NanoSystems

The N3XT 1,000×

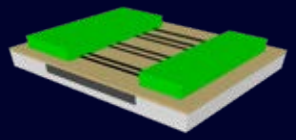
Subhasish Mitra

Department of EE & Department of CS
Stanford University

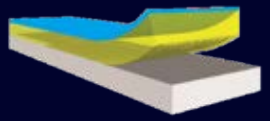
Nano*Systems*

New nanotech

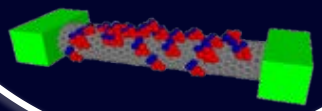
Devices



Fabrication

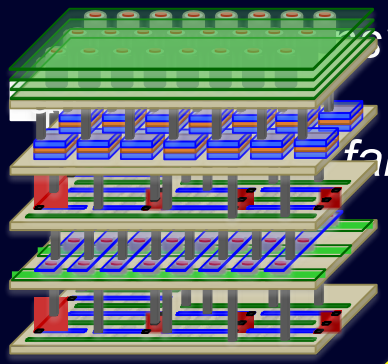


Sensors



New systems

New architectures

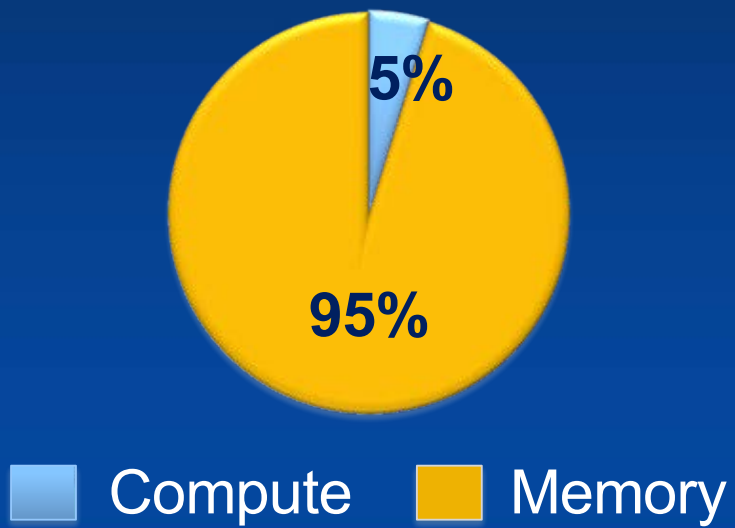


New applications

Application Challenges

Abundant-data apps.

Memory wall



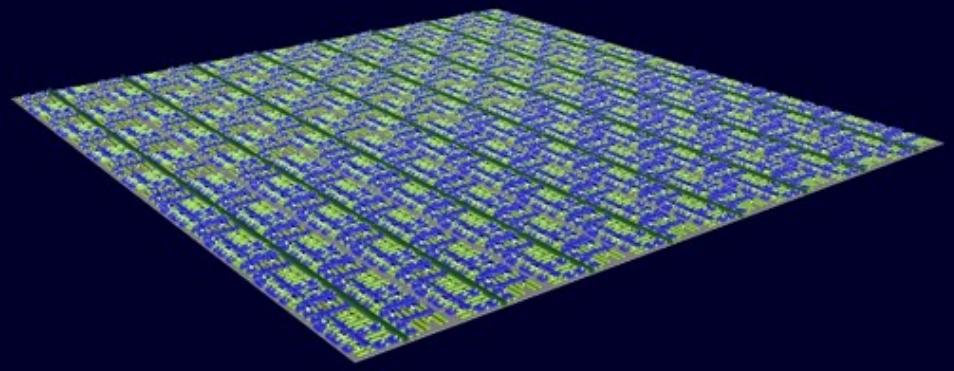
Brain-inspired ⊃ Neural Nets

Chip realization ?

- Compute + memory
- Dense connectivity
- Energy efficiency
- Footprint

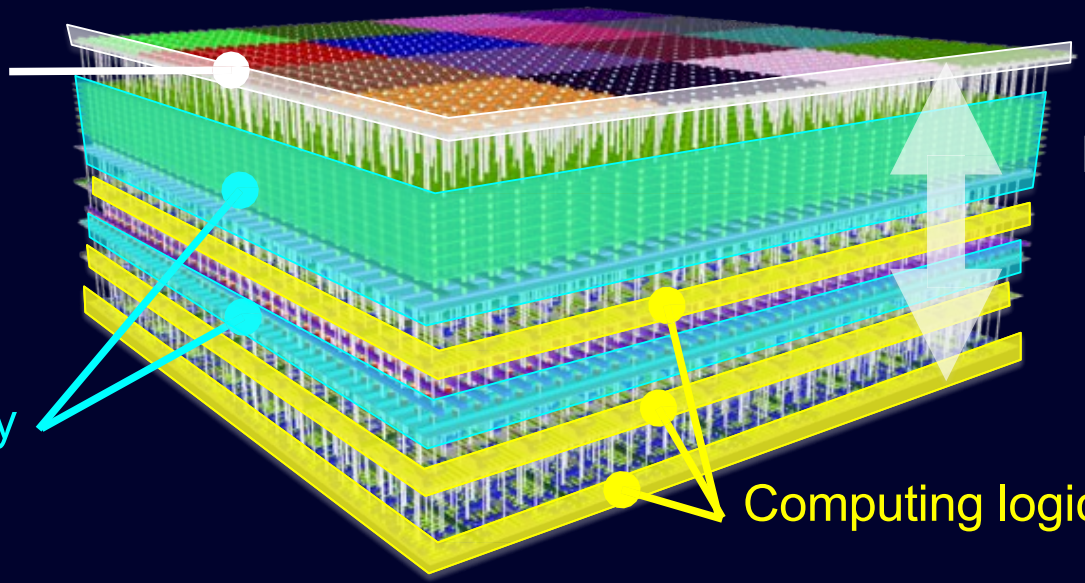
N3XT NanoSystems

Computation immersed in memory



N3XT NanoSystems

Computation immersed in memory



Impossible with business as usual

Nano-Engineered Computing Systems Technology

[Aly IEEE Computer 15, Proc. IEEE 19] Stanford + CMU + MIT + NTU Singapore + UC Berkeley + U. Michigan

N3XT Computation Immersed in Memory

3D Resistive RAM (massive)

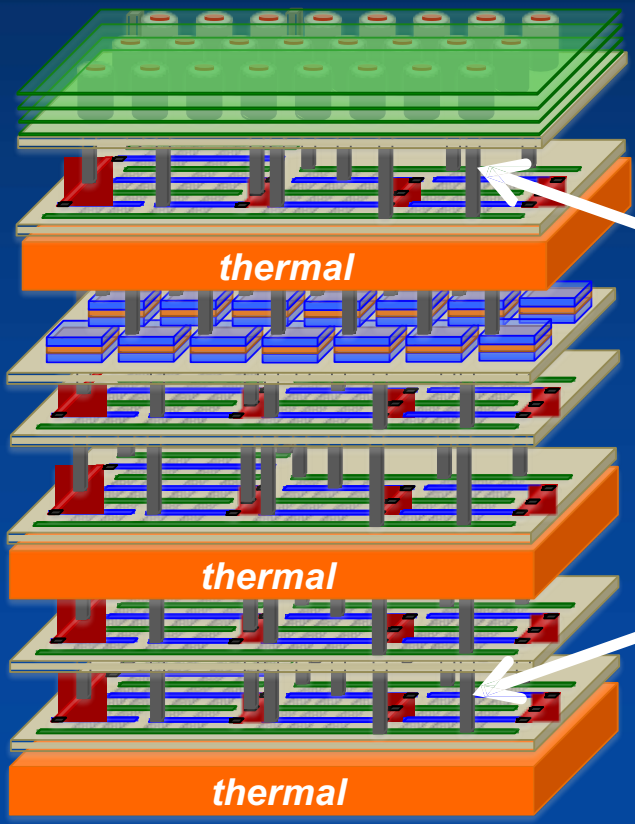
•••

1D CNFET, 2D FET (logic)

