The Future of

Hardware Technologies for Computing

Subhasish Mitra

Department of EE and Department of CS

Stanford University

Thanks: Students, Sponsors, Collaborators

ASML cādence

Meta

SAMSUNG

skywater
SLAC

Stanford Precourt Institute for Energy
Stanford SystemX Alliance
SYNOPSYS®

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"

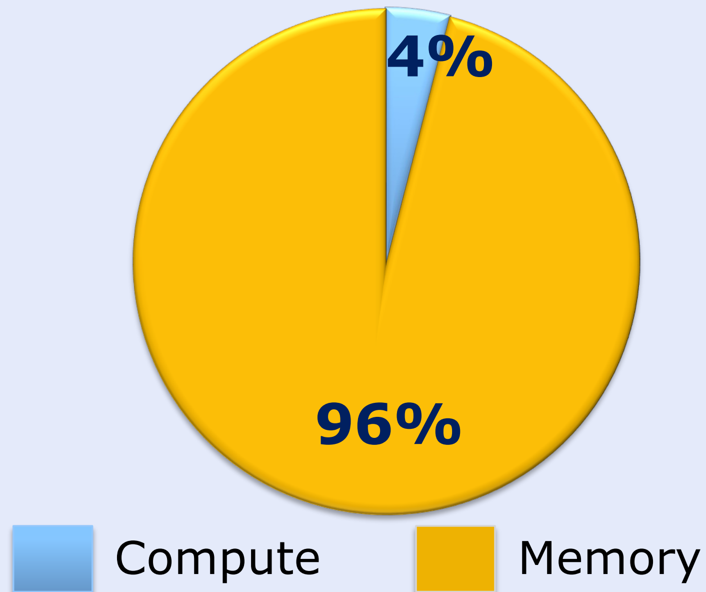
"best [devices] … not made with semiconductors"

"elegant devices messed up with all the other stuff"

Abundant-Data Computing

Many walls simultaneously

Memory Wall



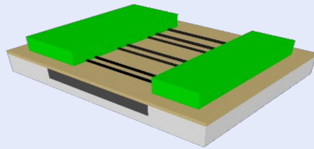
Miniaturization Wall

✚ power wall, cooling wall, resilience wall …

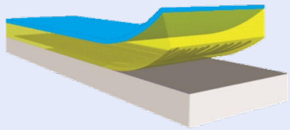
Nano*Systems*

New nanotech

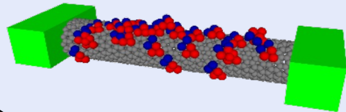
Devices



Fabrication

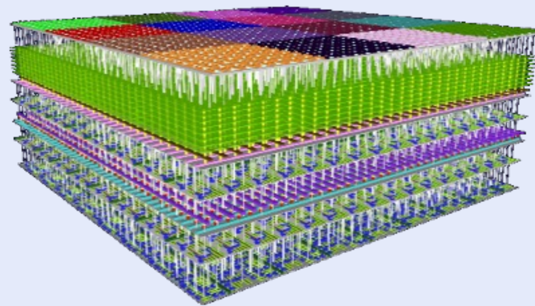


Sensors

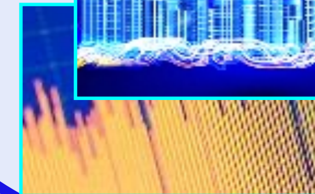


New systems

New architectures



New applications



Data Explosion & Memory Wall

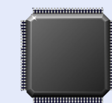
Both Von Neumann & non-Von Neumann architectures

"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"

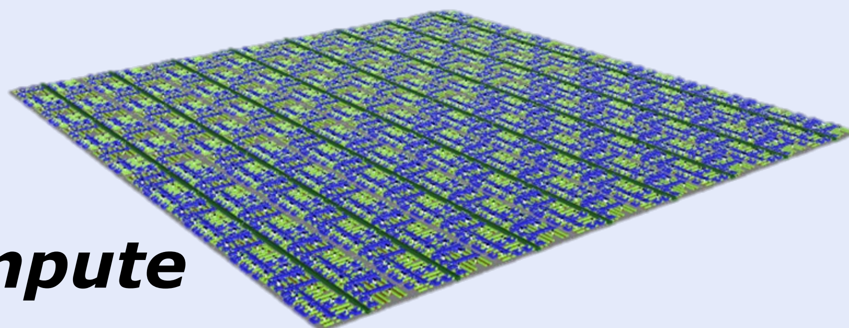
[Burks, Goldstine, Von Neumann, 1946]

Computation immersed in memory

Computing Today



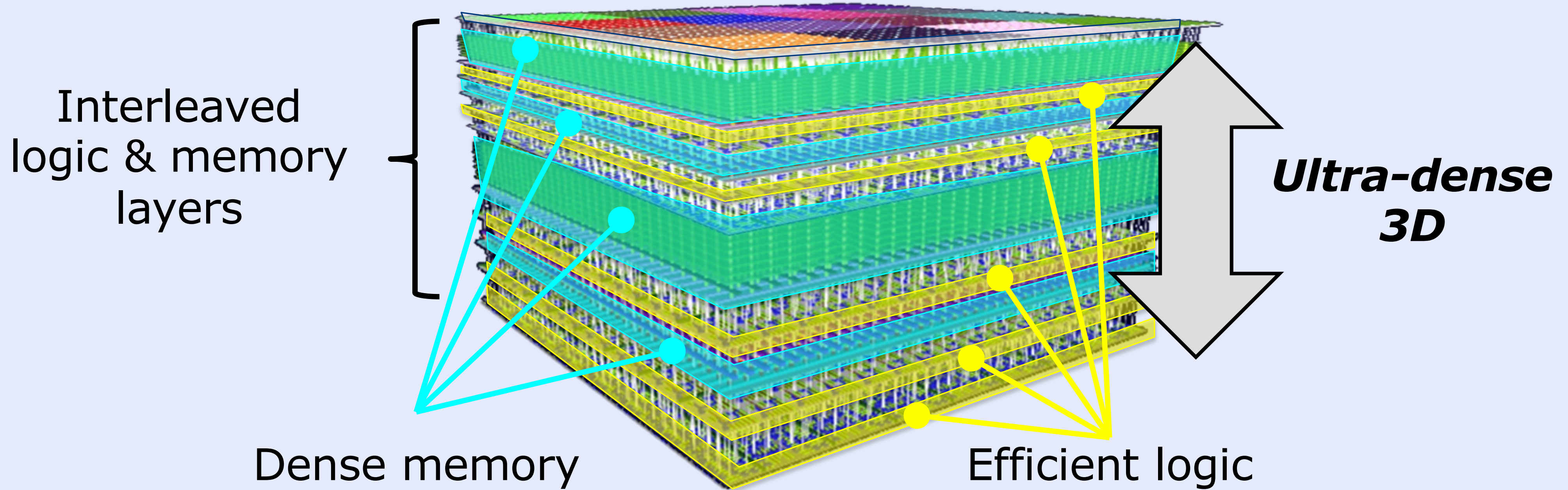
Memory



Compute

N3XT 3D: Computation immersed in Memory

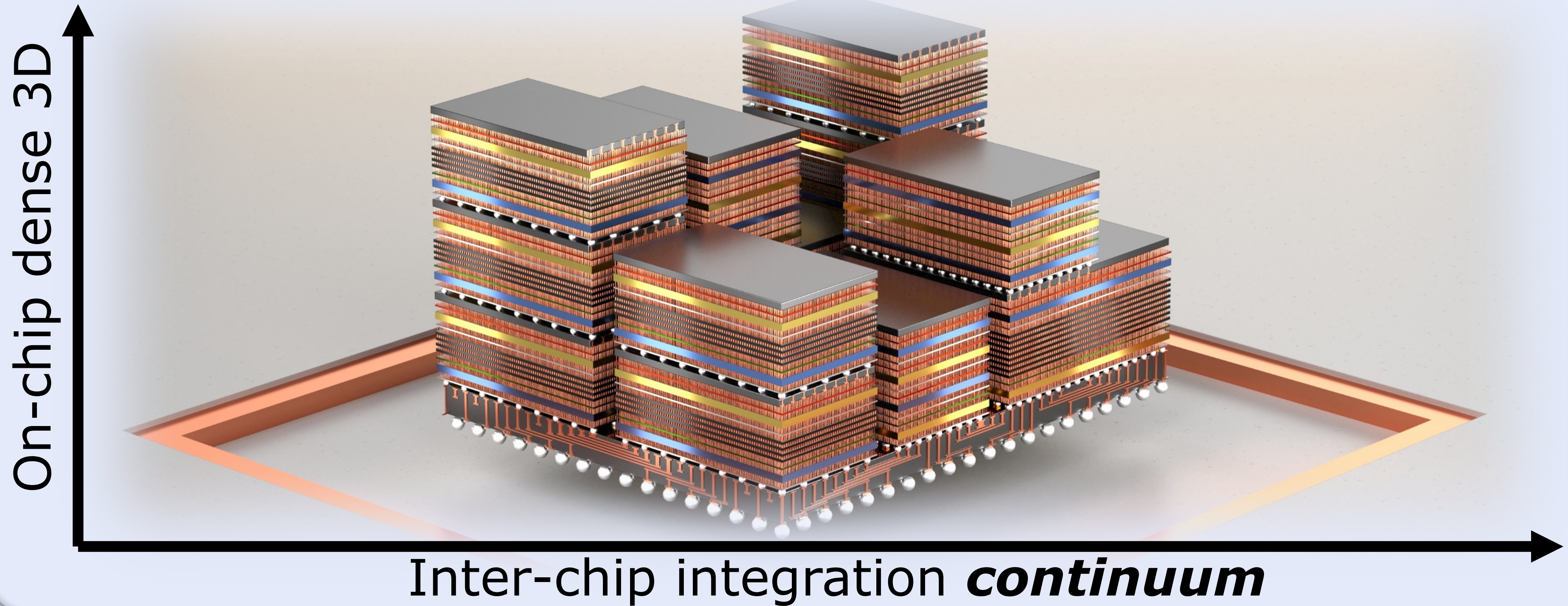
Nano-Engineered Computing Systems Technology



