
The Future of AI Compute Chips

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"*

AI Hardware: Past Decade

Accelerators

**Large one-time gains
plateauing**

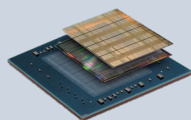
- Arithmetic
- Systolic array
- Custom dataflow

*Caution: Excessive
customization risky*

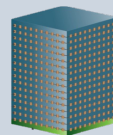
Tech. innovations

**Richer than
process nodes**

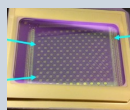
2.5D, 3D



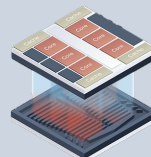
HBM



Optical



Microfluidic

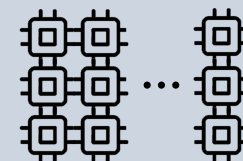


End-to-end systems

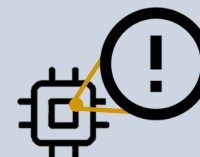
≠ peak TOPS/W



Off-chip



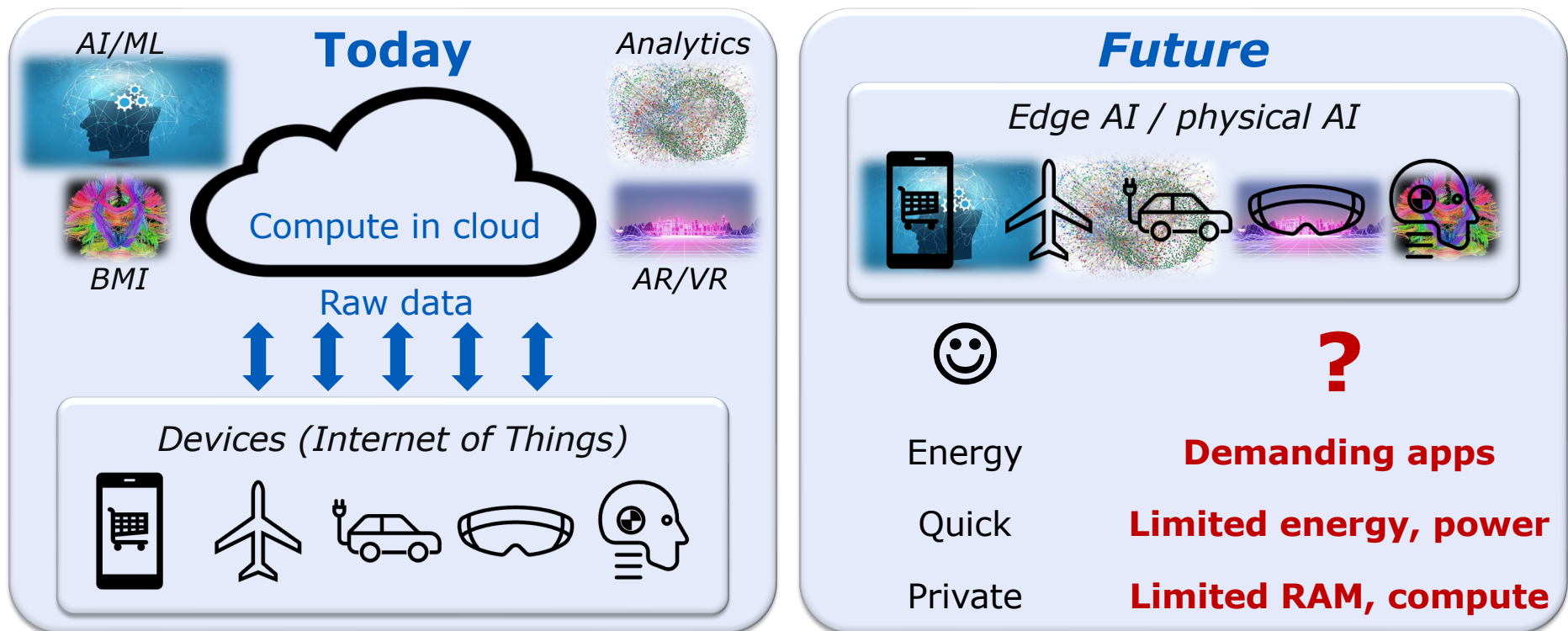
Many chips



Silent Data
Corruption

Swimming in Sensors, Drowning in Data

Abundant data

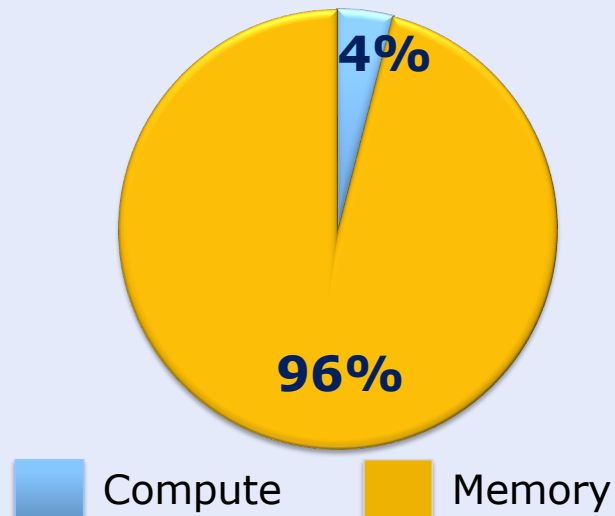


AI/ML = Artificial Intelligence/Machine Learning AR/VR = Augmented/Virtual Reality BMI = Brain Machine Interface

Memory Wall, Miniaturization Wall

Deadly combination!

Memory Wall



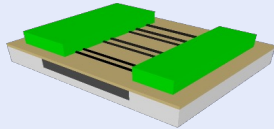
Miniaturization Wall



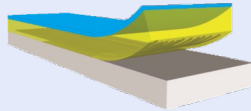
NanoSystems

New nanotech

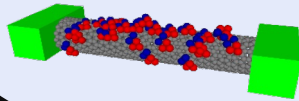
Devices



Integration

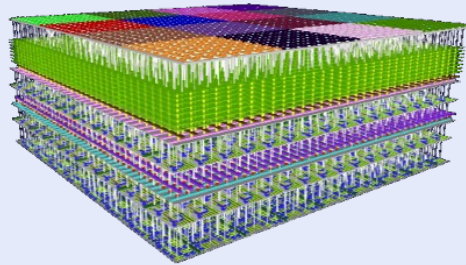


Sensors



New systems

New architectures



New applications



Data Explosion & Memory Wall

von Neumann, non-von Neumann

“Ideally ... desire an **indefinitely large memory capacity** such that **any particular ... word** would be **immediately available**. ... It does not seem possible physically ... We are ... forced ... hierarchy of memories”

[Burks, Goldstine, Von Neumann, 1946]