MRAM (quick access)

1D CNFET, 2D FET (logic)

1D CNFET, 2D FET (logic)



No TSV

Ultra-dense, fine-grained vias

Silicon compatible

DARPA 3DSoC Program

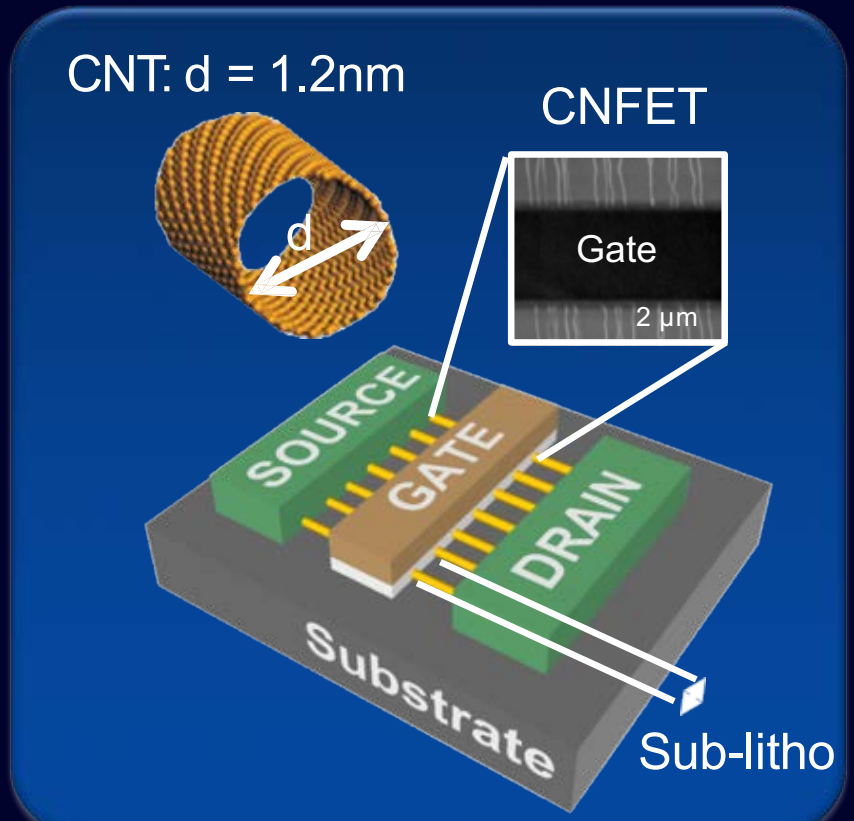
Max Shulaker
Anantha Chandrakasan

Subhasish Mitra
H.-S. Philip Wong, Simon S. Wong

Brad Ferguson
Mark Nelson

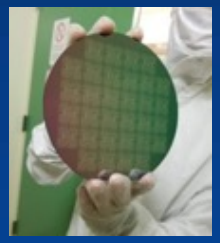
Jefford Humes

Carbon Nanotube FET (CNFET)

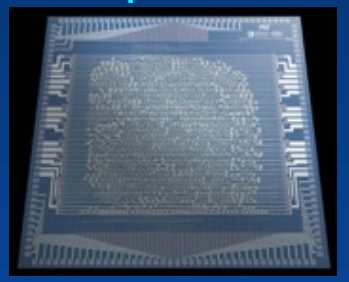


Major progress

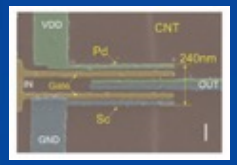
SkyWater



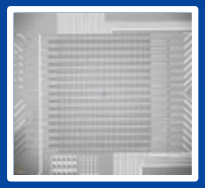
Microprocessor



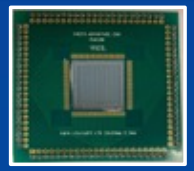
Scaled devices



SRAM arrays



Monolithic 3D imager

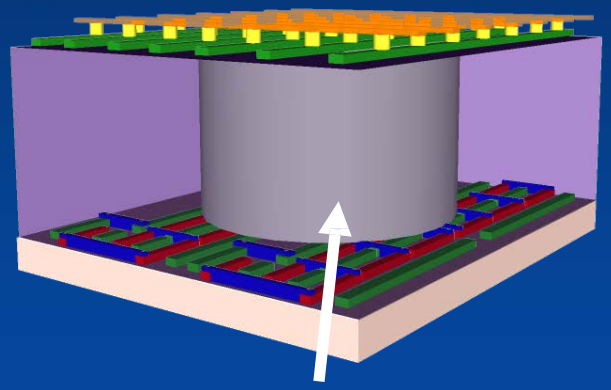


MIT (Shulaker), Peking U. (Peng), Stanford (Mitra, Wong) & others

3D Integration

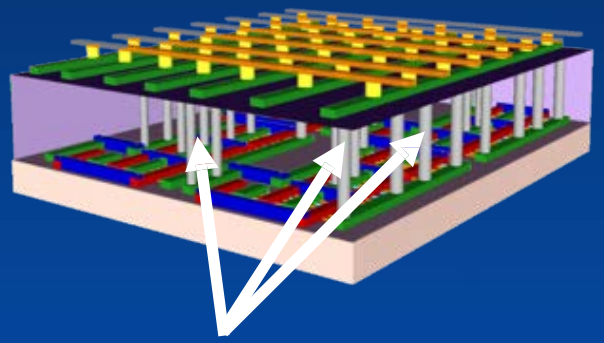
- Massive ILV density >> TSV density

TSV (chip stacking)



Through silicon via (TSV)

