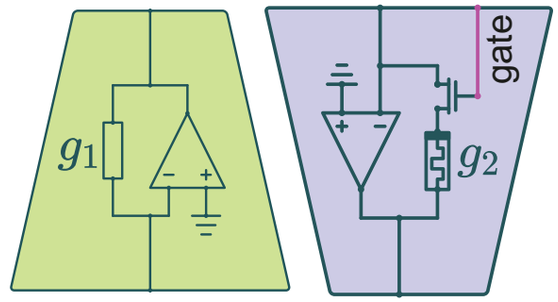
Sun, Zhong, *et al.,* "One-step Regression and Classification with Cross-Point Resistive Memory Arrays", *Science advances*, 2020.

Design: MIMO Detection Module



Highlight:

- MIMO detection in **one-step** (i.e., $O(1)$ complexity).
- Conventional computational complexity is $O(N^3)$.



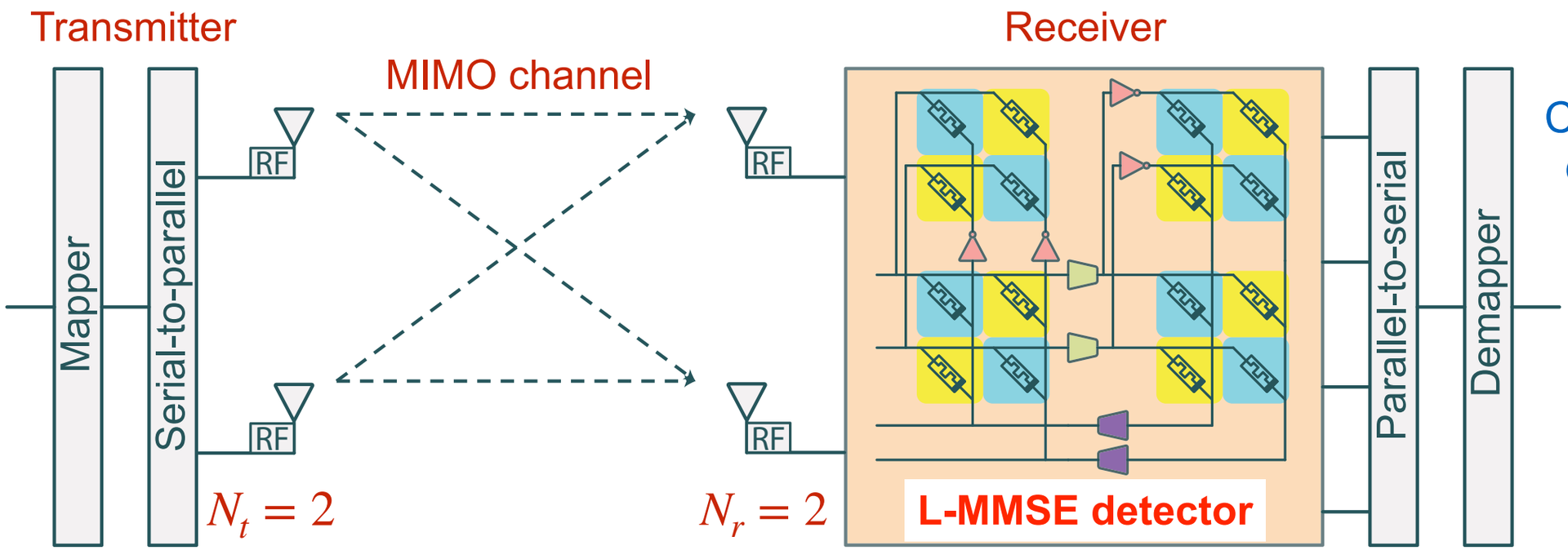
L-MMSE detection:

