N3XT 3D MOSAIC:

3D Thermal Scaffolding, Multi-Chip Illusion

Subhasish Mitra

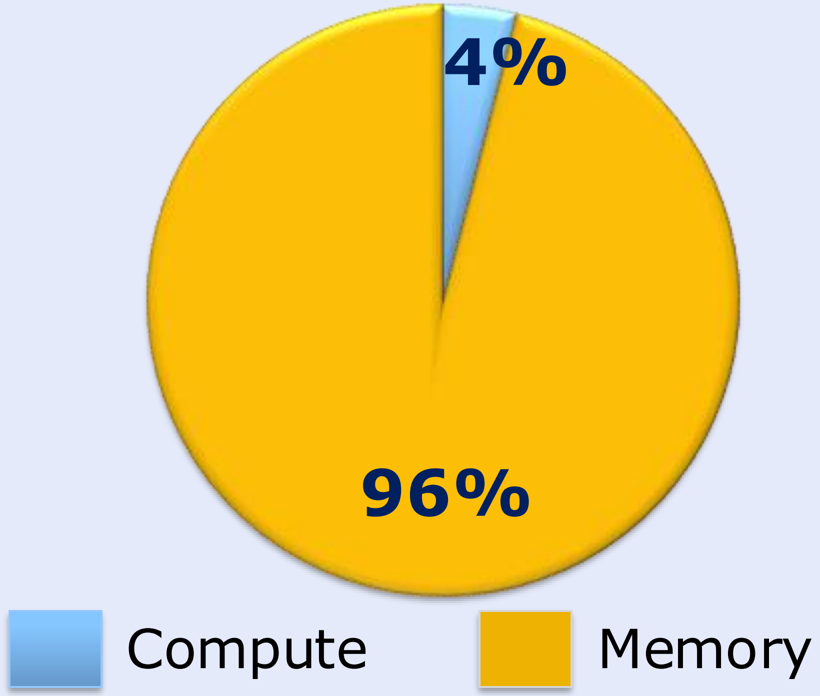
Department of EE and Department of CS

Stanford University

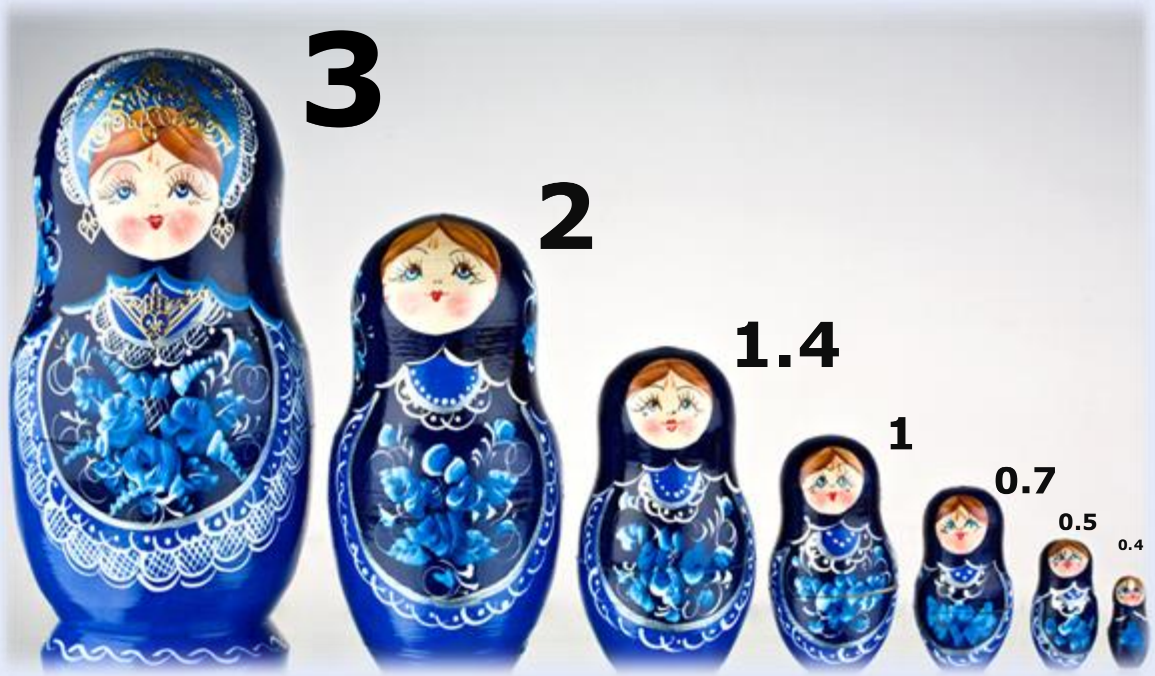
Abundant-Data Computing: e.g., AI/ML

Deadly combination!