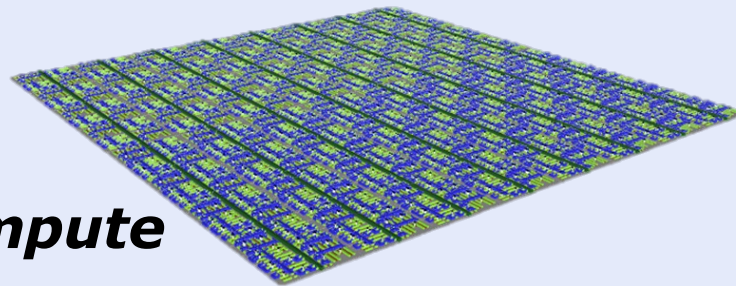
“Dream” Chip

Computing Today



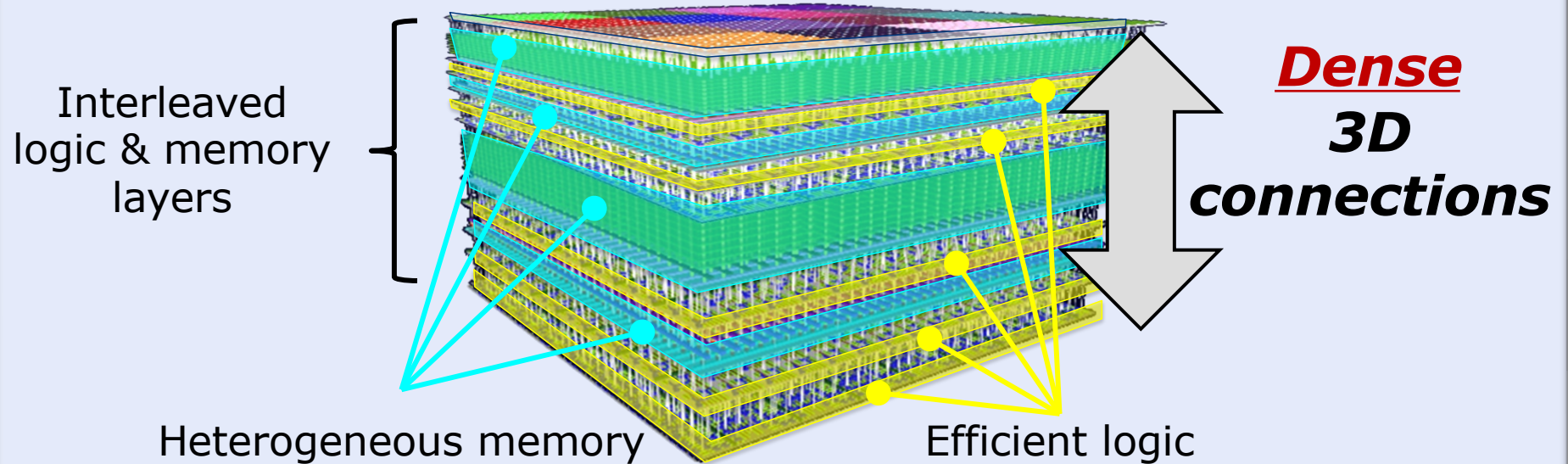
Memory

Compute



N3XT 3D: Dream Chip

Nano-Engineered Computing Systems Technology

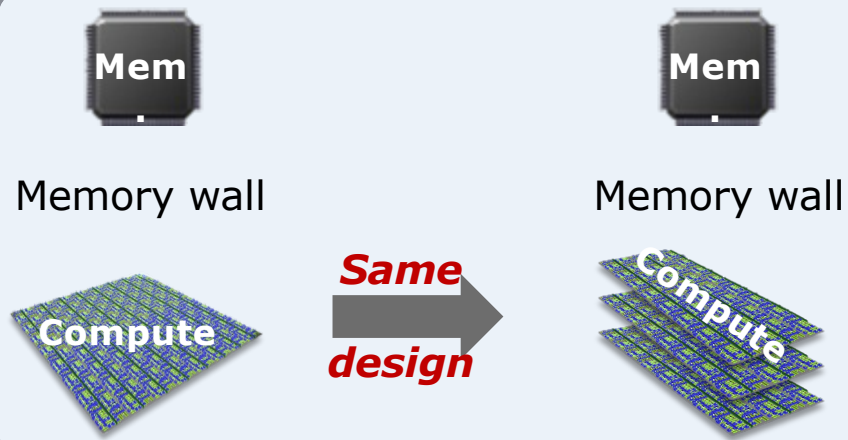


100× – 1,000× Energy Delay Product (EDP) benefits

N3XT 3D: Beyond 3D Folding

3D Folding

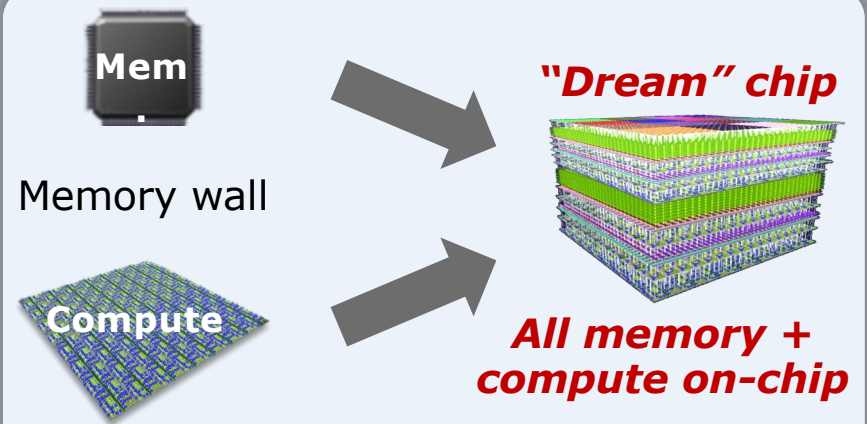
Limited EDP benefits: 1.4×



Some wirelength benefits

N3XT 3D

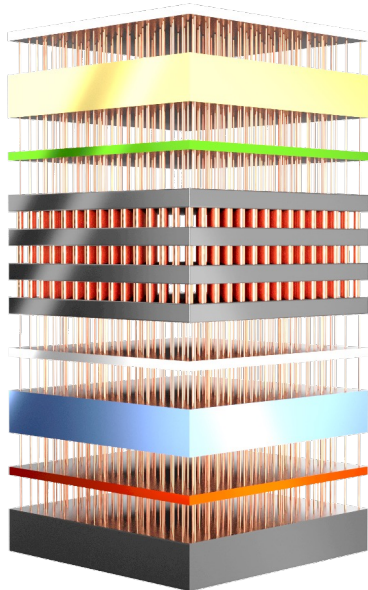
Large EDP benefits: 100-1,000×



New chip architectures

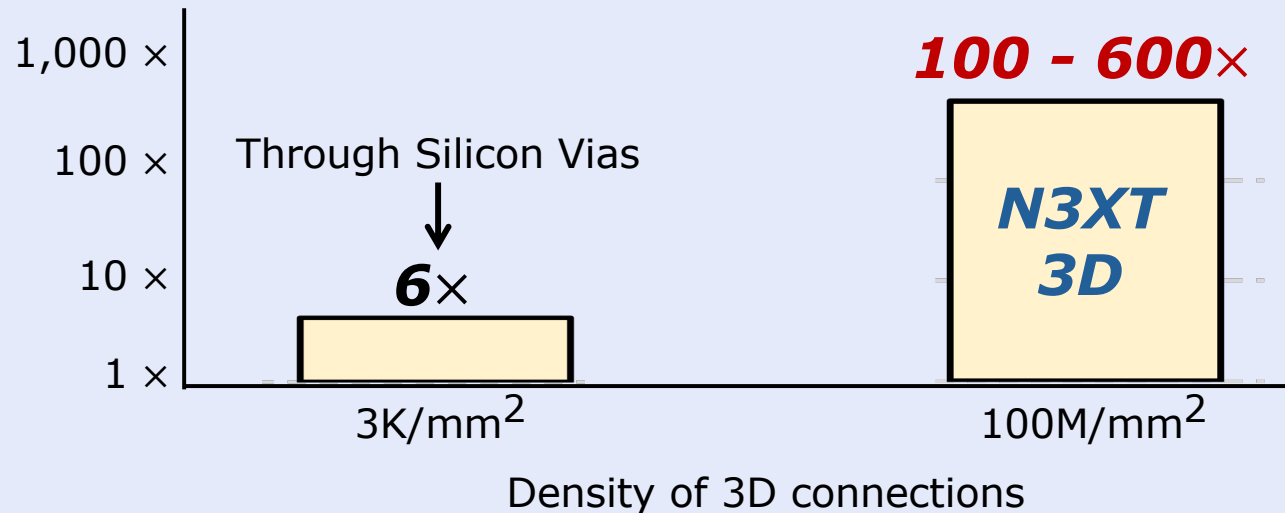
Dense 3D Connections: Large Benefits

N3XT 3D Chip



Multiple logic + **multiple** memory layers in 3D

Energy Delay Product benefits vs. today's packaging

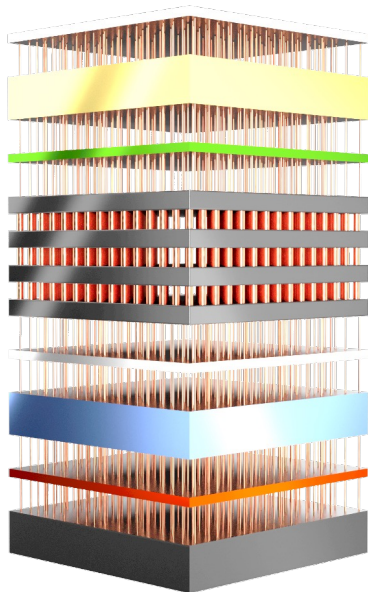


Monolithic 3D: only way today

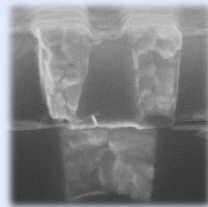
N3XT 3D: Many New Technologies

BEOL compatible $\leq 400^{\circ}\text{C}$ fabrication

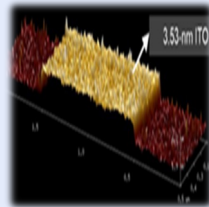
N3XT 3D Chip



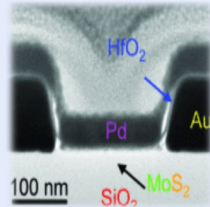
Carbon nanotubes