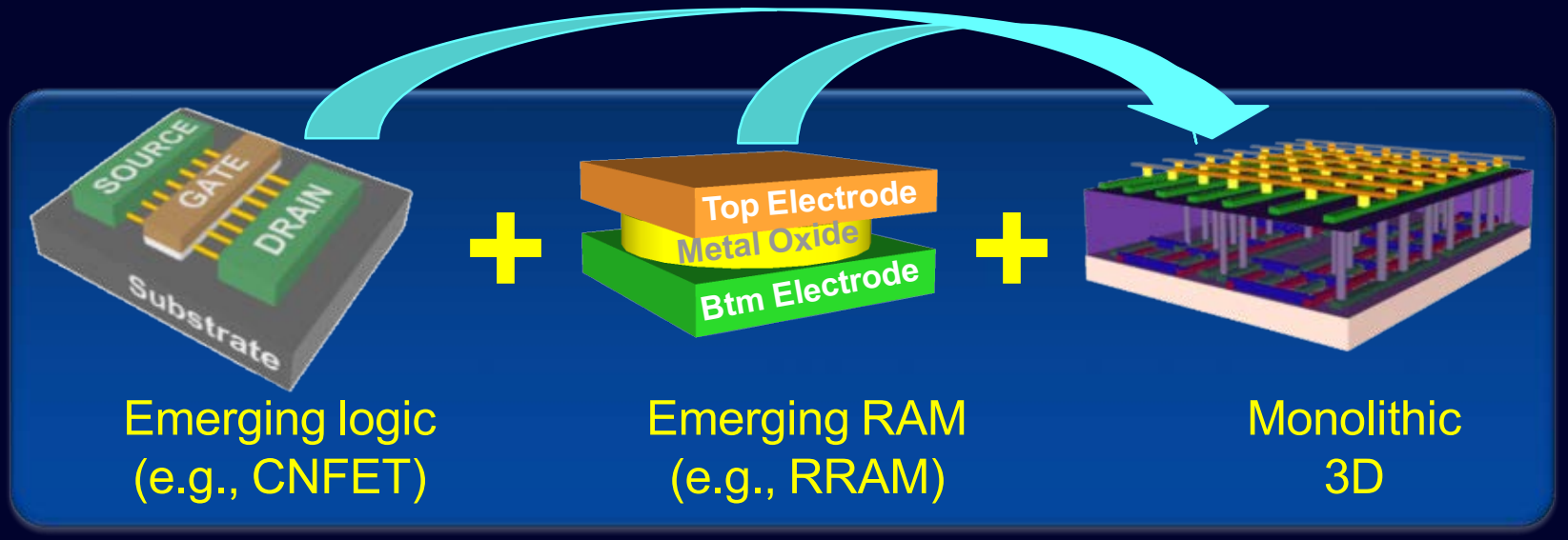
Dense, e.g., monolithic



Nano-scale inter-layer vias (ILVs)

Realizing Monolithic 3D

Device + Architecture benefits



Emerging logic (e.g., CNFET)

Emerging RAM (e.g., RRAM)

Monolithic 3D

Naturally enabled: < 400 °C fabrication

Foundry CNFET + RRAM + Monolithic 3D

First 3D Nanotube and RRAM ICs Come Out of Foundry

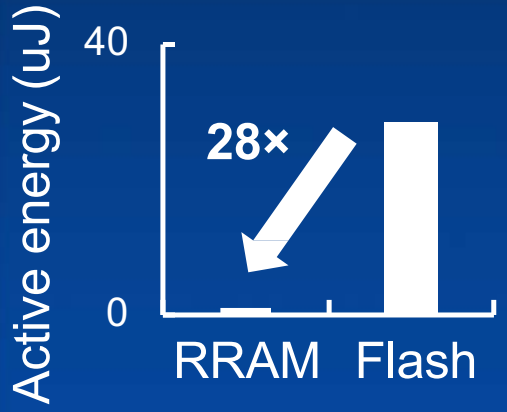
SkyWater Technology Foundry produces first wafers in a drive to match performance of cutting-edge silicon chips