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

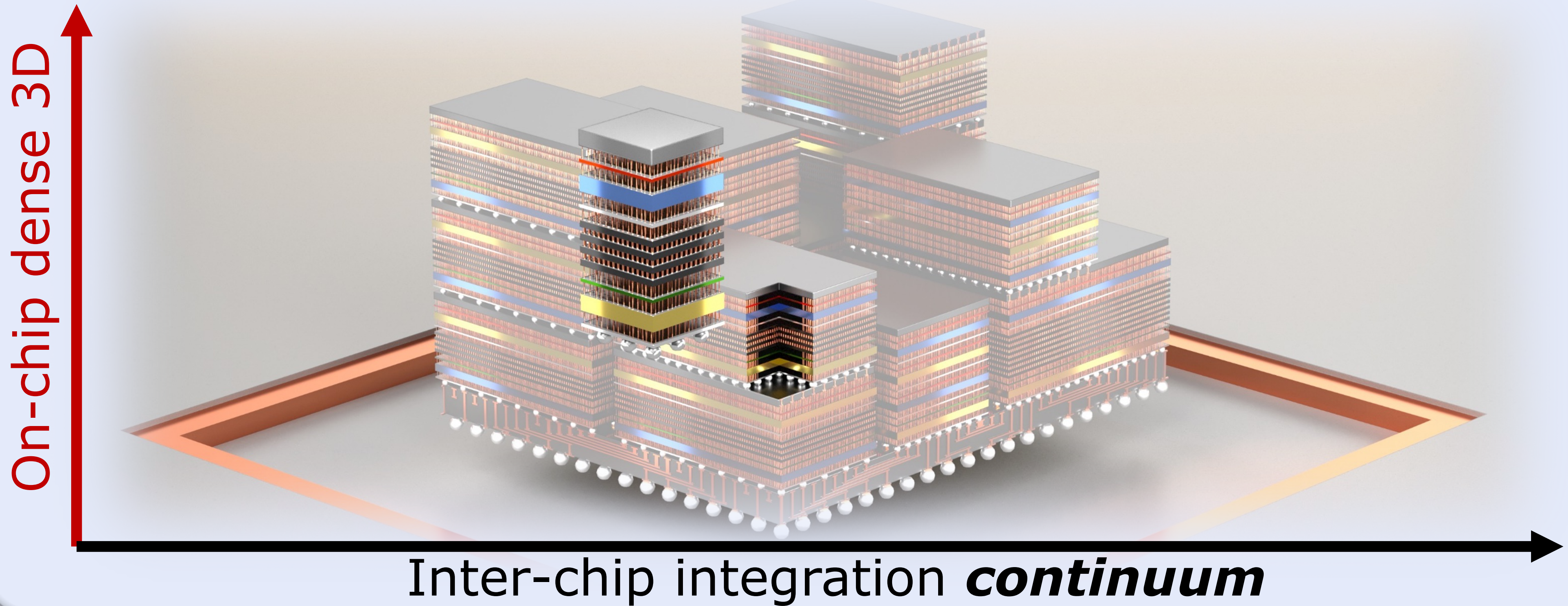
N3XT 3D MOSAIC

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



N3XT 3D MOSAIC

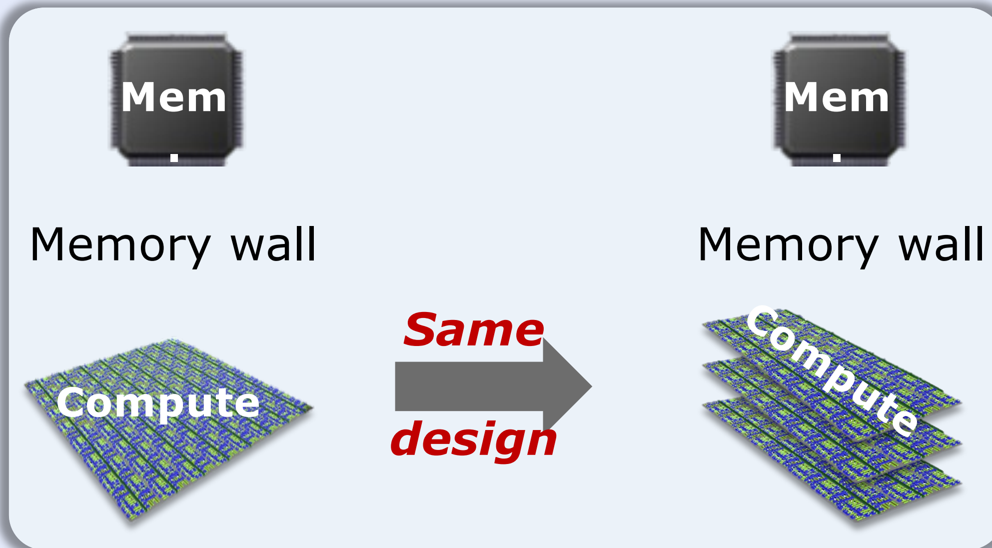
MOnolithic / Stacked / Assembled IC



N3XT 3D ⊃ 3D Folding

3D Folding

Limited EDP benefits: 1.4×

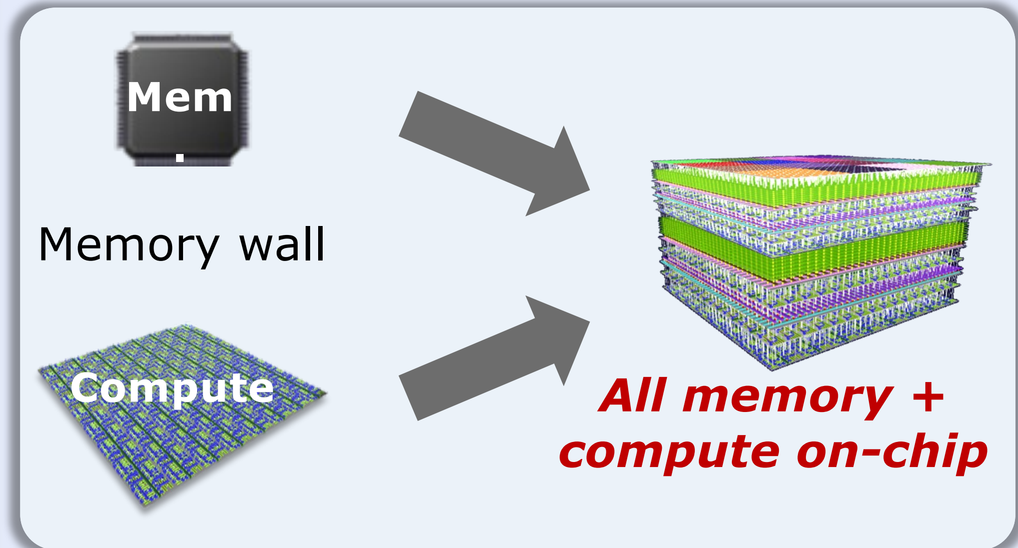


Some wirelength benefits

Memory wall stays

N3XT 3D

Large EDP benefits: 100-1,000×

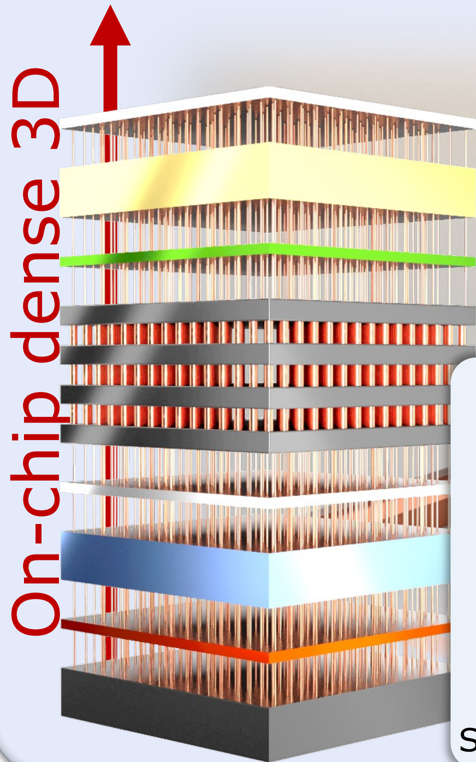


Many concurrent on-chip accesses

New arch. via new 3D physical design

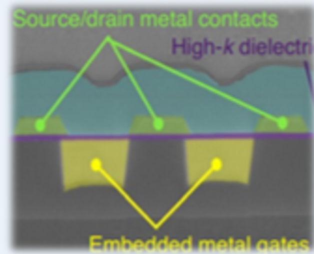
N3XT 3D: Many Technologies

N3XT 3D Chip:



Back-End-Of-Line-Compatible Technologies

Carbon-Nanotube FETs

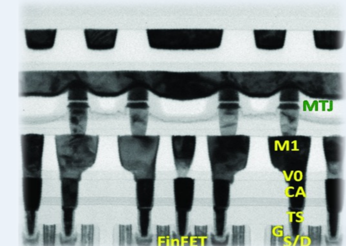


MIT + Skywater 20

Resistive RAM

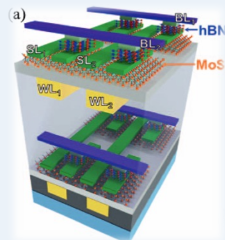
Stanford+SkyWater 21

Magnetoresistive RAM



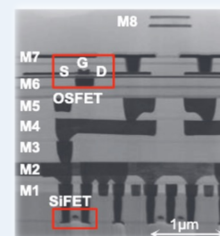
IBM 20

2D FETs



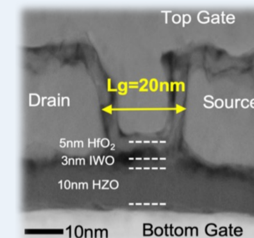
Stanford + Soochow 18

Oxide FETs (2T Gain Cells)



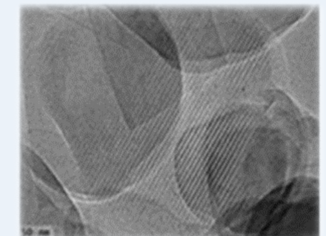
UMC 17

Ferroelectric FETs



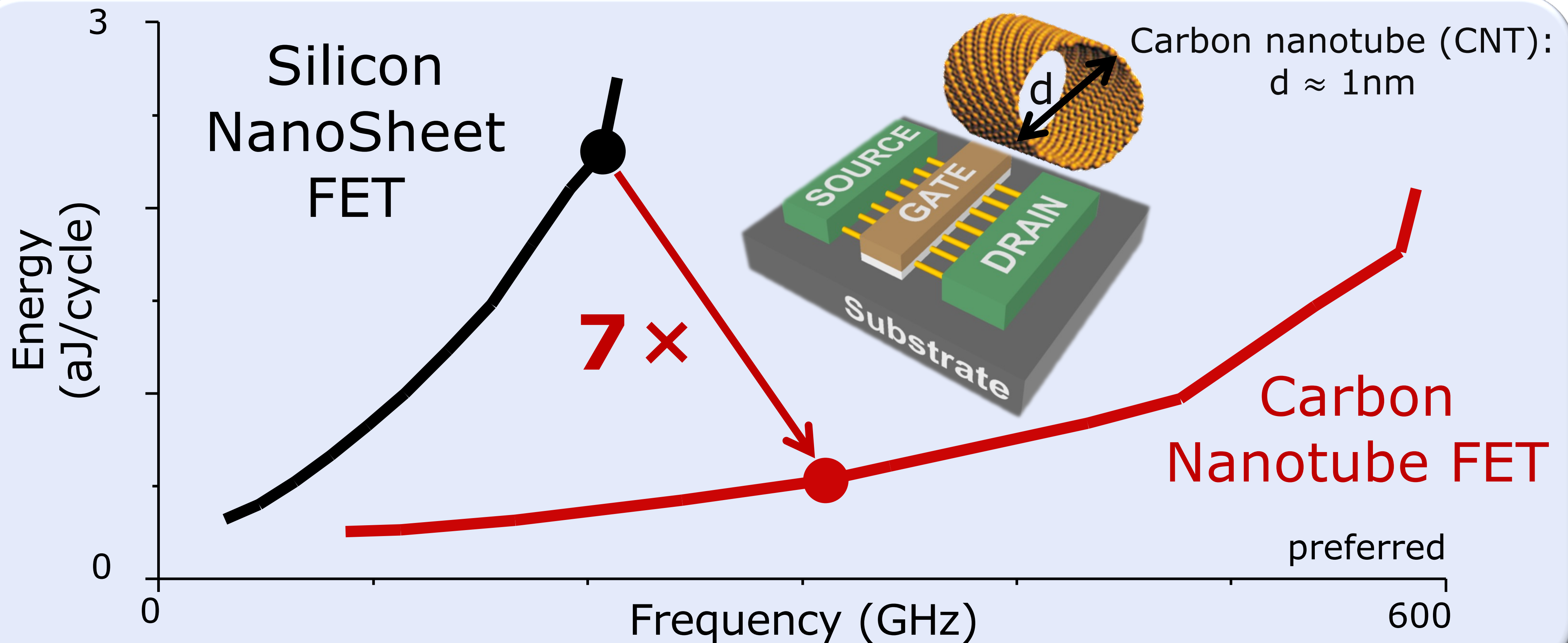
Notre Dame + GA Tech 20

Thermal Management



Shanghai + Chalmers 16

Carbon Nanotube FET (CNFET): Large EDP Benefits



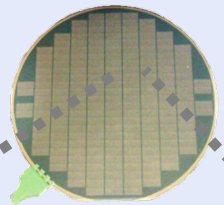
Inverter in a 15-stage inverter-based ring oscillator [Gilardi IEDM 21]

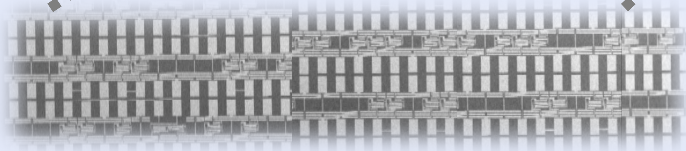
Imperfection-Immune Paradigm → Major Progress

First CNT computer (Stanford)

Stanford student

[Nature 2013]





178 CNFETs: PMOS logic

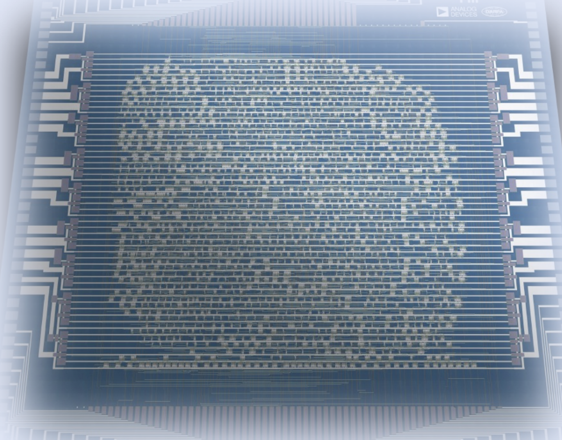
1 instruction (Turing complete)

1-bit data

CNT RISC-V (MIT, Analog Devices)