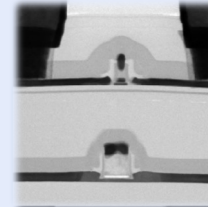
Oxide transistors



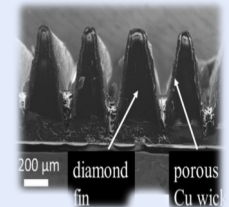
2D materials



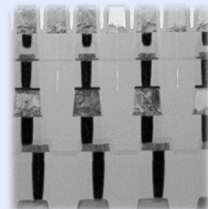
Low-Temp. Silicon



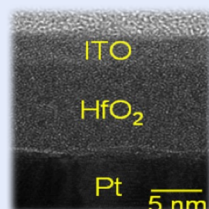
2-Phase Cooling



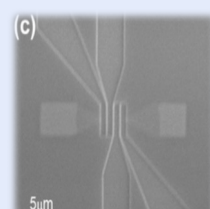
Resistive RAM



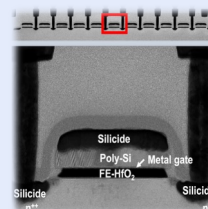
Hybrid Gain Cell



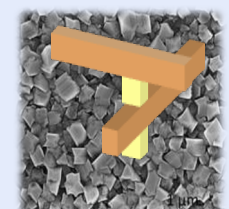
Magnetic RAM



Ferro-electric



3D Thermal Scaffolding



BEOL = Back-End-of-Line

Lab to Fab



Many “firsts” in industry fabs

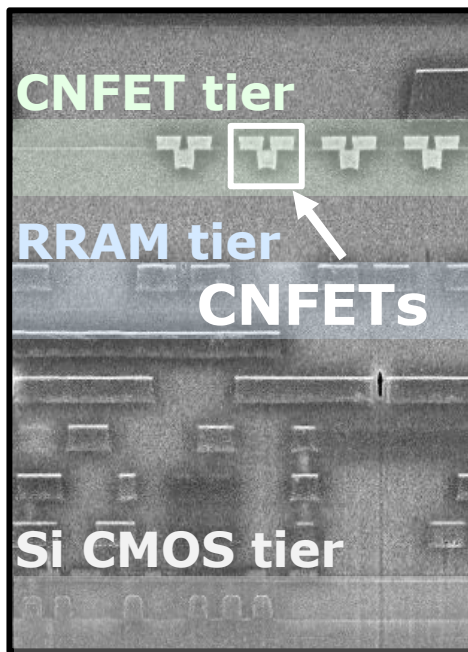
1. Carbon nanotube FETs (CNFETs)
2. U.S. foundry Resistive RAM (RRAM)
3. Dense monolithic 3D (up to 5 tiers):
CNFET+ RRAM + silicon CMOS

4. Product development



✓ Full wafer ✓ All integrated ✓ Design support

Foundry Monolithic 3D: CNFET + RRAM + Silicon CMOS



- M8 – M9
- CNFET
- M6 – M7
- RRAM
- M1 – M5
- Si CMOS

Monolithic 3D CNFET + RRAM (vs. Silicon FET + RRAM)

- ✓ iso-footprint
- ✓ iso-energy/latency
- ✓ iso-reliability/endurance
- ✓ iso-retention
- ✓ multiple bits per cell

