California-Pacific-Northwest AI Hardware Hub

Microelectronics Commons

Berkeley UNIVERSITY OF CALIFORNIA · Stanford University · UCDAVIS UNIVERSITY OF CALIFORNIA · UNIVERSITY of WASHINGTON · UNIVERSITY of HAWAI'I MĀNOA · Oregon State University · Western Digital

AI Hardware: 2013 – 2023

Accelerators *vs. general-purpose*

Large one-time benefits

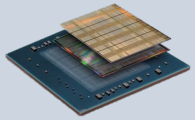
- Arithmetic (e.g., BF16, FP8)
- SIMD
- Systolic arrays
- Custom dataflows

Caution: Excessive customization risky

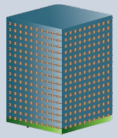
Technology innovations *essential*

Not just advanced nodes

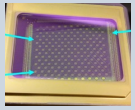
2.5D, 3D



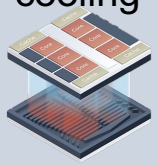
HBM



Optical circuit switch



Microfluidic cooling

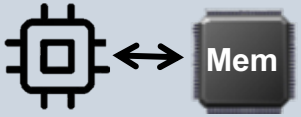


RRAM, MRAM for edge

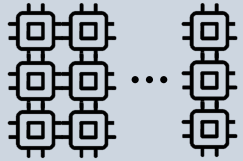
Infineon and TSMC to introduce RRAM

End-to-end systems *crucial*

≠ peak TOPS/W



Off-chip accesses



Many-chip partitioning

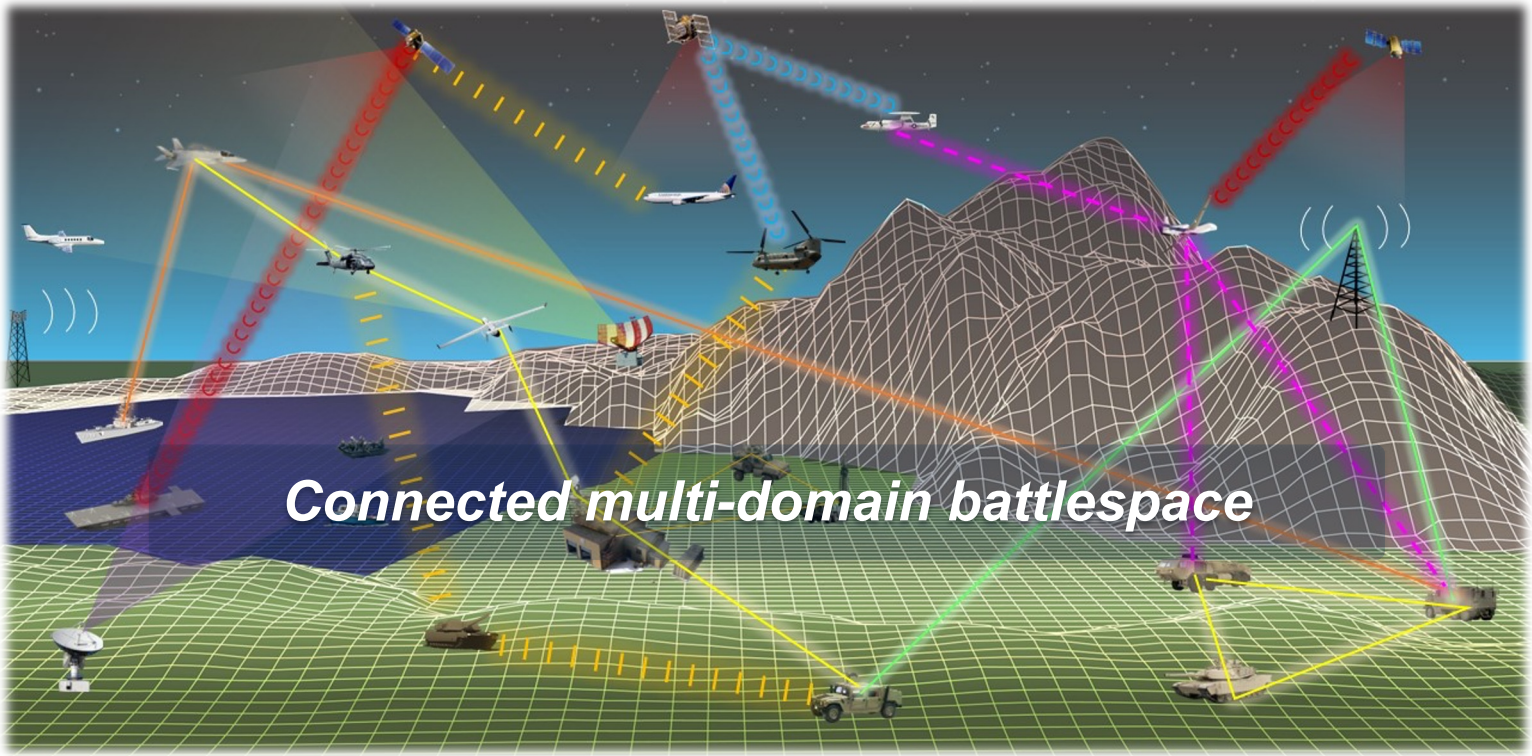
Resilience: silent data corruption

Images from Corintis, Google, Infineon, TSMC, Xilinx, Xperi

BF: Brain Float, FP: Floating Point, SIMD: Single Instruction/Multiple Data, CGRA: Coarse-Grained Reconfigurable Arch, HBM: High-Bandwidth Memory, RRAM: Resistive RAM

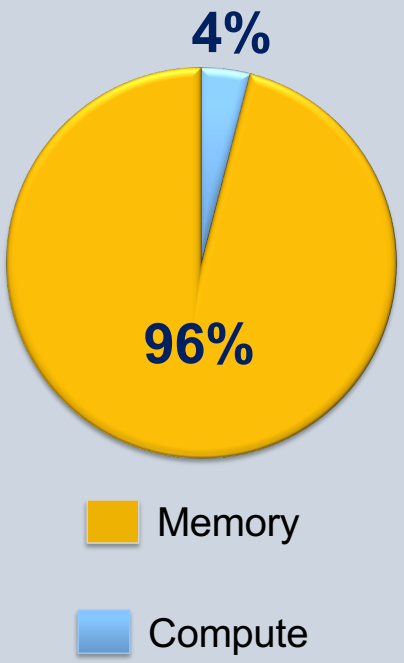
DoD AI: BIG Challenges

Energy & power, real-time, accuracy, continuous learning, harsh (e.g., space)



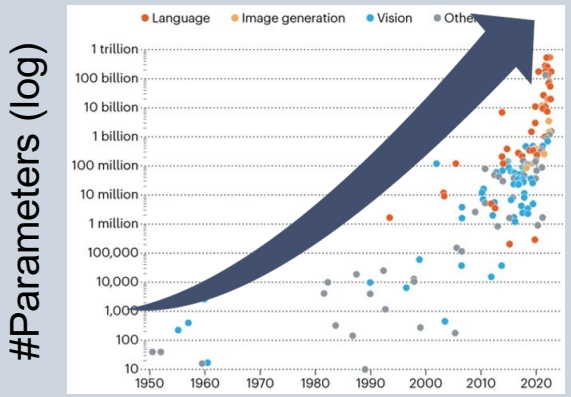
More Challenges Moving Forward

Memory wall



Miniaturization wall

Neural net size explosion



Editorial, Nature, 2023.

