The Future of Hardware Technologies for Computing

N3XT 3D MOSAIC, Illusion Scaleup, Co-design

Subhasish Mitra

Department of EE and Department of CS

Stanford University

Thanks: Students, Sponsors, Collaborators

Relays, Vacuum Tubes, Discrete Transistors, ICs

TURNING POTENTIAL INTO REALITIES:
THE INVENTION OF THE INTEGRATED CIRCUIT

[Kilby Nobel Lecture 2000]

"yields … too low to be profitable"

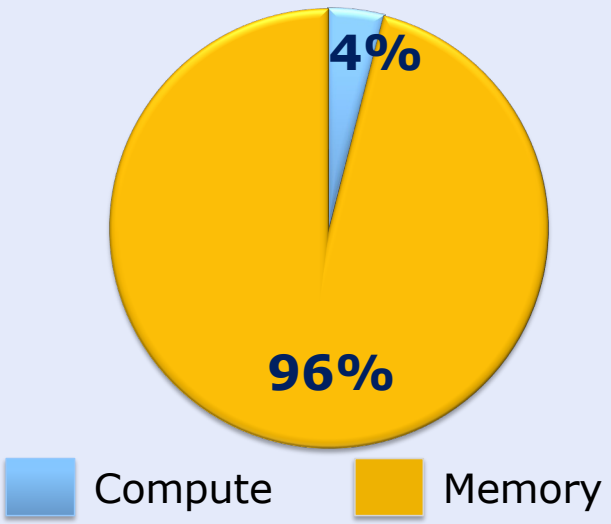
"best [devices] … not made with semiconductors"

"elegant devices messed up with all the other stuff"

Abundant-Data Computing

Many walls simultaneously

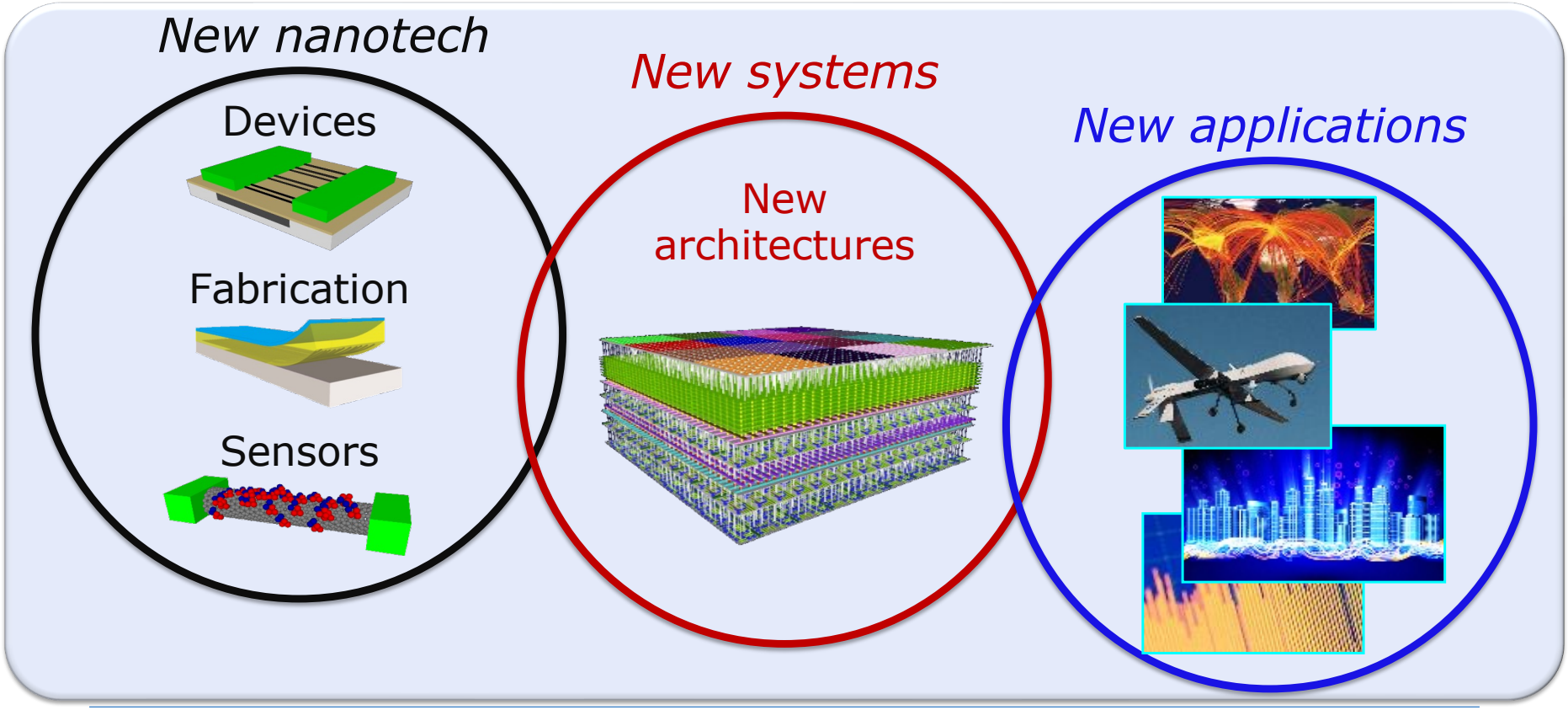
Memory Wall



Miniaturization Wall

✚ power wall, cooling wall, resilience wall …

Nano*Systems*



Data Explosion & Memory Wall

Both Von Neumann <u>&</u> non-Von Neumann architectures

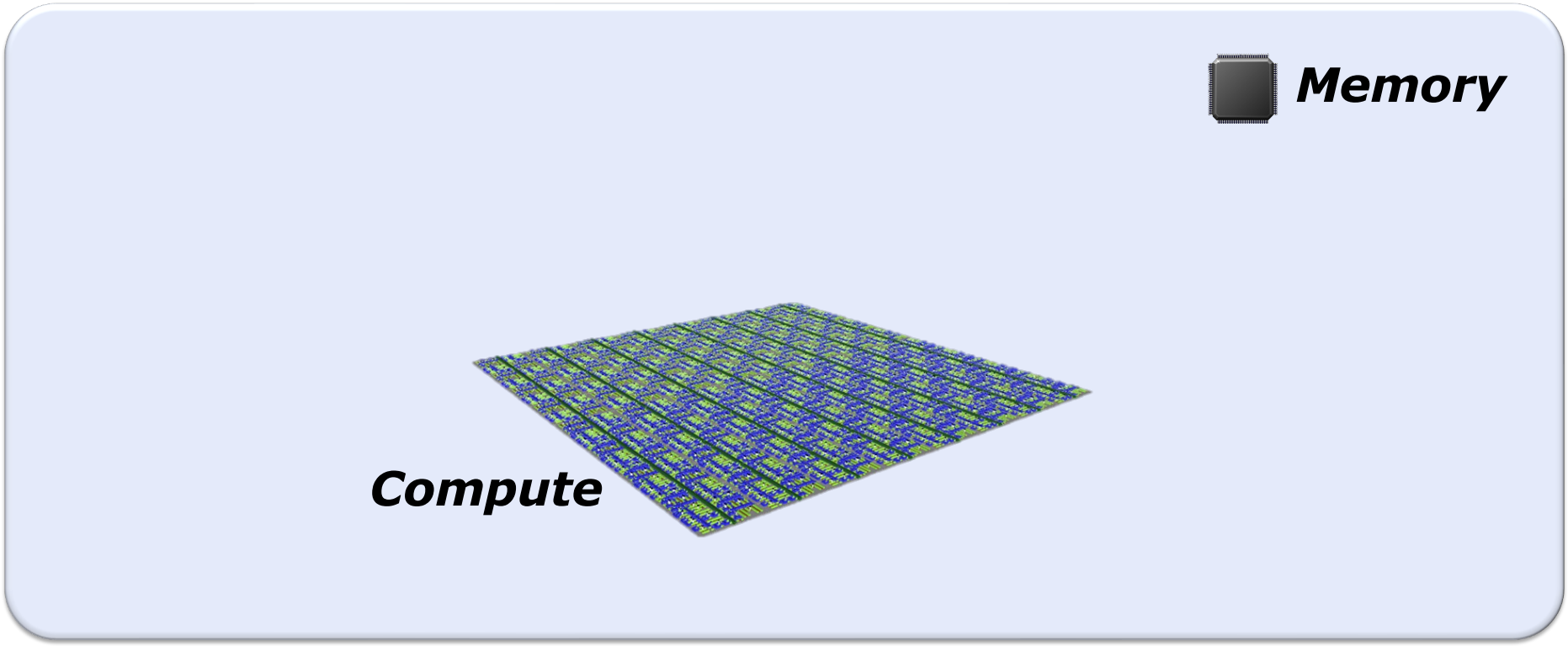
"Ideally … desire an indefinitely large memory capacity such that any particular … word would be immediately available. … It does not seem possible physically to achieve such a capacity. We are therefore forced … hierarchy of memories"

[Burks, Goldstine, Von Neumann, 1946]

Compute-in-X

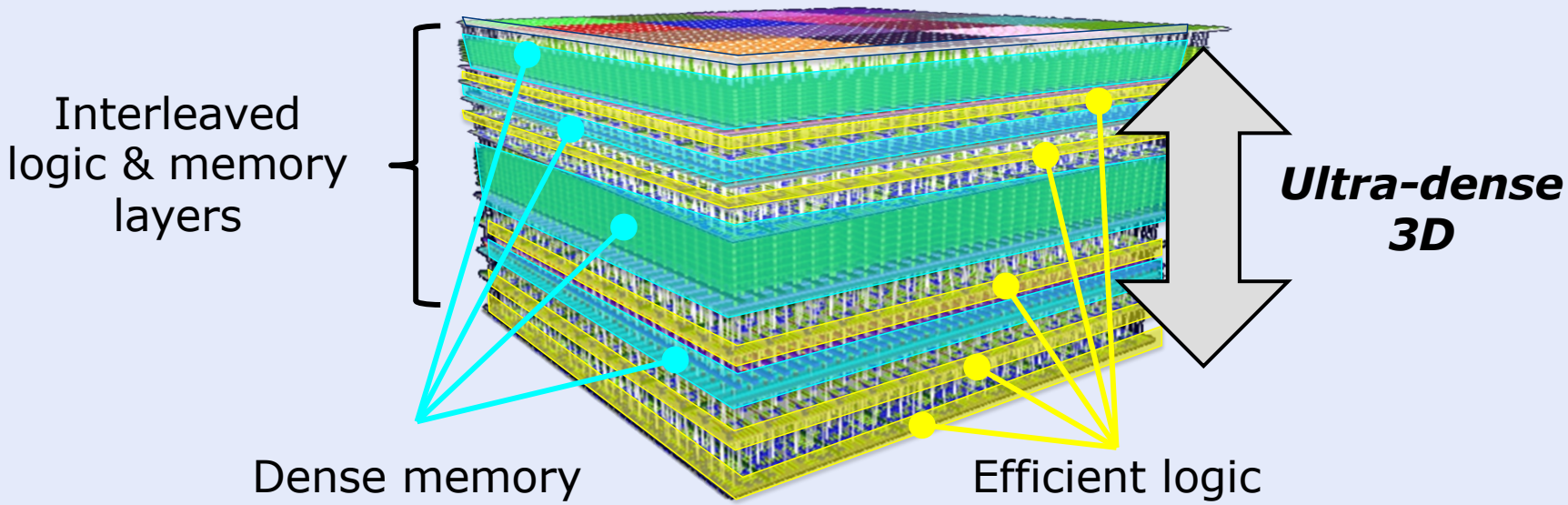
Computation immersed in memory (& sensors)