IEEE SPECTRUM

By **Samuel K. Moore**

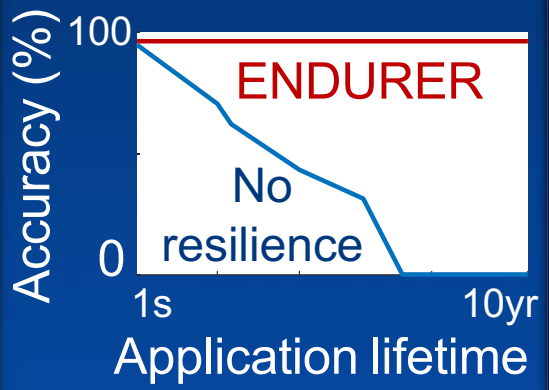
RRAM Advances

Low-energy Operation



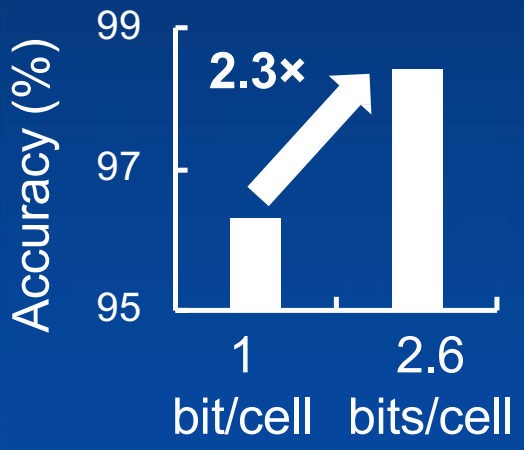
28× Lower vs. Flash

New RRAM Endurance



10-year continuous ML inference

1st RRAM System Multi-bits/cell

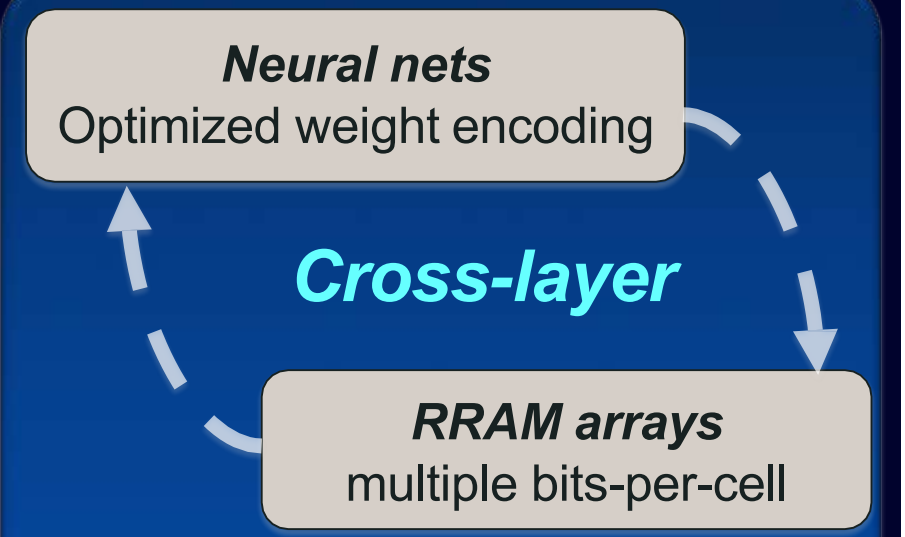


Accurate ML inference

[Wu ISSCC 19] Stanford + CEA LETI + NTU Singapore

First Multiple bits-per-cell RRAM System

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

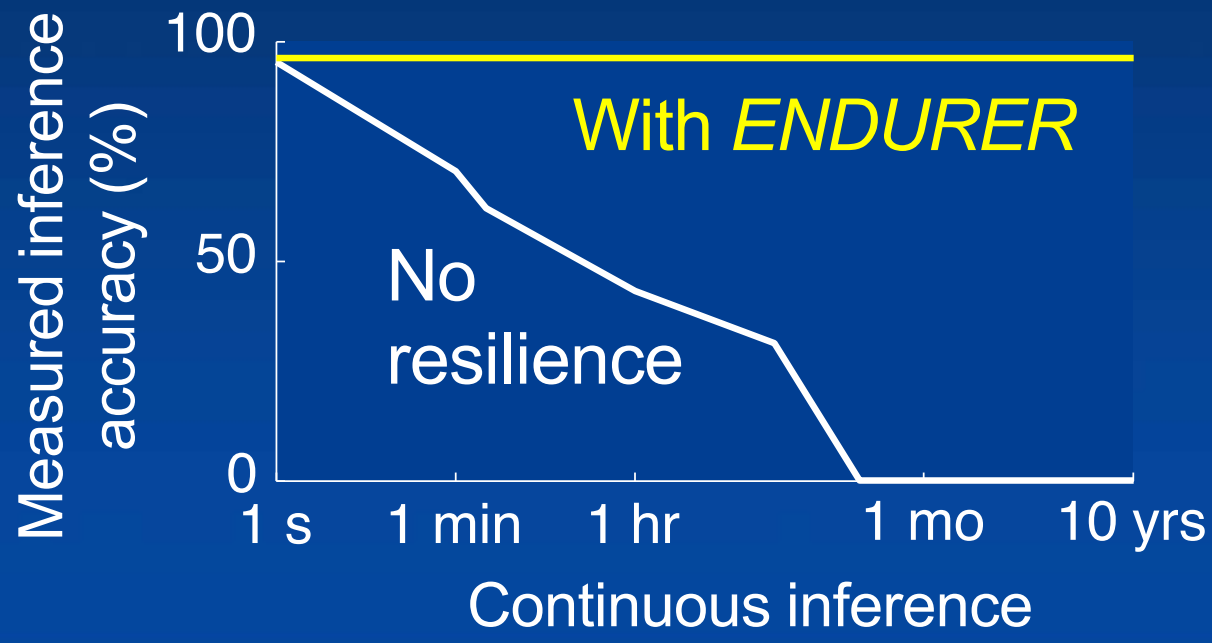


2.3× accurate inference

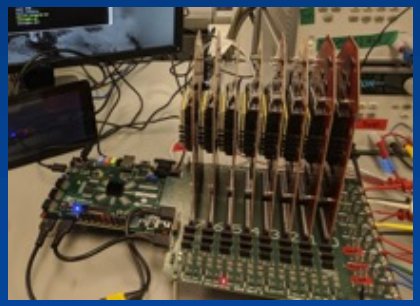
Same hardware, bigger model

ENDURER

10-year lifetime (measured)



System test setup



Non-volatile IoT microcontroller

N3XT Cross-Layer ModSim

Many chips system

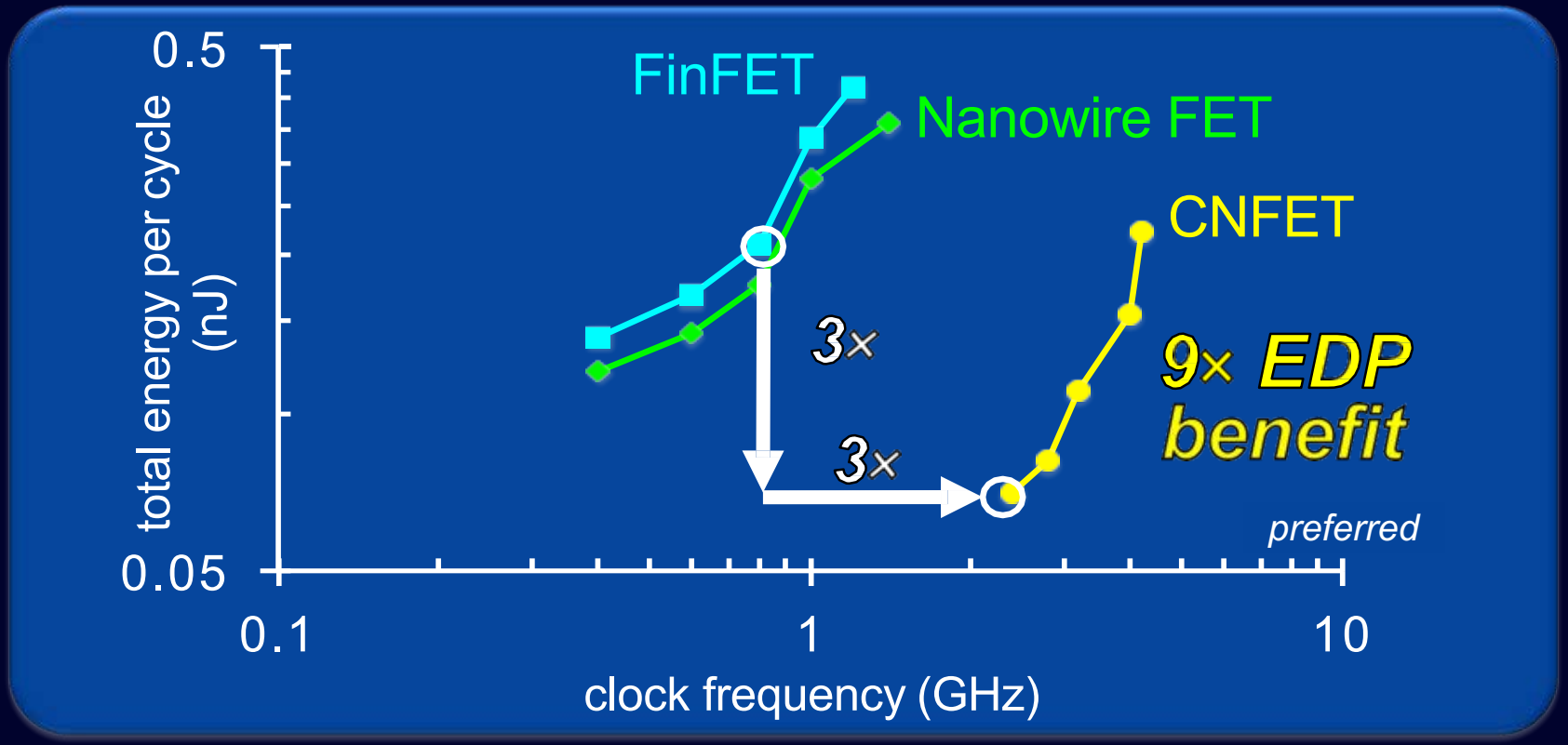
Few chips system

Architecture

Circuit

Device

OpenSPARC T2 Processor Core



[Hills IEEE TNANO 18] Stanford + IMEC + TSMC

Many FAQs Answered

1. Where do benefits come from ?
2. Wires limit performance ? Why research FETs ?
3. CNFET contact resistance ?
4. Aren't FETs good enough already ?
5. CNT variations ?