MIT Professor

[Nature 2019]

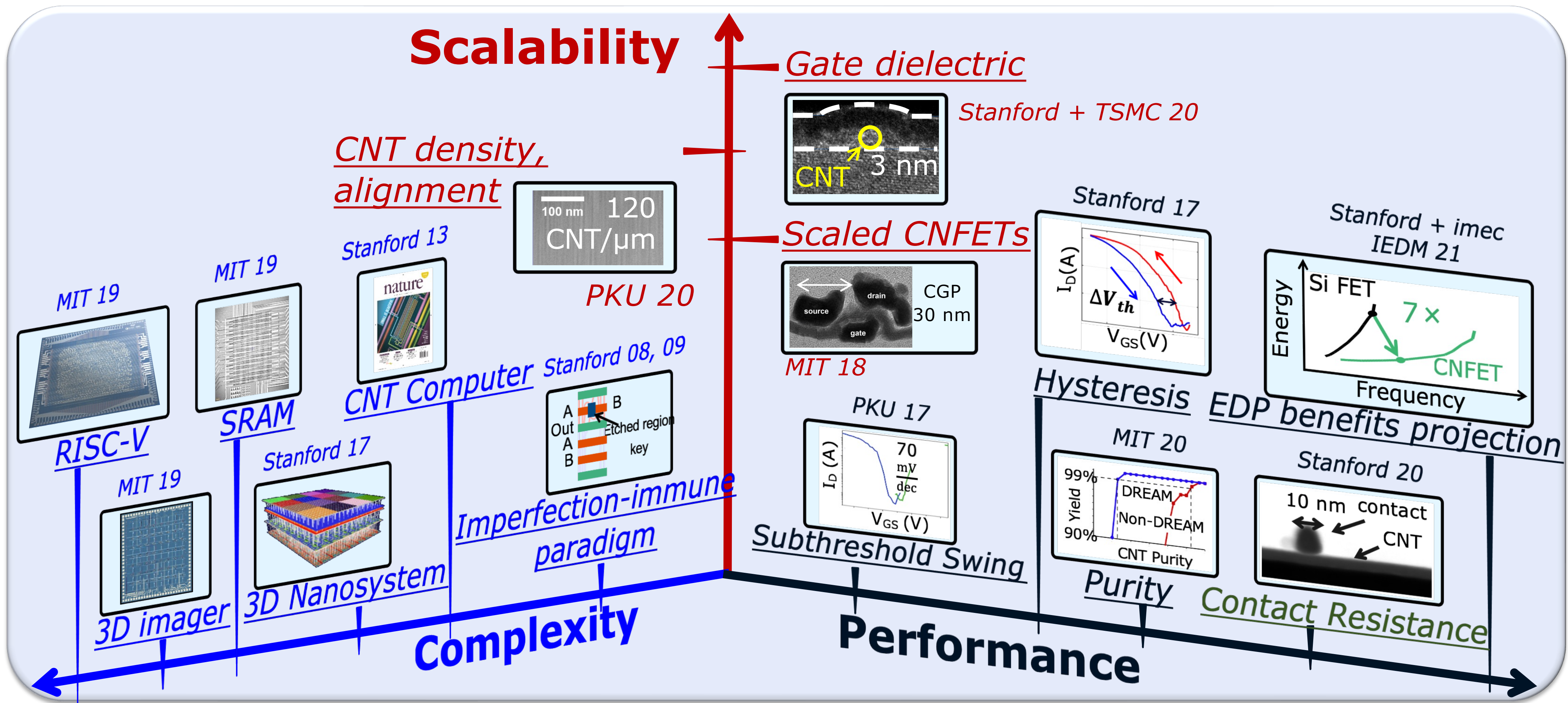


14,702 CNFETs: CMOS logic

All RV32E instructions

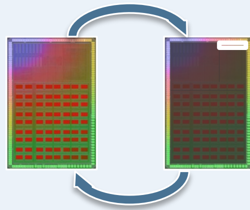
16-bit data

Carbon Nanotube FETs (CNFETs): Many Innovations



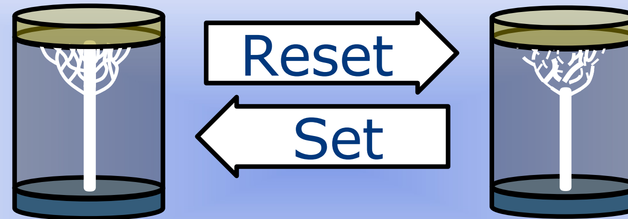
Resistive RAM (RRAM)

Non-volatile computing system



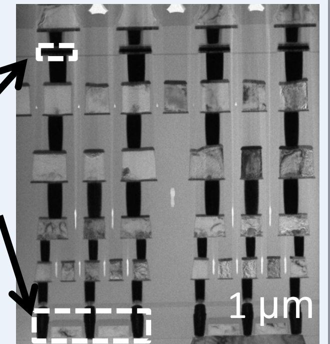
Stanford +
CEA LETI + NTU Singapore 19

Low R High R



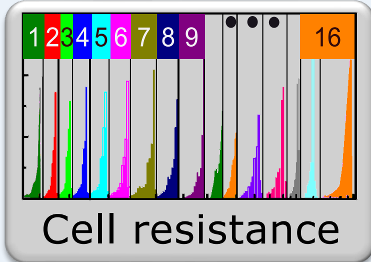
Large on-chip memory

3 Million RRAM cells



Stanford + UCSD
+ Tsinghua + Notre Dame 20

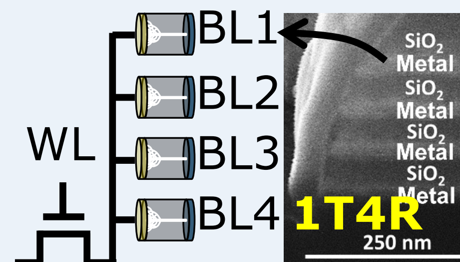
Multi bits/cell arrays



Stanford + CEA LETI 19,
Stanford + SkyWater 21

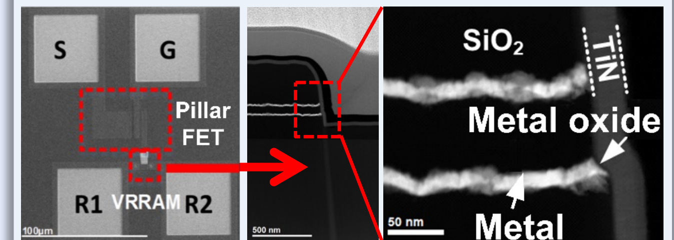
1TnR

Stanford +
SkyWater 21



Stanford 21

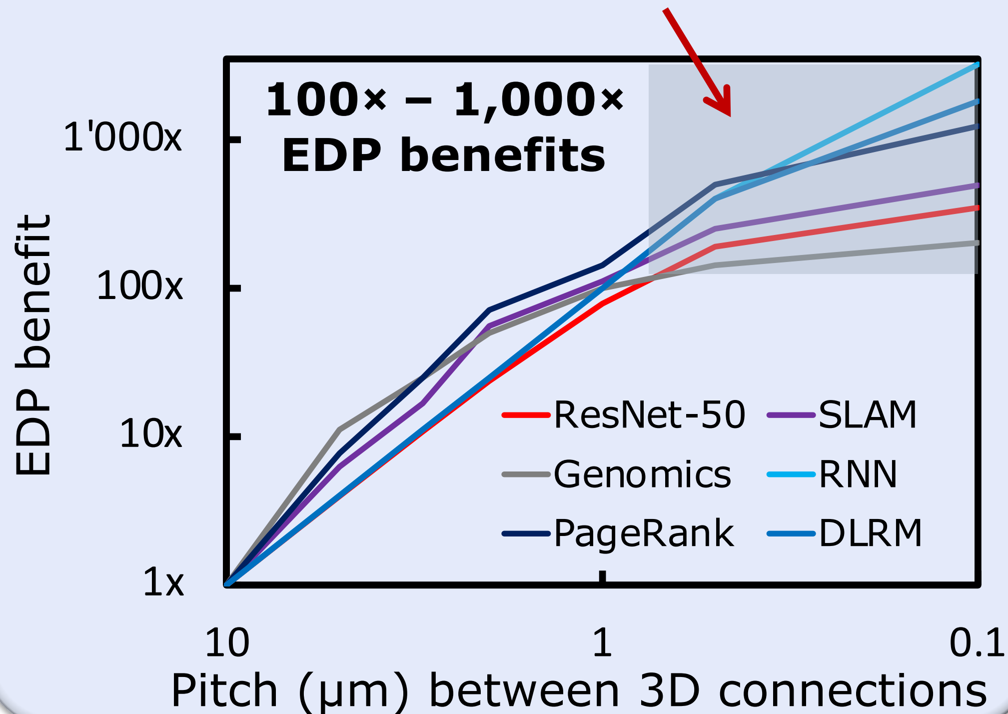
3D Vertical RRAM



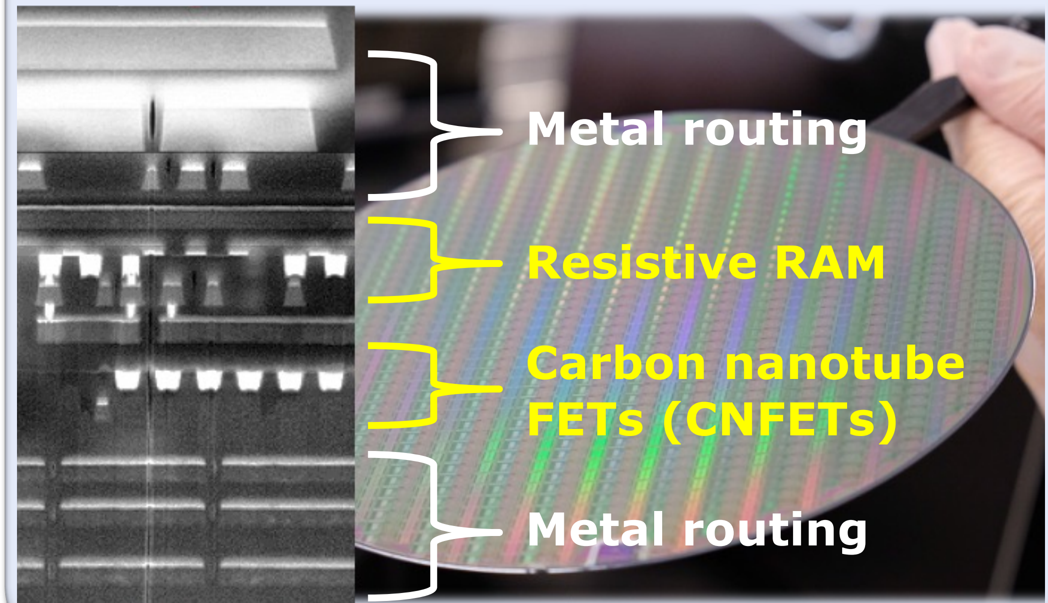
Stanford 21

Ultra-dense Monolithic 3D

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



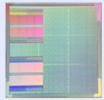
Low-temperature fabrication



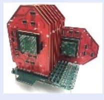
Simulations: 88% accurate vs. hardware

Lab to Fab

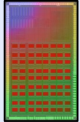
Lab



Edge AI: training + inference
340× better Energy Delay Product

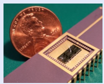


Illusion
Dream Chip: all memory + compute on-chip

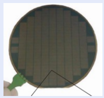


Non-volatile Internet-of-Things
10× battery life vs. embedded FLASH

Hyperdimensional computing
Brain-inspired, one-shot learning



First 3D NanoSystem
Dense monolithic 3D: carbon nanotube + Resistive RAM + silicon