Memory Wall

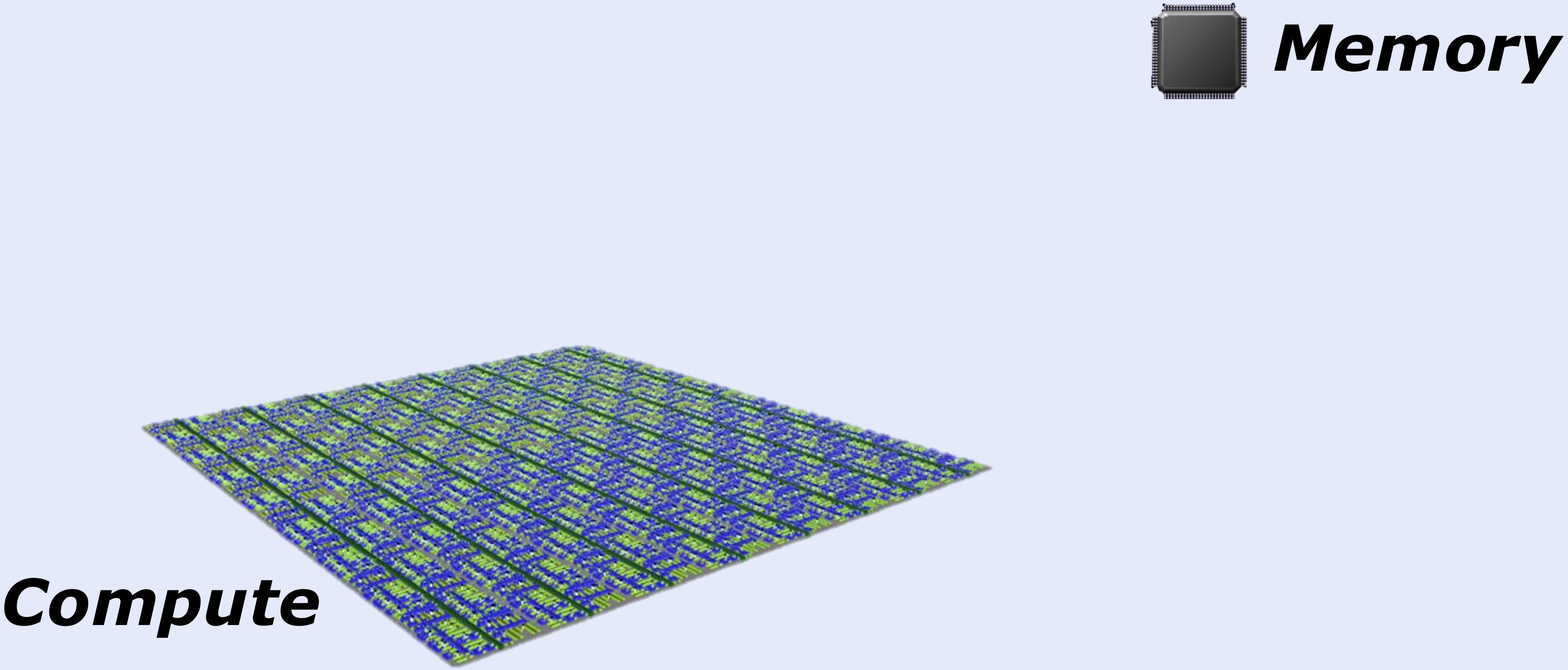


Miniaturization Wall



AI/ML = Artificial Intelligence/Machine Learning

Computing Today

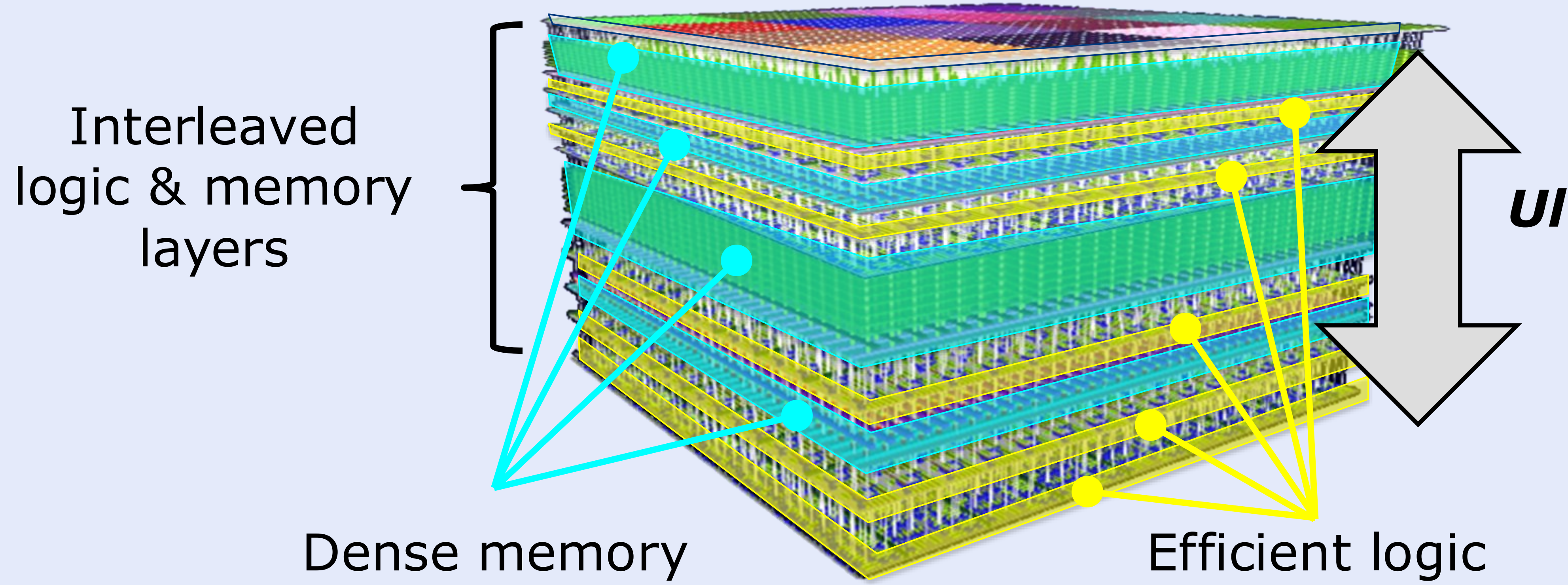


Memory

Compute

N3XT 3D: Computation immersed in Memory

Nano-Engineered Computing Systems Technology

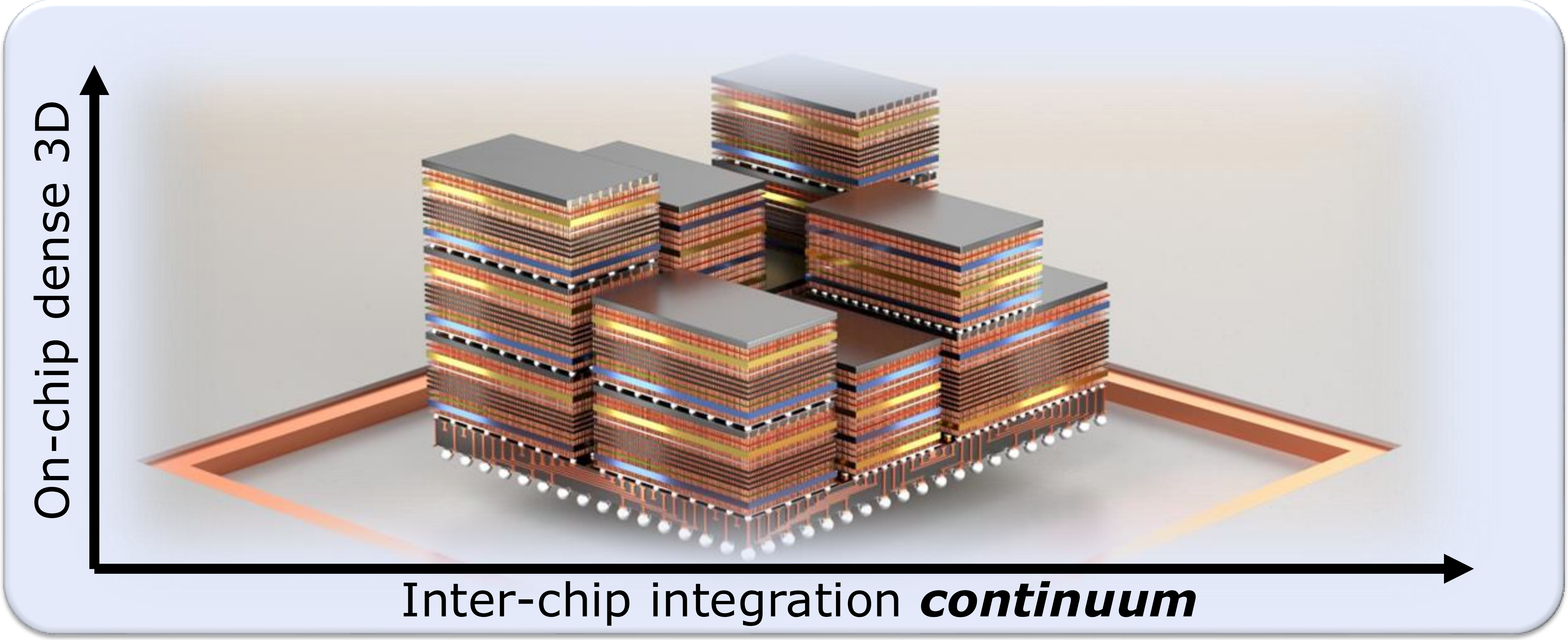


Large Energy Delay Product (EDP) benefits

Collaborators: Prof. H.-S.P. Wong (Stanford) + others

N3XT 3D MOSAIC

MOnolithic / Stacked / Assembled IC



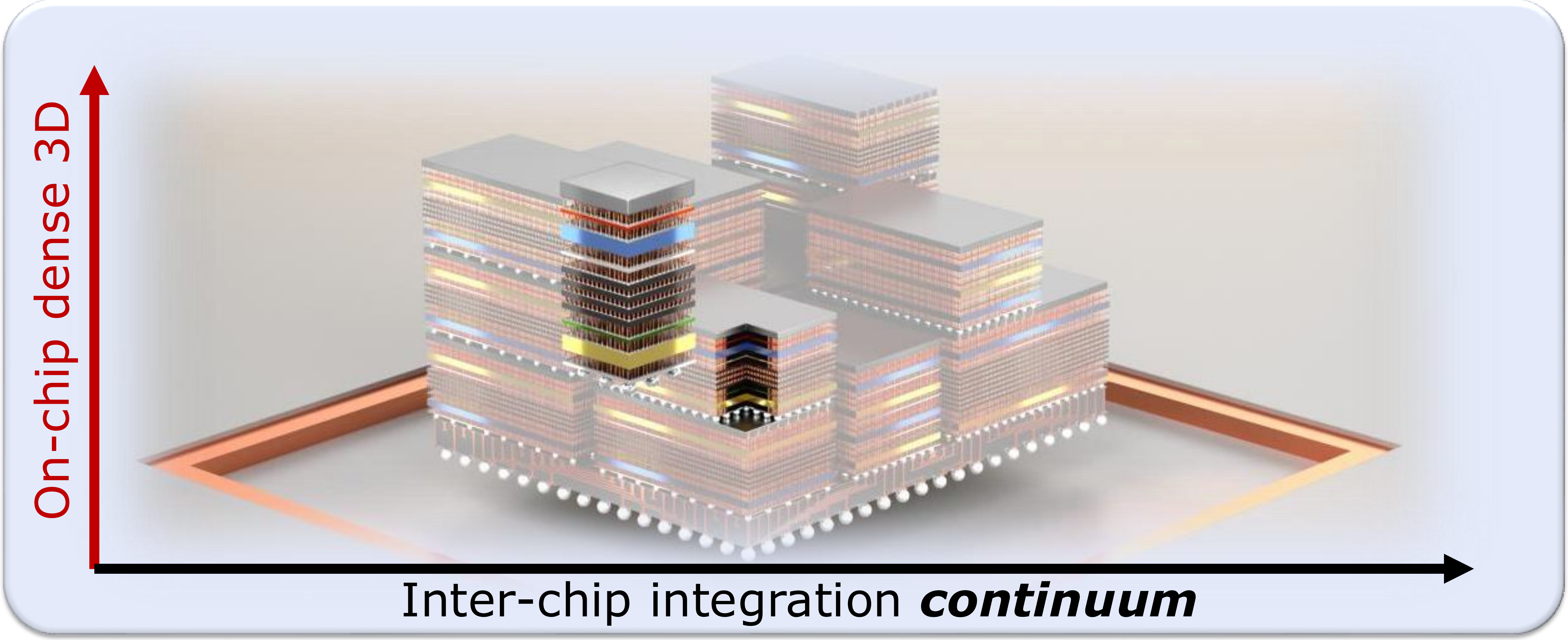
On-chip dense 3D

Inter-chip integration ***continuum***

Collaborators: Prof. H.-S.P. Wong (Stanford) + others

N3XT 3D MOSAIC

MOnolithic / Stacked / Assembled IC

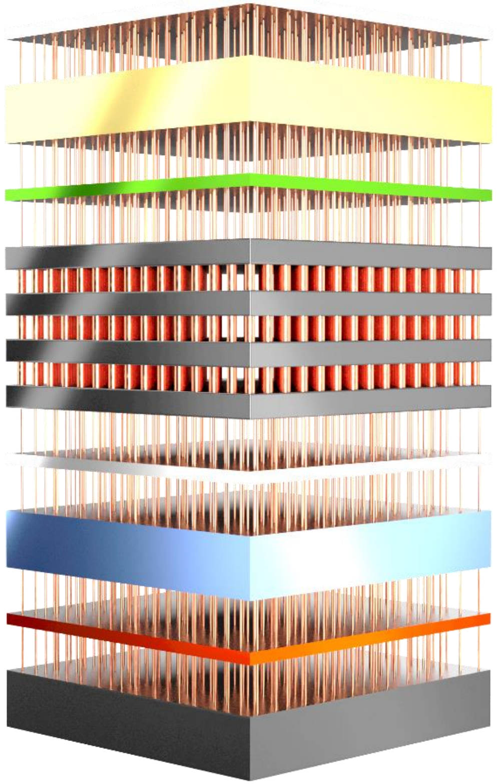


On-chip dense 3D

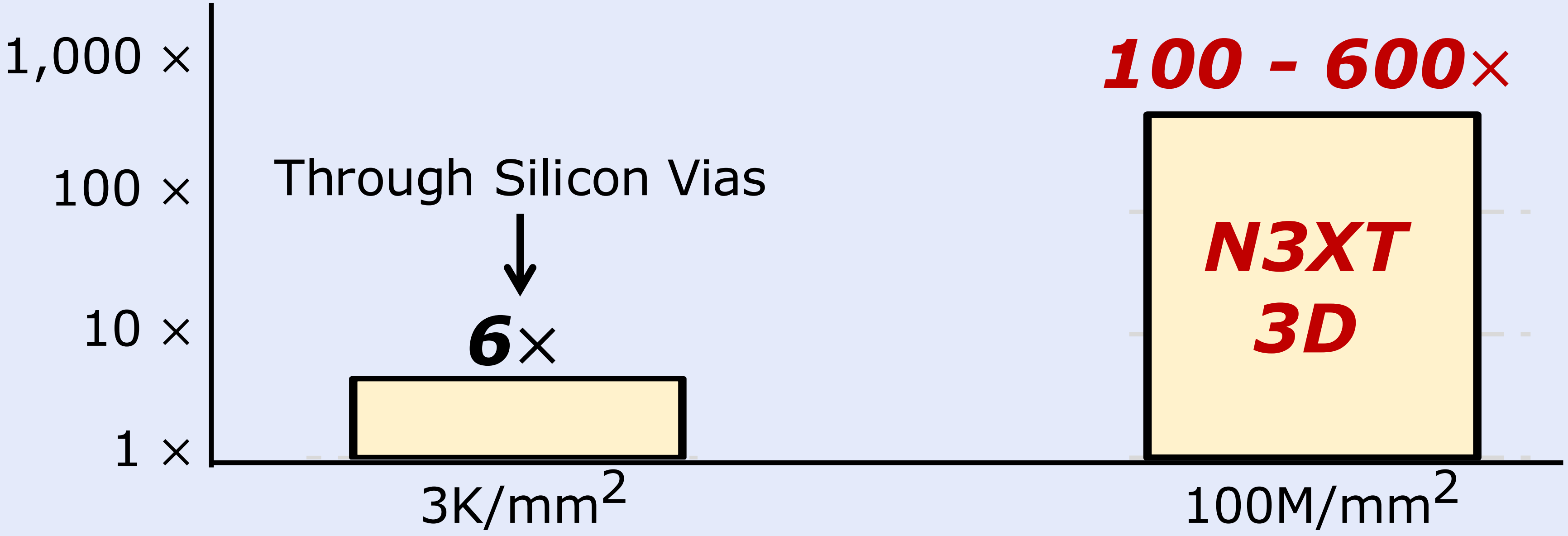
Inter-chip integration ***continuum***

Dense 3D Connections: Large Benefits

N3XT 3D Chip



Multiple logic **&** memory layers in 3D


Energy Delay Product benefits vs. today's packaging

Density of 3D connections

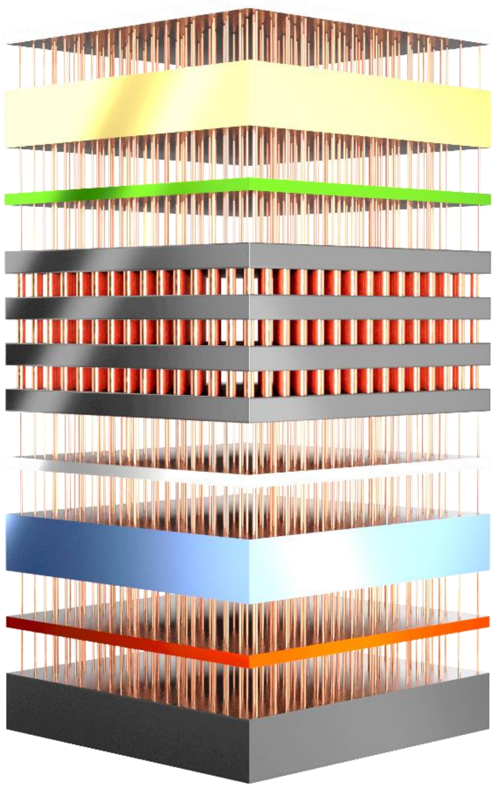
Monolithic 3D: *only* way today

Apps: Crypto, Graph, Genomics, Sparse matrix, Neural nets.