same wafer, same process
Foundry monolithic 3D PDK

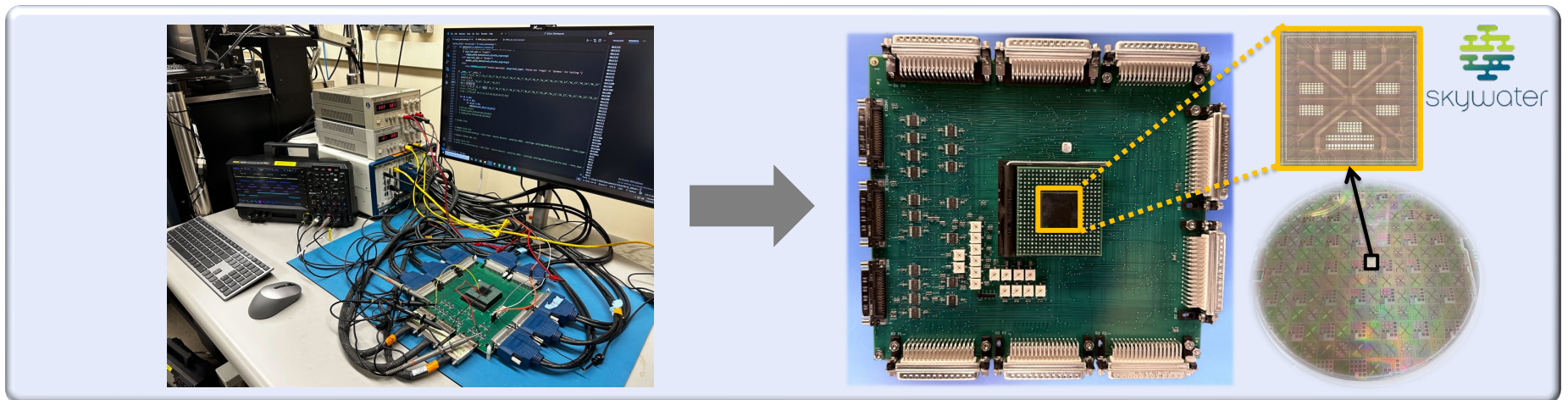
First Foundry Monolithic 3D System

4x throughput in hardware, apples to apples

New design principles: iso-node, iso-footprint, iso-memory capacity

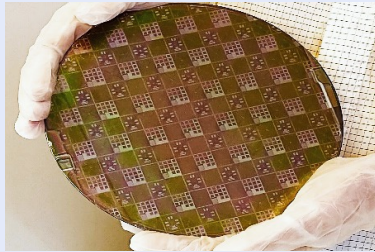
Densest 3D connections: surpass hybrid bonding, TSVs

3 heterogeneous device layers: CNFET + RRAM + Silicon CMOS



New Dense 3D Architectures

More Bandwidth

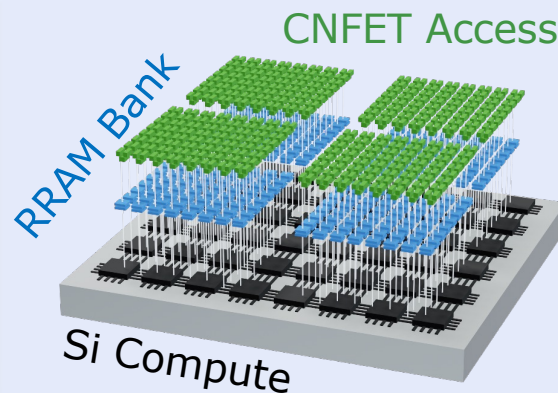


$4\times$ EDP gains in HW

iso-node, iso-footprint,
iso-memory capacity

[Choi IEDM 25]

More Compute

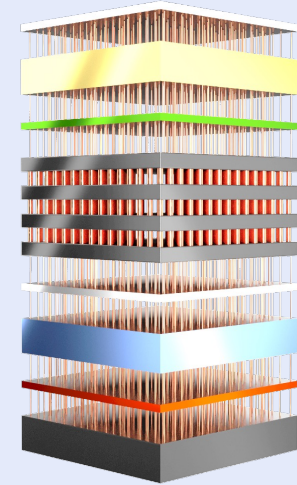


$10\times$ EDP gains

iso-node, iso-footprint,
iso-memory capacity

[Srimani DATE 23]

Ultimate N3XT 3D



**$100 - 1,000\times$
EDP gains**

[Srimani IEDM 23, Aly Proc. IEEE 19]

Beyond-silicon Systems: Large End-to-End Benefits



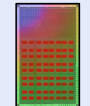
First foundry monolithic 3D system
[IEDM 2025]



MINOTAUR Transformer RRAM Edge AI
[Symp. VLSI 2024]



CHIMERA: First foundry RRAM Edge AI
[Symp. VLSI 2021]



First RRAM IoT microcontroller
[ISSCC 2019]

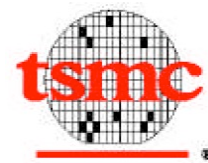


First monolithic 3D NanoSystem
[Nature 2017]



First carbon nanotube computer
[Nature 2013]

Products



*Infinion and TSMC to introduce RRAM
... automotive microcontroller product*

*GlobalFoundries ... non-volatile resistive
RAM ... IoT & 5G applications*

*Weebit Nano ReRAM IP now available in
SkyWater Technology' S130 process*

Beyond-silicon Systems: Large End-to-End Benefits



First foundry monolithic 3D system
[IEDM 2025]



MINOTAUR Transformer RRAM Edge AI

Products



New design principles:

Loosely coupled with technology-specific implementations



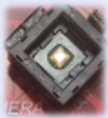
First monolithic 3D NanoSystem
[Nature 2017]



First carbon nanotube computer
[Nature 2013]

Weebit Nano ReRAM IP now available in
SkyWater Technology' S130 process

Co-design: Resistive RAM Write Challenges



CNN incremental training

10-year, 20 images/sec

Low-Rank Training + ENDURER (vs. SGD)	
Weight updates	101× fewer
EDP	340× better
Lifetime	10 years vs. 2 weeks

[CHIMERA Symp. VLSI 21]



Transformer fine-tuning

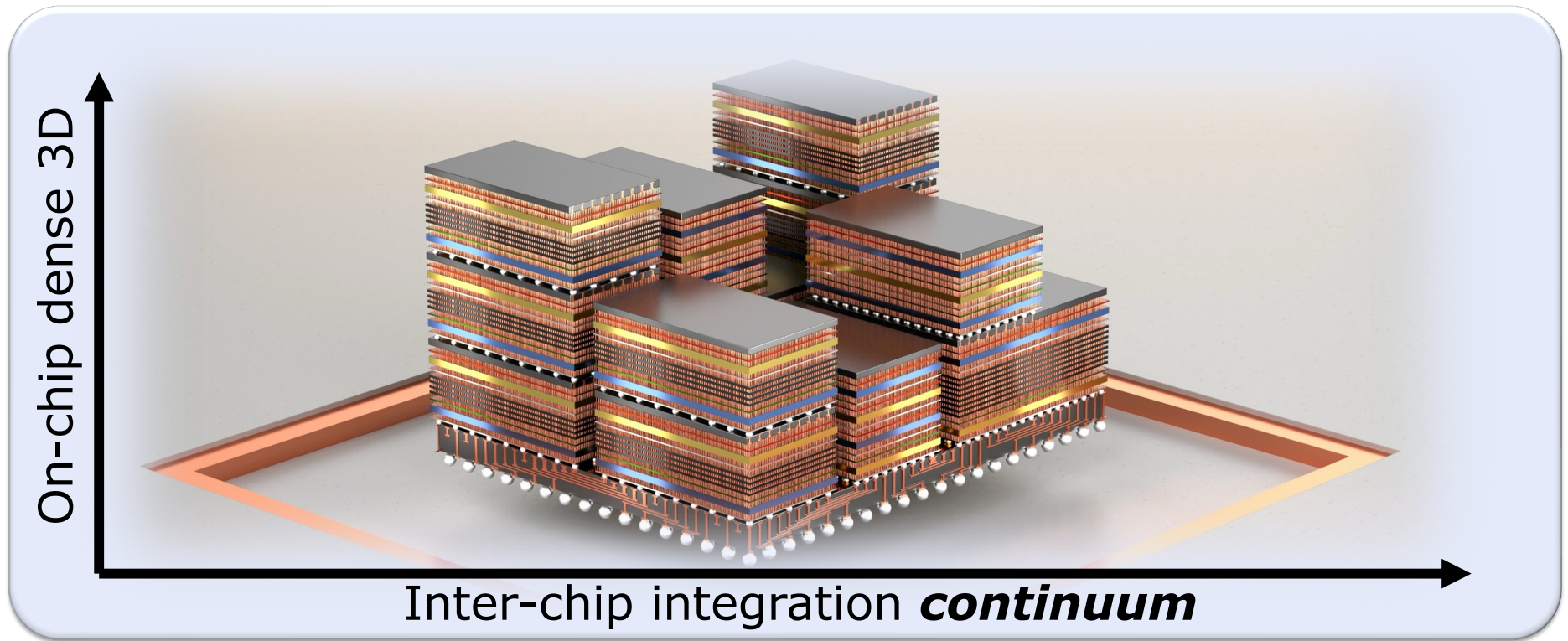
SGD: doesn't fit on-chip

Low-Rank Adaptation-based
Fine-tuning: full on-chip
On-chip: 13.3 MBytes
Weight updates: 229× fewer (vs. SGD)