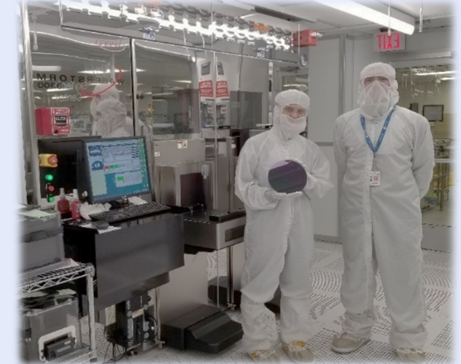


First carbon nanotube computer

Fab

Analog Devices

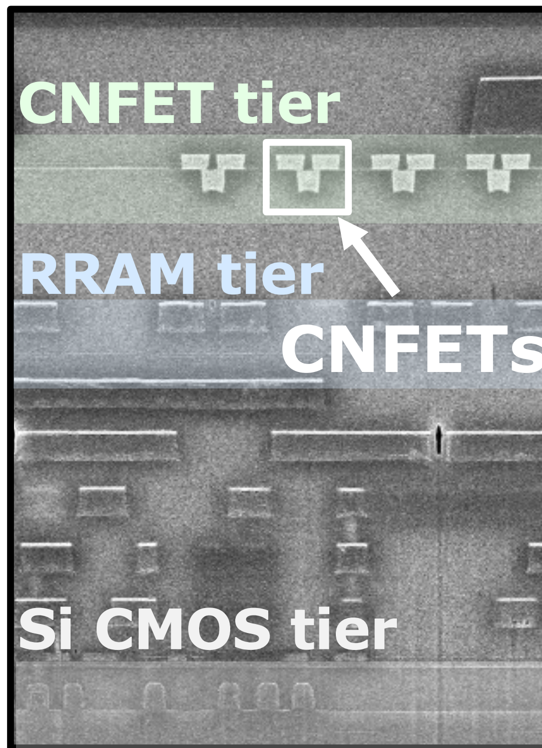
SkyWater

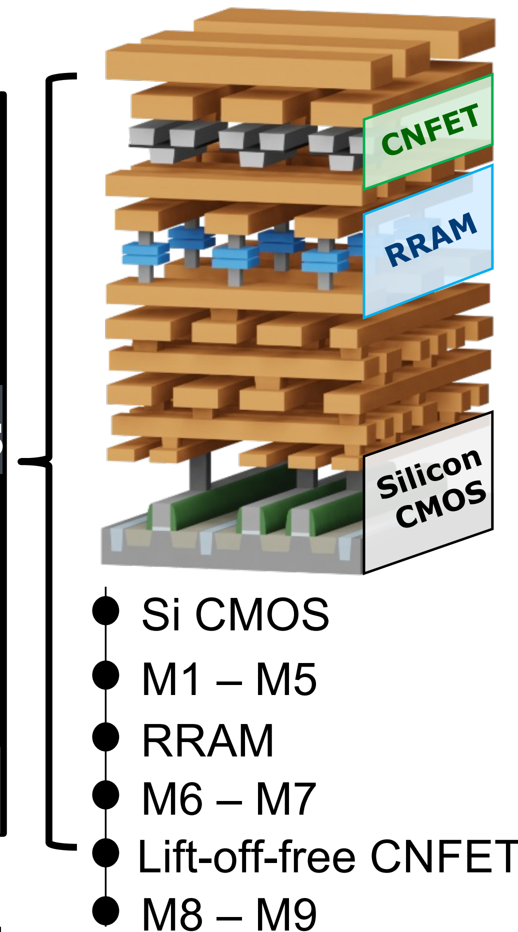


Many firsts in industry fabs

- Carbon nanotube FETs (CNFETs)
- Dense monolithic 3D: CNFET+ RRAM + silicon CMOS
- U.S. foundry Resistive RAM

Foundry Monolithic 3D: CNFET + RRAM + Silicon CMOS





SkyWater [Srimani Symp. VLSI 23]

Monolithic 3D CNFET + RRAM

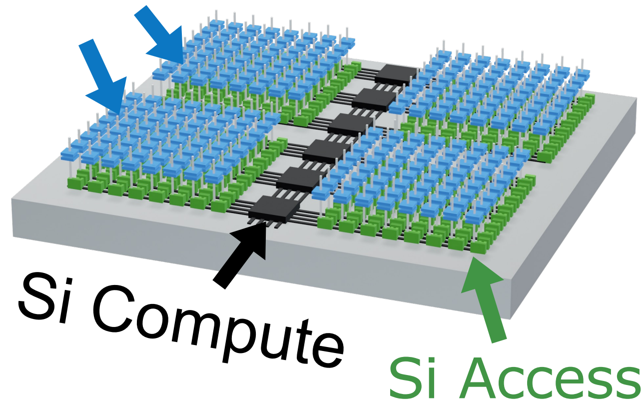
(vs. Silicon + RRAM)

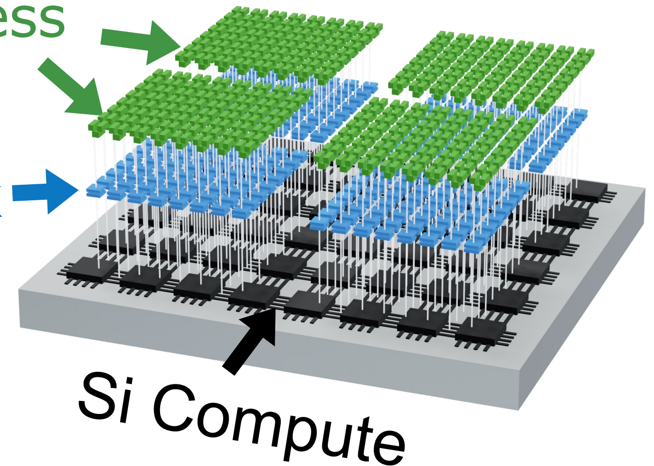
- ☐ Iso-footprint
- ☐ Iso-energy/latency
- ☐ Iso reliability/endurance
- ☐ Iso-retention
- ☐ Multiple bits per cell

Monolithic 3D Physical Design → *New Arch. Design Points*

Si + RRAM

Foundry Monolithic 3D (Si + RRAM + CNFETs)



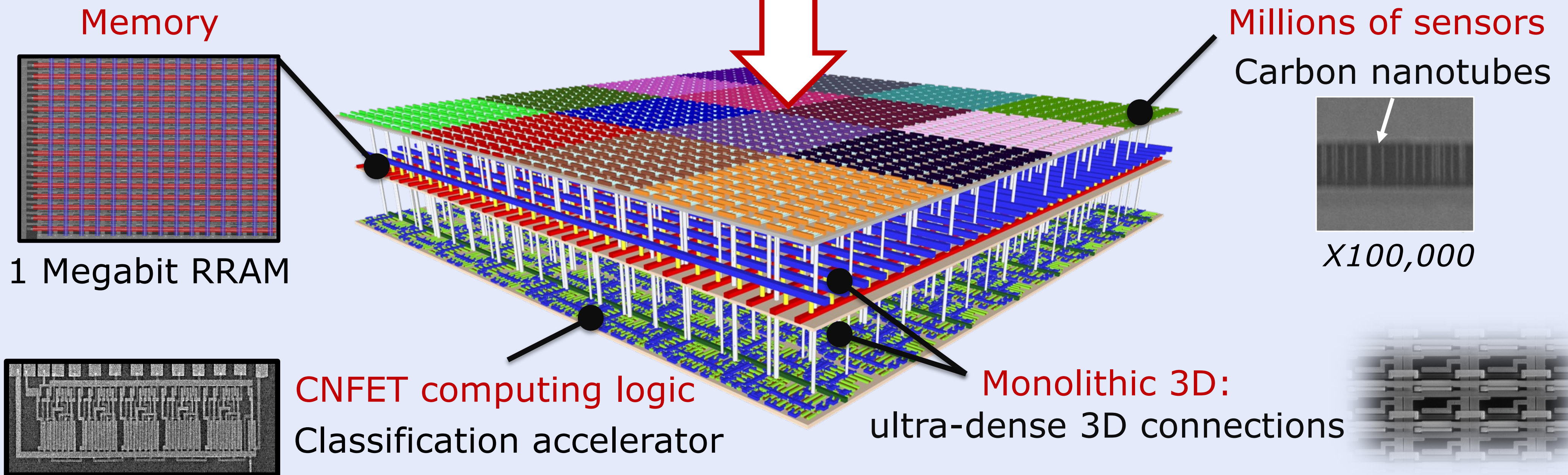


Many more compute: highly parallel

Large EDP benefits: 5× – 10×

[Srimani DATE 23]

3D NanoSystem



[Shulaker Nature 17]

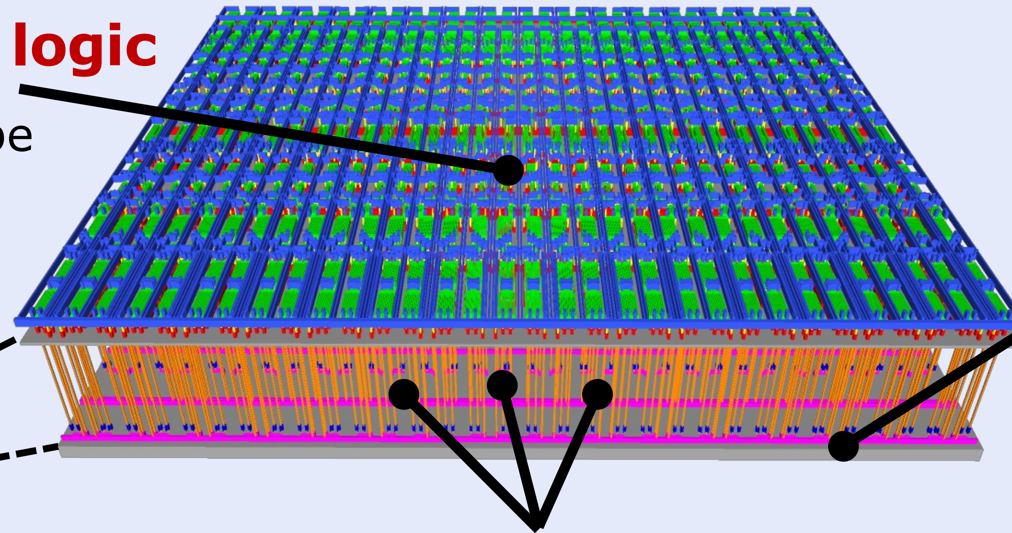
HD Computing: Brain-Inspired ⊃ Neural Nets

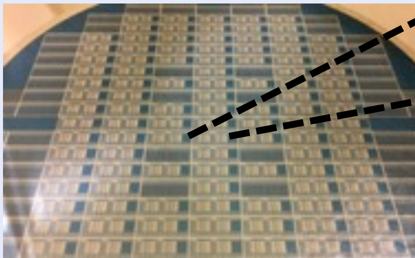
Carbon nanotube logic

(1,952 Carbon Nanotube FETs)

RRAM TCAM

(224 RRAM cells)