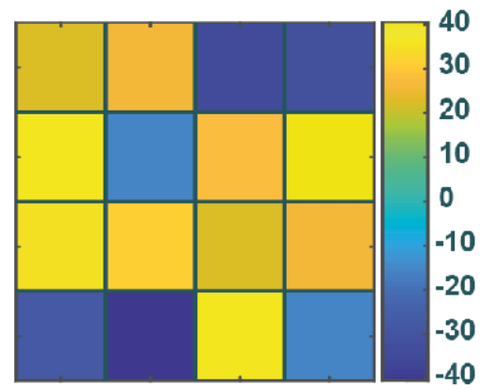
$$\hat{\mathbf{x}} = \left(\mathbf{H}^{\mathsf{H}}\mathbf{H} + \frac{1}{\mathsf{SNR}}\mathbf{I}\right)^{-1}\mathbf{H}^{\mathsf{H}}\mathbf{y}$$

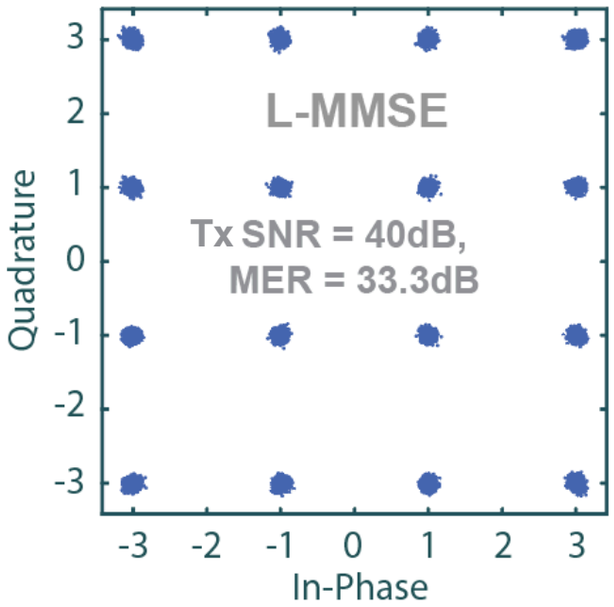
- $\mathsf{SNR} \propto (g_1 g_2)^{-1}$
- L-MMSE ⇒ ZF by turning off the transistors

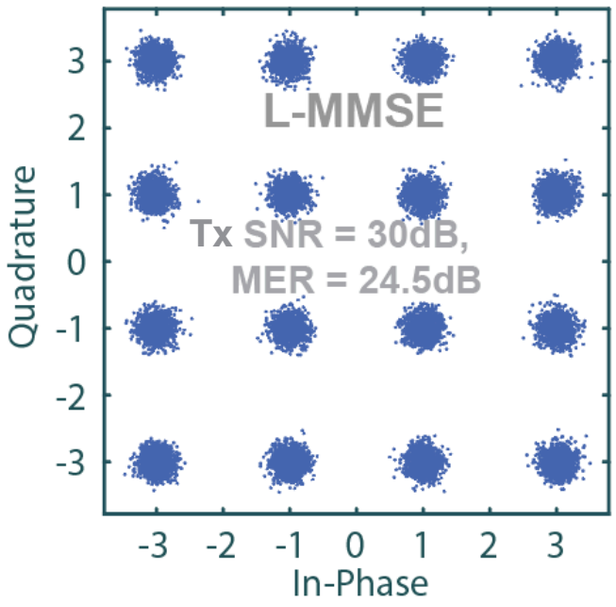
Validation: MIMO System

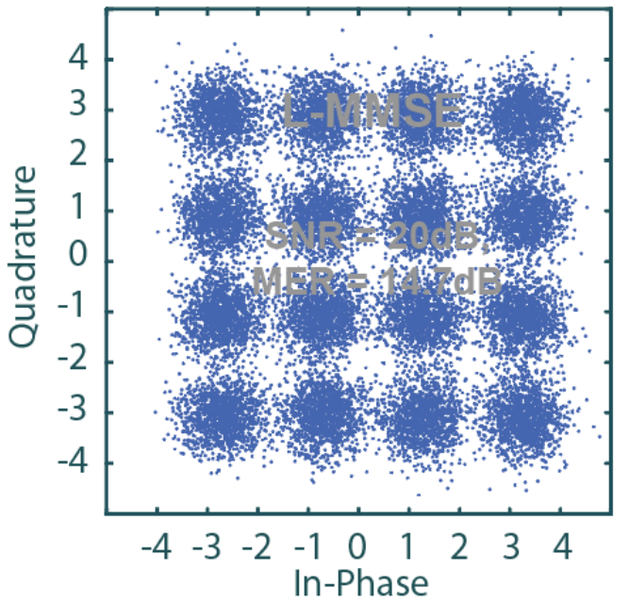
- Hardware implementation of MIMO system:



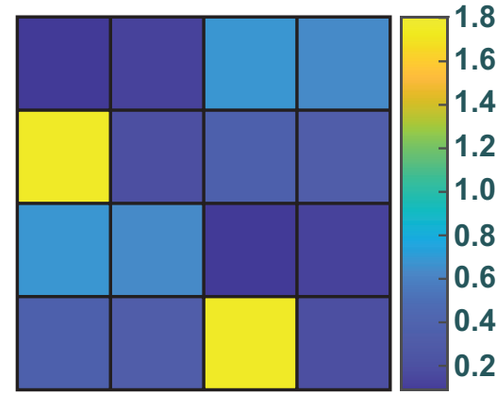








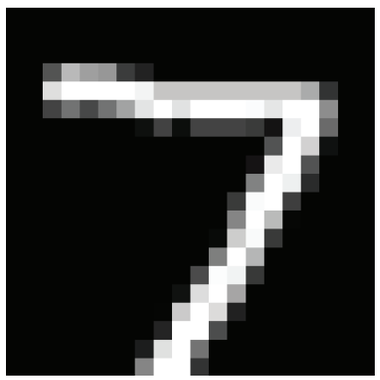
Real mapping of channel matrix (μS)



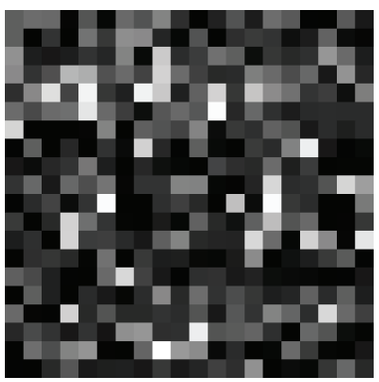
Error matrix (%)

Noiseless channel → **Noisy channel**

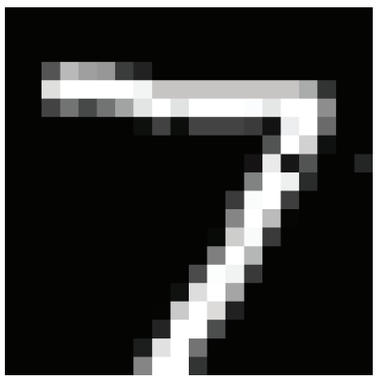
Performance Evaluation: Complete System



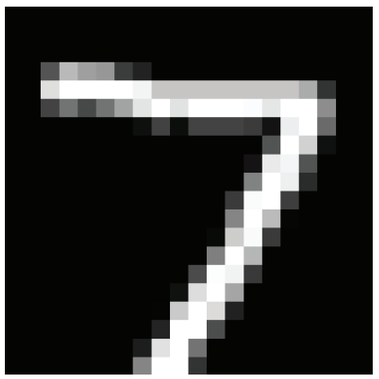
original



write-without-verification



write-with-verification



Digital processor (benchmark)

Parameters:

- 2048 subcarriers for OFDM
- 4 x 4 antennas for MIMO
- Channel SNR = 30dB

RRAM-based baseband processing

Verification is necessary!

System Performance Improvements

100-Time Faster and More Energy Efficient

	Latency (ms)	Energy (mJ)
Qualcomm Snapdragon X65	<10	N/A
Domain Adaptive Processor [1]	28.56	27.71
Combined FFT [2] + MIMO [3]	23.16	22.98
Our RRAM-based processor	**0.2322**	**0.01015**

[1] K.-Y. Chen, *et al*., "A 507 GMACs/J 256-Core Domain Adaptive Systolic-Array-Processor for Wireless Communication and Linear-Algebra Kernels in 12nm FINFET", *Proc. VLSI Techn. Circuits*, 2022.