NanoSystem Design Kit

Available: nanohub.org

N3XT Cross-Layer ModSim

Many chips system

Few chips system

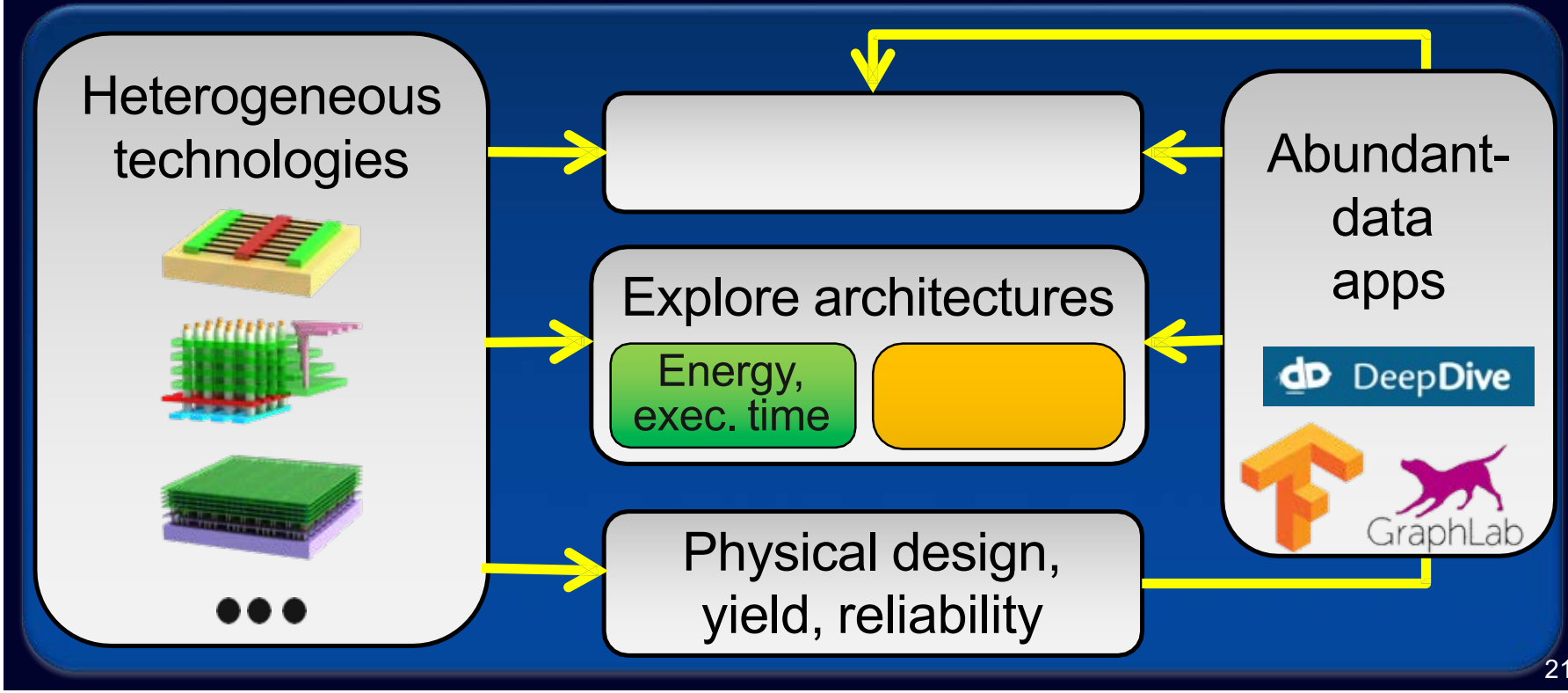
Architecture

Circuit

Device

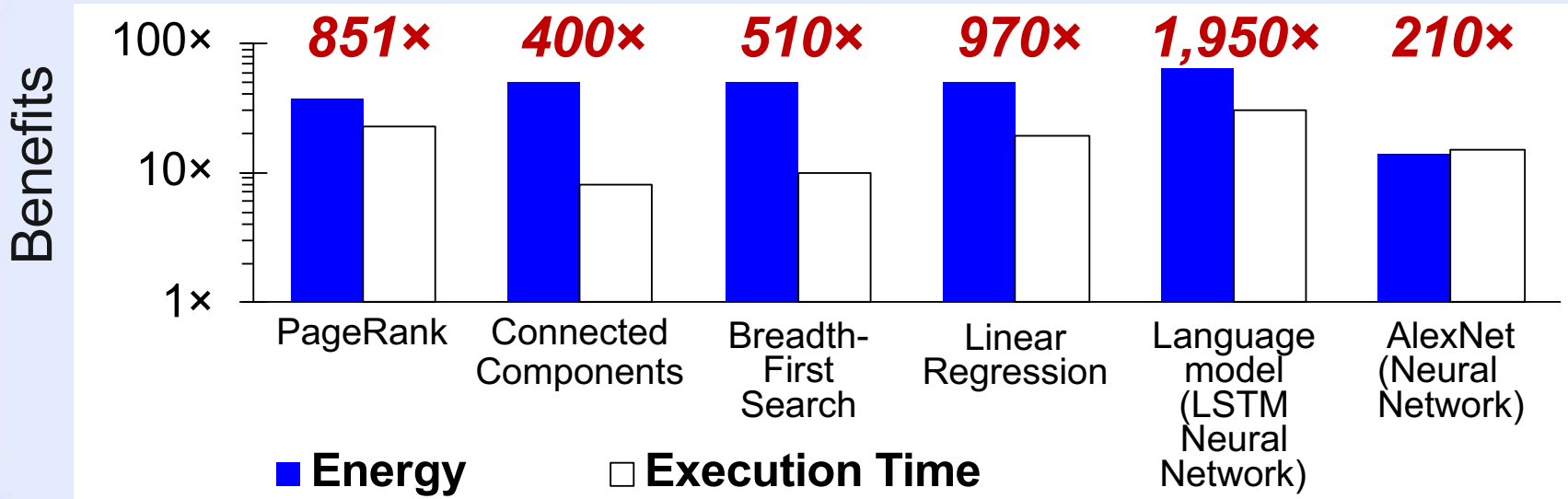
N3XT Simulation Framework

Joint technology, design & app. exploration



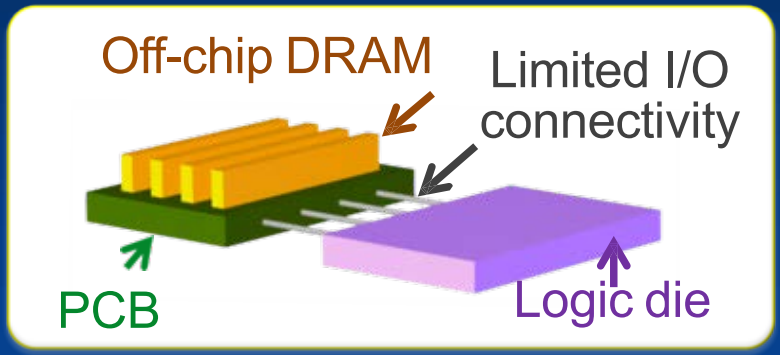
Massive Benefits
Deep Learning, Graph Analytics, …