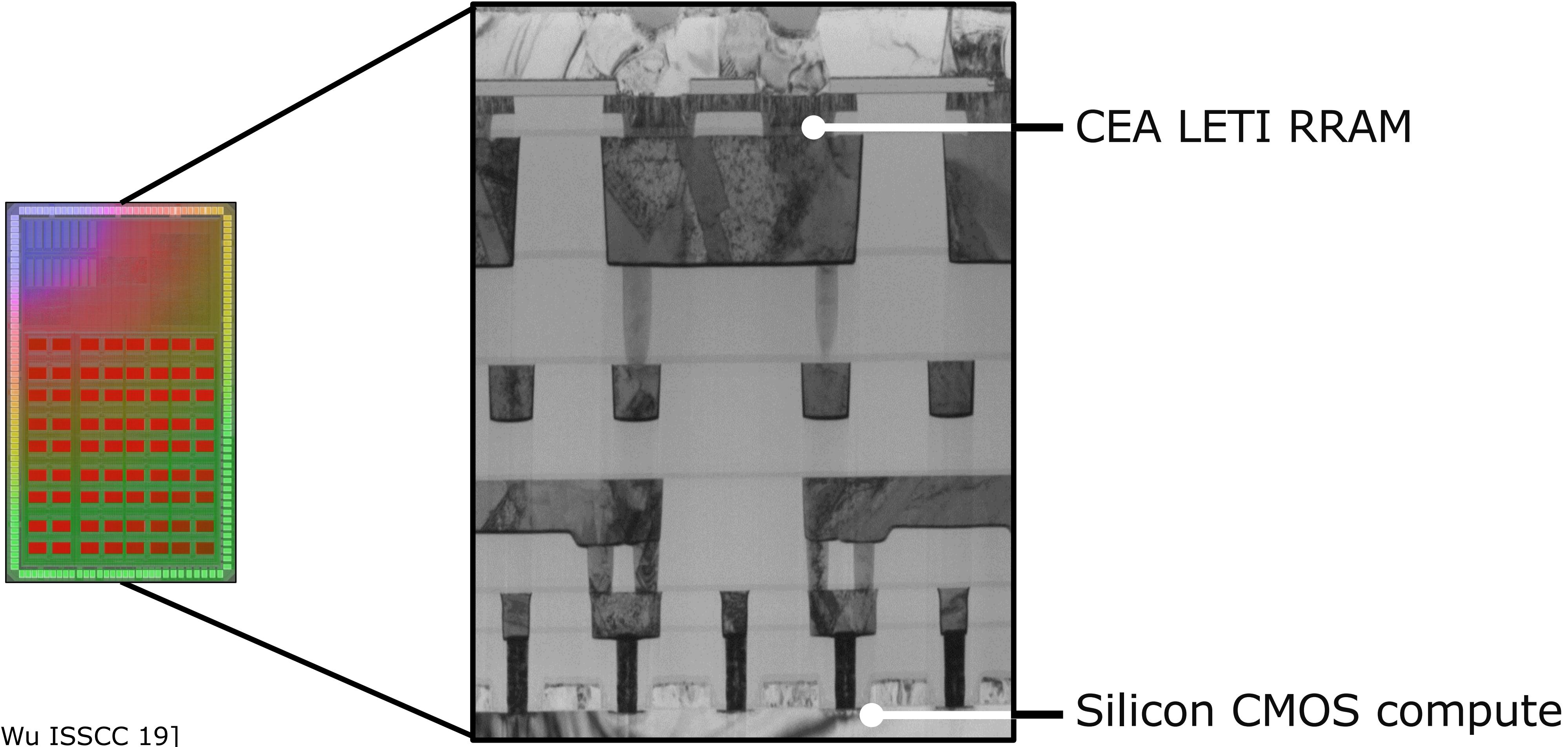
Monolithic 3D: dense 3D connections

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

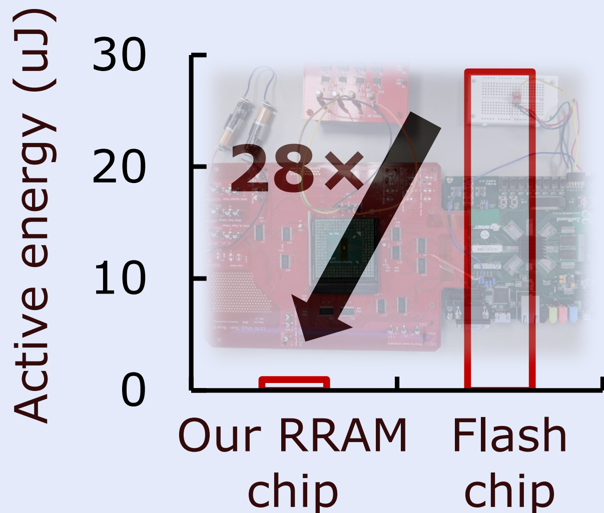
Non-volatile Computing System: RRAM + Silicon CMOS



[Wu ISSCC 19]

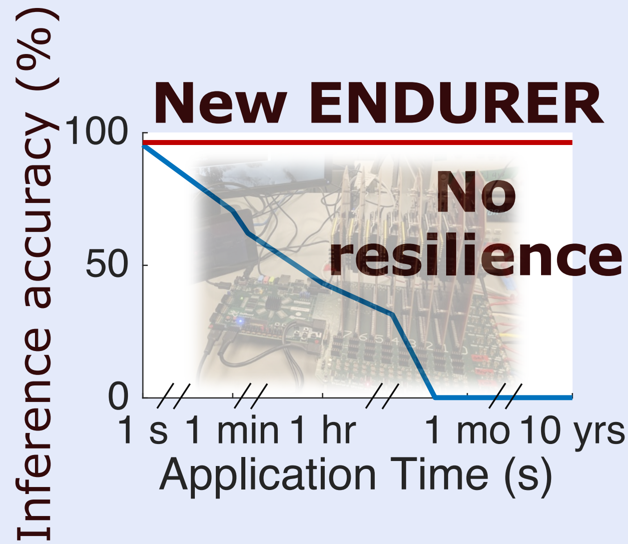
Non-volatile Computing System: RRAM + Silicon CMOS

Fine-grained temporal power gating



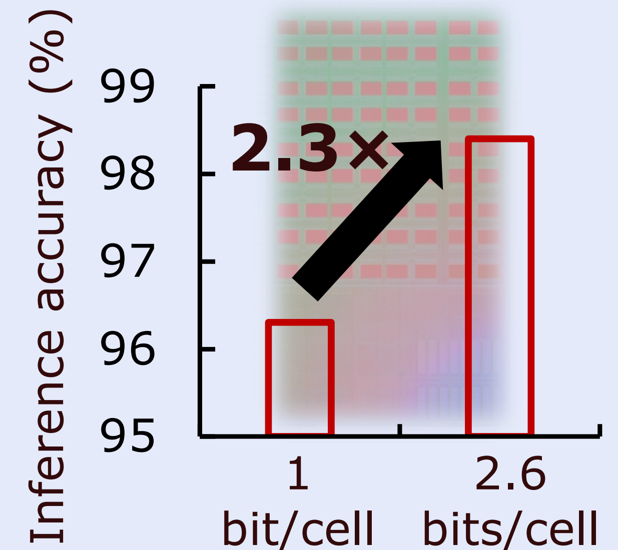
10× battery life vs. Flash chip

New RRAM Endurance



10-year continuous AI inference

1st RRAM System: Multiple bits/cell



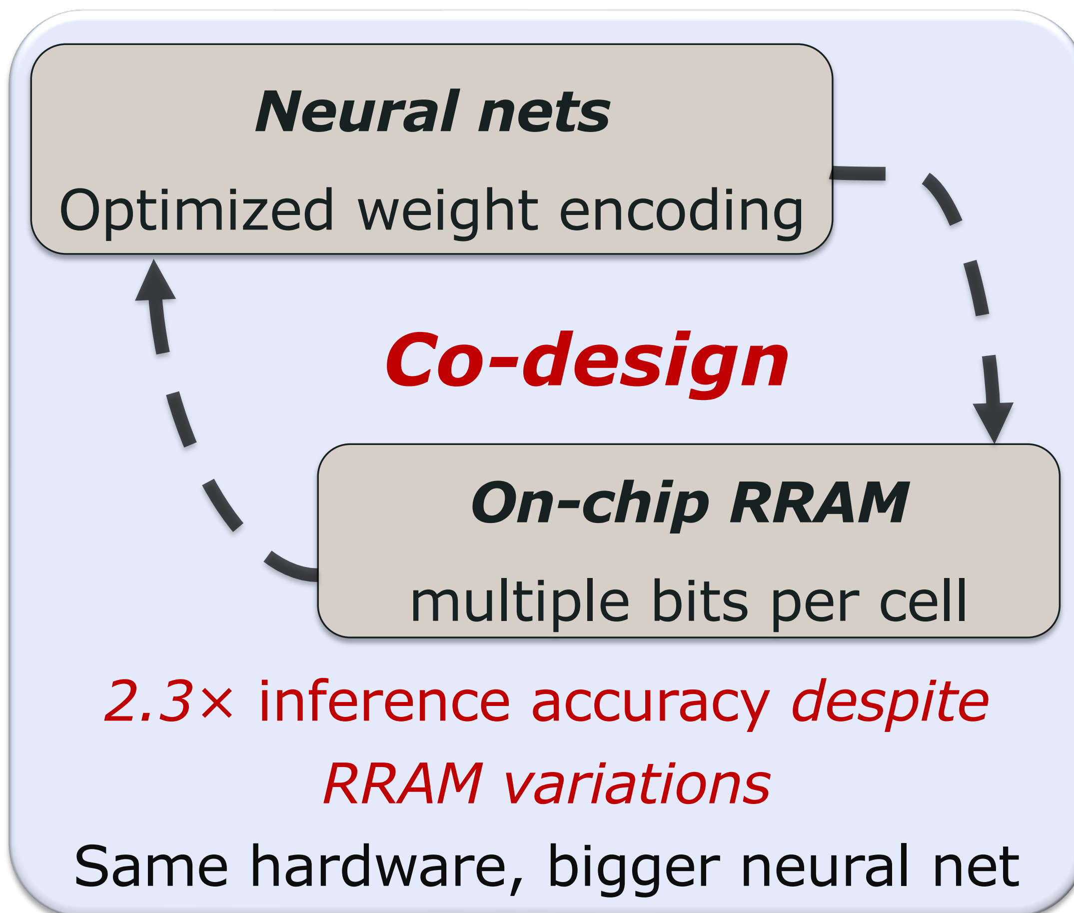
New RRAM algorithms key

[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]



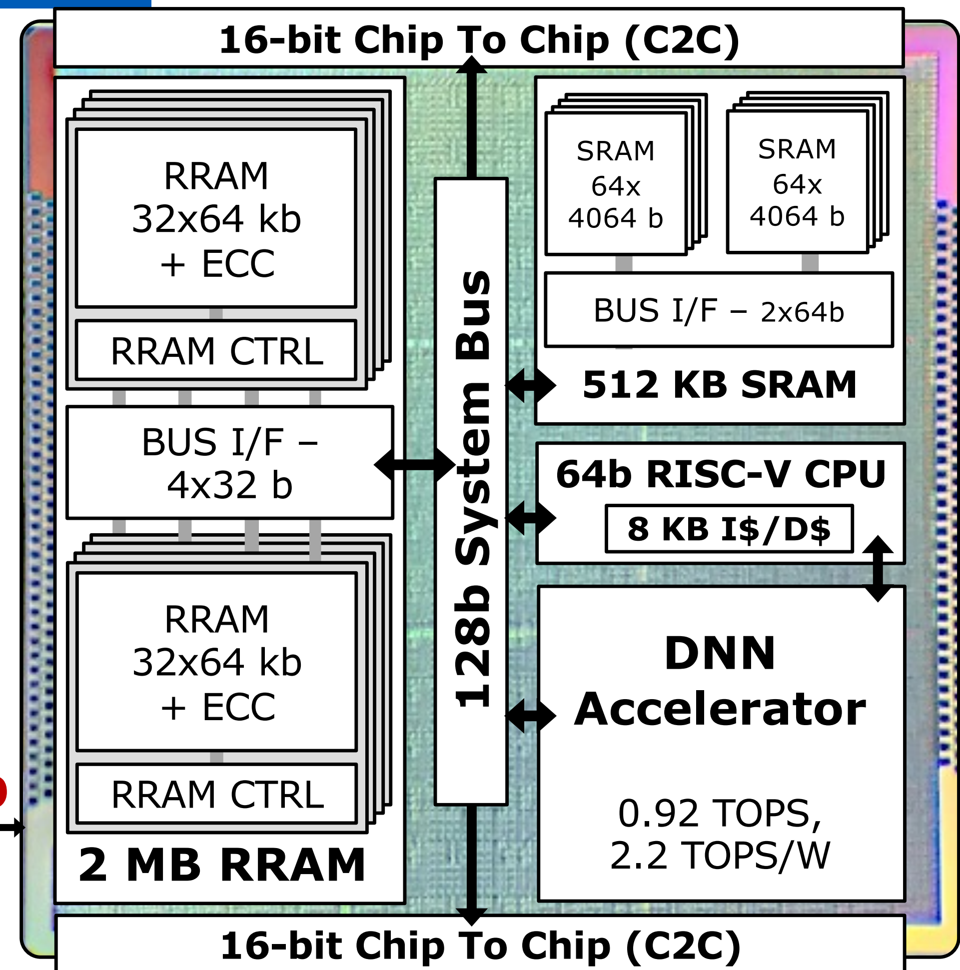
[Wu ISSCC 19]

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]

Infineon + TSMC: RRAM Announcement

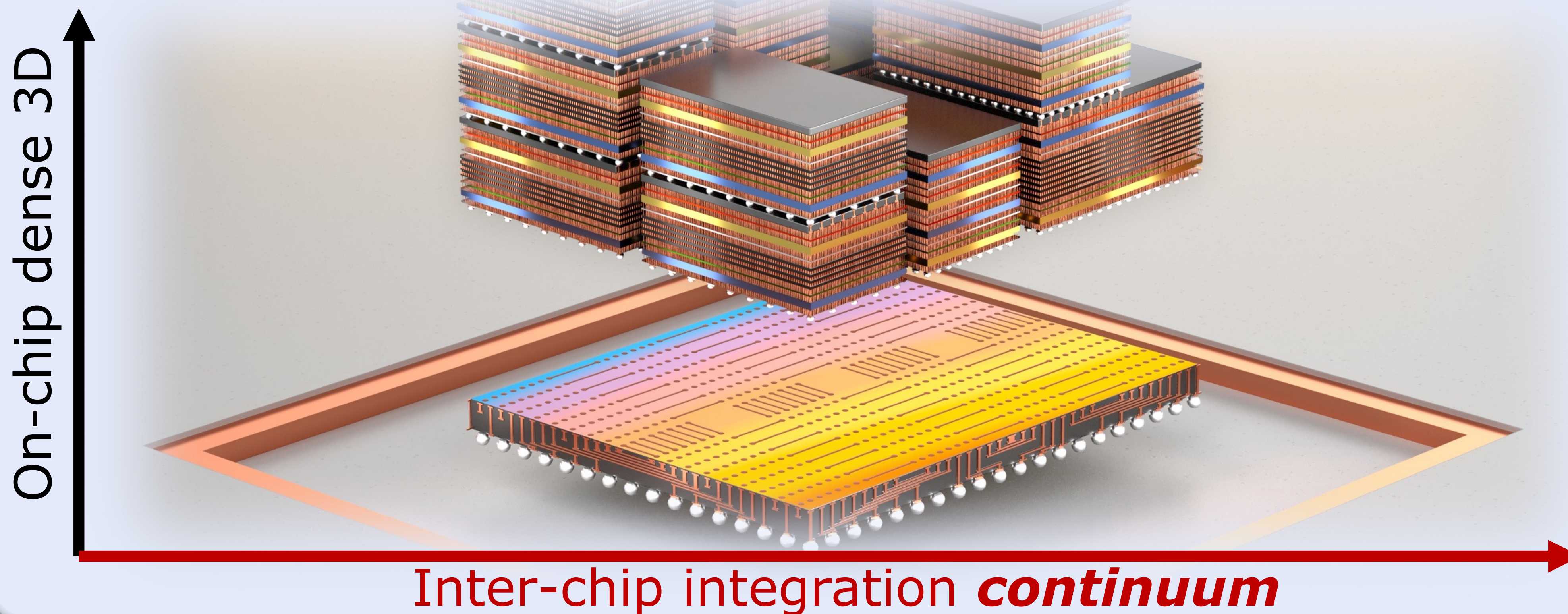
Infineon and TSMC to introduce RRAM technology for automotive AURIX™ TC4x product family

Nov 25, 2022 | Market News

"RRAM technology creates a significant potential for performance expansion, power consumption reduction, and cost improvement."