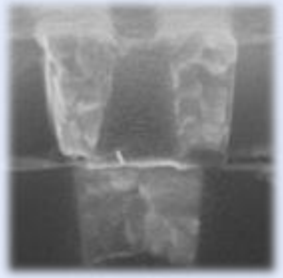
N3XT 3D: Many Technologies

BEOL-compatible: ≤ 400°C fabrication

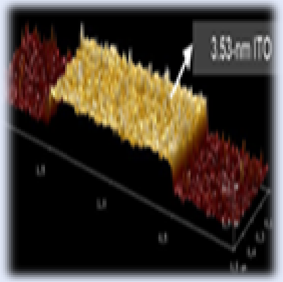
N3XT 3D Chip:



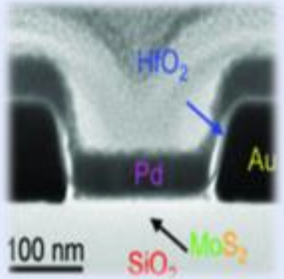
Carbon nanotubes



Oxide FETs

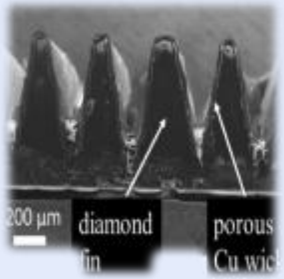


2D materials



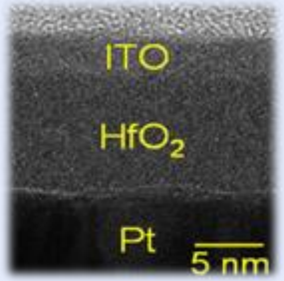
Low-Temp. Si

2-Phase Cooling

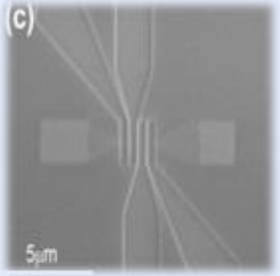


Resistive RAM

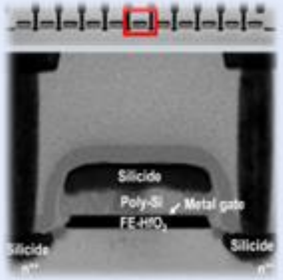
Hybrid Gain Cell



MRAM



Ferro-electric



3D Thermal Scaffolding

BEOL = Back-End-of-Line. CNFETs = Carbon Nanotube Field-Effect Transistors. MRAM = Magneto-resistive RAM.

Many Activities On-going

Lab-to-fab

Industry fabs: many firsts

ANALOG DEVICES skywater

1. Carbon nanotube FETs (CNFETs)

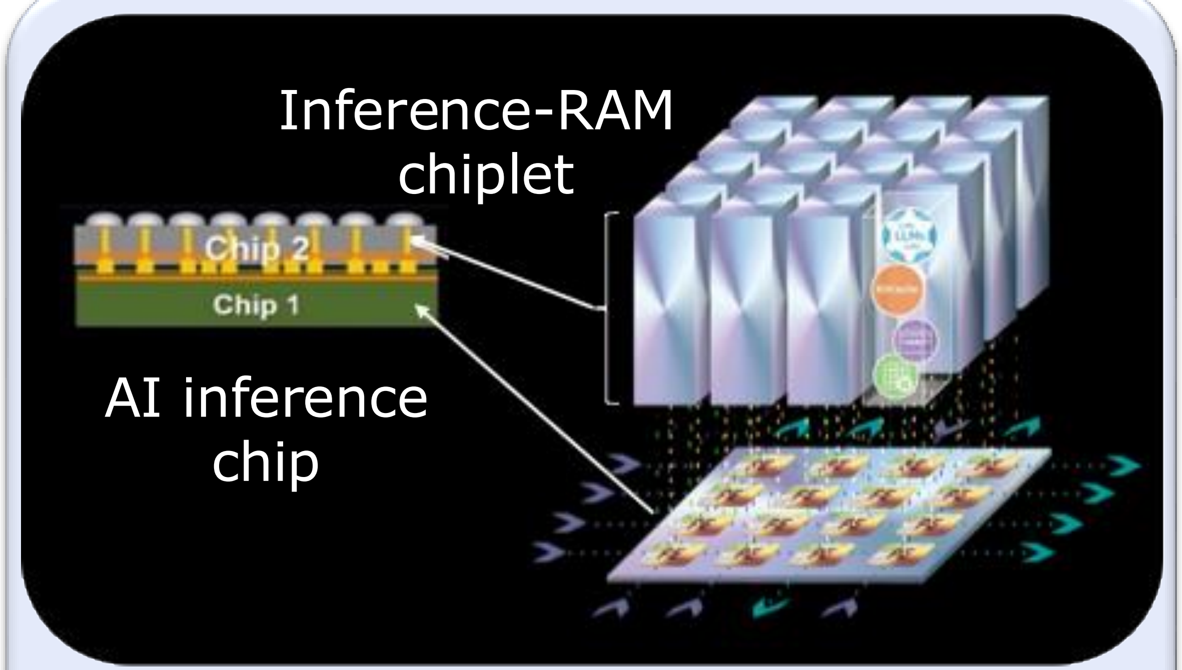
2. U.S. foundry Resistive RAM (RRAM)

3. Ultra-dense monolithic 3D:

   CNFET+ RRAM + silicon CMOS

4. Product development: e.g., ANALOG DEVICES

EMD collaboration



Inference-RAM chiplet

AI inference chip

Inference-RAM:

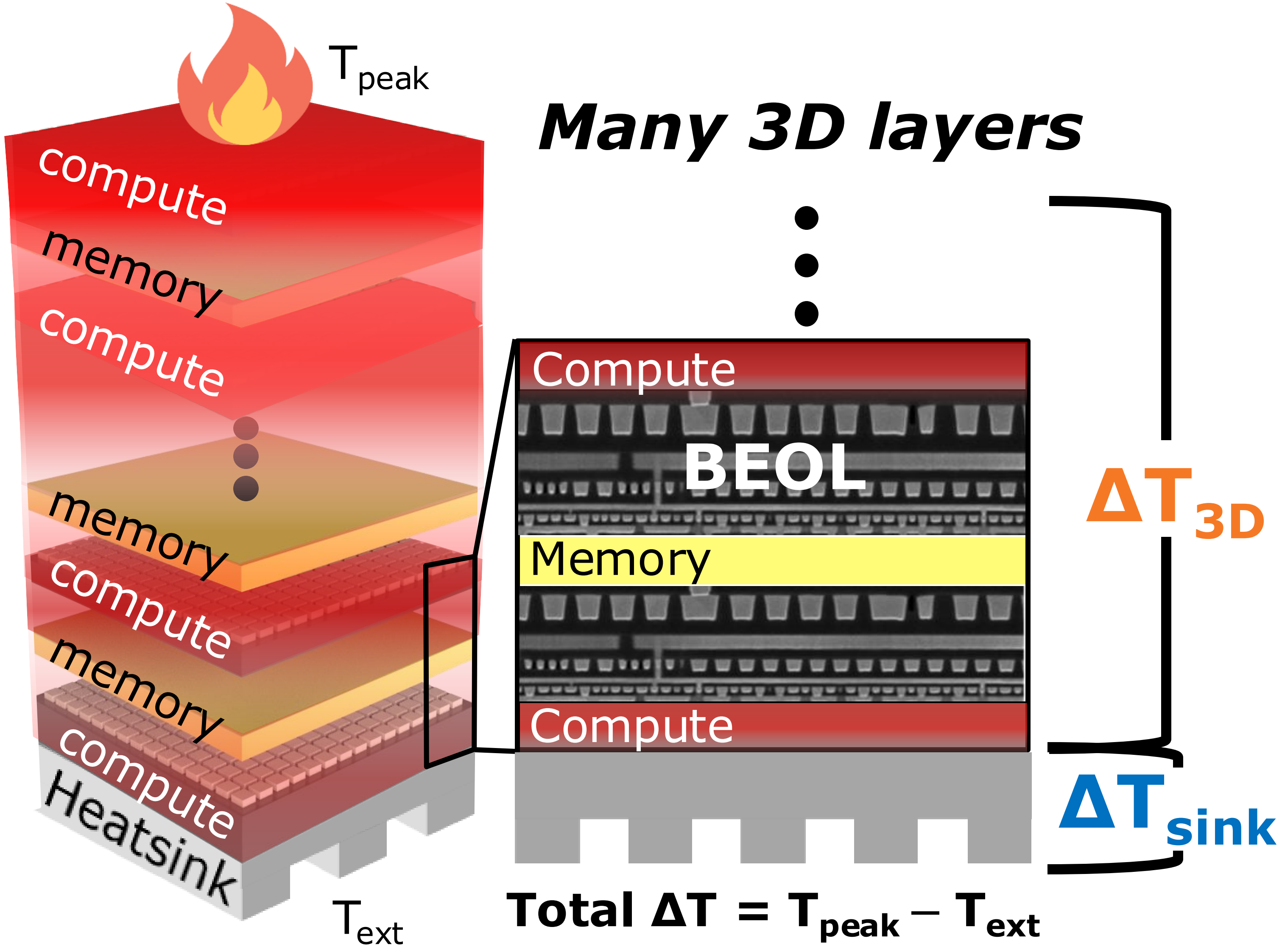
3D, dense, quick & low-energy read, write ability sufficient for KV cache