[MINOTAUR Symp. VLSI 24]

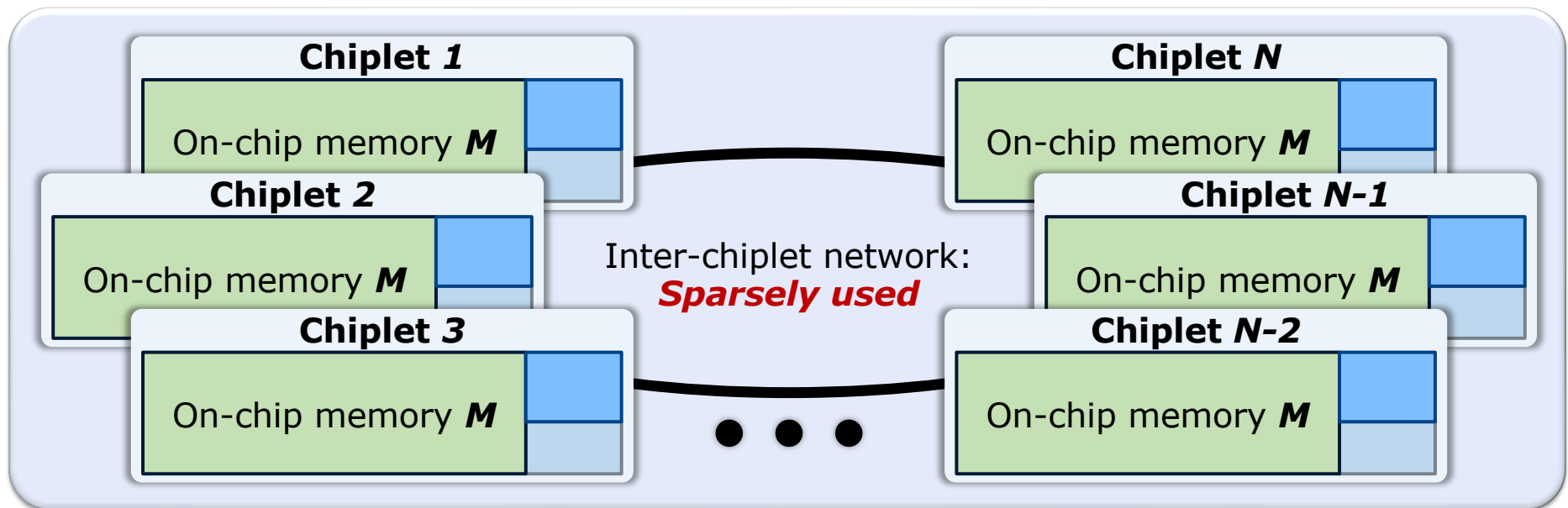
BUT, Dream Chip Infeasible: Moving Target

N3XT 3D MOSAIC



“Illusion” of Dream Chip

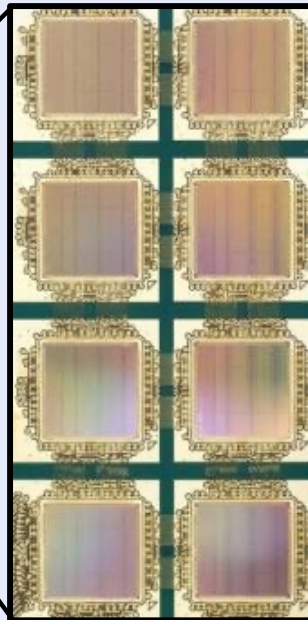
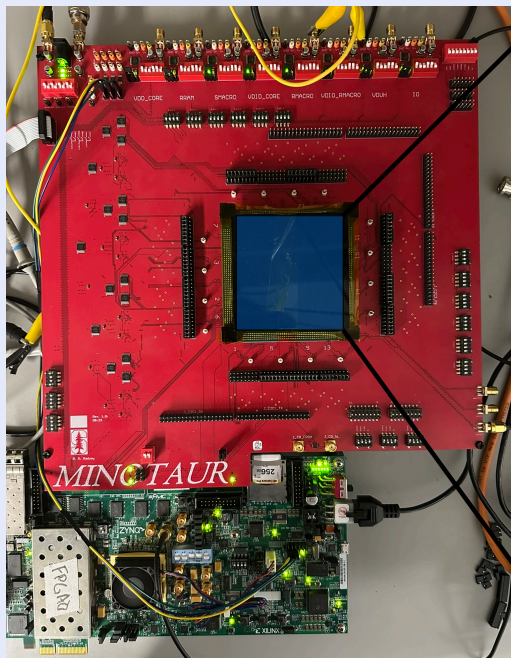
Enough on-chip mem. + **Fine-grained** power management



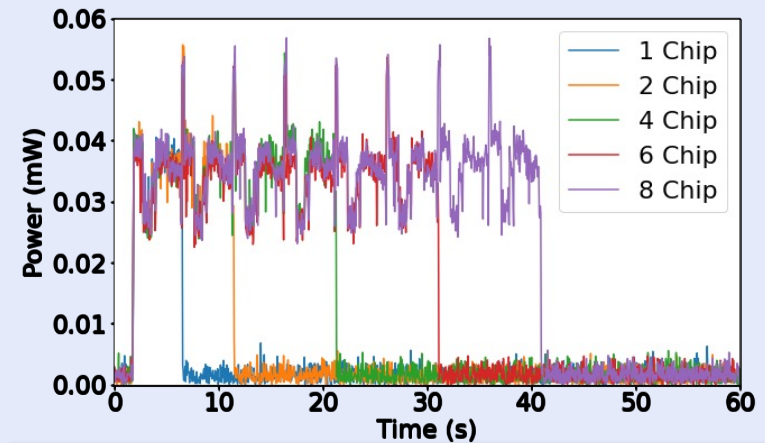
Special mapping: e.g., $1.1\times$ of Dream energy & throughput