[2] S. Liu, *et al.*, "A high-flexible low-latency memory-based FFT processor for 4G, WLAN, and future 5G", *IEEE Trans. VLSI Syst.*, vol. 27 no. 3, pp. 511-523, 2018.

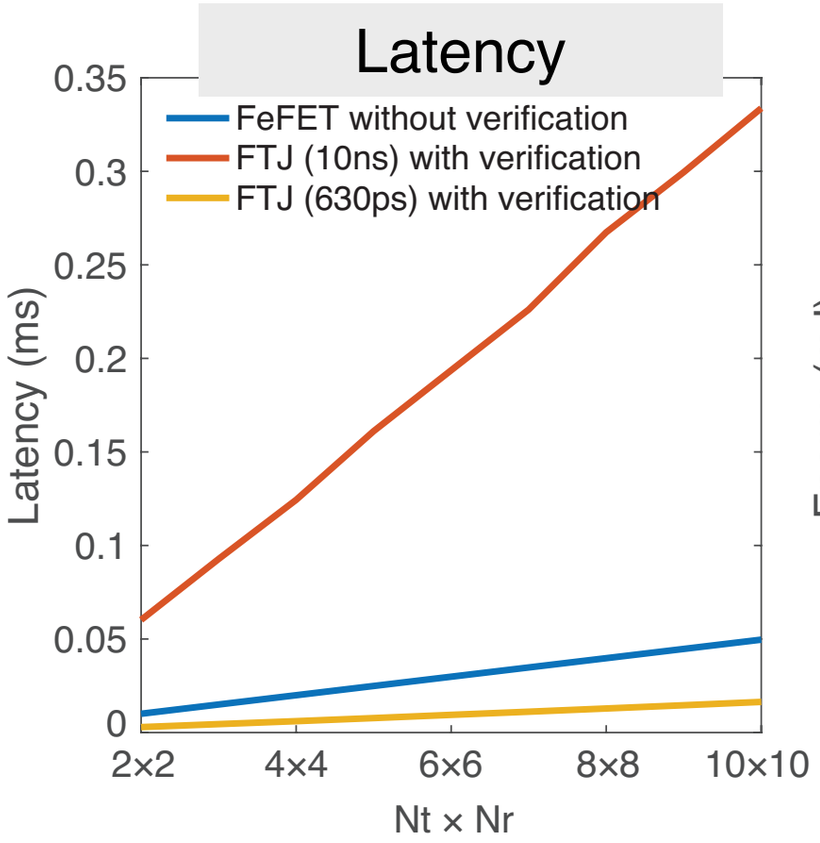
[3] W. Tang, *et al*. "A 2.4-mm^2 130-mW MMSE-Nonbinary LDPC Iterative Detector Decoder for 4×4 256-QAM MIMO in 65-nm CMOS." *IEEE J. Solid-State Circuits*, vol. 54, no. 7, pp. 2070-2080, 2019.