Computing Today



N3XT 3D: Computation immersed in Memory

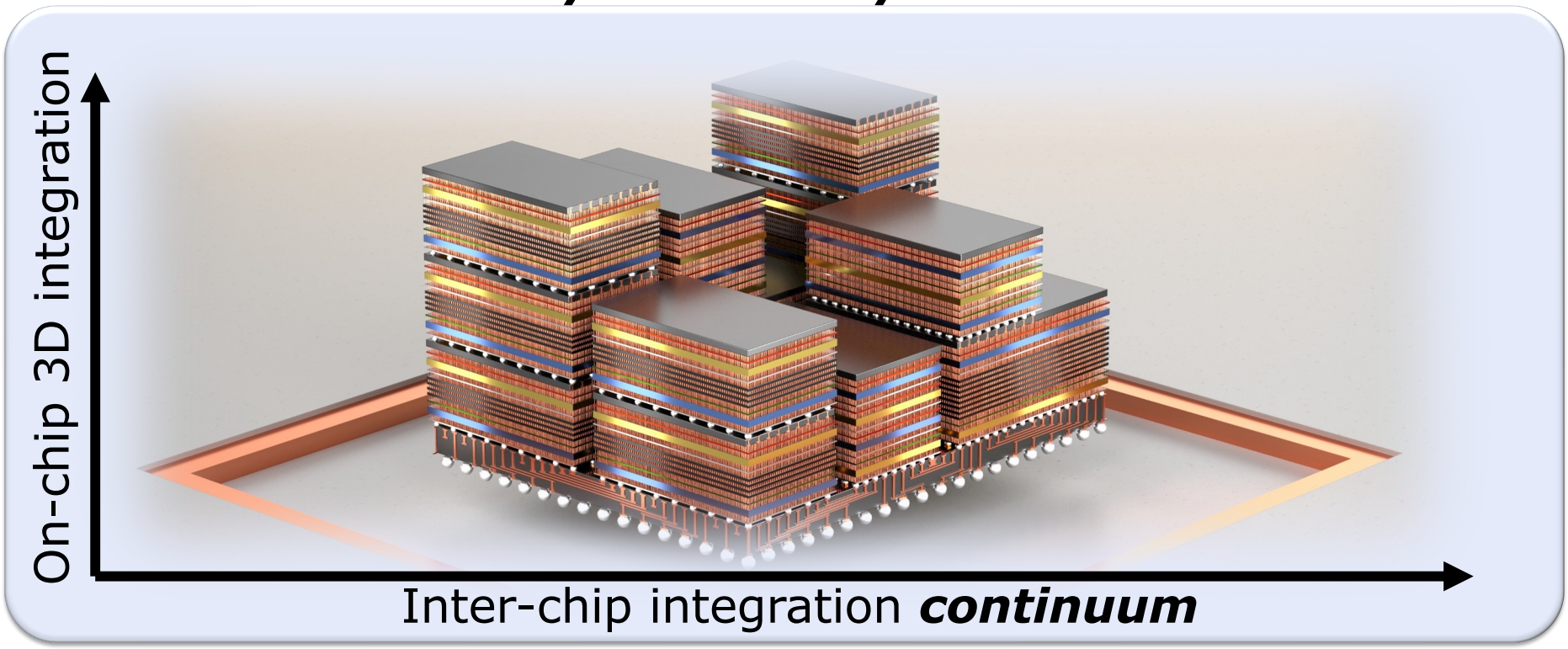
Nano-Engineered Computing Systems Technology



100× – 1,000× Energy Delay Product (EDP) benefits

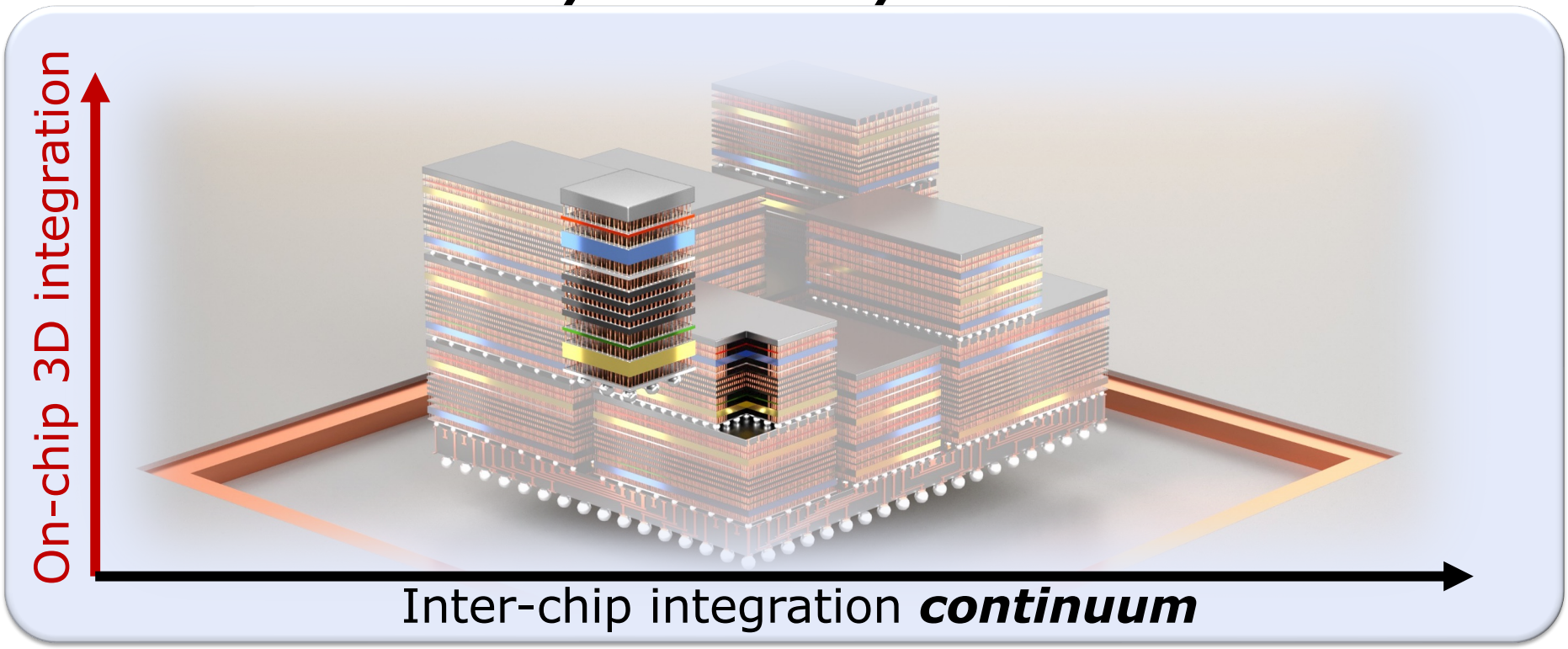
N3XT 3D MOSAIC

MO**nolithic / **S**tacked / **A**ssembled **IC



N3XT 3D MOSAIC

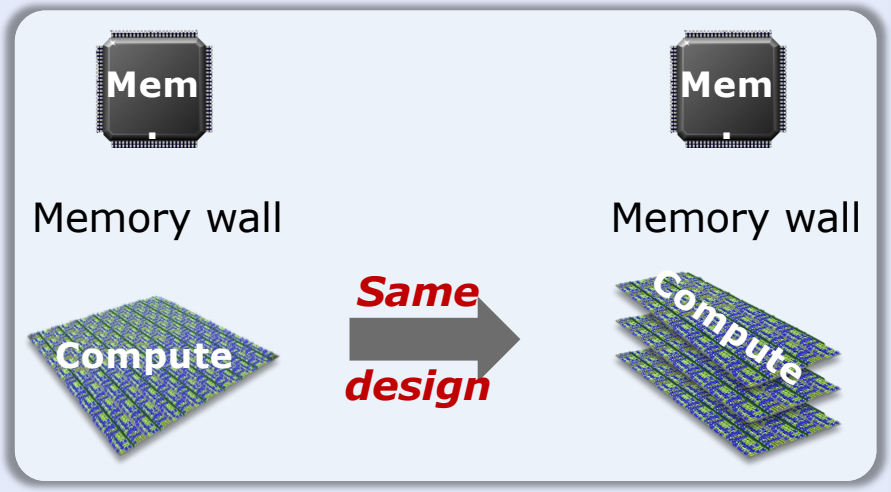
MO**nolithic / **S**tacked / **A**ssembled **IC