Courtesy: S. Mukund (EMD Electronics)

•••

N3XT 3D Thermal

Many 3D layers

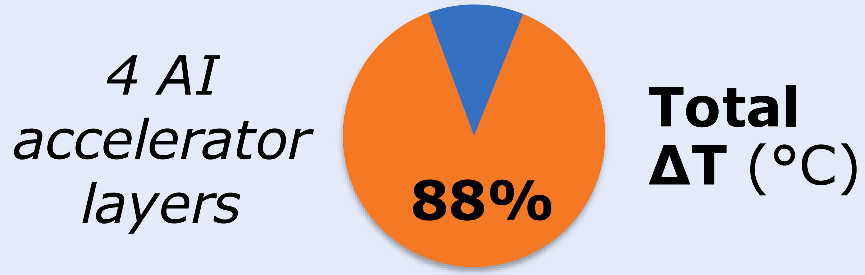


Total $\Delta T = T_{peak} - T_{ext}$

*Today's cooling **inadequate***
even with advanced heatsinks

AI accelerator layers	3	12
Area overhead	5%	78%

ΔT_{3D} dominates



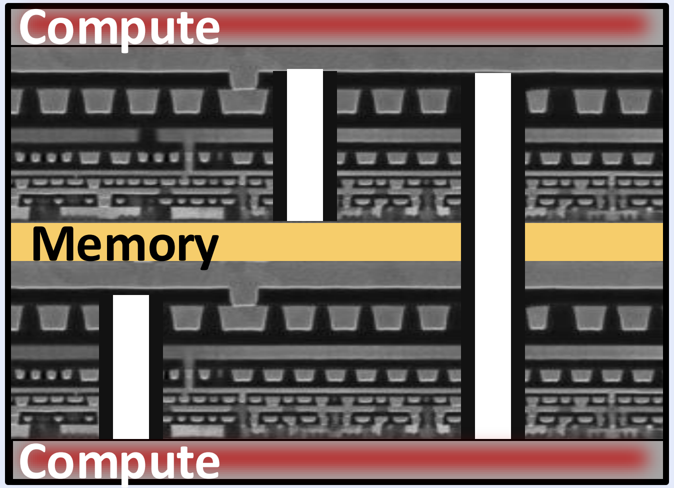
Thermal vias/power delivery network, floorplanning, scheduling, …

J. Cong et al., *Proc. Int. Conf. Comput.-Aided Design.*, 2004. H. Wei et al., *IEDM* 2012. S. K. Samal et al., *DAC* 2014. J. Li et al., *ACM Trans. Embedd. Comput. Syst.* 2013.

[Rich DAC 23] Copper boiling heatsink = 10^6 W/m^2/K [C. Zhang Adv. Funct. Mater. 18]

3D Thermal Scaffolding

Today's
metal vias



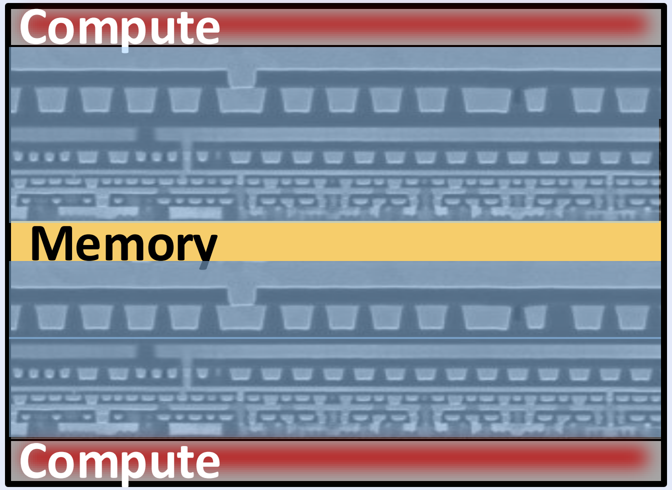
☹ **Area overhead**

☺ High vertical TC

TC =
Thermal conductivity

+

Today's
Inter Layer Dielectric



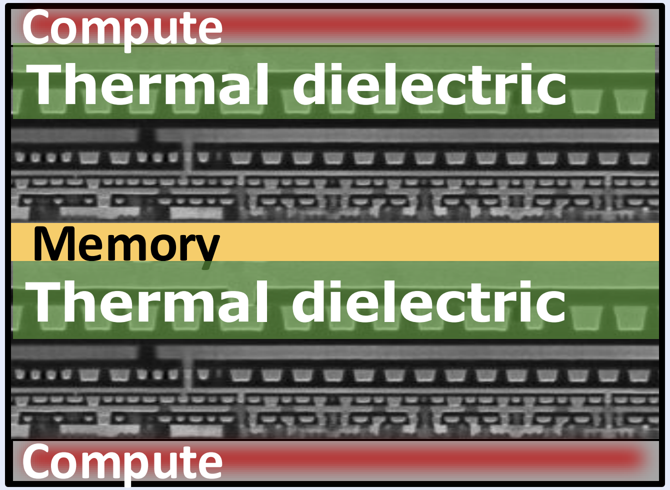
☹ **Low TC**

☺ Ultra-low κ

κ =
Dielectric constant

+

**New Thermal
Inter Layer Dielectric**



☺ **High lateral TC**

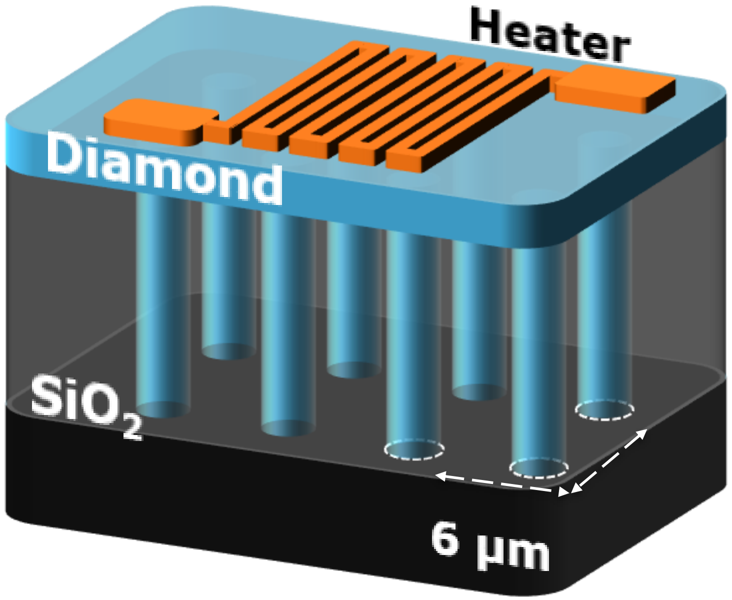
😐 Moderately low κ

Selectively placed:
co-placement algorithms

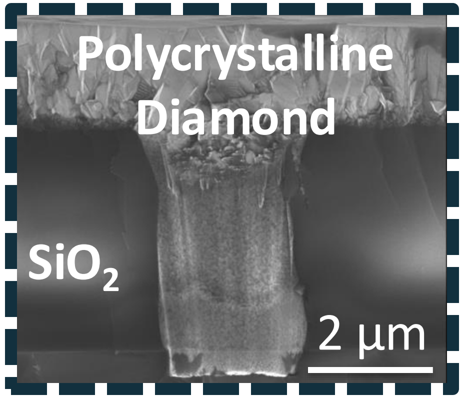
[Rich DAC 23]