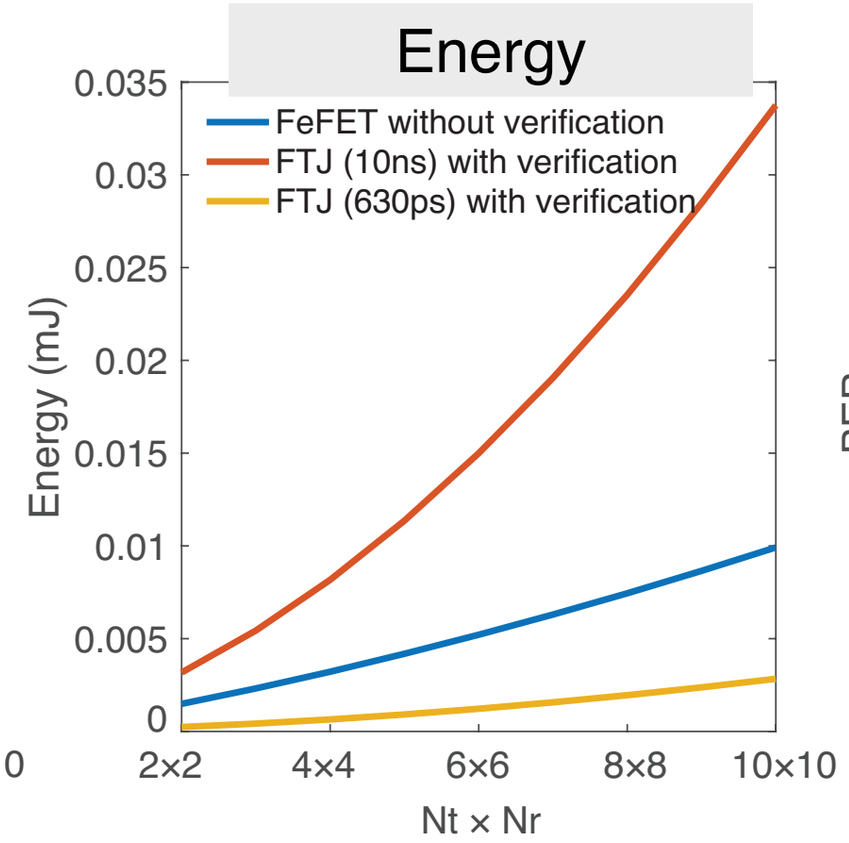
System Performance Improvements

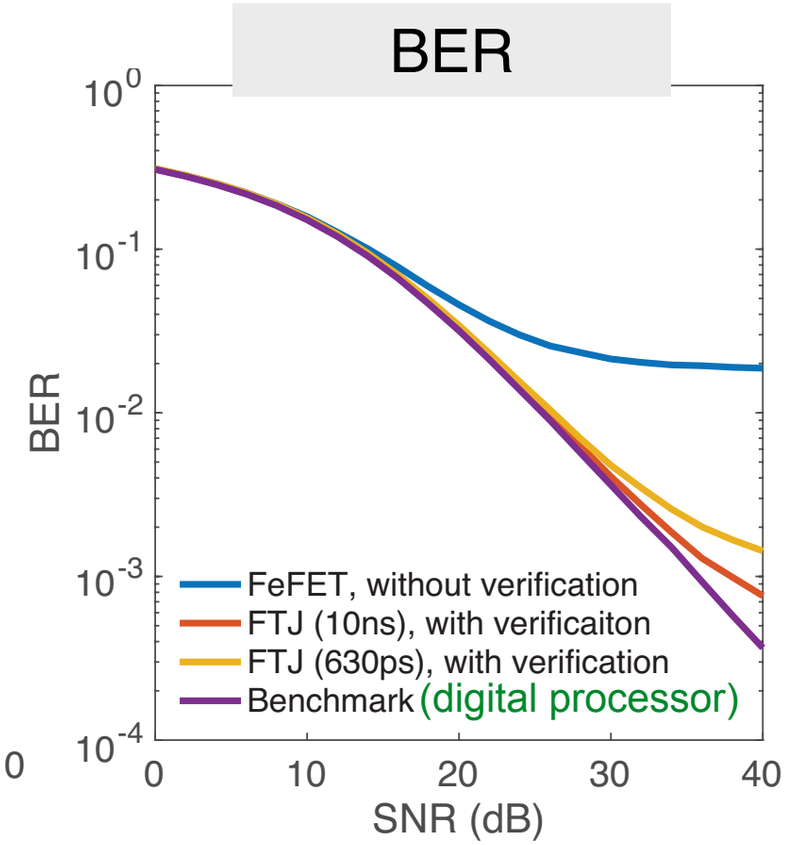
- Latency — Several microseconds (μs)
- Energy — Several micro-Jules (μJ)
- Performance - Approach digital baseband

Memristor Models:

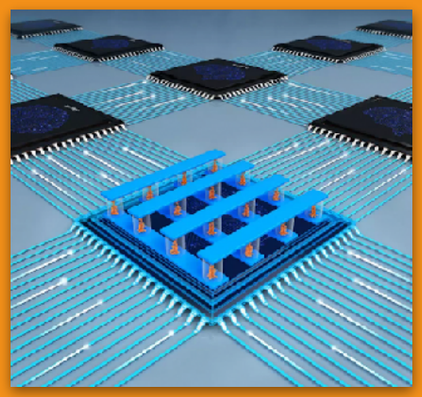
- Ferroelectric field-effect transistor (FeFET)
- Ferroelectric tunnel junction (FTJ)