N3XT 3D ⊃ 3D Folding

3D Folding

Limited EDP benefits: 1.4×

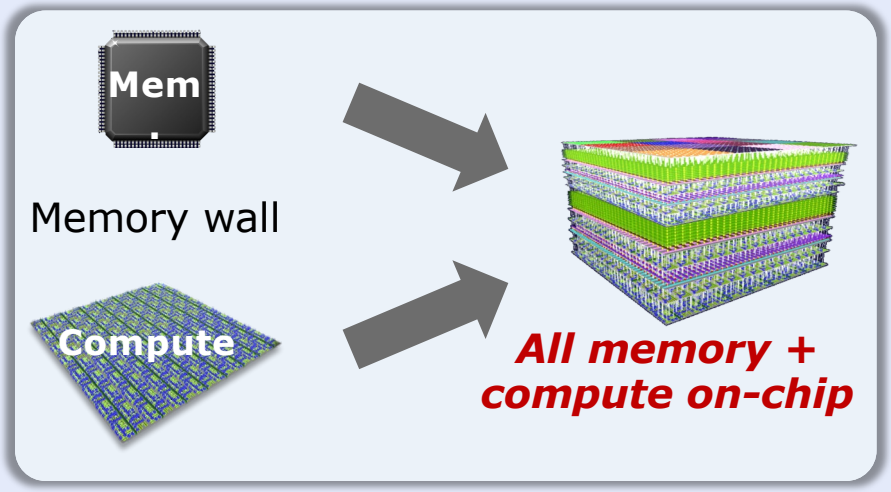


Some wirelength benefits

Memory wall stays

N3XT 3D

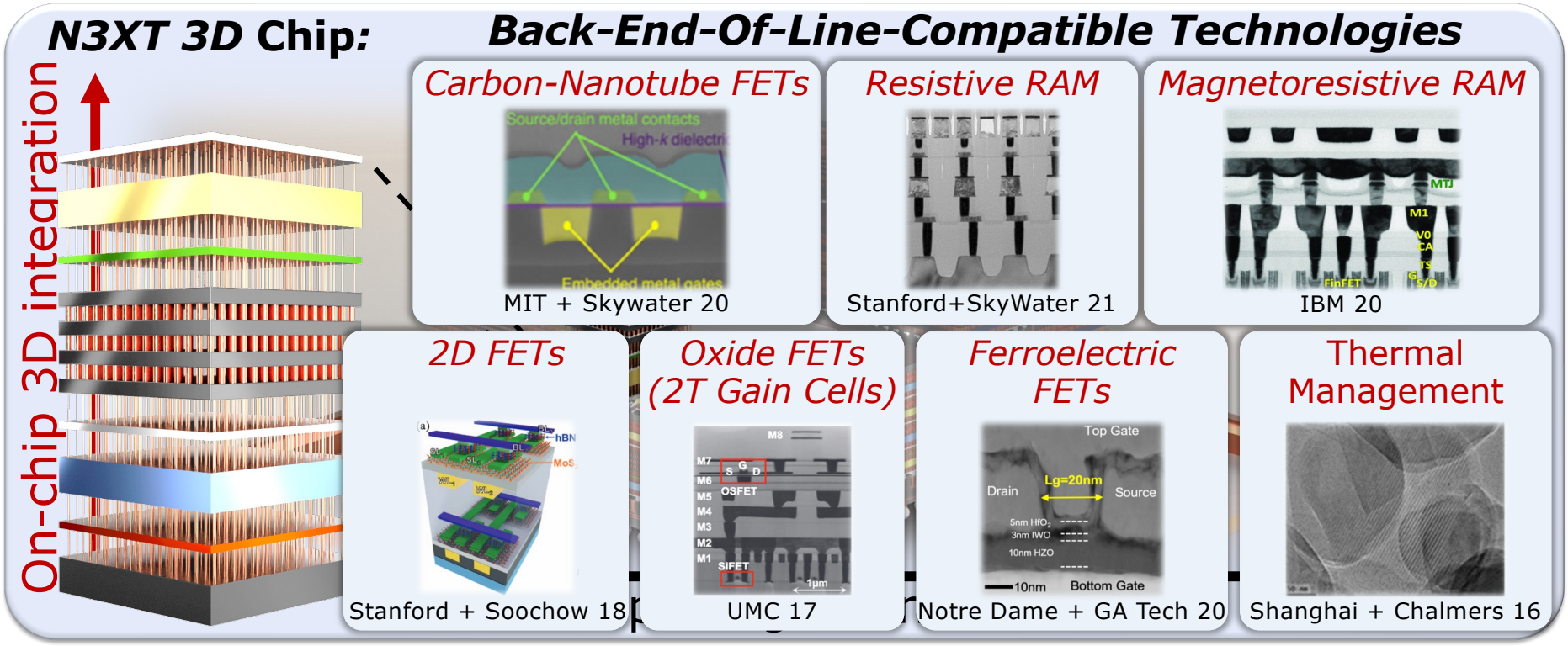
Large EDP benefits: 100-1,000×



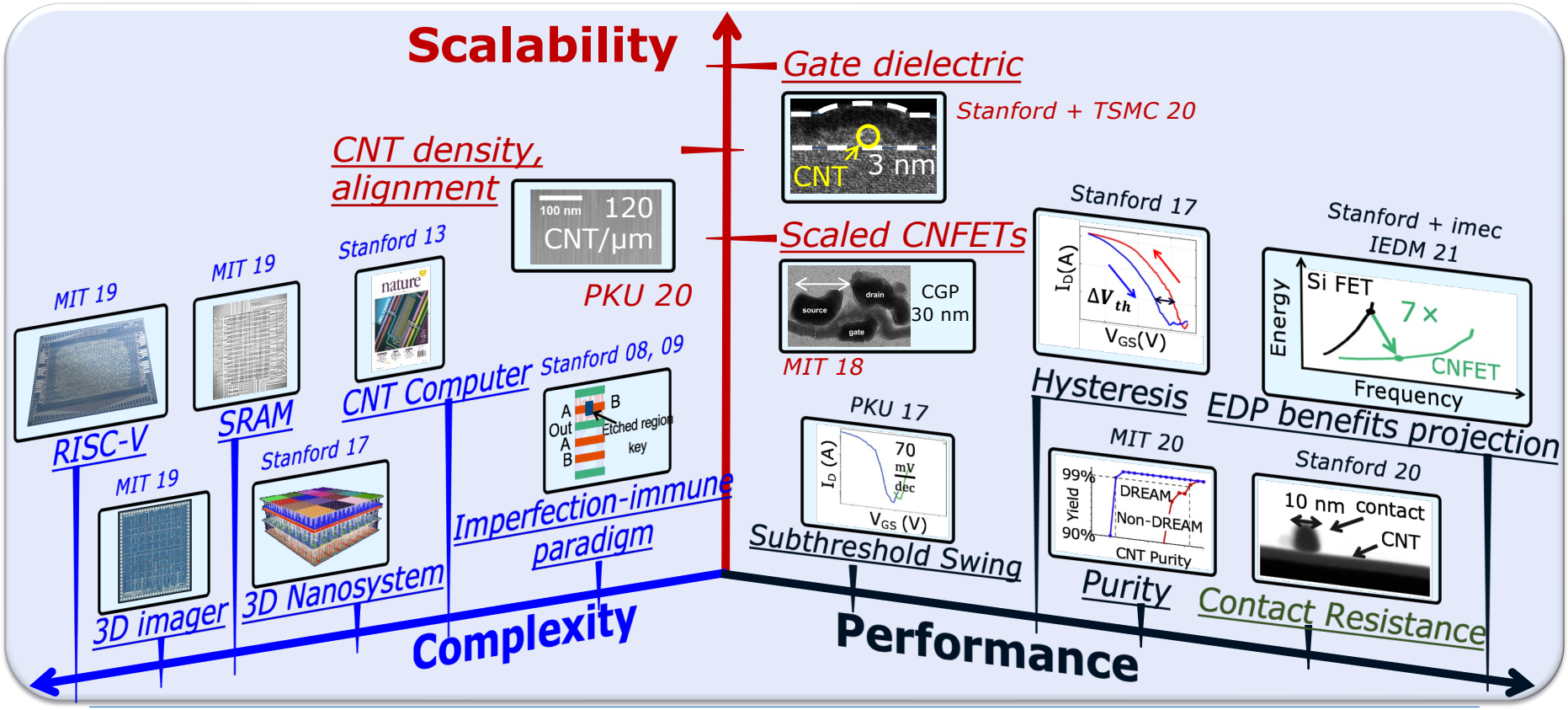
Many concurrent on-chip accesses

New arch. via new 3D physical design

N3XT 3D: Many Technologies

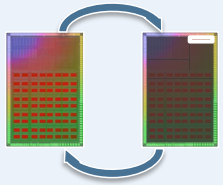


Carbon Nanotube FETs (CNFETs): Many Innovations

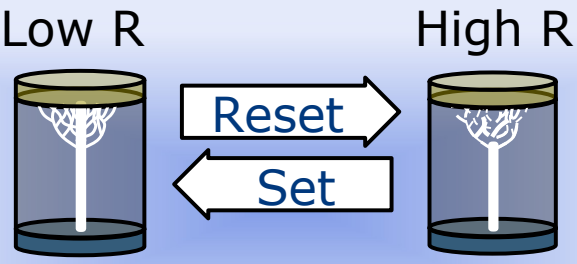


Resistive RAM (RRAM)

Non-volatile computing system

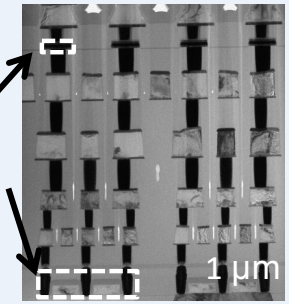


Stanford +
CEA LETI + NTU Singapore 19



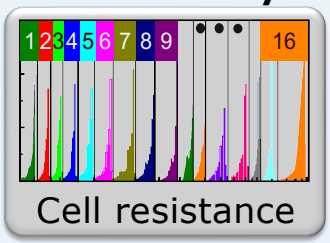
Large on-chip memory

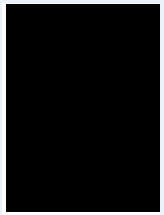
3 Million RRAM cells



Stanford + UCSD
+ Tsinghua + Notre Dame 20

Multi bits/cell arrays

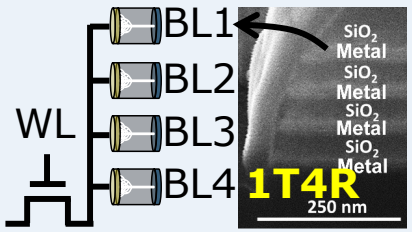




Stanford + CEA LETI 19,
Stanford + SkyWater 21

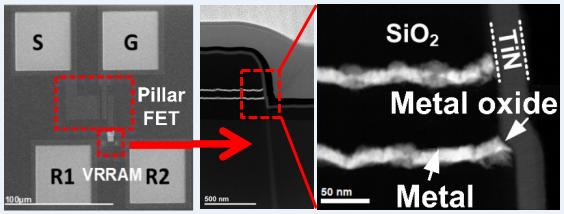
1TnR

Stanford +
SkyWater 21



Stanford 21

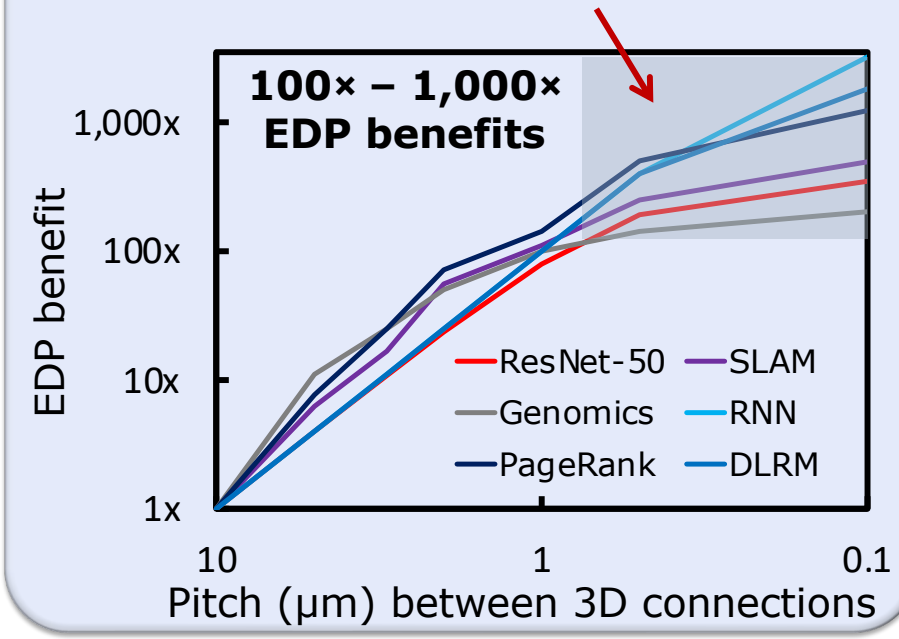
3D Vertical RRAM



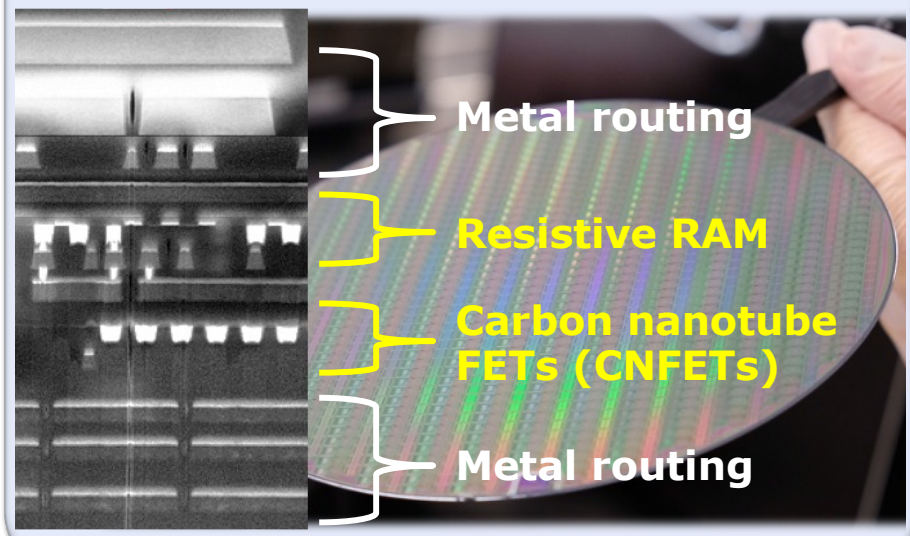
Stanford 21

Ultra-dense Monolithic 3D

Ultra-dense (e.g., monolithic) 3D crucial

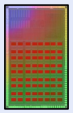


Low-temperature fabrication



Simulations: 88% accurate vs. hardware

Lab Nano*Systems* Hardware

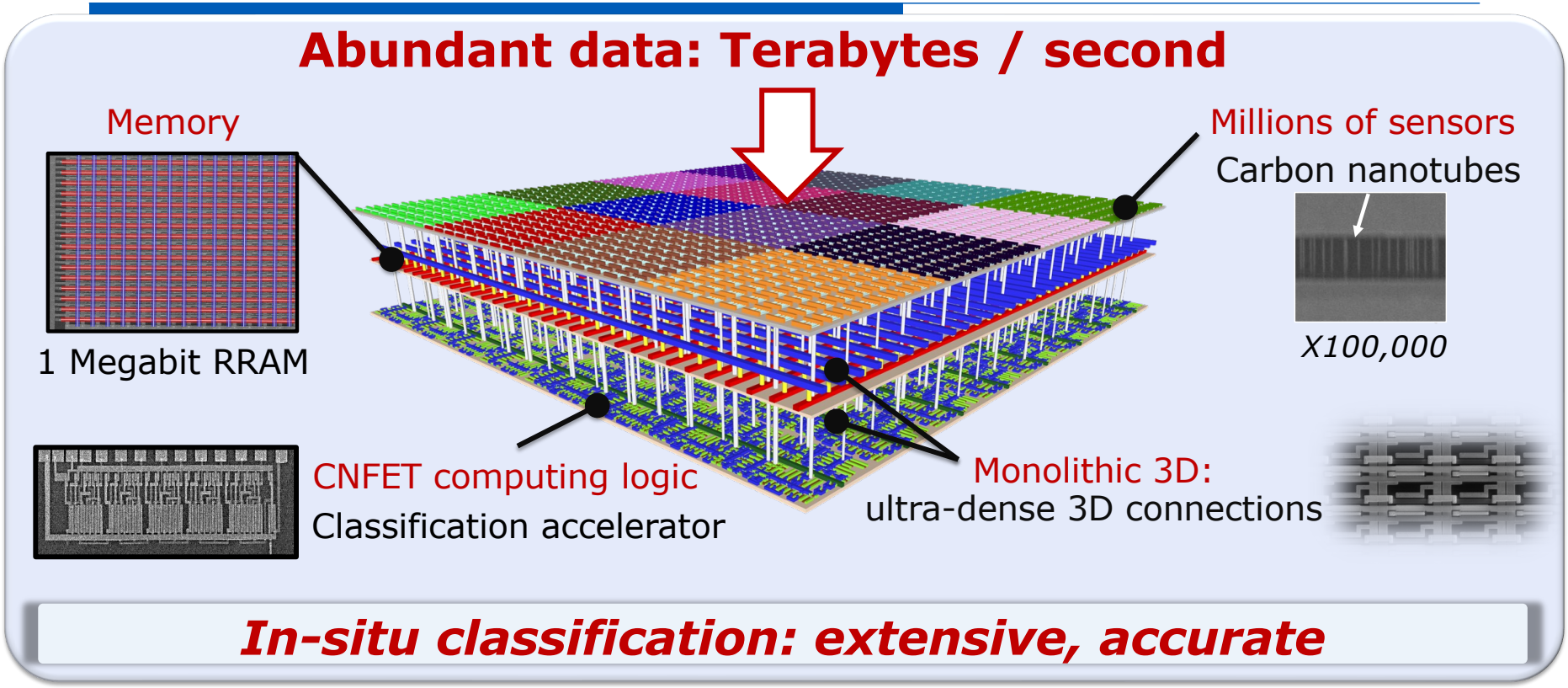
 Edge AI: training + inference [VLSI 21] Silicon, foundry RRAM	340× better EDP vs. Stochastic Gradient Descent
 Illusion [Nature Electronics 21] Multi-chip system: Silicon, RRAM	Illusion of all memory + compute on-chip
 CNFET RISC-V [Nature 19] CNFET	14,702 CNFETs (CMOS)
 Non-volatile IoT [ISSCC 19] Silicon, RRAM 2.3 bits per cell	10× better energy vs. embedded Flash
 Hyperdimensional [ISSCC 18] CNFET, RRAM, monolithic 3D	Brain-inspired, one-shot learning
 3D NanoSystem [Nature 17] Silicon, CNFET, RRAM, monolithic 3D	First heterogeneous monolithic 3D NanoSystem
 CNFET computer [Nature 13] Carbon nanotube FET (CNFET)	First CNFET processor

Lab to Fab

Analog Devices

SkyWater Tech. Foundry

3D NanoSystem



[Shulaker Nature 17]