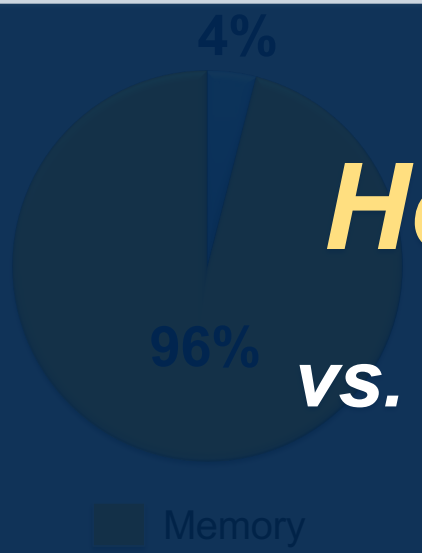
More Challenges Moving Forward

Northwest AI HUB

Memory wall

Miniaturization wall

Neural net size explosion

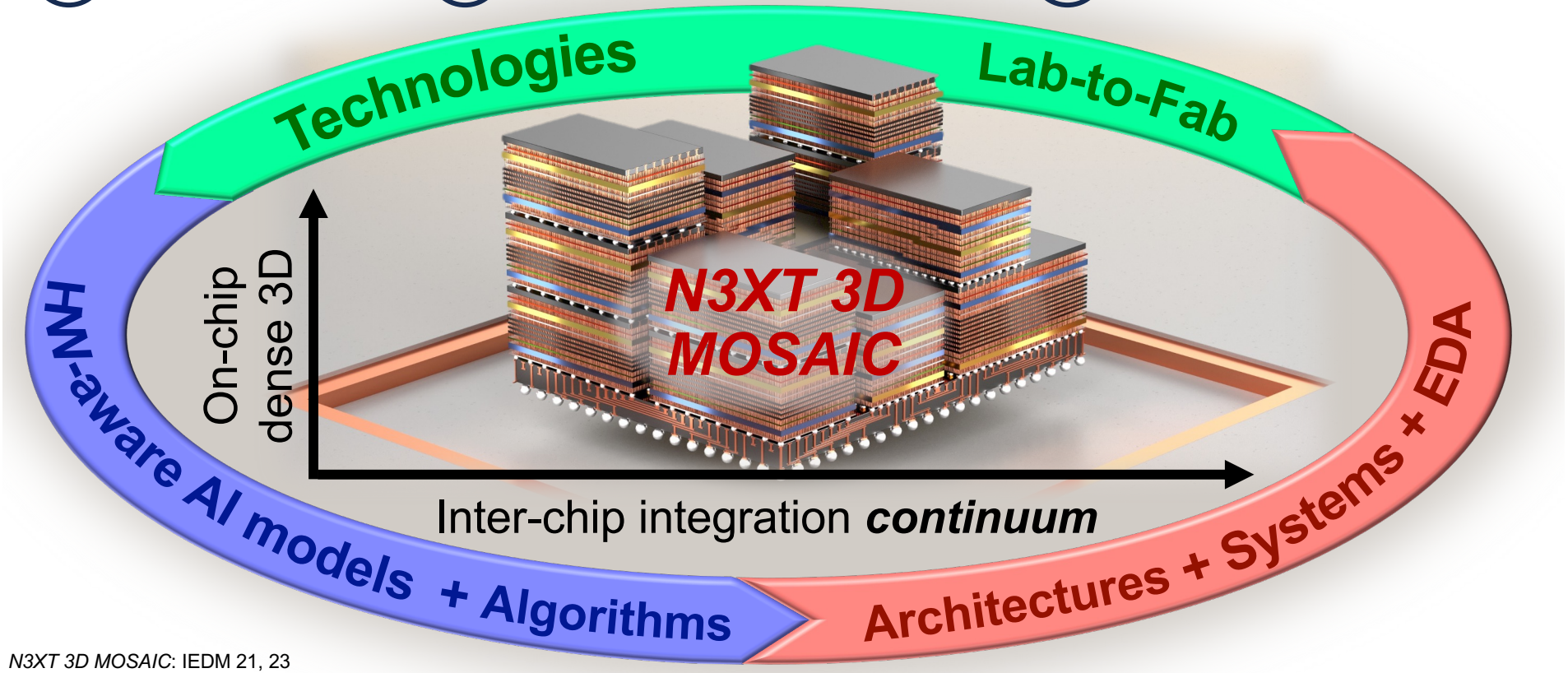


How Next 1,000×?