N3XT 3D MOSAIC

MOnolithic / Stacked / Assembled IC



Dream: *All Memory* + Compute On-chip

Dream Chip: infeasible, ***moving target***

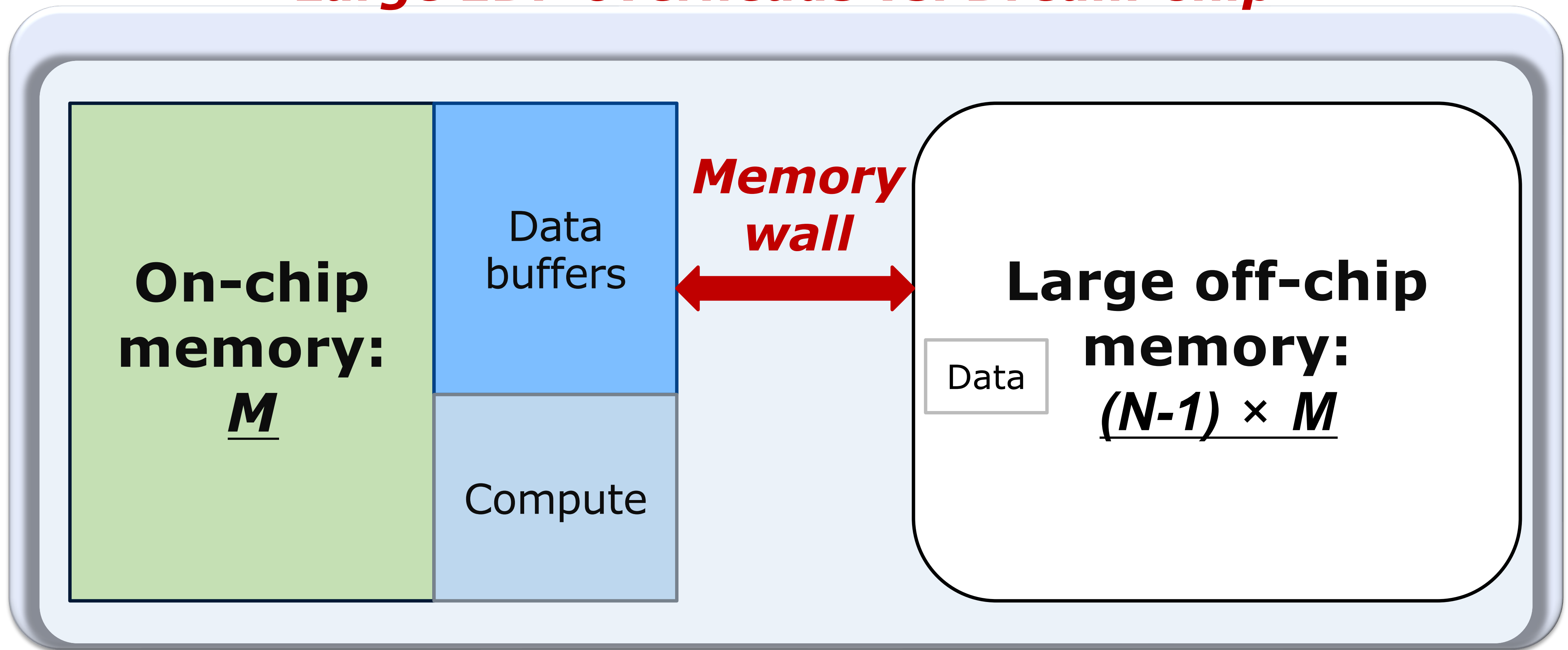
Massive on-chip memory: *N × M*

(Full workload fits in on-chip memory)

Data buffers

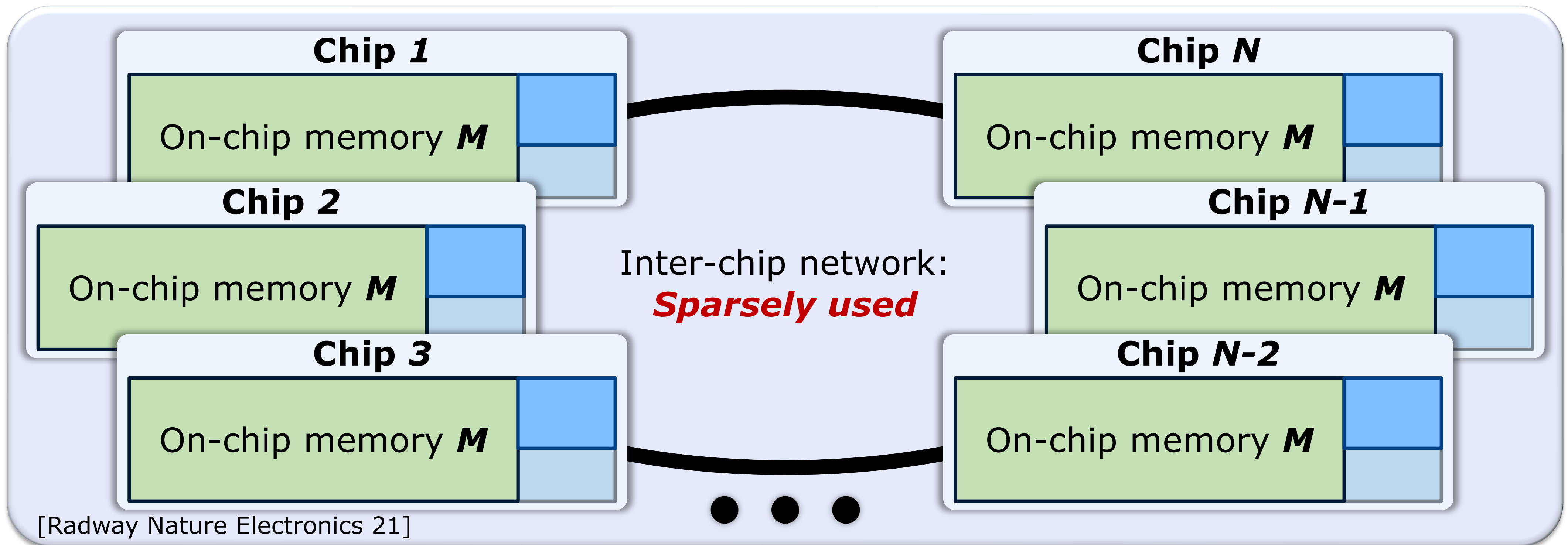
Compute

Off-Chip Memory Accesses Costly



Illusion System

Enough on-chip mem. + ***Quick*** chip ON/OFF = ***Special*** mapping



Illusion mapping ⊃ traditional parallelization

Illusion Ideal for AI

Illusion ≈ Dream

1.1 × Dream EDP

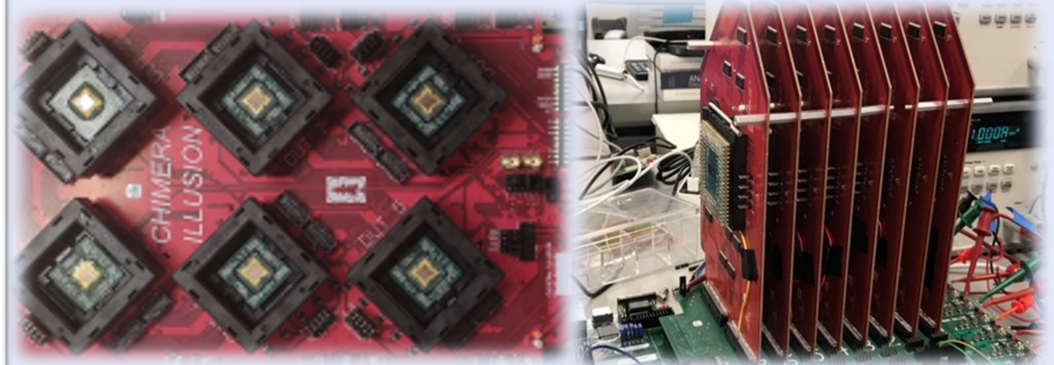
Illusion Energy
≤ 1.05 ×
Dream Energy

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

Illusion Mapping 1: *373 KByte* messages

Chip	Mapping 1 (in time order)
Chip 1	Layers 1-11
Chip 2	12-14
Chip 3	14-15
Chip 4	15-16
Chip 5	16-17
Chip 6	17-18

→Time

Illusion Mapping 2: *1.5 MByte* messages

Chip	Mapping 2 (in time order)
Chip 1	Layers 1-3 (Chips 1-3); 17-18
Chip 2	Layers 1-3 (Chips 1-3); 16-17
Chip 3	Layers 1-3 (Chips 1-3); 15-16
Chip 4	4-5; 15
Chip 5	6-7; 13-14
Chip 6	8-13

→Time

Compute | Message | Shutdown

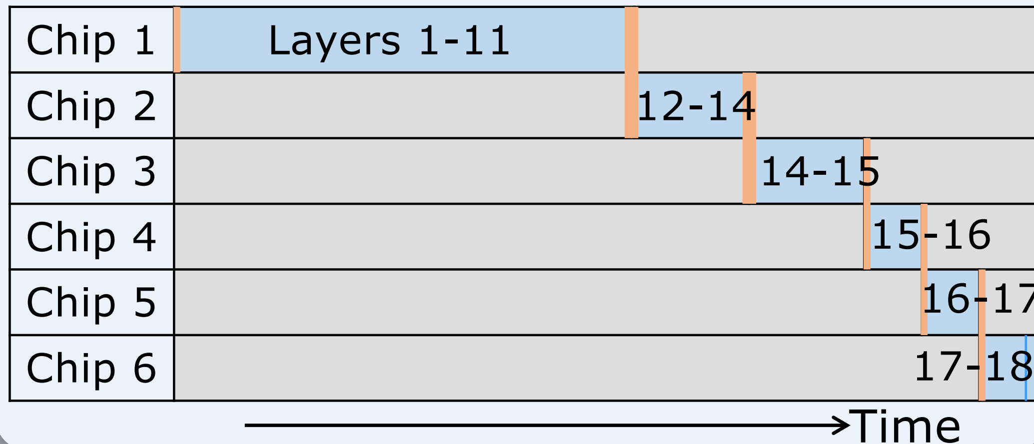
Both mappings ≤ 1.1× Dream EDP

[Giordano Symp. VLSI Circuits 21]