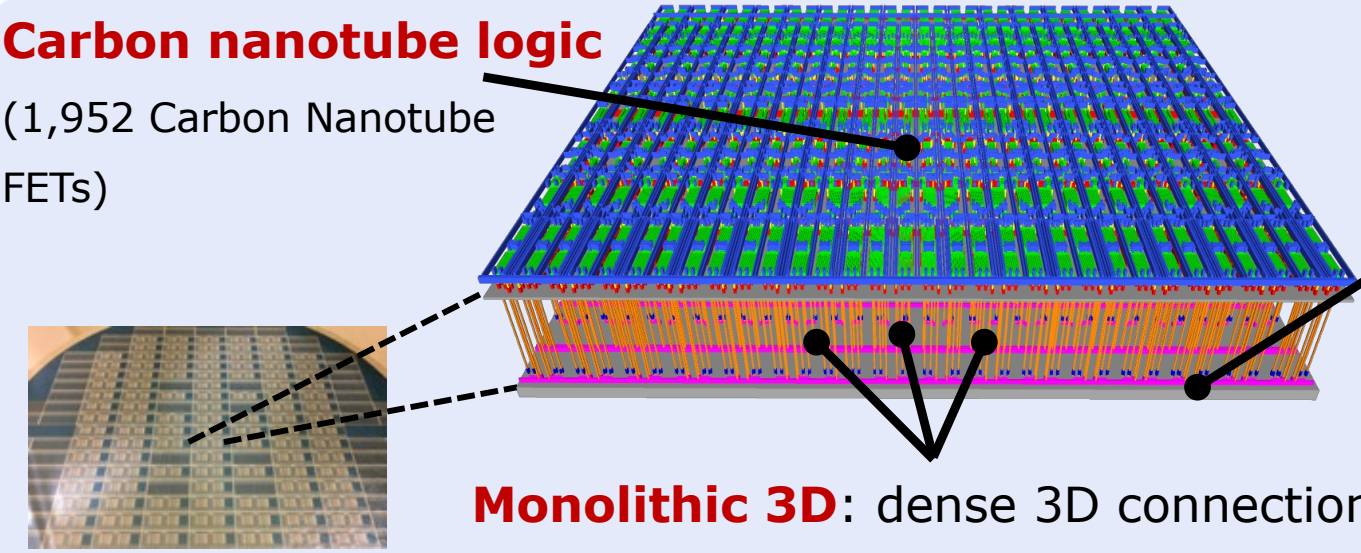
HD Computing: Brain-Inspired ⊃ Neural Nets

Carbon nanotube logic

(1,952 Carbon Nanotube FETs)

RRAM TCAM

(224 RRAM cells)



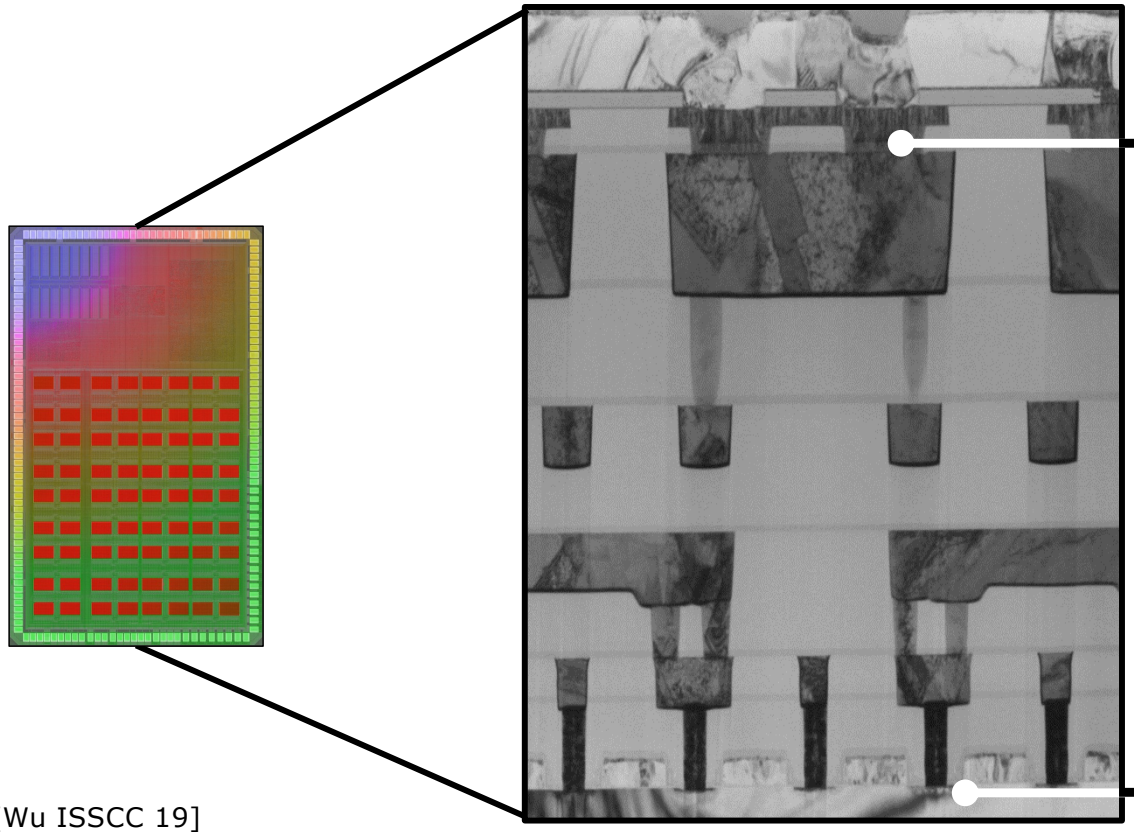
Monolithic 3D: dense 3D connections

Exploit: inherent variations, RRAM gradual Reset, application resilience

Language classification, one-shot learning

[Wu ISSCC 18, IEEE JSSC 18] HD = Hyperdimensional TCAM = Ternary Content Addressable Memory

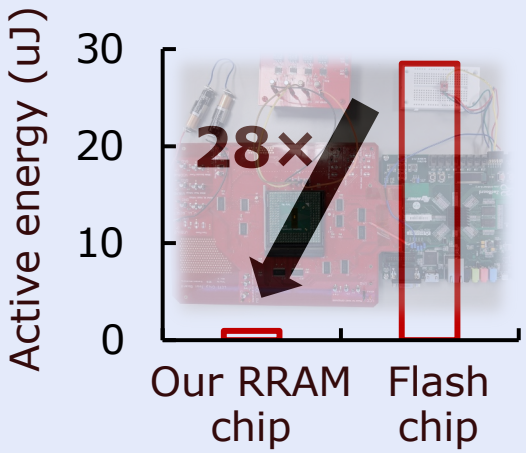
Non-volatile Computing System: RRAM + Silicon CMOS



[Wu ISSCC 19]

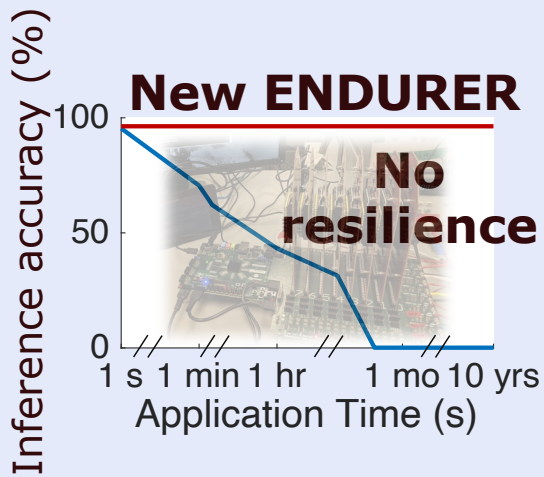
Non-volatile Computing System: RRAM + Silicon CMOS

Fine-grained temporal power gating



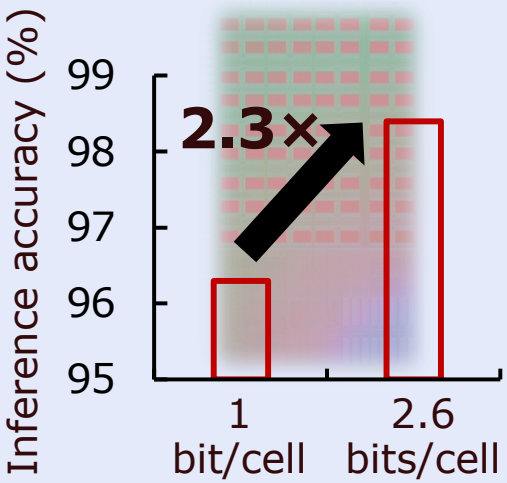
10× battery life vs. Flash chip

New RRAM Endurance



10-year continuous AI inference

1st RRAM System: Multiple bits/cell



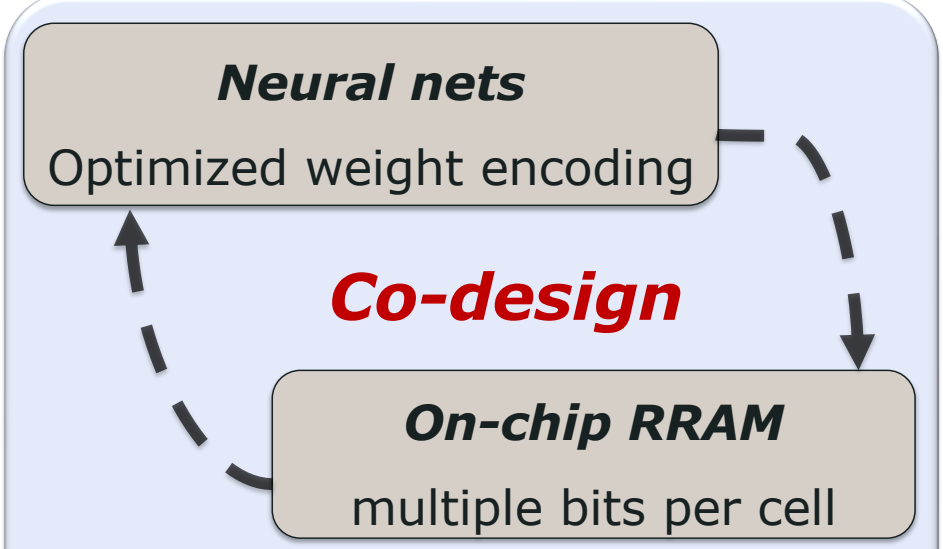
New RRAM algorithms key

[Wu ISSCC 19]

Multiple bits-per-cell RRAM System

	Bits per cell	Cells measured
Our work *new algorithms*	3, 4	Arrays
Prior work *ad hoc*	2-6.5	Single cell, hand-picked cells

[Le IEEE TED 19, IEEE TED 21]



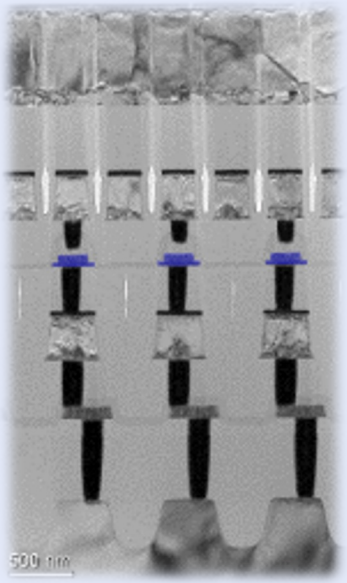
2.3× inference accuracy *despite RRAM variations*

Same hardware, bigger neural net

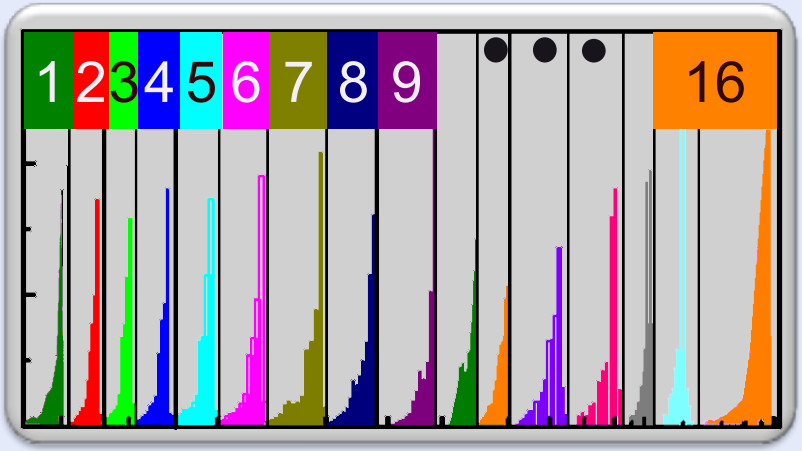
[Wu ISSCC 19]