vs. today's best AI hardware

Our Approach

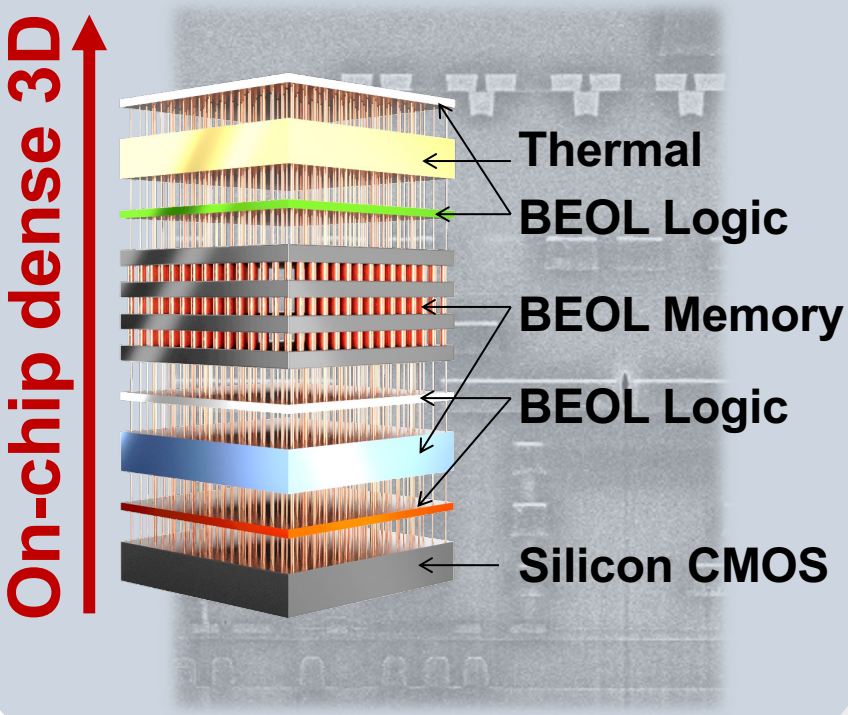
1 Application-driven 2 Diverse & specialized functions 3 Multi-chip 3D NanoSystems



N3XT 3D MOSAIC: IEDM 21, 23

N3XT 3D: Dense 3D & CMOS + X

CMOS + X: Many X's



Lab-to-Fab

ANALOG DEVICES

skywater

Many firsts in industry fabs

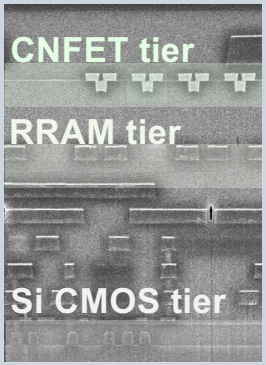
- Carbon nanotube FETs (CNFETs)
- Dense monolithic 3D:
  CNFET + RRAM + Si CMOS
- U.S. foundry Resistive RAM (RRAM)