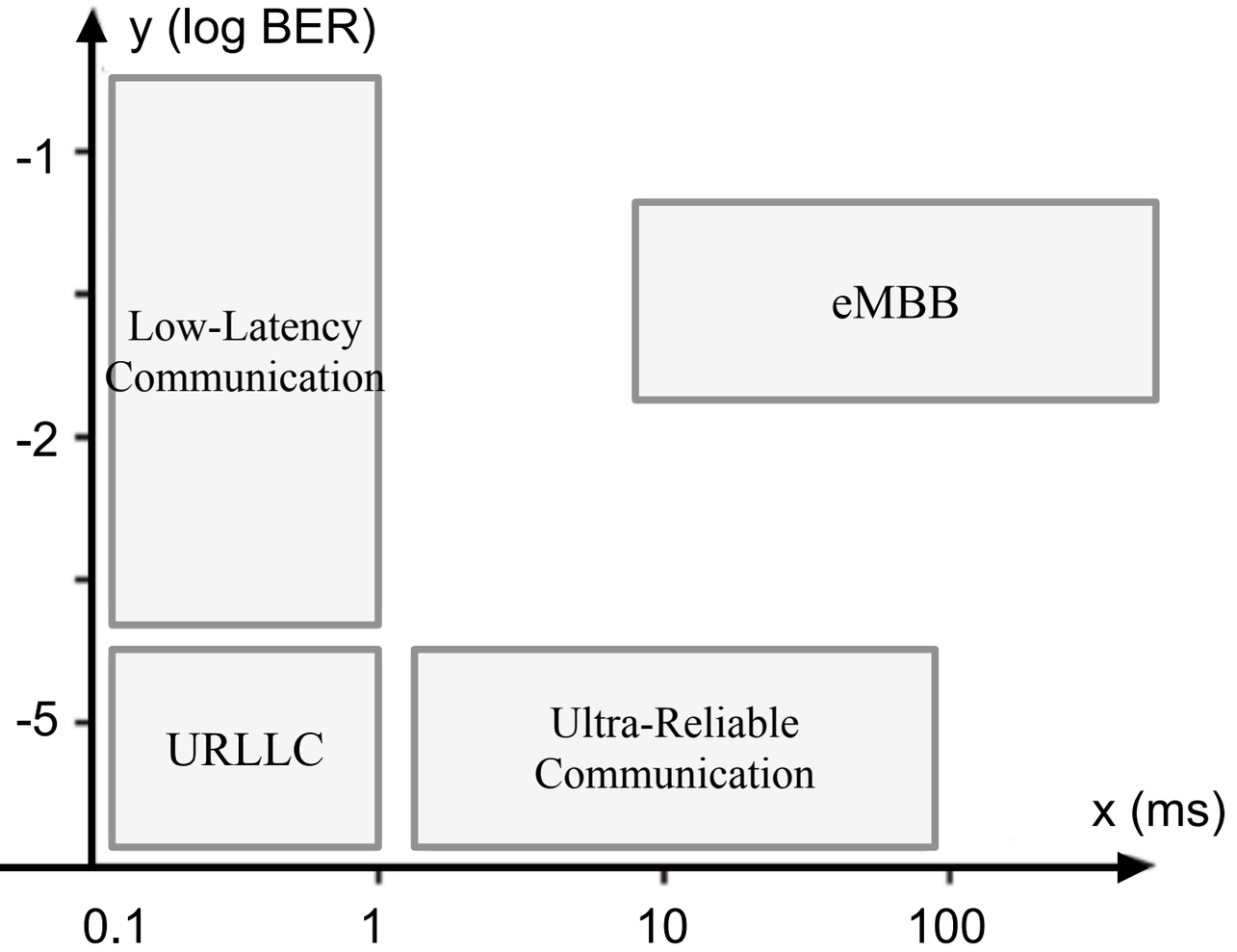


Analog Communication and Computing Are not Dead



Analog Communication, Baseband and Neuromorphic Computing

?



Thank You