1T*n*R × Multiple bits-per-cell RRAM

1T8R RRAM arrays



4 bits/cell 1T8R arrays



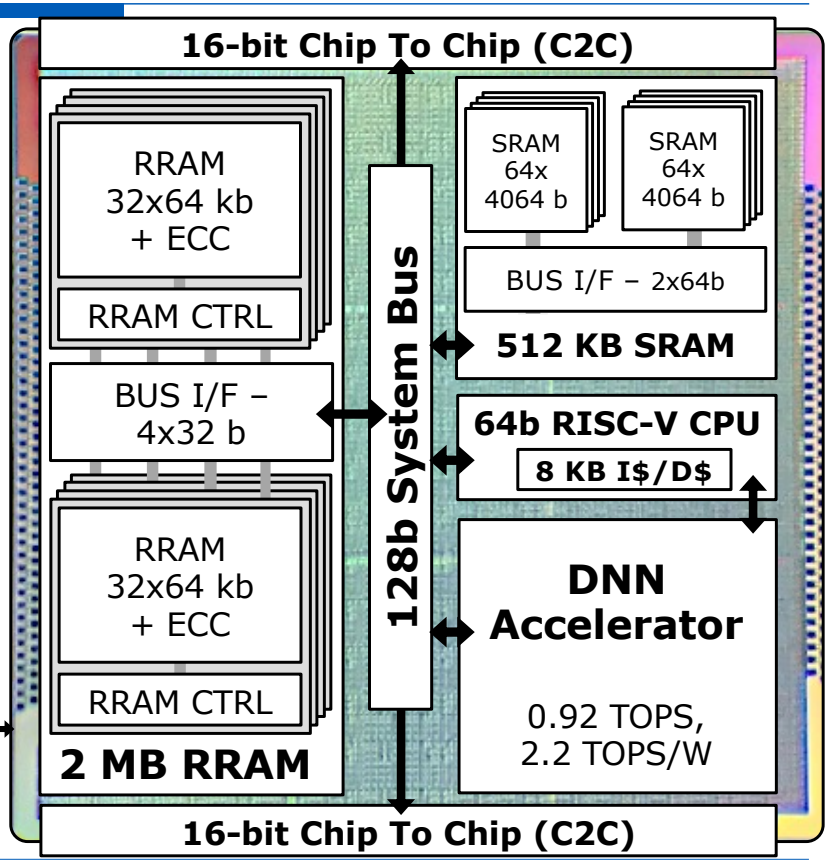
RRAM cell resistance

[Hsieh IEDM 19, IEEE EDL 21]

CHIMERA: RRAM Edge AI Inference & Training

- 2 MBytes foundry RRAM
  - Neural net weights, CPU instructions
- 512 KBytes SRAM
  - Neural net activations
- Deep neural net (DNN) accelerator
  - 0.92 TOPS, 2.2 TOPS/W
- Chip-to-chip (C2C) Links
  - Multi-chip systems

~~Off-chip memory~~ ⟷ **No off-chip memory**

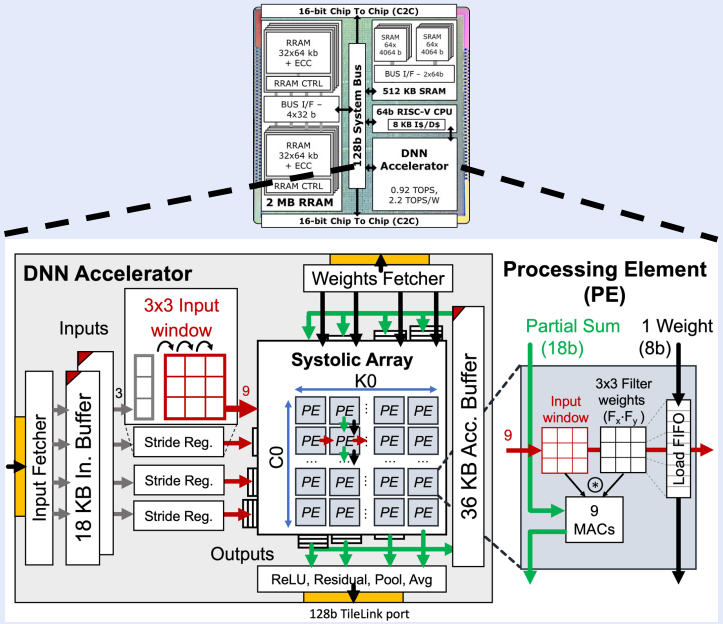


[Giordano Symp. VLSI Circuits 21]

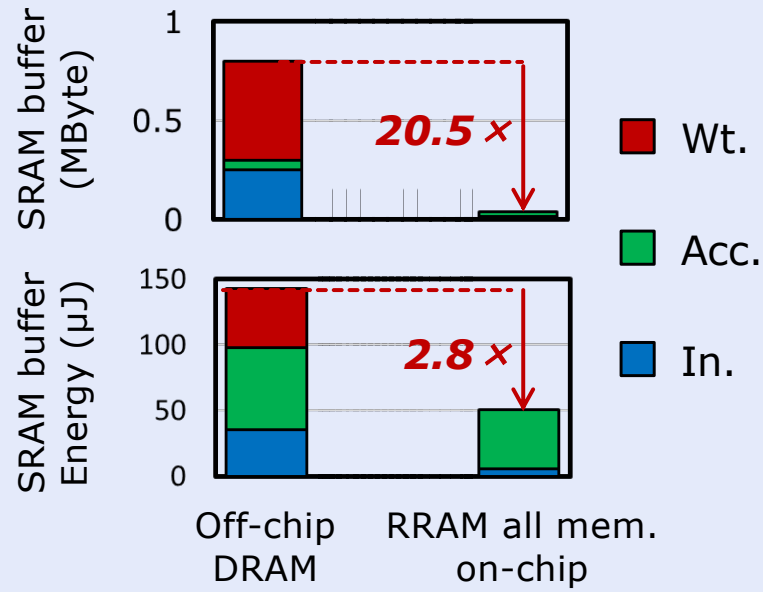
CHIMERA: RRAM All Memory On-chip Dataflow

New dataflow: large benefits

CHIMERA DNN Accelerator



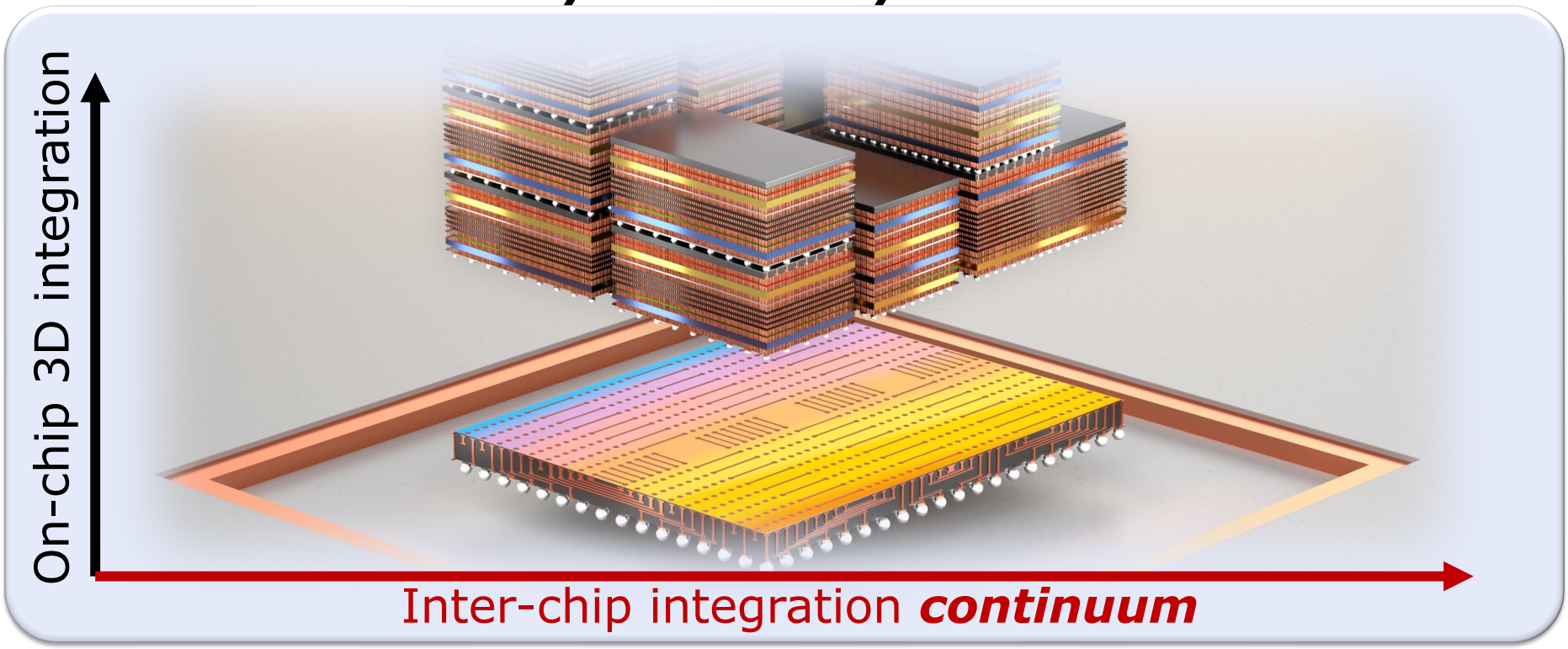
Dataflow Benefits



Data Buffers: Wt. = Layer weight, Acc.= Layer partial sum accumulation, In.= Layer Input

N3XT 3D MOSAIC

MO**nolithic / **S**tacked / **A**ssembled **IC



Dream: *All Memory* + Compute On-chip

Dream Chip: infeasible, *moving target*

Massive on-chip memory: *N* × *M*

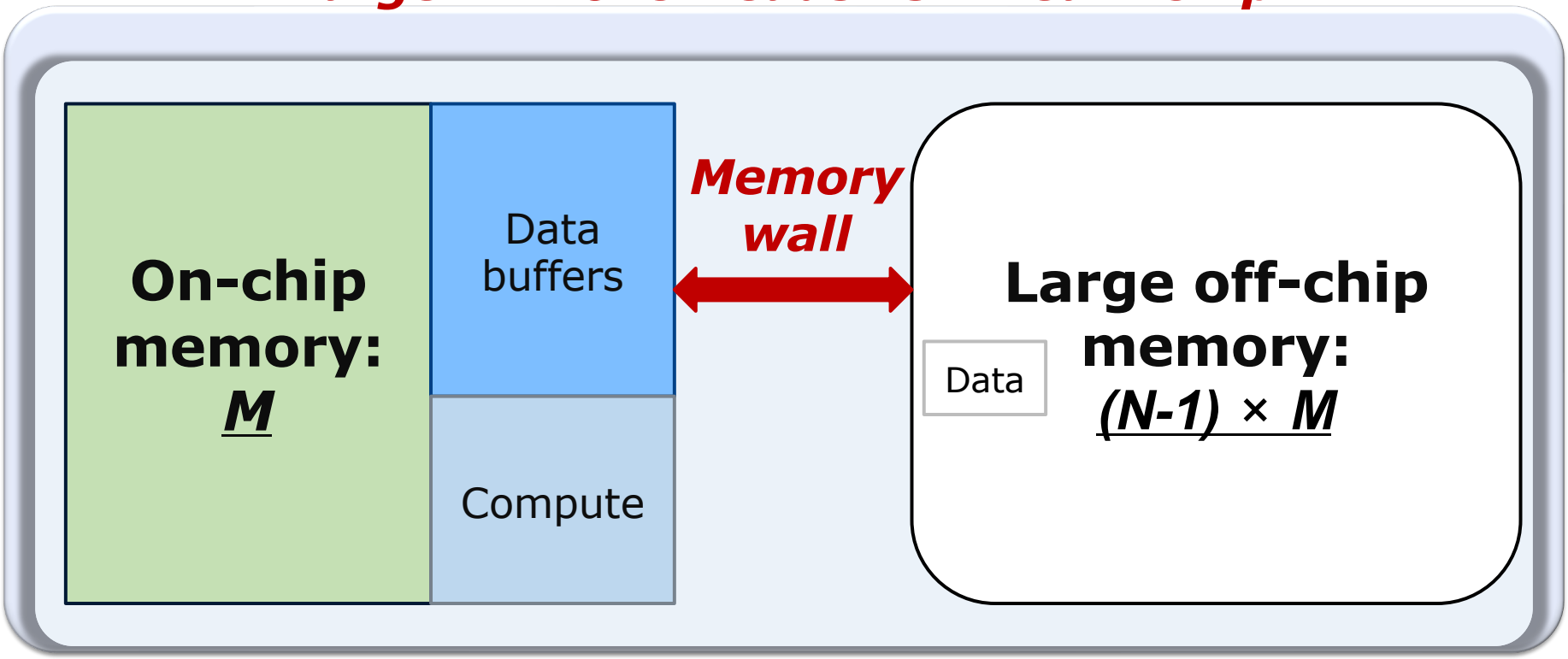
(Full workload fits in on-chip memory)