Polycrystalline Diamond Thermal Dielectric

Hardware test structure



BEOL-compatible:
≤ 400°C fab



	TC	κ
Today’s dielectric	**0.2**	**2**
Polycrystalline diamond	**105**	**4**

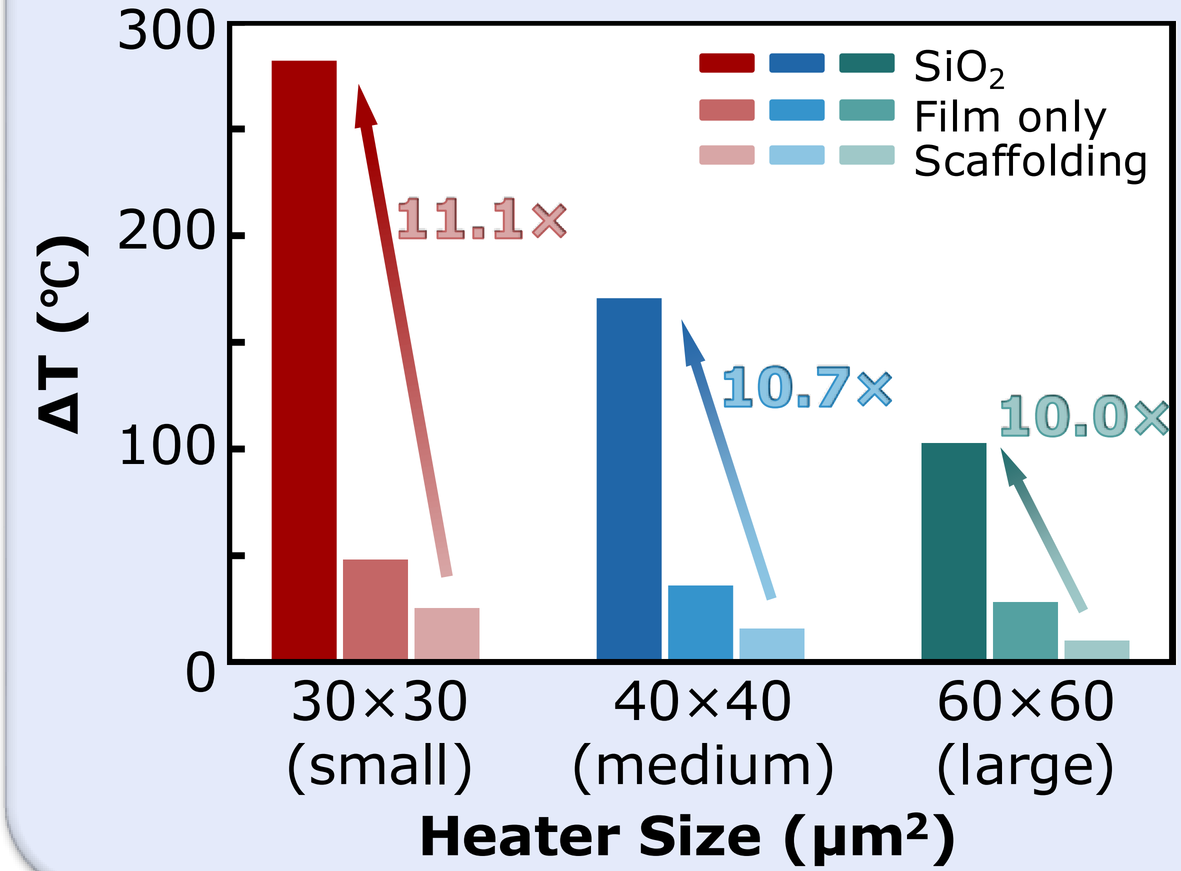
500× better

TC =
Thermal conductivity

κ =
Dielectric constant

Large benefits

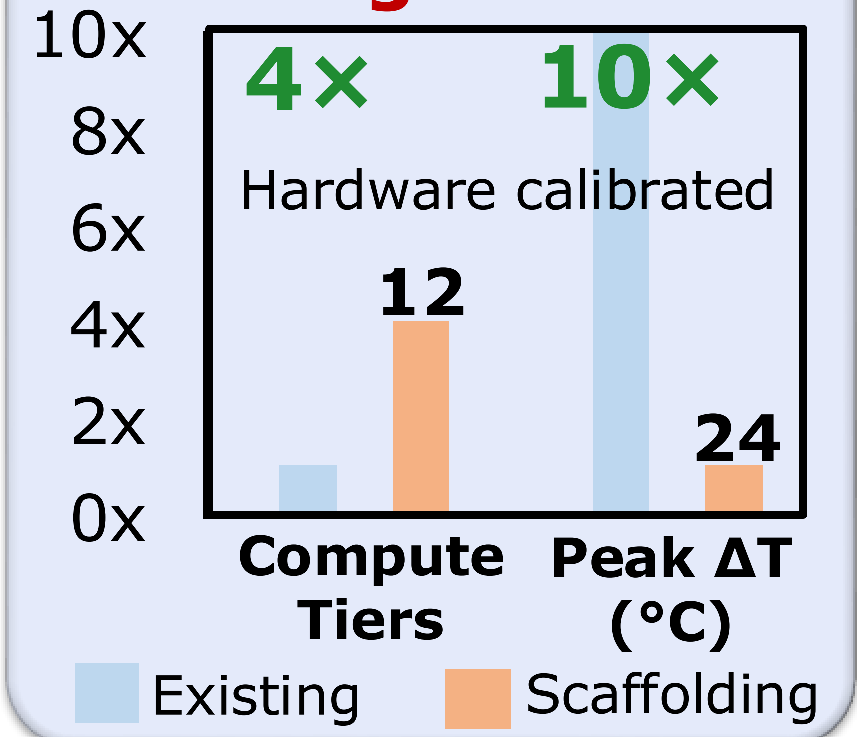
(hardware test structure)



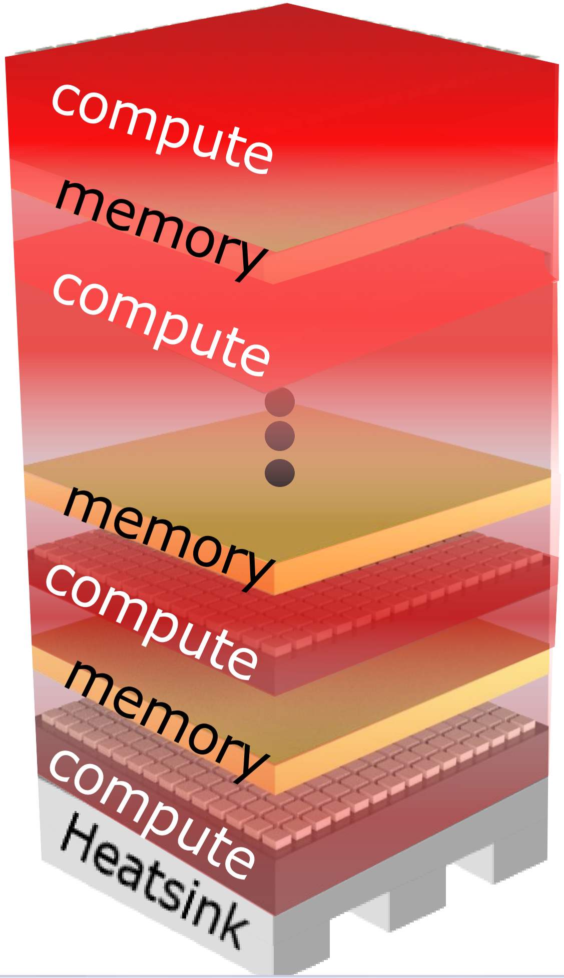
[Lyu IEDM 2024] Prof. S. Chowdhury (Stanford) BEOL = Back-end-of-line

Co-Placement Matters: 3D Thermal Scaffolding Results

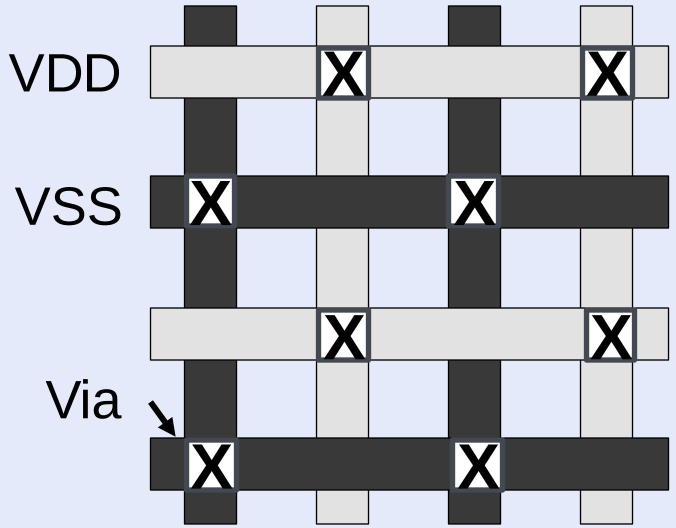
Cooling Benefits



Peak ΔT = $T_{peak} - T_{ambient}$



Power Delivery Benefits



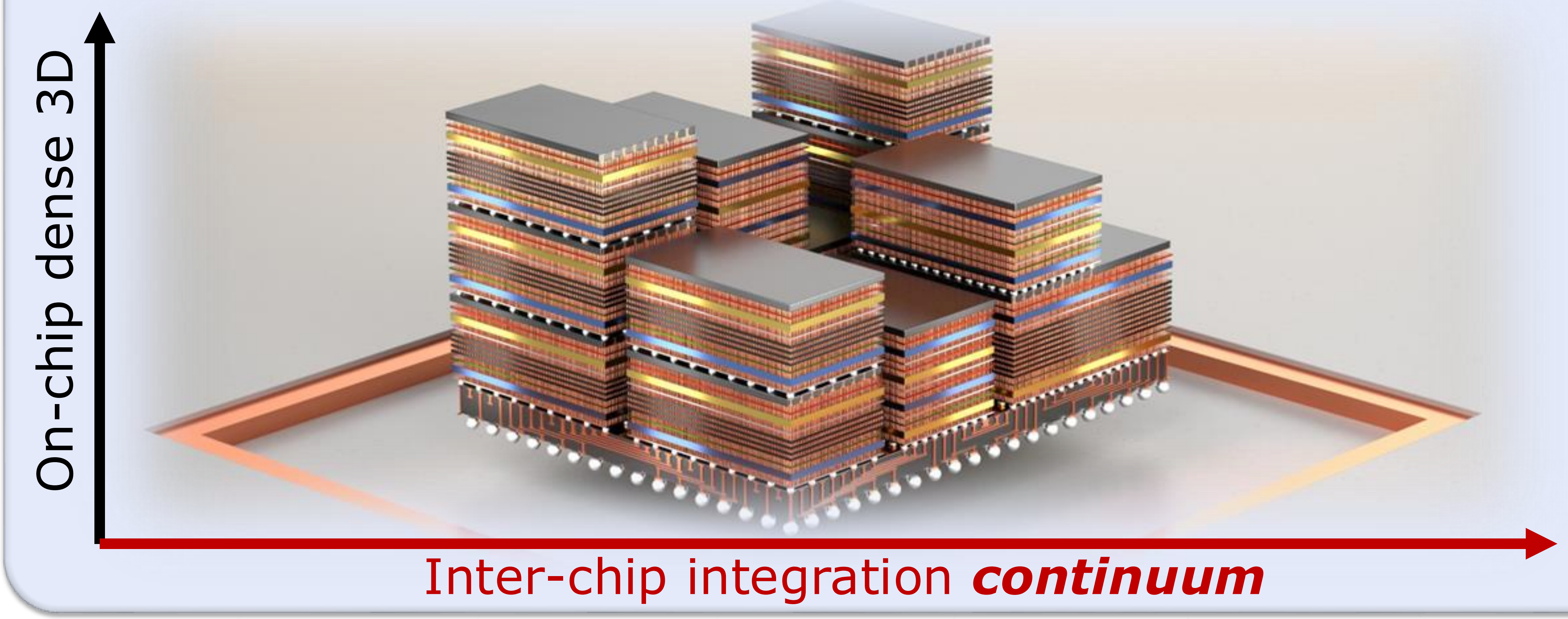
Scaffolding vias for power

only 5.5% extra footprint area

[Rich ICCAD 24]