BEOL: Back End of Line

CMOS + X: Many X's

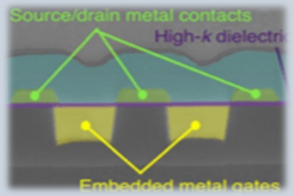
Northwest AI HUB

Monolithic 3D



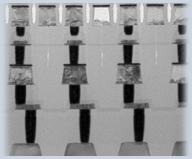
Stanford + MIT + Skywater 23

CNFETs



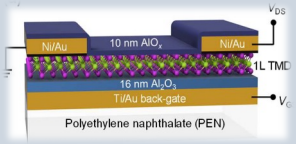
MIT + Skywater 20

Resistive RAM



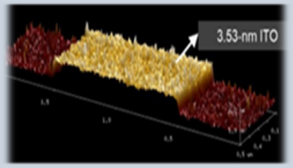
Stanford + SkyWater 21

2D FETs



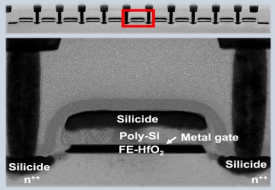
Stanford 23

Oxide FETs



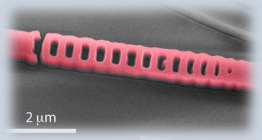
Stanford 23

Ferroelectric FETs



Stanford + GF 23

Inverse-Designed Photonics

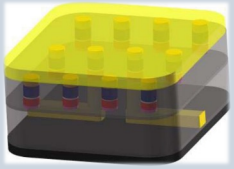


Stanford 22

Codesigned Photonics

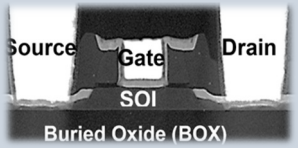
UC Davis 23

Magneto-resistive RAM



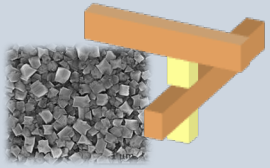
WD 22

NCFETs



Berkeley + MITLL 22

3D Thermal Scaffolding



Stanford 23

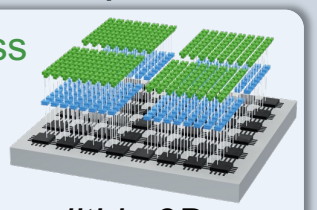
N3XT 3D MOSAIC

Architectures + Systems + EDA

Architectures

Single-chip

CNFET access
RRAM banks
Si compute

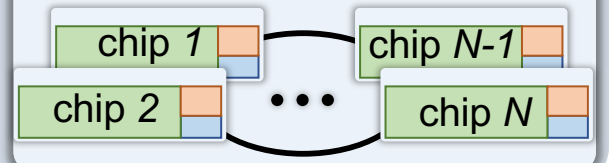


Foundry monolithic 3D:
large EDP benefits

DATE 23, VLSI 23

Multi-chip Illusion

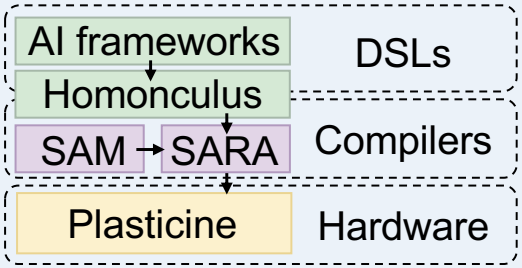
Enough 3D mem. + ***Quick***
ON/OFF = ***Special*** mapping



Nature Electronics 21

Systems

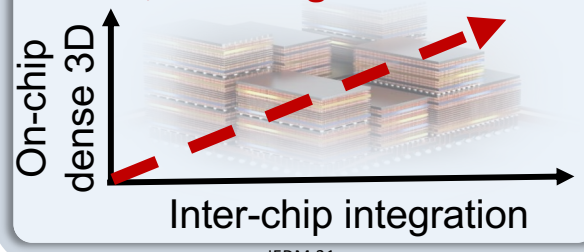
Full-stack dataflow



ASPLOS 23, ISCA 17, 21

Multi-chip Illusion scaleup

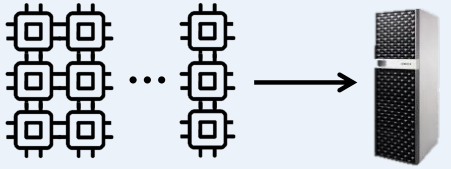
1,000× larger AI models



IEDM 21

EDA

Multi-chip AI systems



partition, map, emulate in mins.
+ 3D thermal, power, noise

Resilience

DATE 08

CNFET: Carbon nanotube FET, RRAM: Resistive RAM, EDP: Energy Delay Product, DSL: Domain-Specific Language

HW-Aware AI Models & Algorithms

New AI algorithms

FlashAttention

I/O-aware Transformer training

- Fast (3×), less mem. (10×)
- Exact attention
- Longer sequences

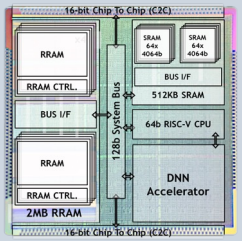
Neural net training theory

Globally optimal training
Quantized activations
Convex optimization

ICLR 23

RRAM-aware training

CHIMERA RRAM Edge AI

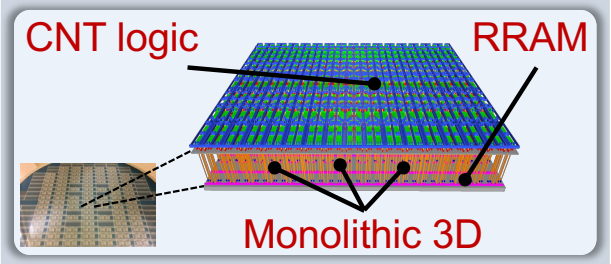


VLSI 21

	vs. SGD
RRAM weight update steps	101× fewer
EDP	340× better
Lifetime (20 images/min)	10 years vs. 2 weeks