Data buffers

Compute

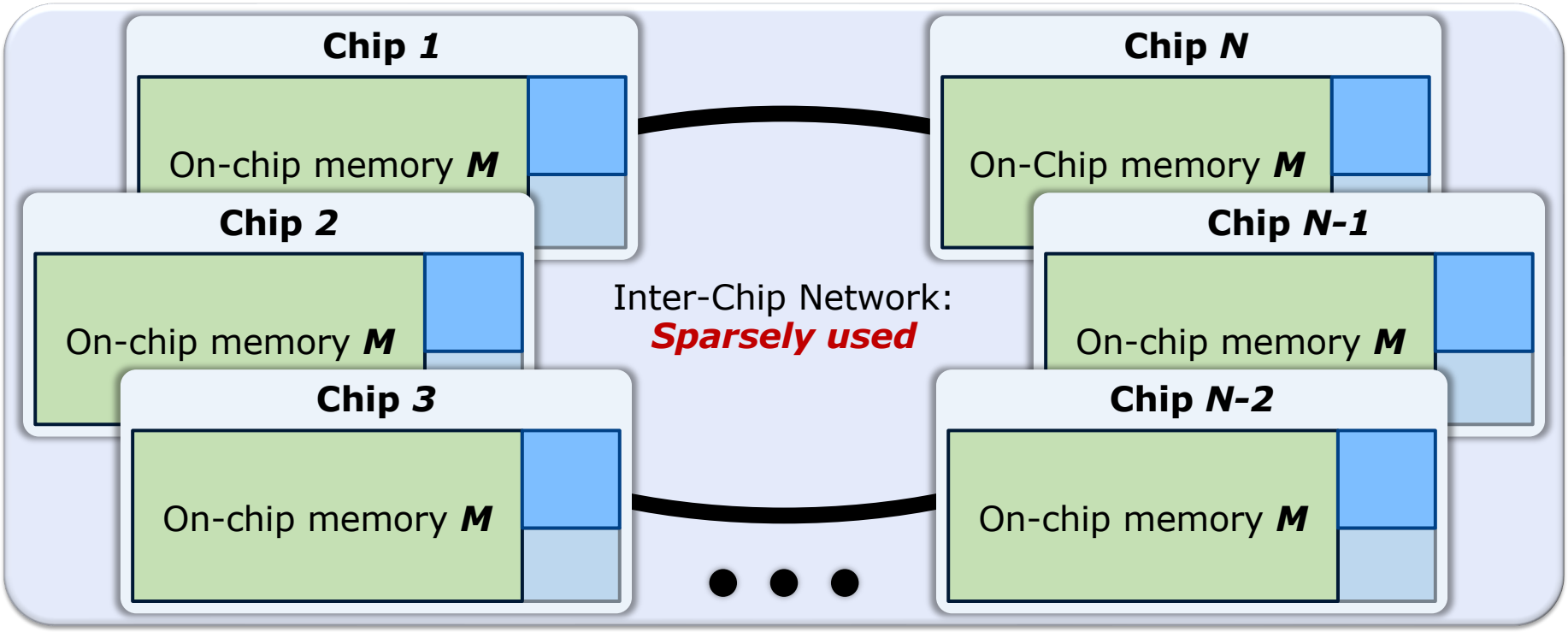
Off-Chip Memory Accesses Costly

Large EDP overheads vs. Dream Chip



Illusion System

1. *Enough* on-chip mem. **2. *Quick*** chip ON/OFF **3. *Special*** mapping



[Radway Nature Electronics 21]

Illusion Ideal for AI

Illusion ≈ Dream

1.1 × Dream EDP

Illusion Energy

≤ 1.05 ×

Dream Energy

Illusion Exec. Time

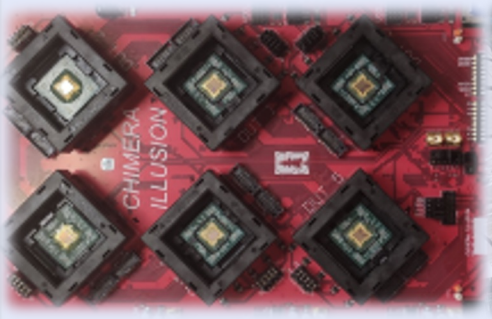
≤ 1.05 ×

Dream Exec. Time

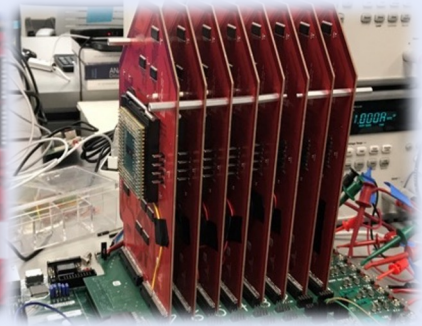
(measured for AI inference)

Hardware-proven

backed by theory



6-CHIMERA chip
Illusion system

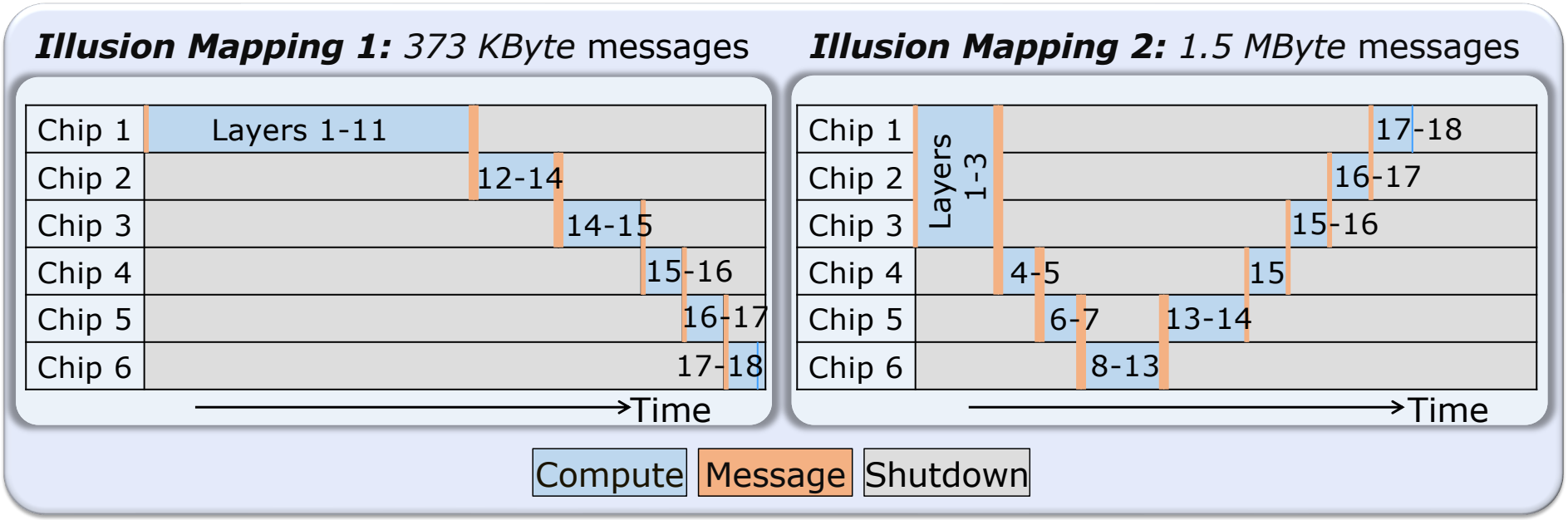


8-chip
Illusion system

[Giordano Symp. VLSI Circuits 21, Radway Nature Electronics 21]

Many Illusion Mappings: Energy/Exec. Time Tradeoffs

Illusion system using 6 CHIMERA chips: ***12 MByte*** ResNet-18



Both mappings ≤ 1.1× Dream EDP

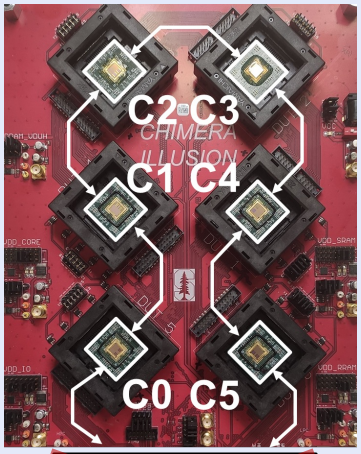
[Giordano Symp. VLSI Circuits 21]

CHIMERA Illusion Benefits vs. Off-chip Flash

12 MByte ResNet-18 Edge AI: 1 inference per second

(e.g., security, activity monitoring, healthcare, wildlife)

6-CHIMERA chip Illusion System



~~No off-chip Flash~~

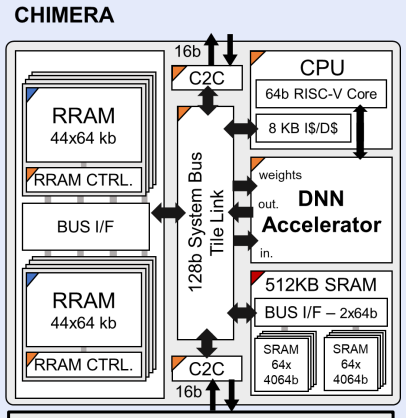
6 CHIMERA chips

12 MBytes RRAM overall

Minimal inter-chip messages

CHIMERA with off-chip Flash

(Non-volatility at the edge)