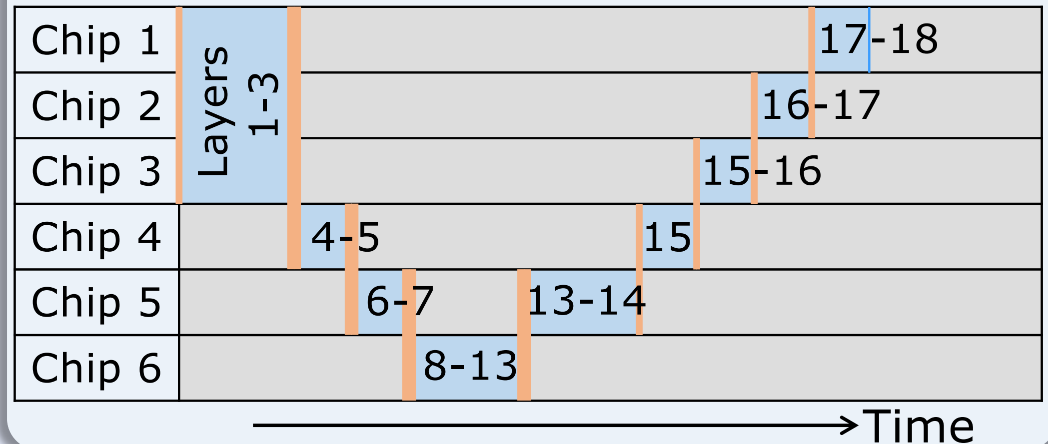
Illusion ⊃ Traditional Parallel Computing

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

Illusion Mapping 1: *373 KByte* messages



Illusion Mapping 2: *1.5 MByte* messages

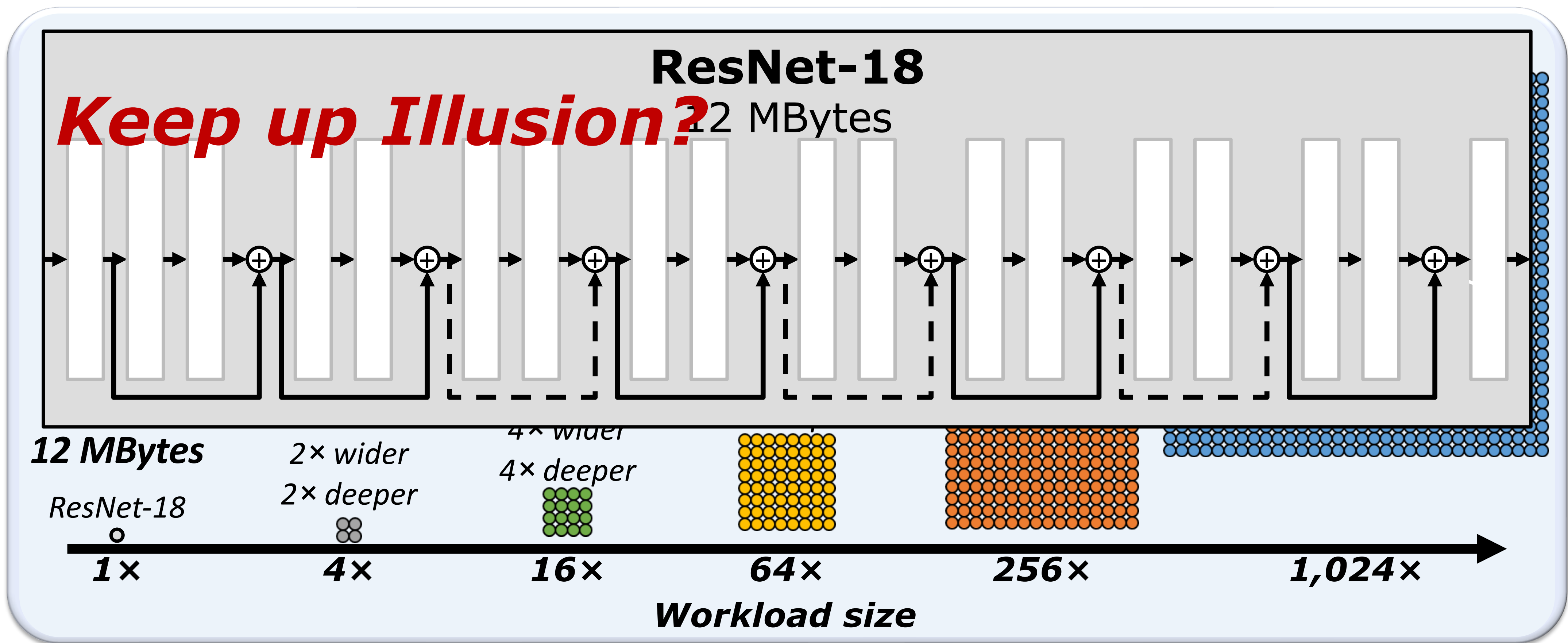


Compute | Message | Shutdown

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

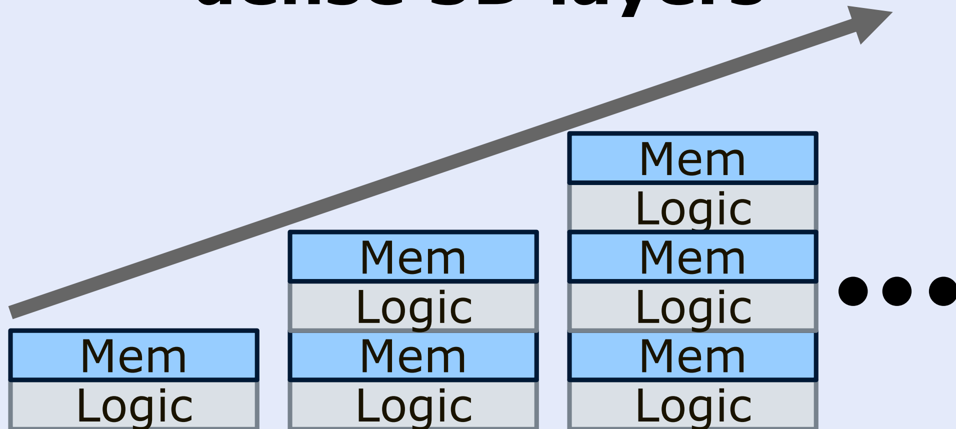
Traditional parallel: e.g., [Zimmer Symp. VLSI Circuits 19, Shao MICRO 19]

1,024× Workload Growth



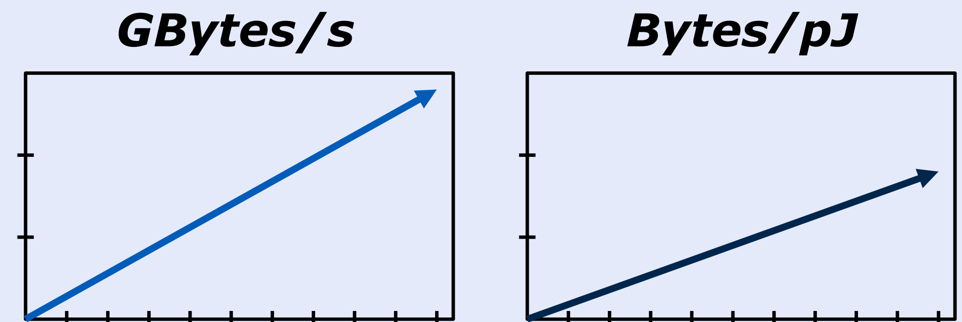
Illusion Scaleup

Linearly **increase dense 3D layers**



Reduce *message counts*

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

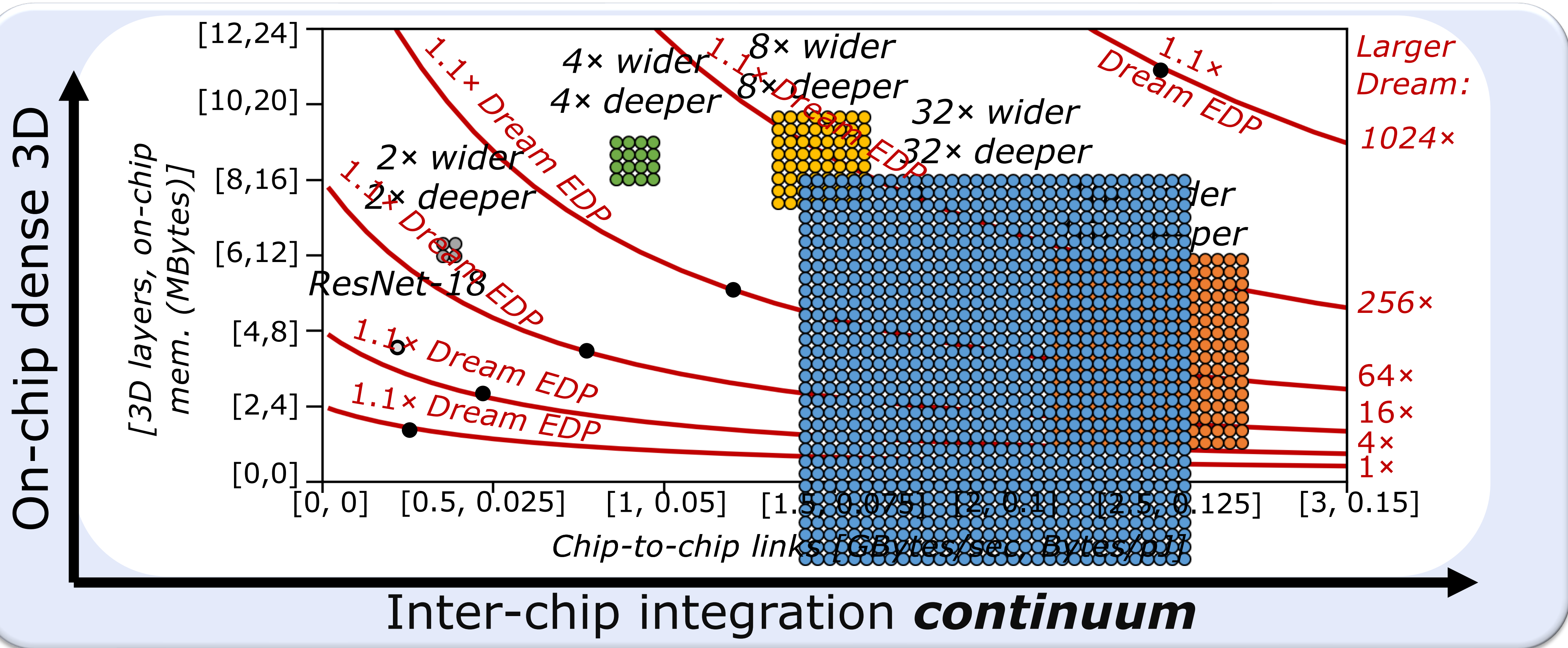


Reduce *per-message cost*

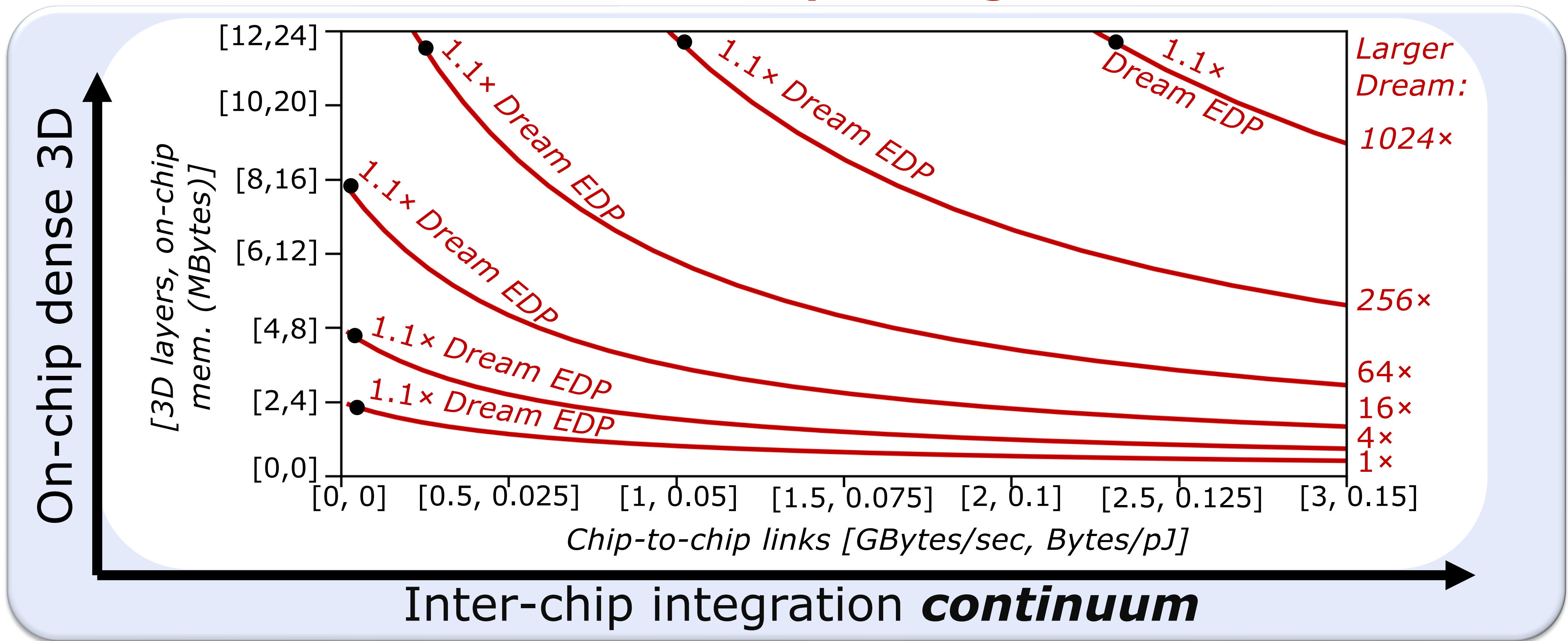
Quadratically **reduce Illusion *total message cost***