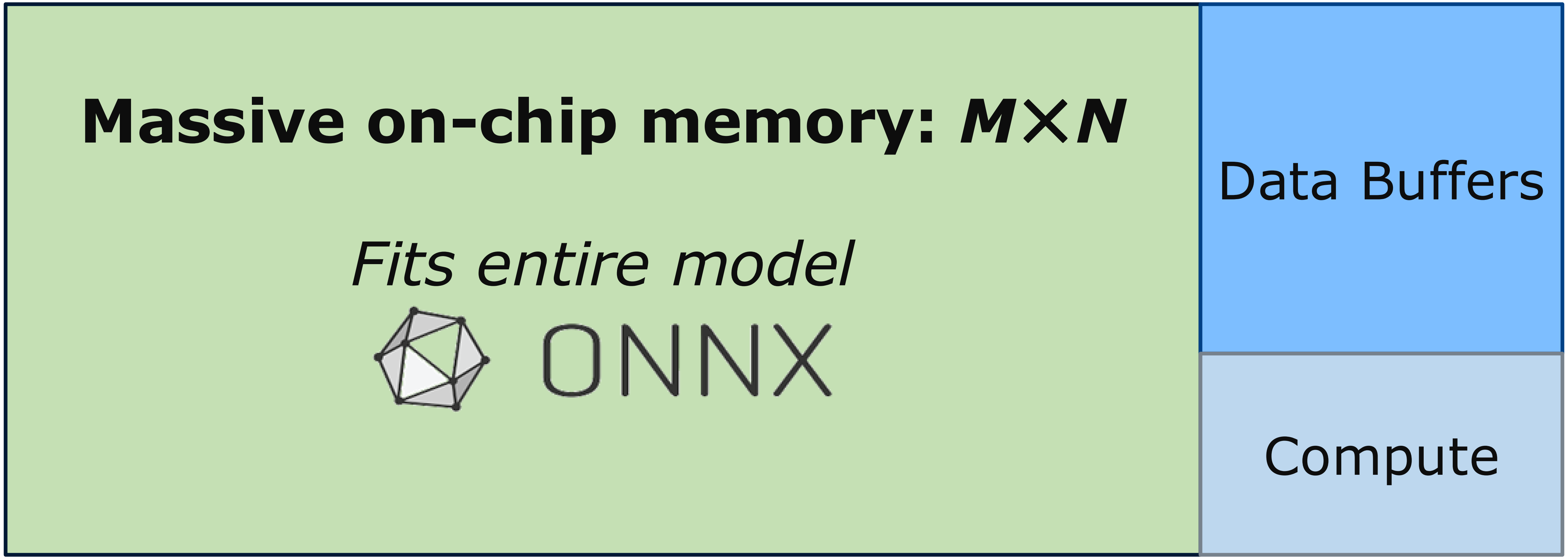
N3XT 3D MOSAIC

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



“Dream” Chip: All Memory + Compute On-Chip

Infeasible, Moving Target



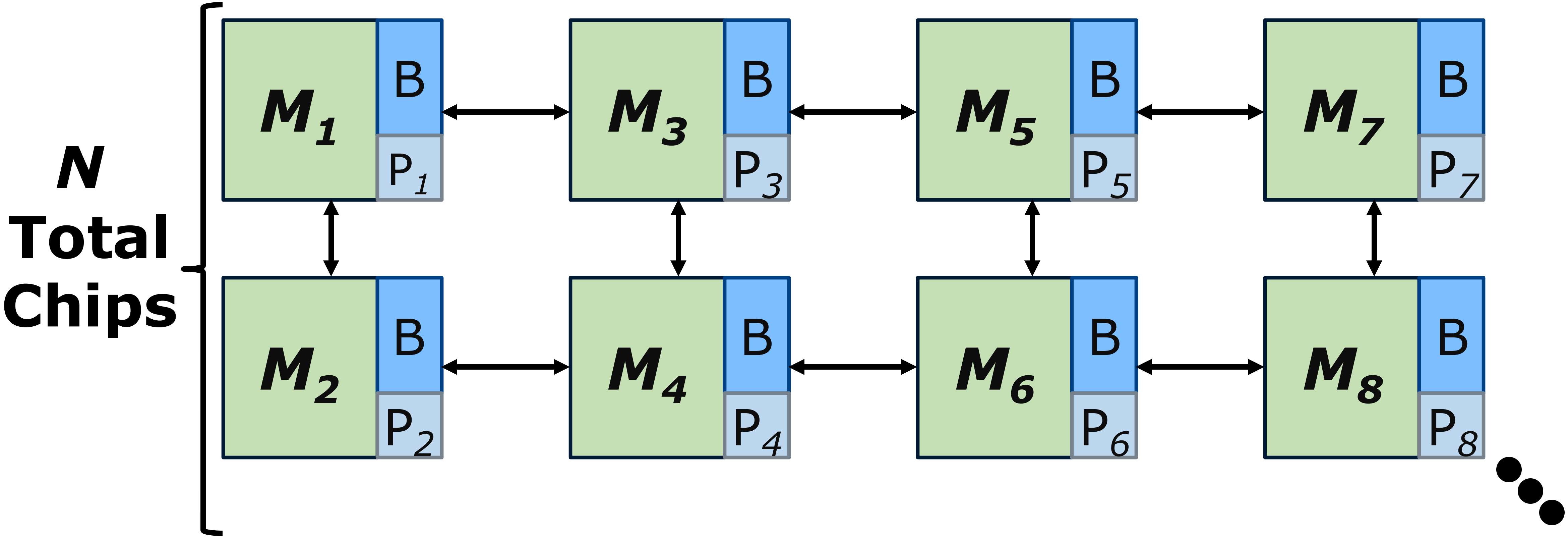
Massive on-chip memory: *M*×*N*

Fits entire model

Same dream as [Burks, Goldstein, Von Neumann, 1945]

Illusion Multi-Chip System

Target: within 10% end-to-end EDP of "Dream" chip



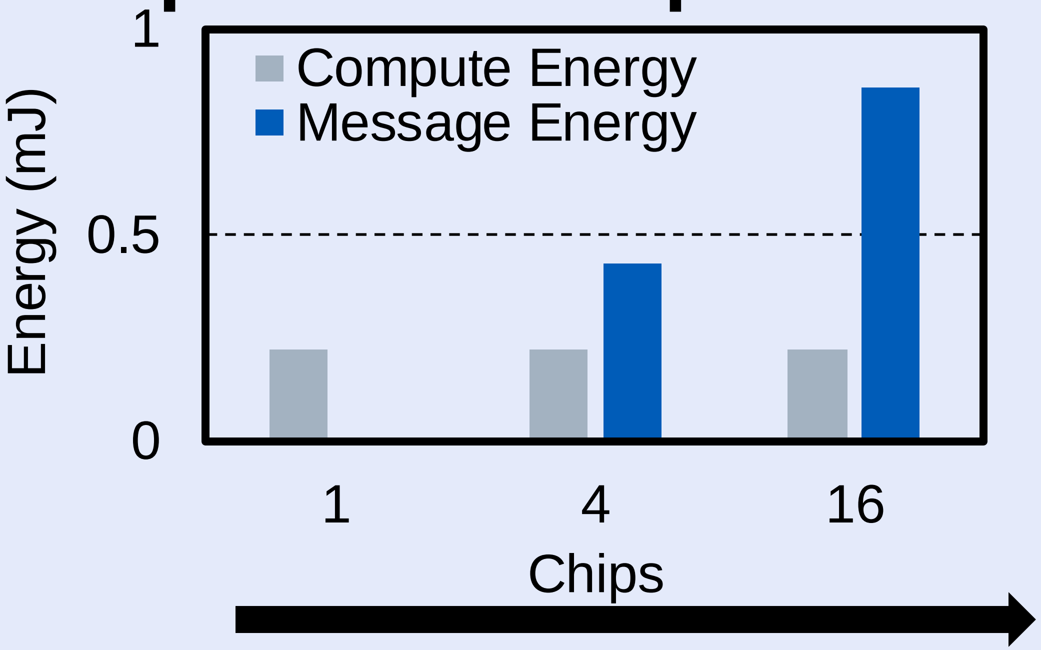
Optimize *N*, *M's*, *P's*, integration, mapping

[Radway Nature Electronics 21] Example network: 2D mesh shown for simplicity

1. “Enough” Memory per Chip, **Message Costs**

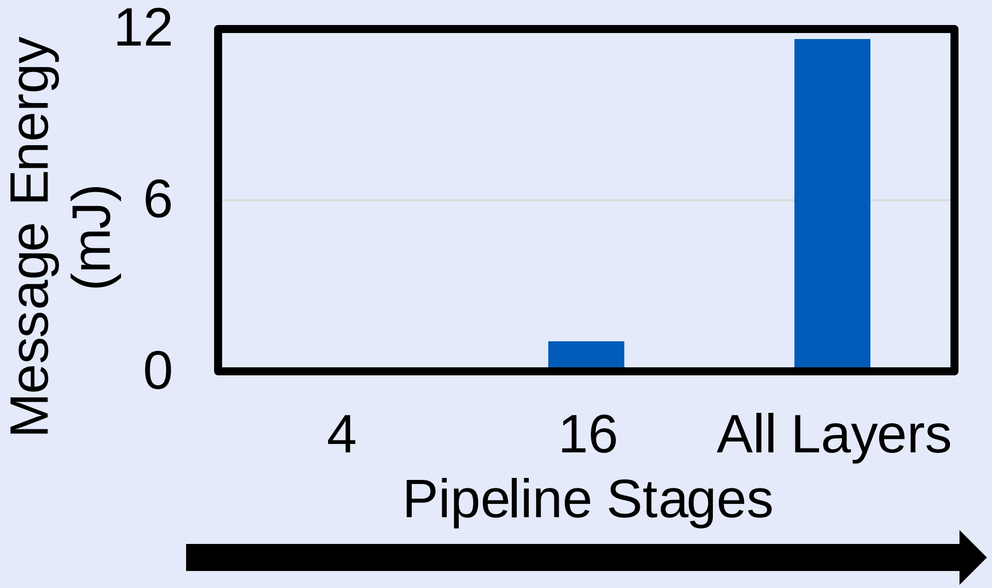
Must achieve target Message Costs

Excessive intra-layer parallelism expensive



Message latency & energy dominate

Naïve inter-layer pipelining expensive



Message latency can be hidden, message energy dominates

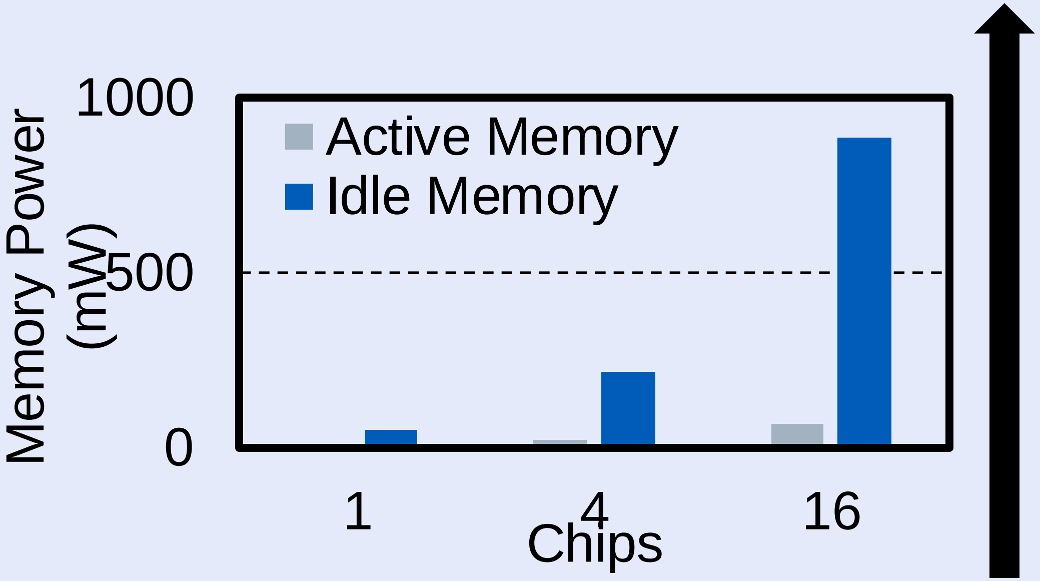
(Left) ResNet-18 1.1.Conv2. Parallelism [Shao Micro 19], no sync cost. (Right) 16-chip 187 MByte ResNet