Illusion: Hardware-Proven, Backed by Theory

96-MByte Transformers

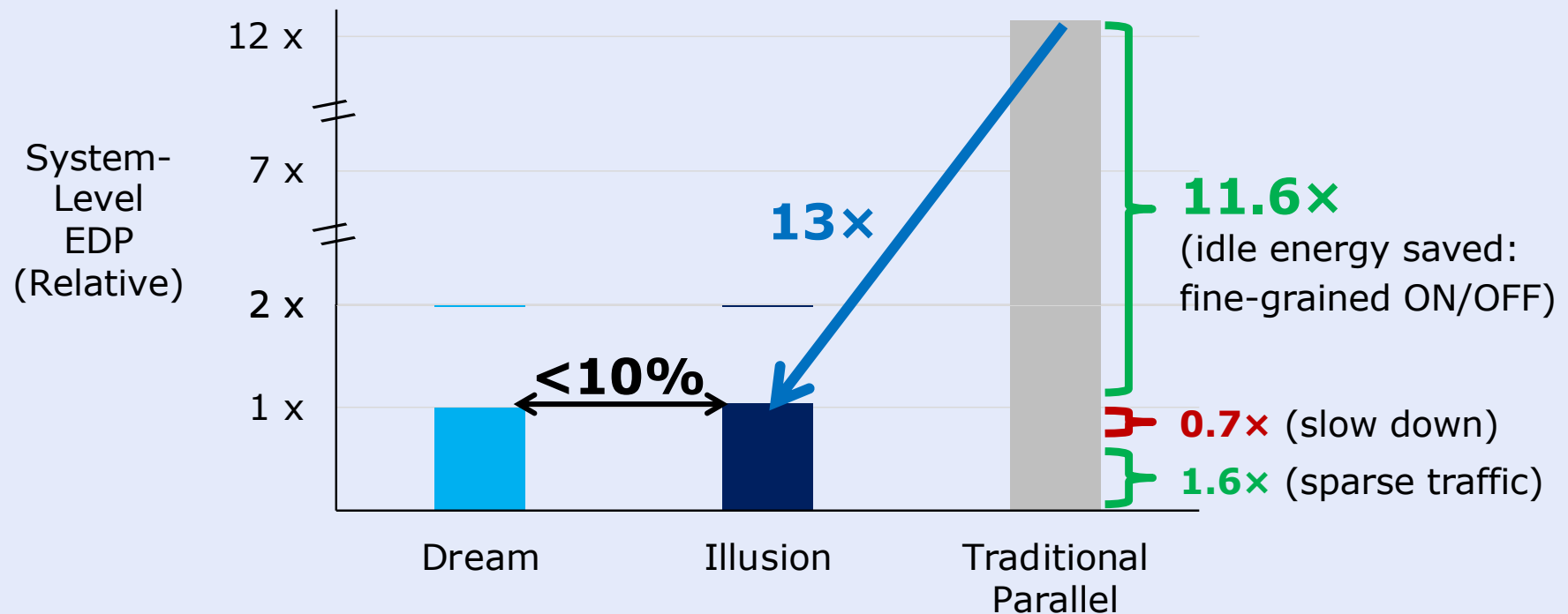


BERT Encoders scaled to chiplet count



**Within 10% of Dream
energy & execution time**

Illusion vs. Traditional Parallel



16 chips, 16 encoder 186 MByte MobileBERT. Traditional parallel applied iso-hardware using approach in [Shao MICRO 19]. EDP relative to Dream Chip.

More Topics

- ☐ *N3XT 3D* thermal
- ☐ Illusion scaleup
- ☐ Role of photonics
- ☐ Analog computing
- ☐ ...
- ☐ **Cost \$\$\$: Design cost, yield, reliability**

Chip Design: Deep Expertise, Manual Effort, Many Months

Verification is Key!

14%

designed correctly
first time

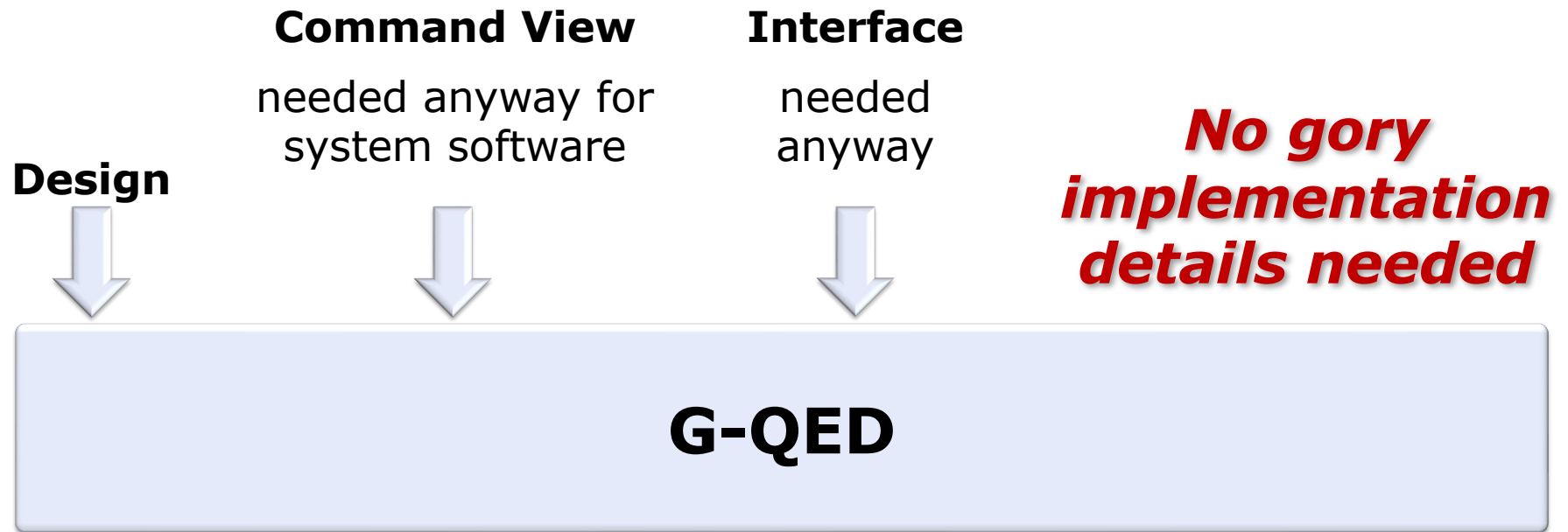
>\$10M~100M

Massive "re-spin" cost
(Bug fix after manufacture)

2~3 verification engineers per design engineer

Design & verification: **~70%** overall effort

New G-QED Verification: AI-Ready



Solid theory: all bugs (up to BMC Bound), **no false fails**

BMC: Bounded Model Checking

New G-QED Verification: **AI-Ready**

Only Technique:

- All bugs up to BMC bound
- Large accelerators
250M gates, 300K-deep bugs
- No false fails
New design-for-verification
- AI-ready: drastic productivity