***~1,000×* benefits, existing software**

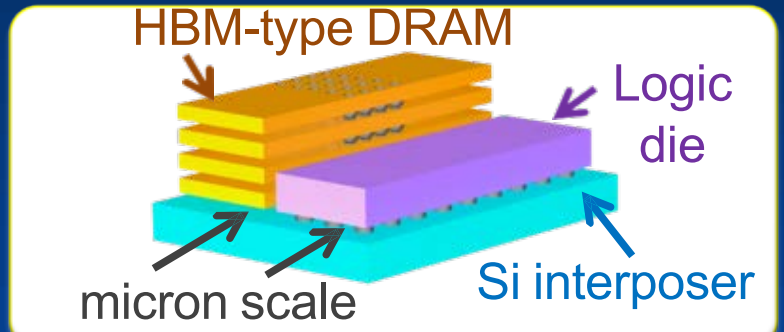


Compute + Memory Integration

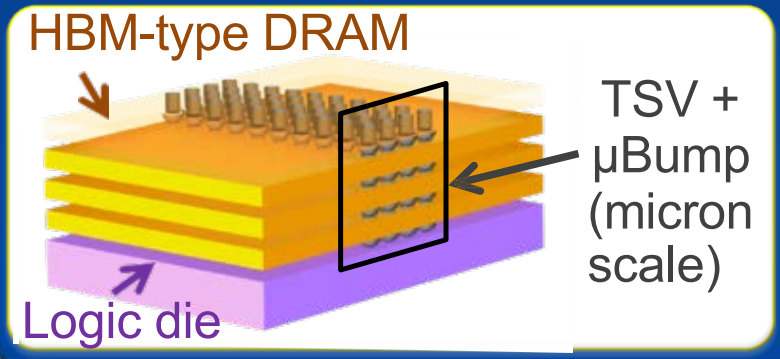
Traditional 2D system



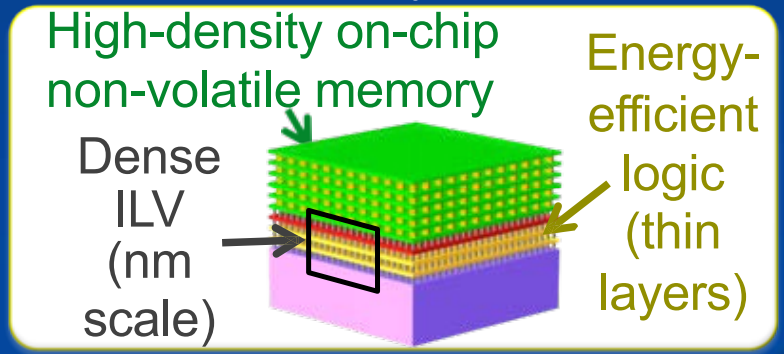
2.5D system



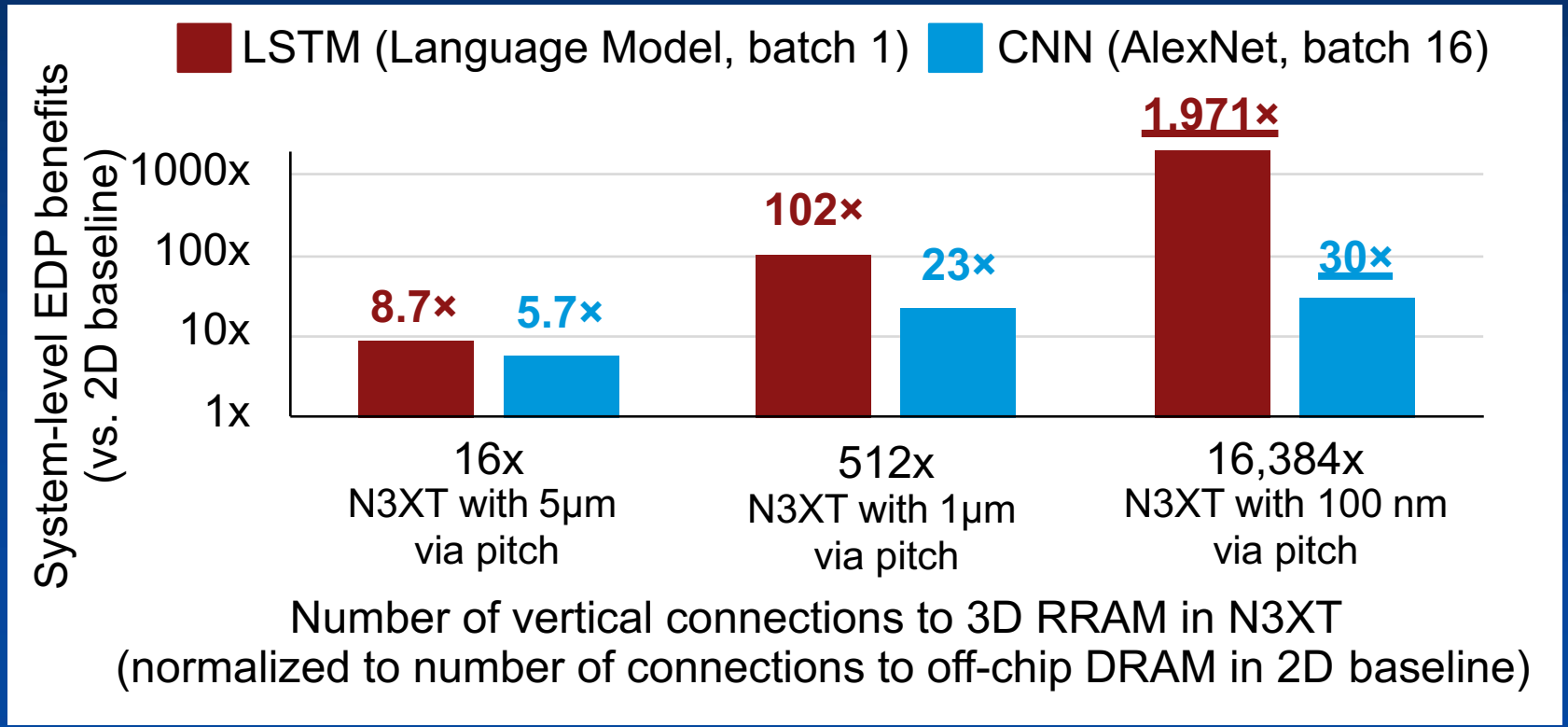
3D TSV system



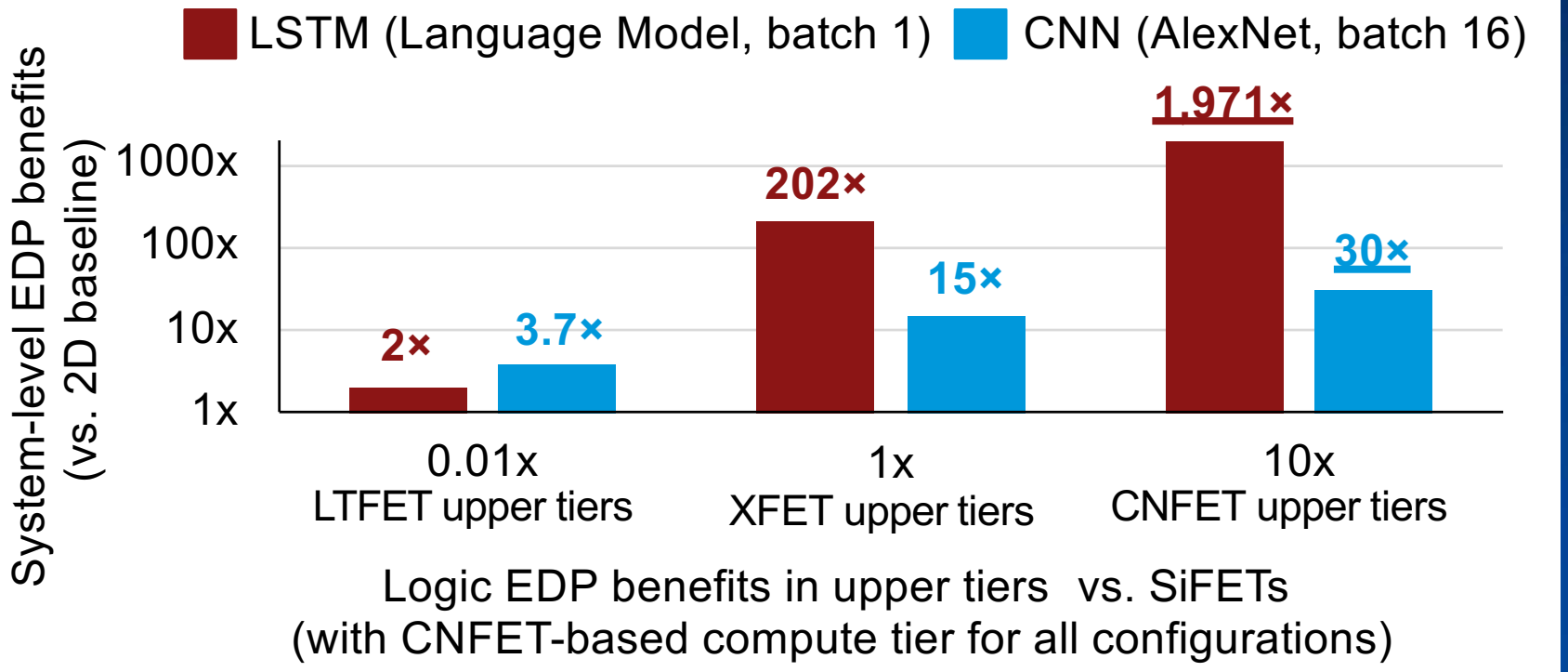
N3XT system