2. Spatiotemporal Fine-grained Power-Gating

Idle energy overheads must be ~0

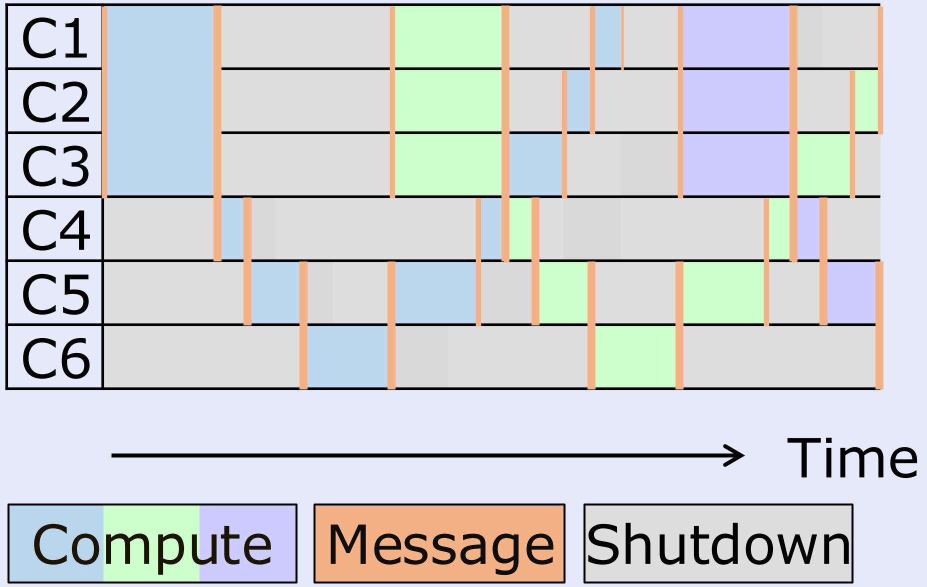
(validated on our MINOTAUR multi-chip system hardware)

In Space



Idle memory power quickly dominates

In Time



Avoid (grey) idle intra- and inter-input

Illusion Mapping

AI/ML Model

ONNX

Chip Assignment:

$i = 1,$ $i = 2,$ $i = 3, \ldots$

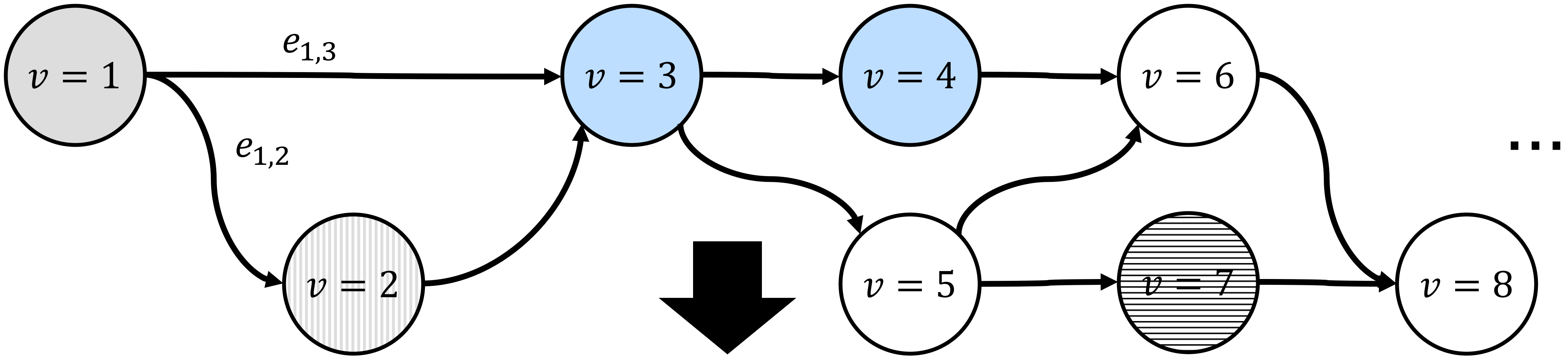
PE Assignment:

$j = 1,$ $j = 2,$ $j = 3, \ldots$

Operator Graph G

Vertices = Tensor Operations

Edges = Data Dependencies



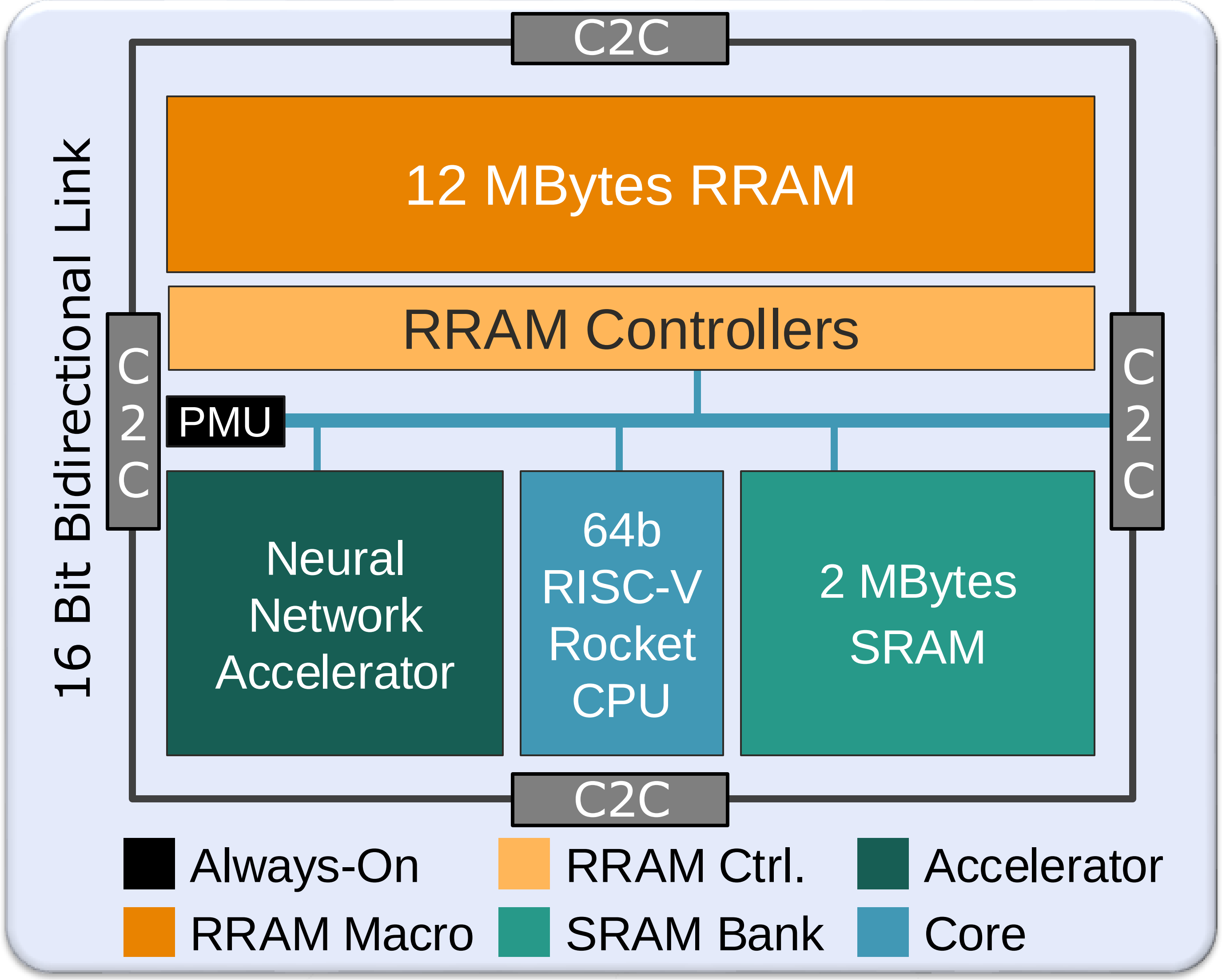
Mixed Integer Quadratic Programming (MIQP)

Gurobi Solver

Illusion MIQP vs. Existing Approaches

	Shao MICRO '19	**Narayanan SOSP '19**	**Unger OSDI '22**	**Tarnawski NeurIPS '20**	**Wang ICML '24**	**Our**
Method	Pre-defined Communication Patterns	Dynamic Programming (DP)	Graph Search	Mixed Integer Linear Programming (MILP)	MILP + DP	**MIQP**
True Minimum	N/A	No		Yes	No	**Yes**
Computational Cost Model	Measured	Performance Profiler		Architectural Model		**Cycle-Accurate Simulation + Emulation + HW Validation**
Target	Latency / Throughput	Training Speedup / Throughput		Latency / Throughput		**Energy-Delay Product or Energy or Latency or Throughput**
Interconnect Topology	Fixed	Fixed	Variable	Fixed	Fixed	**Arbitrary**
Runtime	N/A	Profiling Time	Minutes	Seconds	Hours	**Minutes for 64× larger models**