Wide range of designs

Hot AI hw startup

New **chip-killer** bug + more

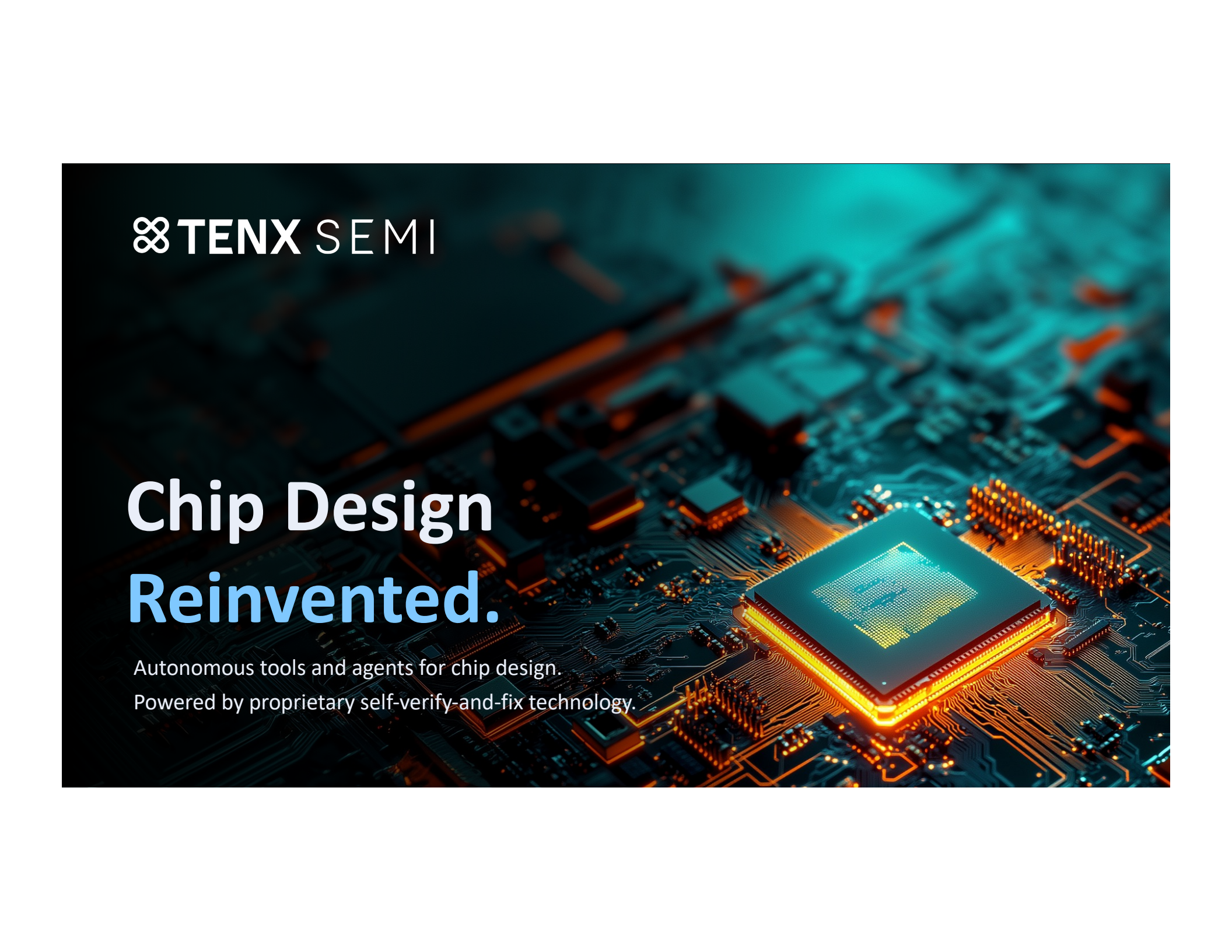
Major AI hw company

New **sw-unfixable** bug + more

Top auto hw company

9 new critical bugs + rest

3 weeks vs. 1 year

The background of the slide is a close-up, high-angle photograph of a complex electronic circuit board. A central square integrated circuit (chip) is the focal point, with its surface reflecting light in a grid-like pattern. The board is covered in intricate copper-colored traces and various smaller components. The lighting is dramatic, with a strong blue/cyan tint on the left and a warm orange/yellow glow emanating from the central chip and its immediate surroundings. The overall aesthetic is high-tech and futuristic.

 **TENX SEMI**

Chip Design Reinvented.

Autonomous tools and agents for chip design.
Powered by proprietary self-verify-and-fix technology.

Manufacturing Defects

Silent Data Corruption by 10× Test Escapes
Threatens Reliable Computing



Subhasish Mitra¹, Subho Banerjee, Martin Dixon, Rama Govindaraju², Peter Hochschild, Eric X. Liu, Bharath Parthasarathy, Parthasarathy Ranganathan

Manufacturing test escapes

5,000 DPM or more

Industry target: 100-500 DPM

CPUs, GPUs, Accelerators, Google, Meta, Microsoft, not just hyperscalers!

Most detected in field: hours of system-level testing

Manufacturing Test Foundations Broken Today

Inadequate test metrics



e.g., cell-aware testing

Test floor constraints



Manufacture time

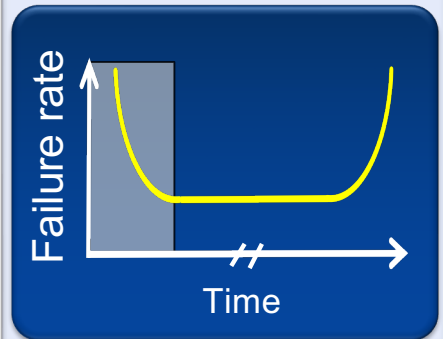
System-level tests



Lack solid foundation

Burn-in challenges

ELF: ~29%

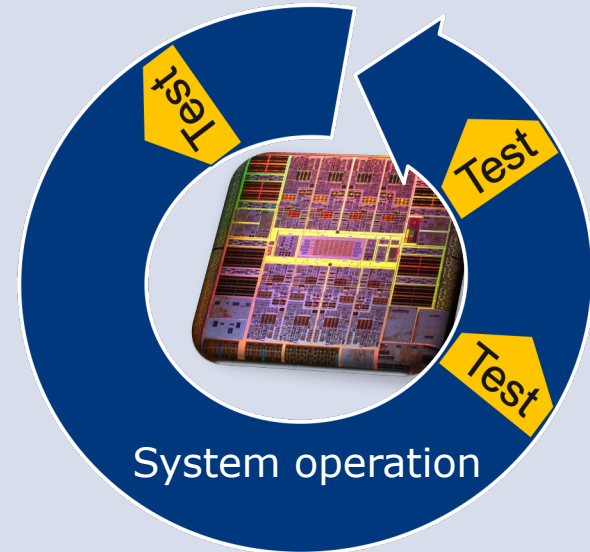


Early-**L**ife
Failures

Big Idea whose Time has Come

CASP: Concurrent Autonomous Stored Patterns [Li DATE 08]