Compute + Memory Integration

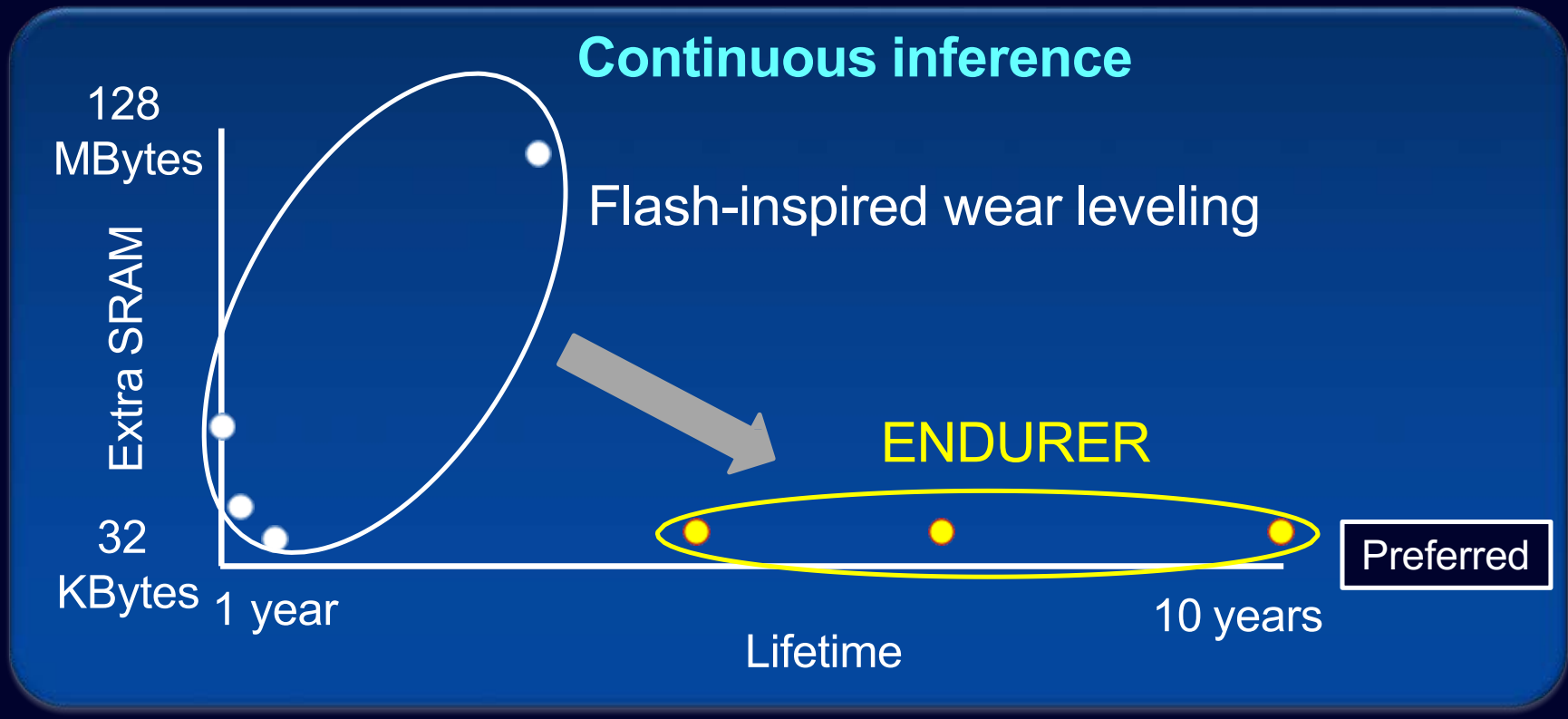


Compute + Memory Integration



Logic EDP benefits in upper tiers vs. SiFETs
(with CNFET-based compute tier for all configurations)

Resilience

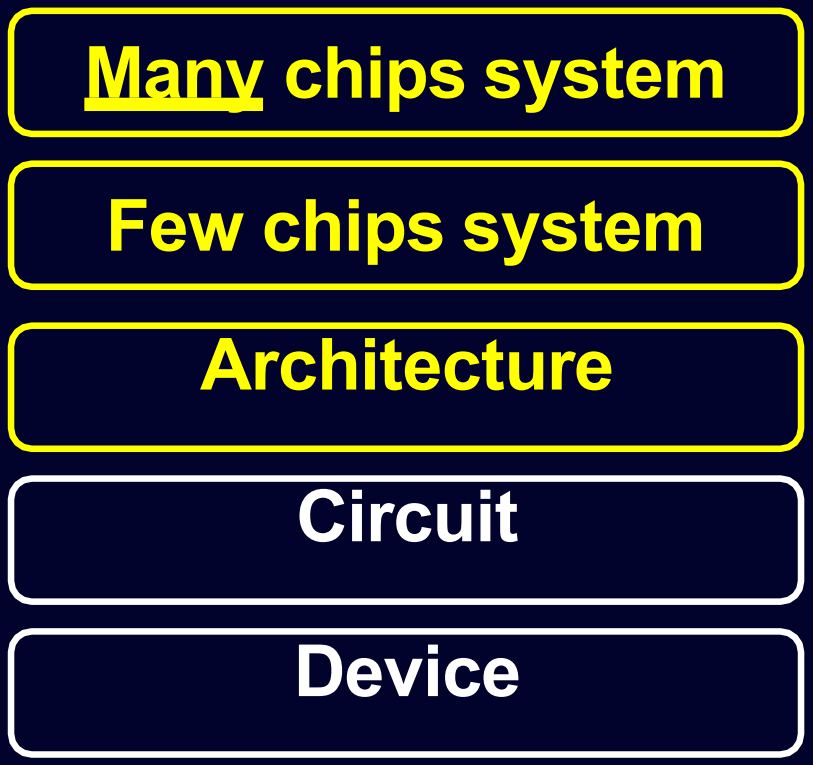


[Grossi IEEE TED 19, Stanford + CEA LETI] Machine learning accelerator + 4 GBytes RRAM