10 MBytes off-chip Flash (DNN weights)

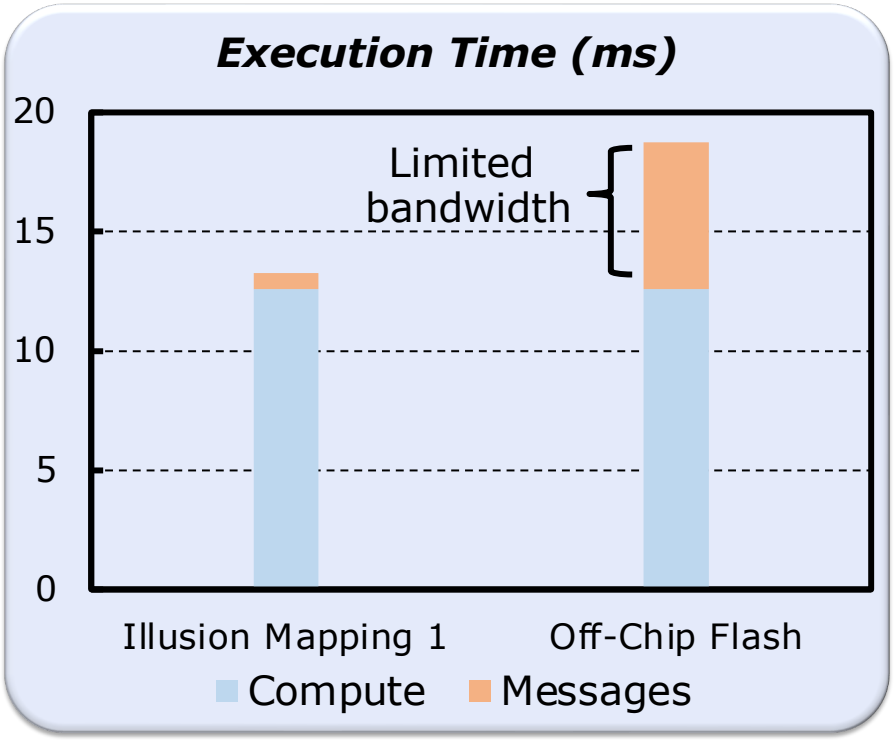
2 MBytes on-chip RRAM

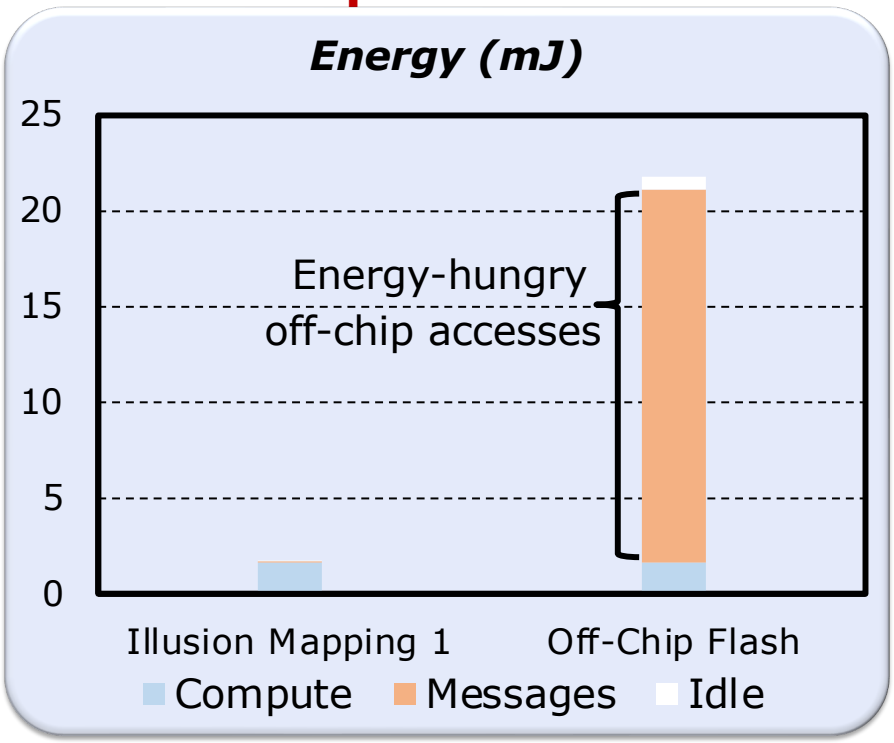
10 MBytes off-chip Flash

Read off-chip weights

CHIMERA Illusion Benefits vs. Off-chip Flash

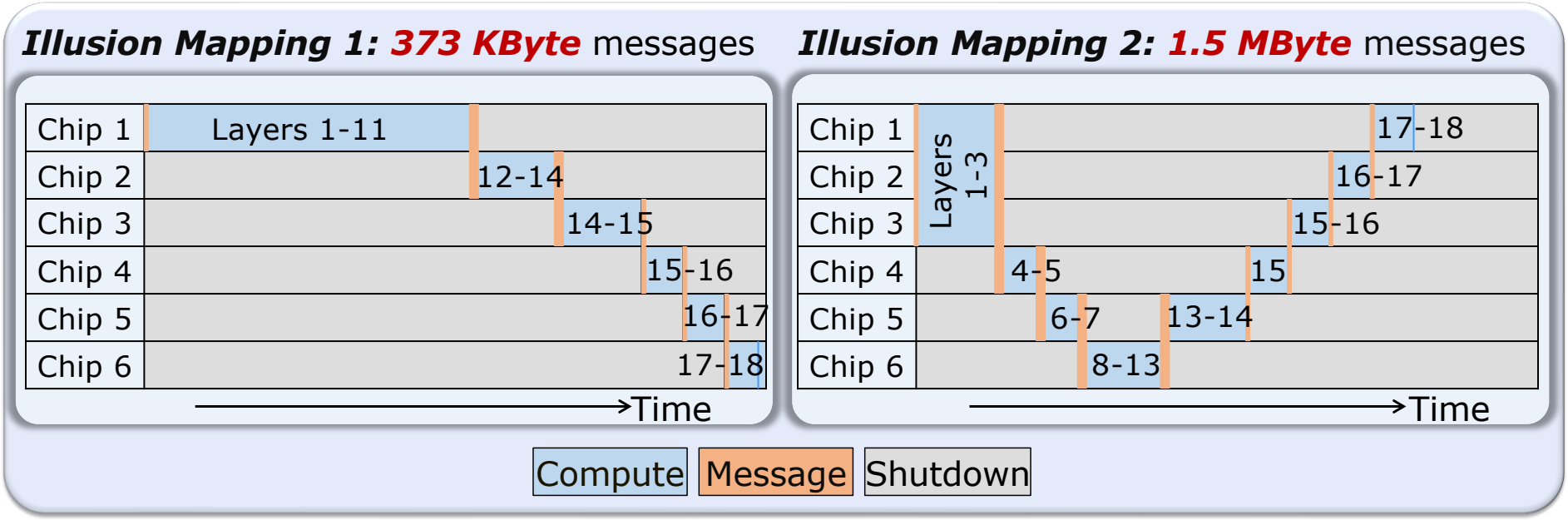
Illusion: *18 ×* lower EDP vs. off-chip Flash





Based on [Giordano, Symp. VLSI Circuits 2021], off-chip Flash: 0.6 GBytes/s, 2 nJ/Byte

Illusion ⊃ Traditional Parallel Computing



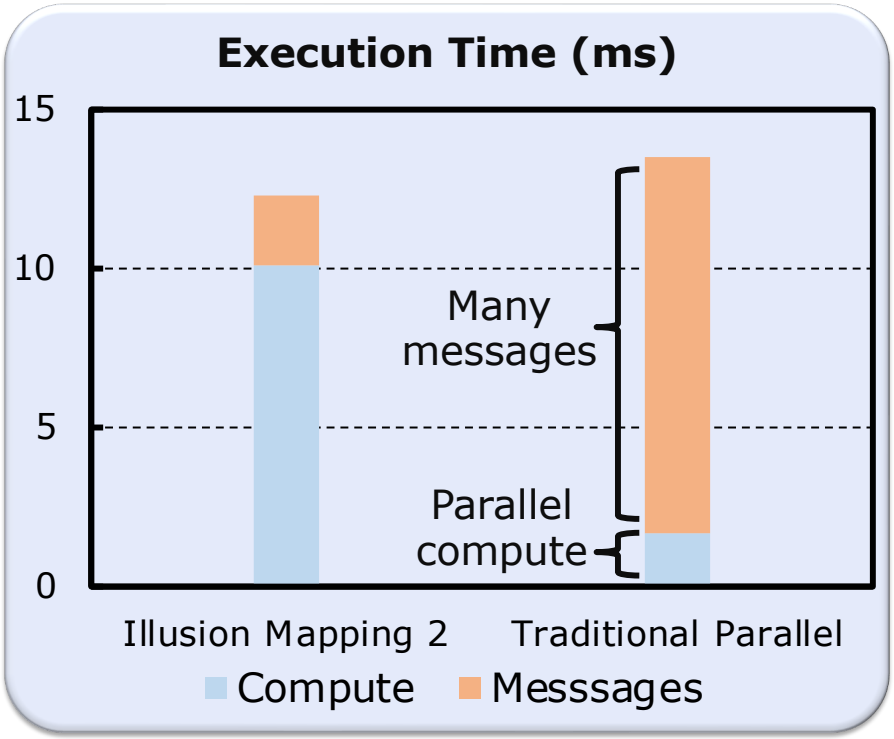
Traditional parallel: *10 MByte* chip-to-chip messages

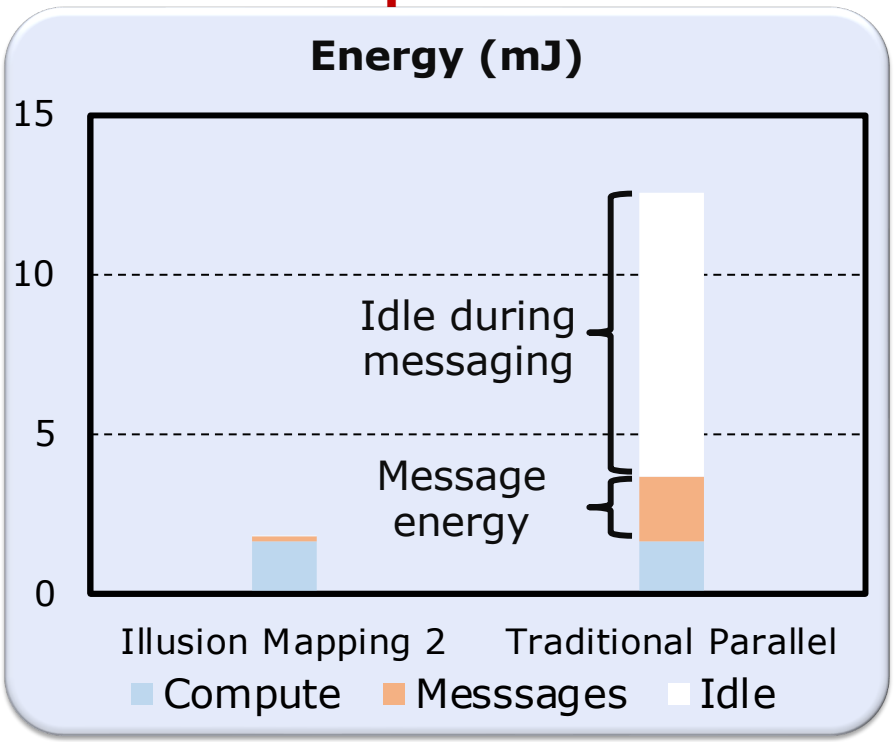
(*12 MByte* RestNet-18)

Traditional parallel: e.g., [Zimmer Symp. VLSI Circuits 19, Shao MICRO 19]

CHIMERA Illusion Benefits vs. Traditional Parallel

Illusion: *8 ×* lower EDP vs. traditional parallel



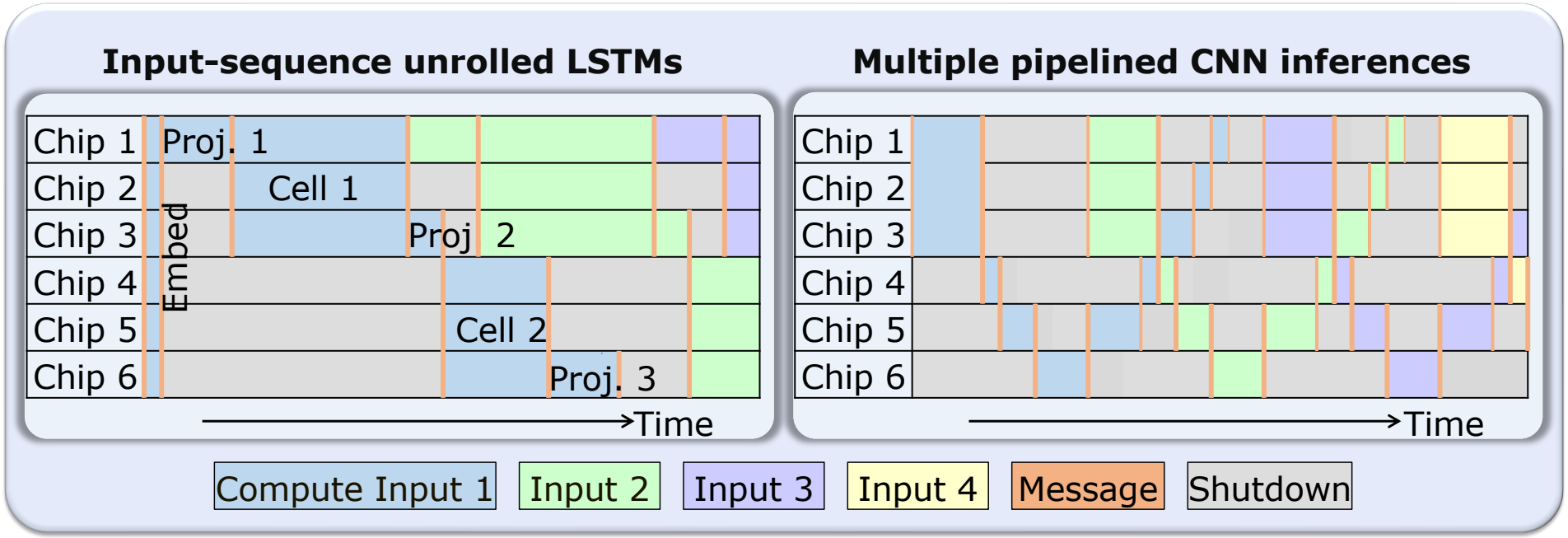


Illusion: [Giordano, Symp. VLSI Circuits 2021], traditional parallel: [Shao MICRO 2019]