Illusion Scaleup

Maintain 1.1× Dream EDP despite growing Dream Chips



Illusion Scaleup

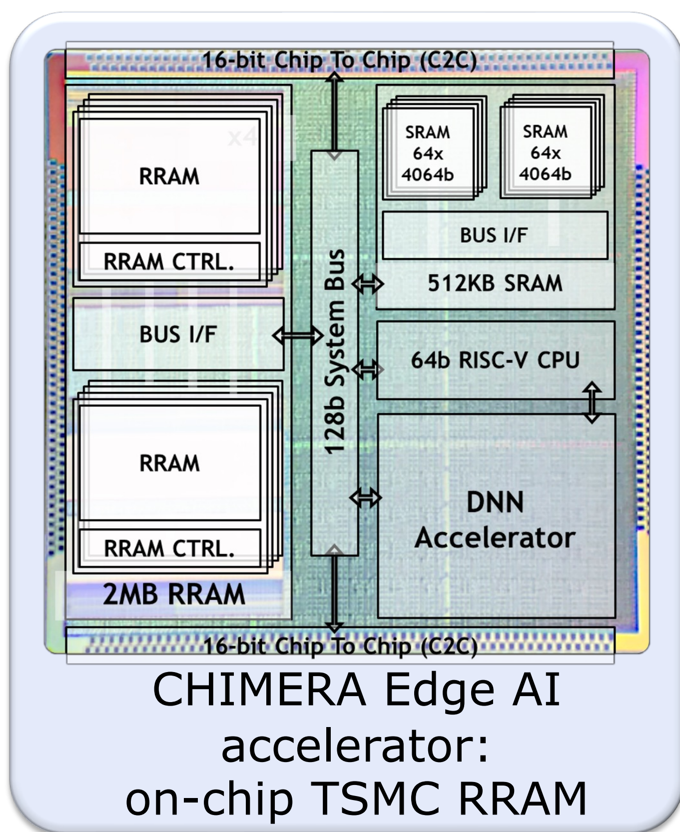


Many NanoSystems Opportunities

- **Co-design: device + circuit + arch. + algorithm**
  - **Multiple** layers **cooperate** for **large** benefits
- Dense compute + thermal + power delivery
- New software optimizations

RRAM Edge AI Incremental Training

New Low-Rank Training (LRT)



CHIMERA Edge AI accelerator: on-chip TSMC RRAM

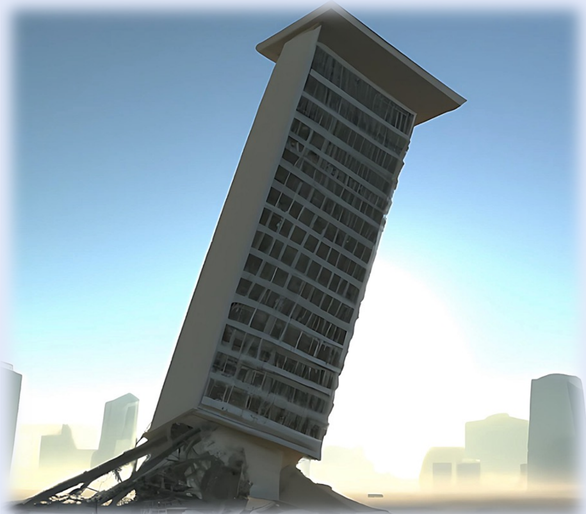
	LRT hardware results: iso-accuracy vs. Stochastic Gradient Descent (SGD)
RRAM weight update steps	**101× fewer vs. SGD**
Energy Delay Product	**340× better vs. SGD**
Endurance (20 samples/min.)	**10 years (LRT + ENDURER*) vs. 2 weeks (SGD)**

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

CO$$$T ?

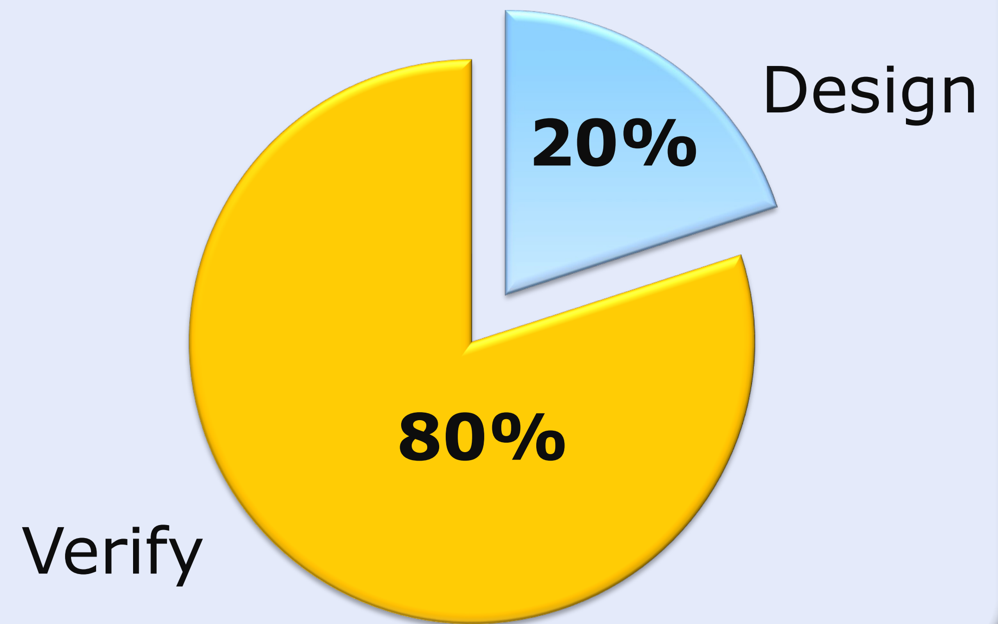
Today's Design Verification Inadequate

Critical bugs escape



Lacks solid foundation

Major bottleneck



Getting worse: custom hardware, complexity, security

Revolutionize Verification: QED

Pre-silicon

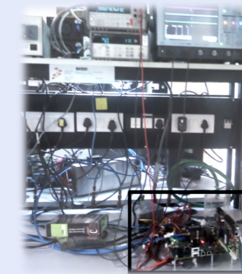
G-QED: Drastic benefits
(Industrial AI chips for cars)

Industry flow	G-QED: solid theory
1 year	**3 weeks**
Critical bugs missed	**New critical bugs (+ rest)**

[Chattopadhyay DAC 23]

Post-silicon

From months to seconds
(NXP SoC)



Industrial	QED
15 Billion	**9**

[Lin IEEE TCAD 14]

Test & Reliability: ***Enormous*** Challenges

Silent Data Corruption

facebook

Silent Data Corruptions at Scale

Harish Dattatraya Dixit, Facebook, Inc., hdd@fb.com
Sneha Pendharkar, Facebook, Inc., spendharkar@fb.com
Matt Beadon, Facebook, Inc., mbeadon@fb.com
Chris Mason, Facebook, Inc., clm@f
Tejasvi Chakravarthy, Facebook, Inc., teju@fb.com
Bharath Muthiah, Facebook, Inc., bharathm@fb.com
Sriram Sankar, Facebook Inc., sriramsankar@fb.com

22 Feb 2021

ABSTRACT

Silent Data Corruption (SDC) can have negative impact on large-scale infrastructure services. SDCs are not captured by error reporting mechanisms within a Central Processing Unit (CPU) and hence are not traceable at the hardware level. However, the data corruptions propagate across the stack and manifest as application-level problems. These types of errors can result in data loss and can require months of debug engineering time.

In this paper, we describe common defect types observed in silicon manufacturing that leads to SDCs. We discuss a real-world

machine learning inferences, ranking and rec
tems. However, it is our observation that com
always *accurate*. In some cases, the CPU can per
incorrectly. For example, when you perform 2x3,
a result of 5 instead of 6 silently under certain
conditions, without an indication of the miscom
event or error logs. As a result, a service utilizing
tially unaware of the computational accuracy an
the incorrect values in the application. This pa
focuses on scenarios where datacenter CPUs exhi

Cores that don't count

Peter H. Hochschild, Paul Turner, Jeffrey C. Mogul, Google, Sunnyvale, CA, US
Rama Govindaraju, Parthasarathy Ranganathan, Google, Sunnyvale, CA, US
David E. Culler, Amin Vahdat, Google, Sunnyvale, CA, US

Abstract

We are accustomed to thinking of computers as fail-stop, especially the cores that execute instructions, and most system software implicitly relies on that assumption. During most of the VLSI era, processors that passed manufacturing tests and were operated within specifications have insulated us from this fiction. As fabrication pushes towards smaller feature sizes and more elaborate computational structures, and as increasingly specialized instruction-silicon pairings are intro-

MI, USA. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3458336.3465297

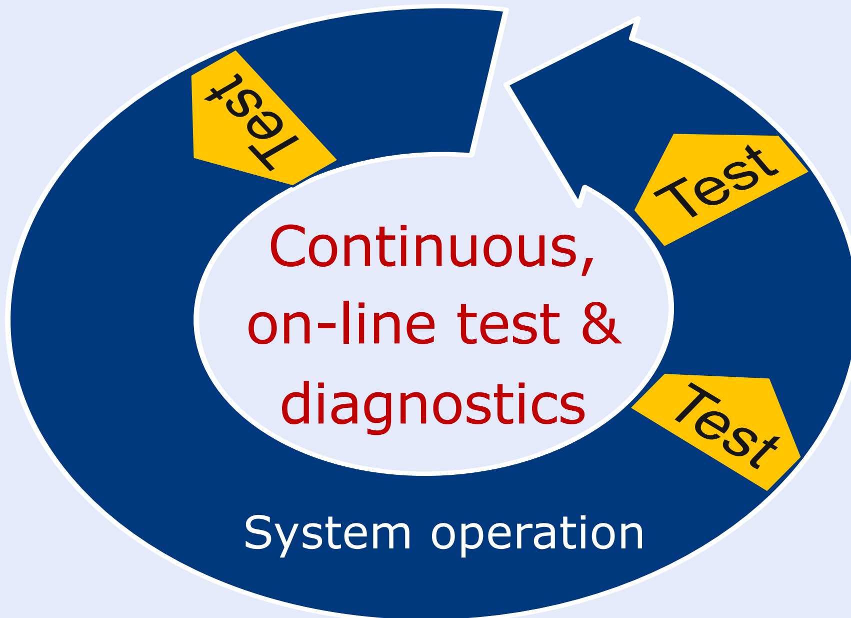
1 Introduction

Imagine you are running a massive-scale data-analysis pipeline in production, and one day it starts to give you wrong answers – somewhere in the pipeline, a class of computations are yielding corrupt results. Investigation fingers a surprising cause: an innocuous change to a low-level library. The change itself was

In-Field Test & Resilience

CASP

Concurrent, Autonomous, Stored Patterns



[Li DATE 08 & subsequent papers]

CASP derivatives

Data center, Automotive (ISO26262)

Intel Sapphire Rapids "in-field scan"
Nvidia, Renesas, TI, …

Intel In-Field Scan "IFS" Poised For Linux 5.19 To Help Spot Faulty Silicon

Written by Michael Larabel in Intel on 13 May 2022 at 06:53 AM

Comment

Deterministic ATPG-Based Runtime Test for Realizing Functional Safety in ISO 26262

DART: Dependable VLSI T

Anatomy of an in-die tester for infield testing

Sreejit Chakravarty

Tester-On-Chip: An in-field system-test interface for heterogeneous IPs

Intel In-Field Scan Feature in Sapphire Rapids Pinpoints Per-Core Errors

By Mark Tyson published March 02, 2022

A Linux driver release uncovered this feature ahead of its announcement.

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