Many More (ModSim) Opportunities

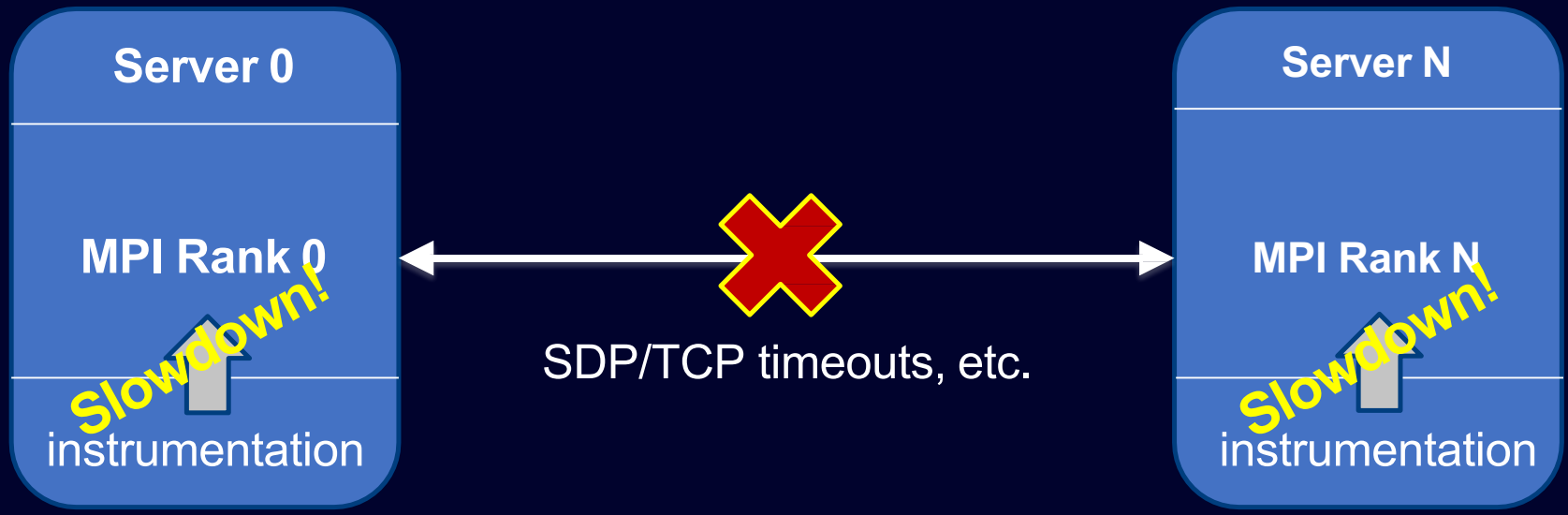
- Software mapping
- New micro-architectures & memory hierarchy
- Dense compute + thermal

N3XT Cross-Layer ModSim



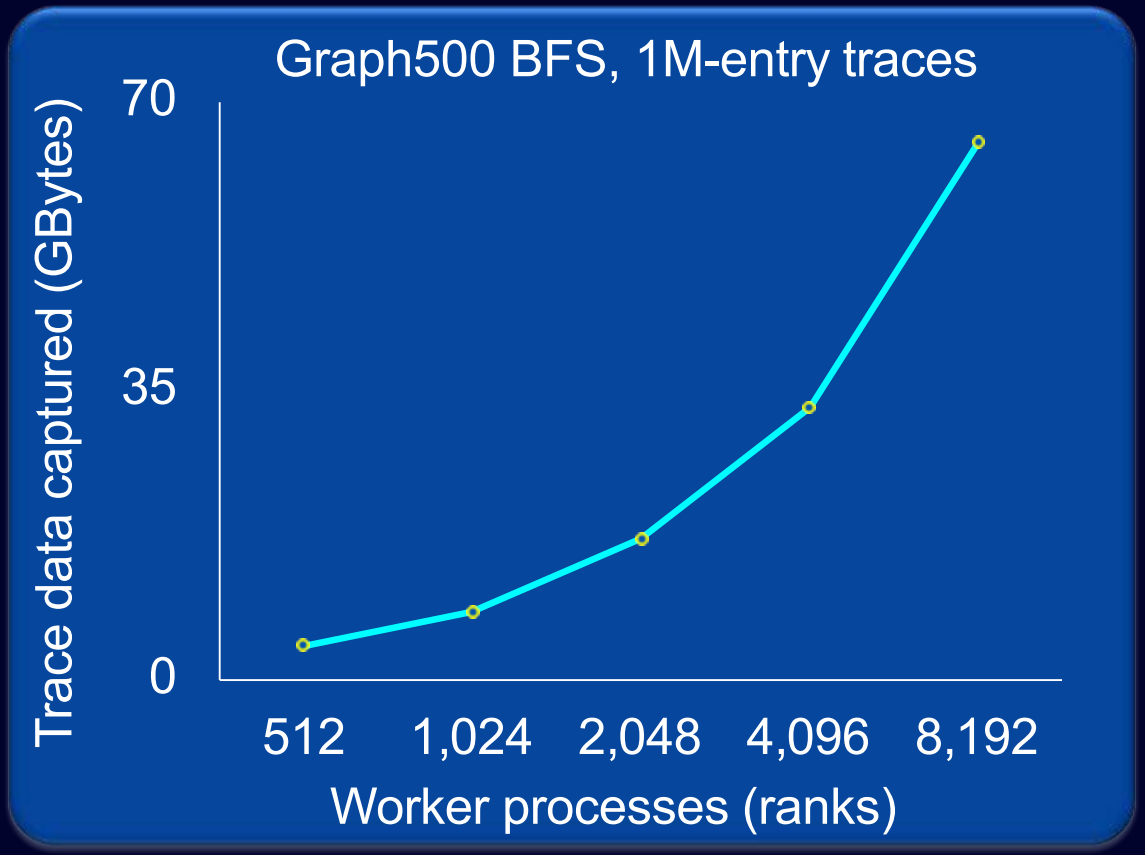
Collaborators:
LBL, ORNL

Distributed Simulation Challenges



- Instrumentation slows execution
- Remote nodes can't reply before timeout

Ongoing: Decoupled Tracing + ModSim



ORNL TITAN cluster
512 nodes,
8,192 ranks

***Instruction-level* traces**

Thanks: Students, Sponsors, Collaborators

Conclusion

- **NanoSystems today**
- **RRAM, CNFET, dense (monolithic) 3D**
  - In fabs now (from labs)
- **Many cross-layer opportunities**