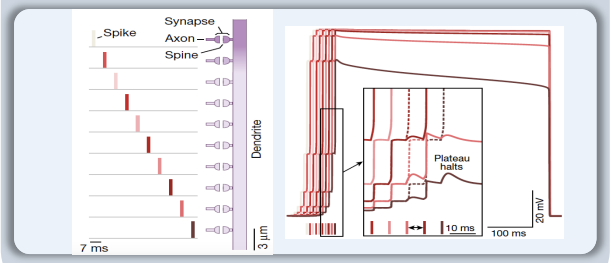
Beyond neural nets

Hyperdimensional



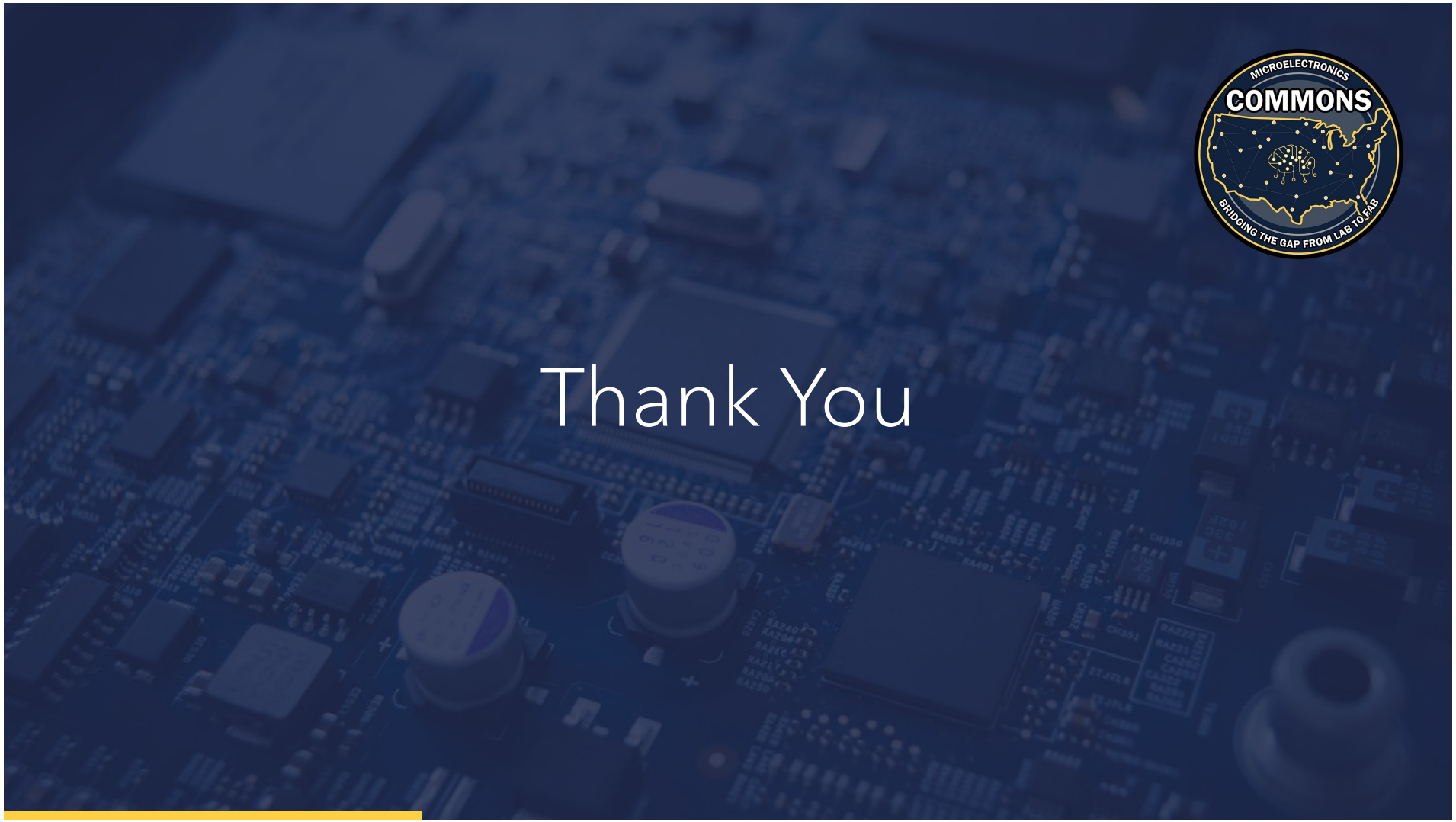
ISSCC 18

Dendritic



Nature 22

RRAM: Resistive RAM, EDP: Energy Delay Product,, SGD: Stochastic Gradient Descent, CNT: Carbon nanotube



Thank You