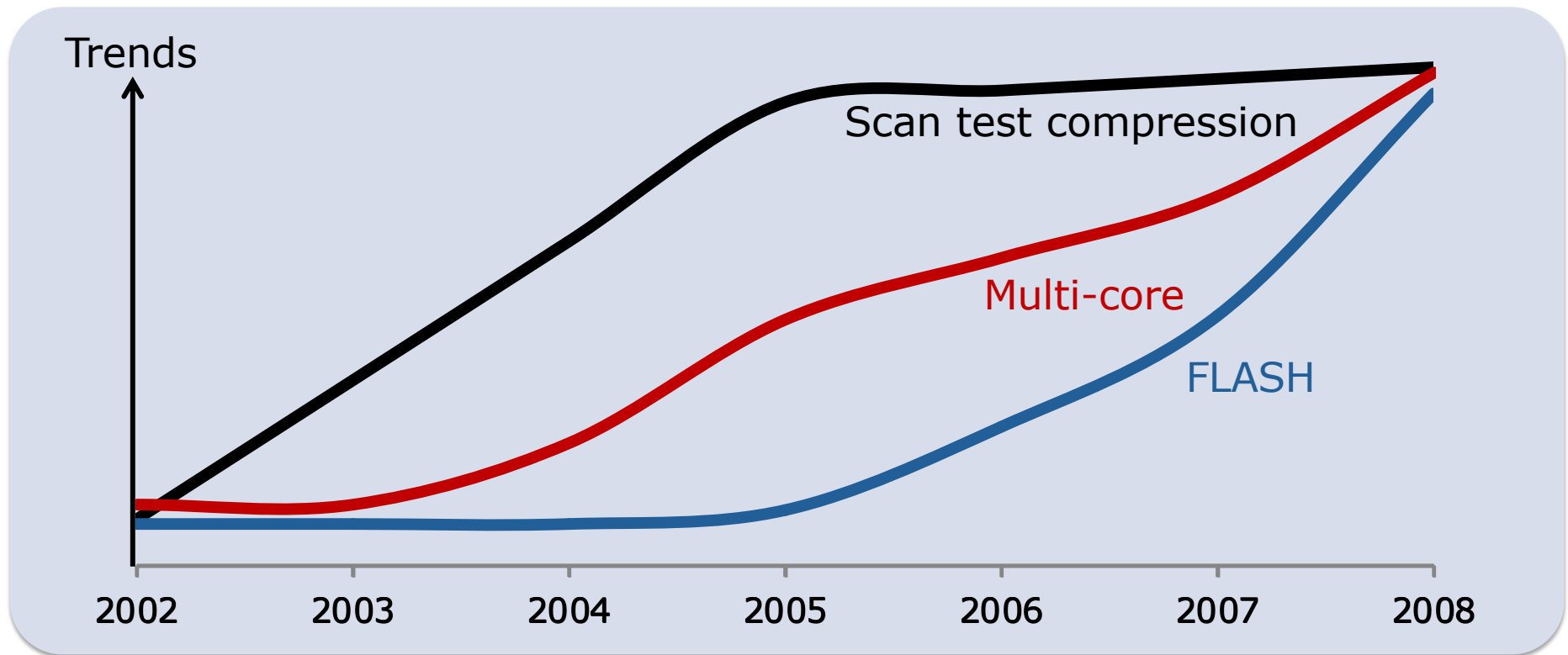
- **On-line or off-line**
- **Beyond** manufacturing tests
 - Flexible: update test content
- Test time $\neq$ test floor constraints



CASP derivatives:

Amazon (In-fleet scan), Intel (In-field scan), NVIDIA (In-system Test), Siemens, Synopsys, ...

Confluence of Hardware Trends *circa* 2007



Confluence of Hardware Trends *circa* 2007

- **Thorough scan tests + updates**

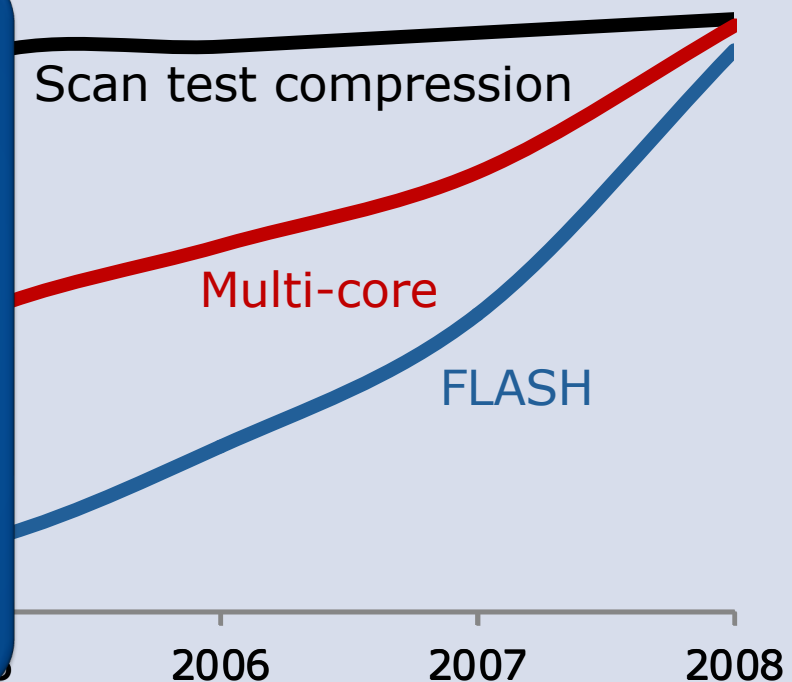
- Off-chip FLASH + scan test compression

- **Scan test patterns = “data”**

- High-speed I/Os $\neq$ scan pins

- **No visible system downtime**

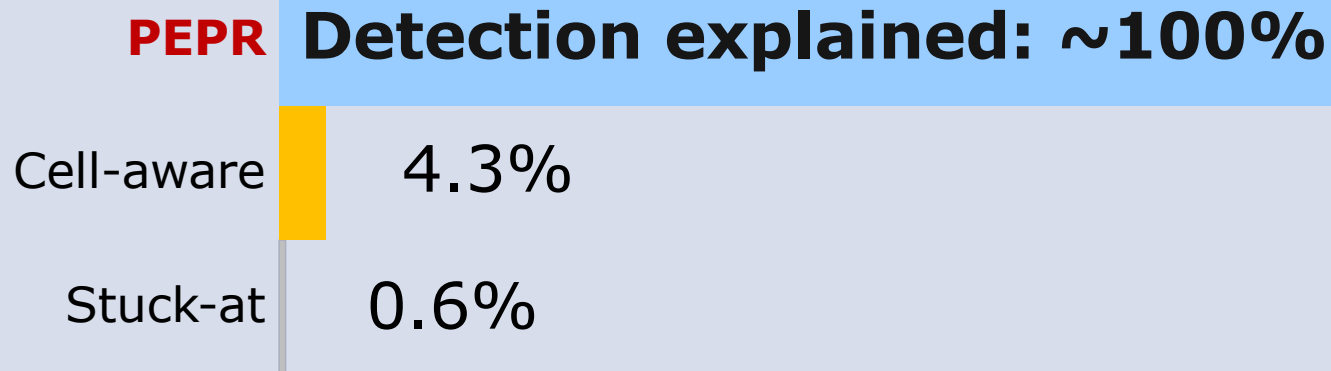
- Multi-core



PEPR: New Coverage Metric, Super-Thorough

Pseudo-Exhaustive Physically-Aware Region Testing

Defective chips



Inspired by [McCluskey, IEEE Trans. Computers, 81]

Large PEPR pattern count: suited for CASP

Resilient Systems

Circuit failure prediction	<i>Eventual</i> failure detection	Classical error detection
+ Before errors occur	Depends	– After errors occur
+ Low cost: CASP + ELF signatures	+ Low cost: CASP + QED	– High cost: Escapes, hidden costs
Early-life failures (ELF)	+ Permanent failures	+ General

Future of AI Compute Chips

- Dream chip: **multiple** logic + memory layers + **Dense 3D**
 - ***N3XT 3D MOSAIC + Illusion:*** new principles
 - Widely applicable for many technologies
 - Large gains, practical: lab-to-fab hardware
- Drastic design productivity: **G-QED** verification + **ABCD**
- Effective reliability: **CASP** + **PEPR** + **QED**