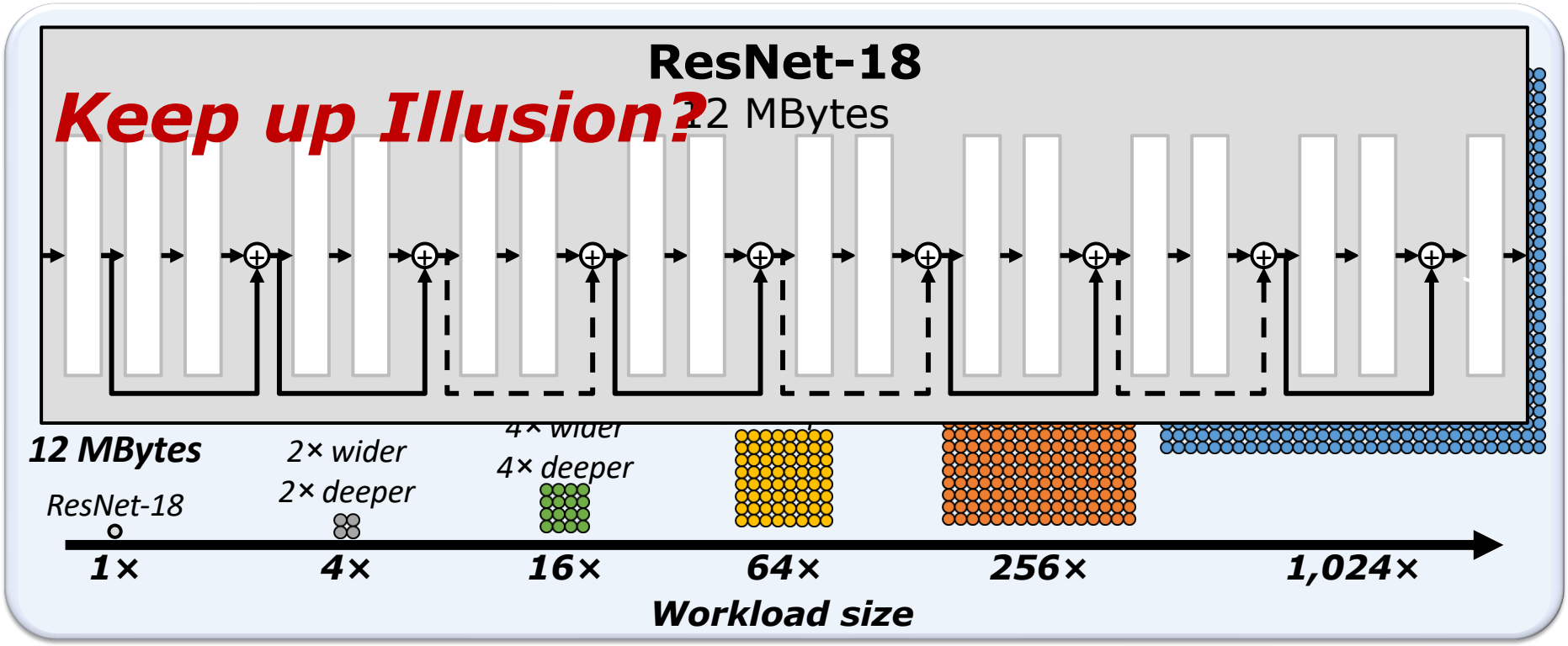
Illusion Ideal for AI

Wide Variety of AI tasks

CNNs, DLRMs, LSTMs, Transformers, single/multiple inferences, training, …

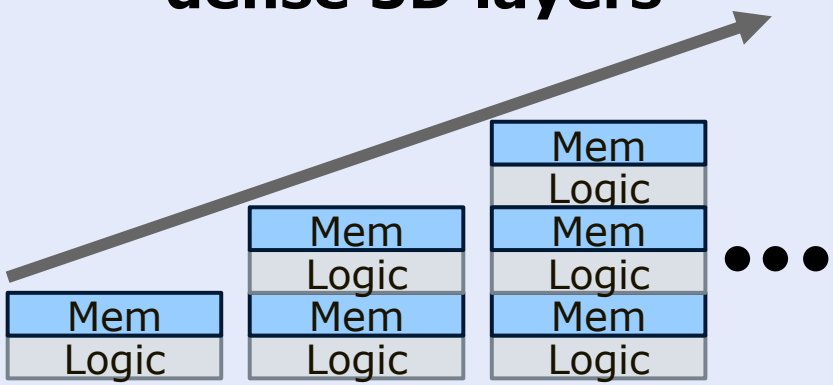


1,024× Workload Growth



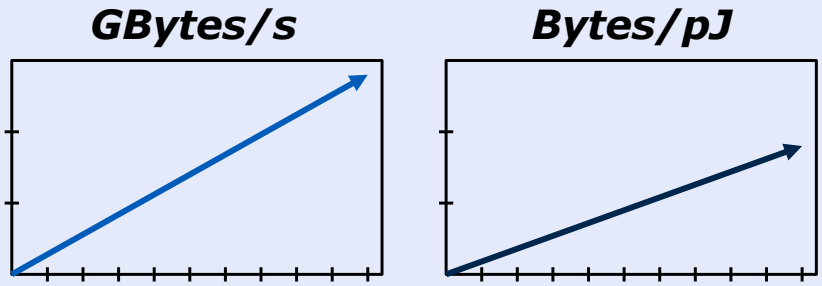
Illusion Scaleup

Linearly **increase dense 3D layers**



Reduce ***message counts***

Linearly **improve chip-to-chip links**

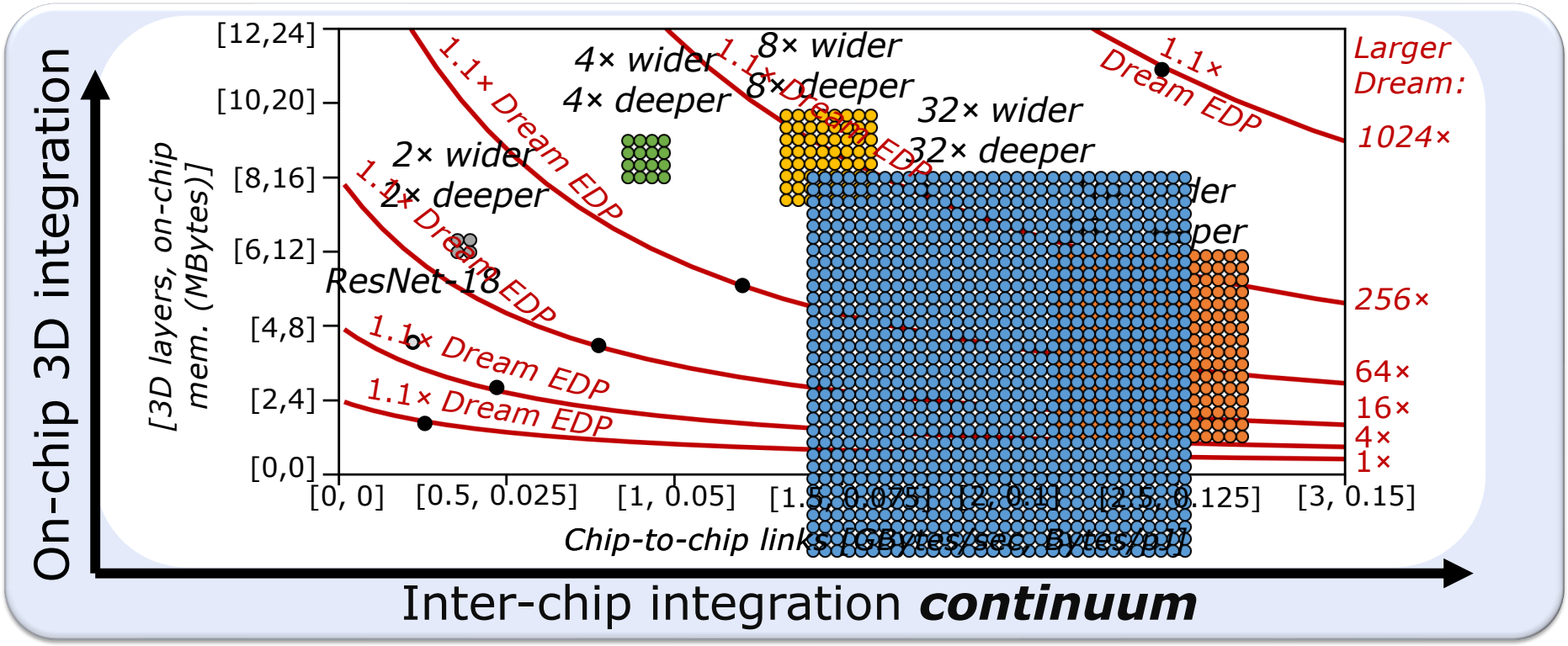


Reduce ***per-message cost***

Quadratically **reduce Illusion** ***total message cost***

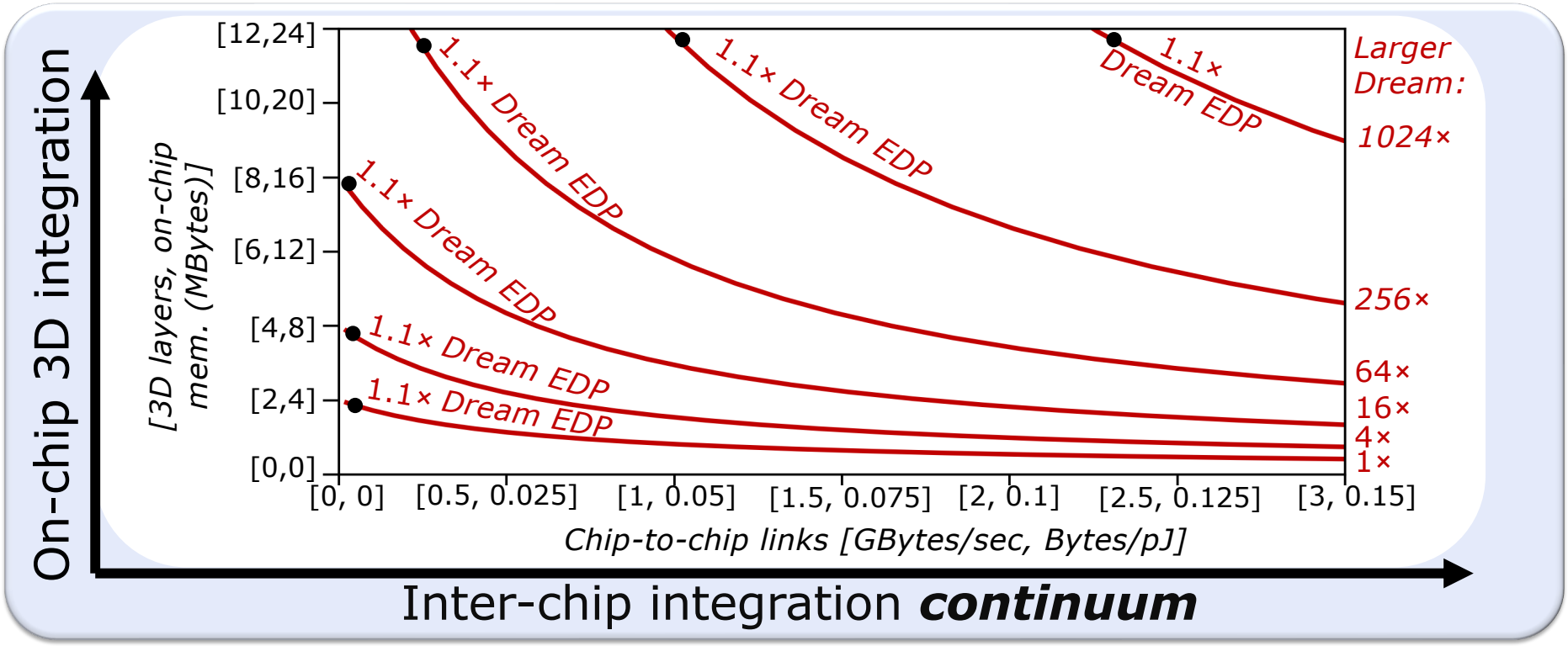
Illusion Scaleup

Maintain 1.1× Dream EDP despite growing Dream Chips



Illusion Scaleup

Illusion Scaleup is Fungible



Many NanoSystems Opportunities

- **Co-design: device + circuit + arch. + algorithm**
- Dense compute + thermal
- New software optimizations

Co-design Examples

multiple *abstraction layers* ***cooperate*** *for* ***large*** *benefits*

- **Illusion scaleup**: tech + circuit + arch + algorithm
- **RRAM edge AI training**: tech + arch + algorithm
- **Efficient hyperdimensional computing**: tech + circuit + arch + algorithm
- **RRAM all memory on-chip dataflow:** tech + arch
- **1T*n*R + multiple bits-per-cell RRAM**: tech + circuit
- **Imperfection-immune carbon nanotube VLSI:** tech + circuit

Edge AI Training

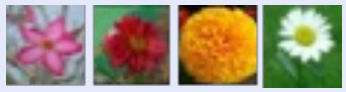
Pre-deployment

Generic

Full ImageNet

+

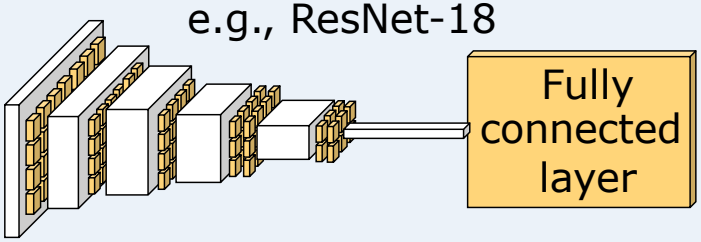
Domain-specific



1 image/class

Flowers102 subset

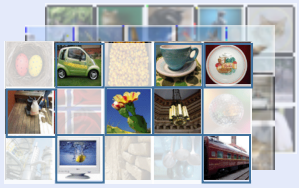
e.g., ResNet-18



All layer weights

Edge *incremental* training

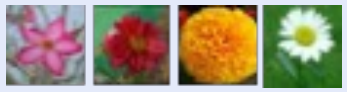
Generic



ImageNet subset

+

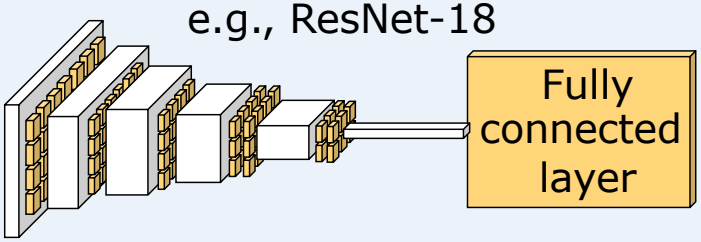
Domain-specific



60 images/class

Full Flowers102

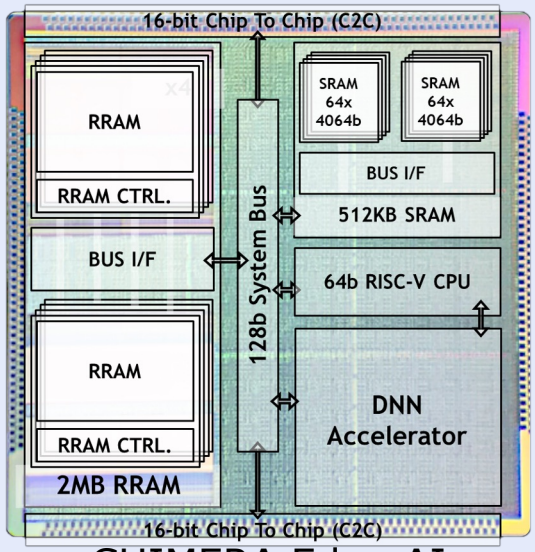
e.g., ResNet-18



Fully connected layer only

RRAM Edge AI Incremental Training Challenging

New Low-Rank Training (LRT)



CHIMERA Edge AI accelerator on-chip foundry RRAM

	LRT hardware results: iso-accurate vs. Stochastic Gradient Descent (SGD)
RRAM weight update steps	**101× fewer vs. SGD**
Energy Delay Product	**340× better vs. SGD**
Endurance (20 samples/min.)	**10 years (LRT + ENDURER*) vs. 2 weeks (SGD)**
On-chip SRAM capacity	**37× smaller vs. SGD**

[Giordano Symp. VLSI Circuits 21] *ENDURER: [Aly Proc. IEEE 19, Wu ISSCC 19]

Conclusion

- **NanoSystems today**
  - Industrial fabs: Carbon nanotube FETs + RRAM + monolithic 3D
- ***N3XT 3D MOSAIC* + Illusion scaleup key**
  - Computation immersed in memory
  - Large benefits over growing problem sizes, ideal for AI
- **Co-design of the "right" kind**
  - Big opportunities for NanoSystems