MINOTAUR: Transformer NVM Edge AI Inference & Training

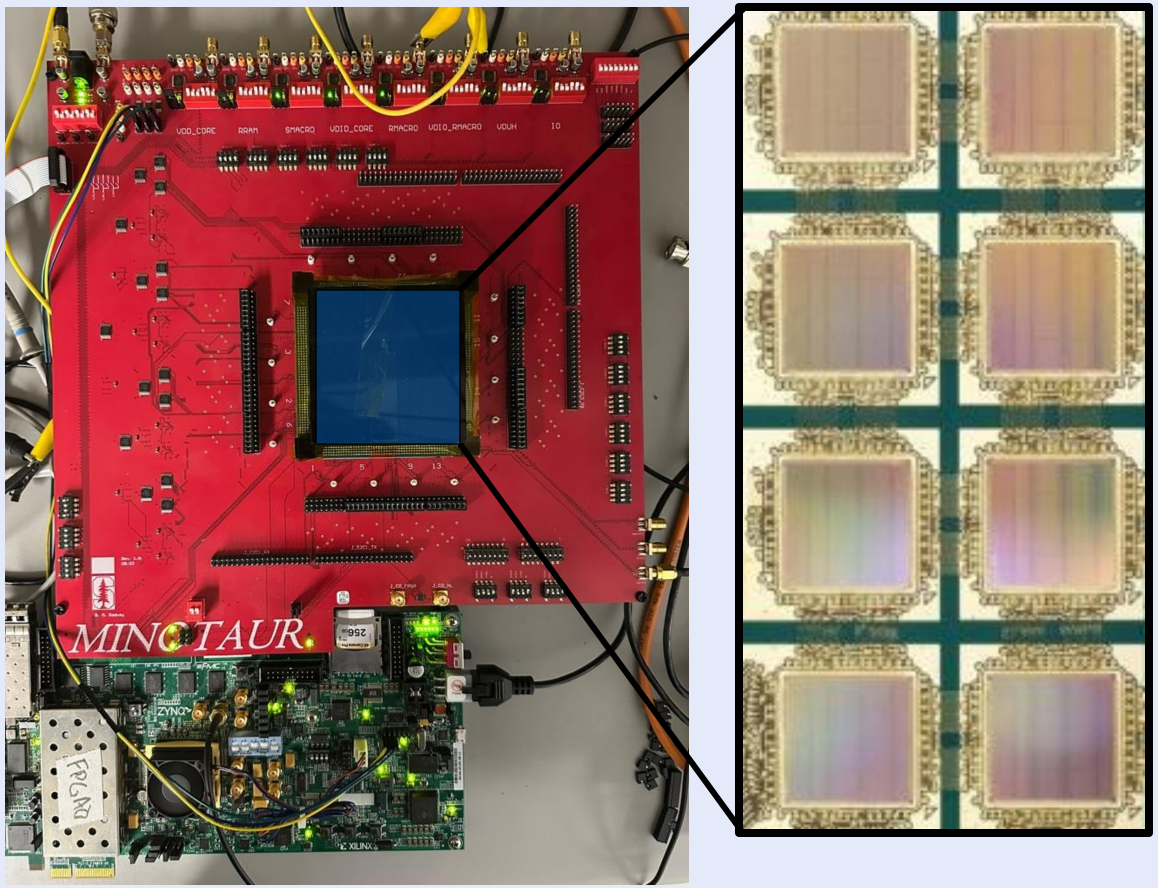


Transformers & CNNs	
Utilization	**93%**
Static memory power	**19× lower** (vs. foundry SRAM)
Active memory power	**3.4× lower** (vs. no fine-grained power management)
On-chip training	**Yes: new algorithm**

[K. Prabhu VLSI 2024] NVM = Non-Volatile Memory (RRAM in this case). Collaborator: Prof. P. Raina (Stanford)

Illusion ***In Hardware***

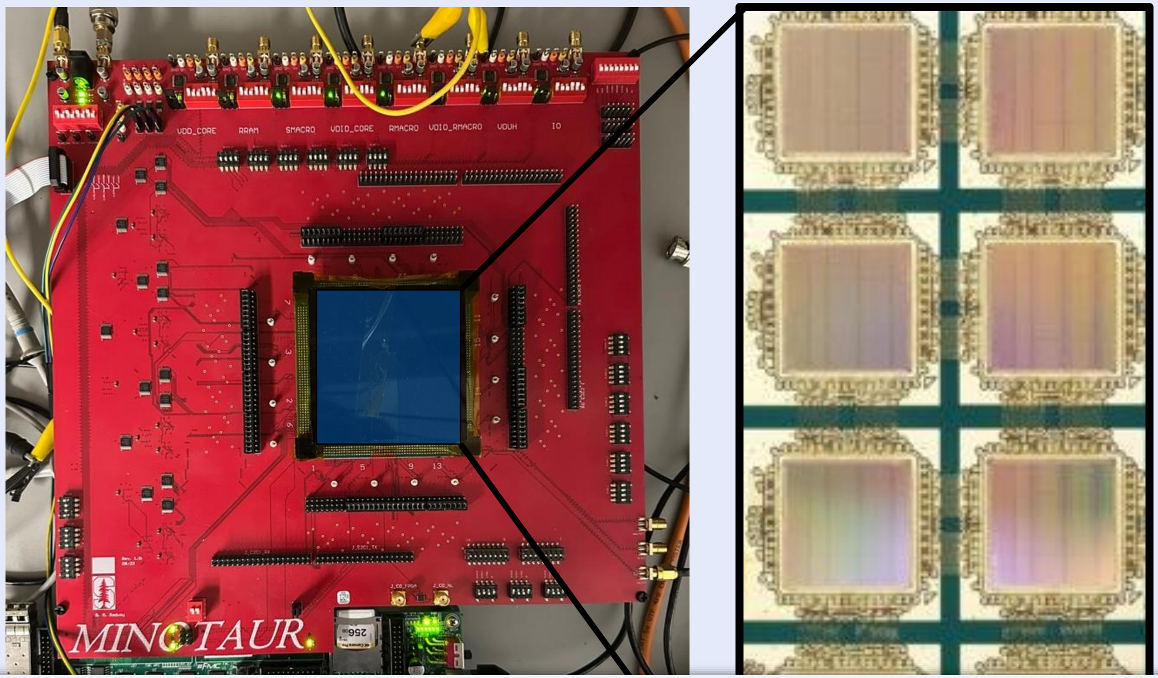
MINOTAUR Illusion: 96-MByte Transformers



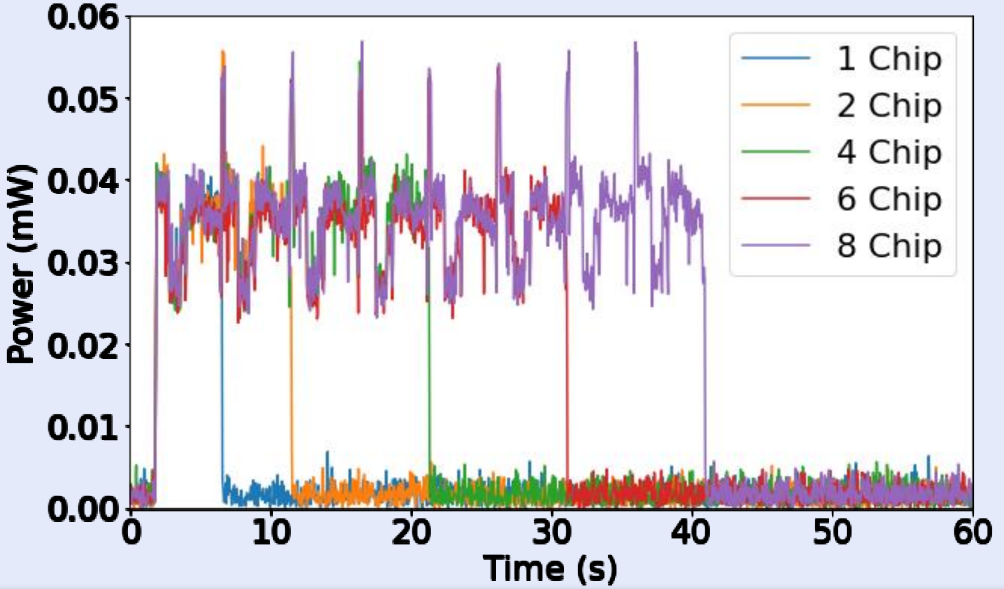
Chips	8 MINOTAUR
Total RRAM	Up to 96 MB
Total SRAM	Up to 16 MB
C2C Links	4 TX/RX per chip
Networks	CNNs, Transformers

Illusion ***In Hardware***

MINOTAUR Illusion: 96-MByte Transformers



BERT Encoders scaled to chip count

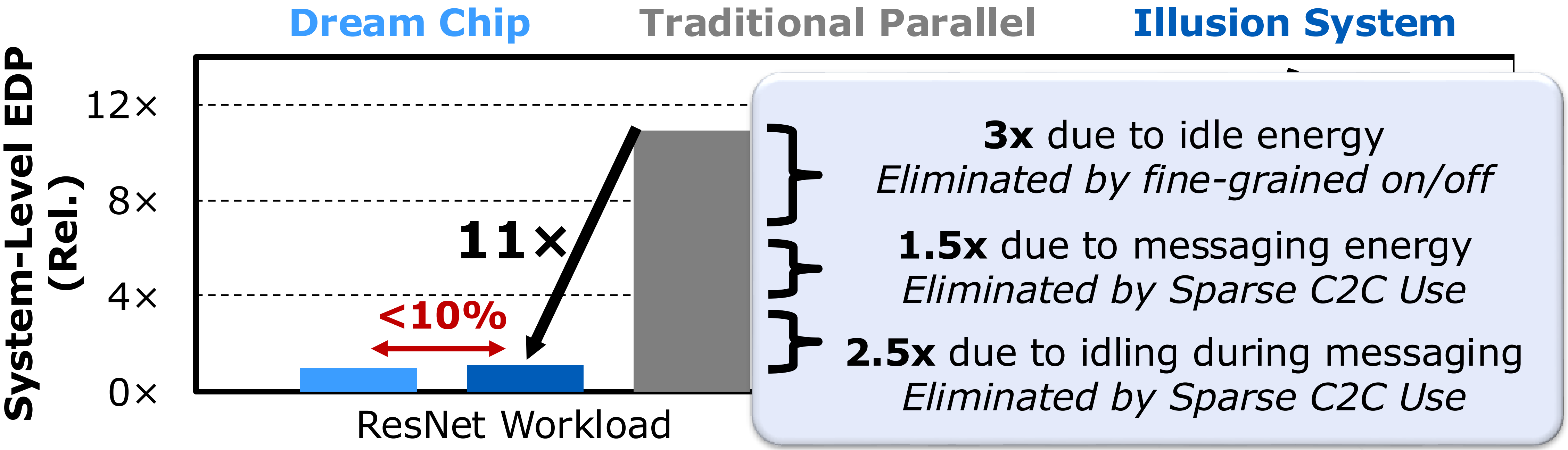


<10% of Dream energy and execution time

Demonstrated on MINOTAUR with BERT scaled from 1-8 Chips

Traditional Parallel vs. Illusion System

	Traditional Parallel	Illusion System
Memory/compute	Always on	Fast, fine-grained on/off
Chip-to-chip	Saturated	Sparse



3x due to idle energy
Eliminated by fine-grained on/off

1.5x due to messaging energy
Eliminated by Sparse C2C Use

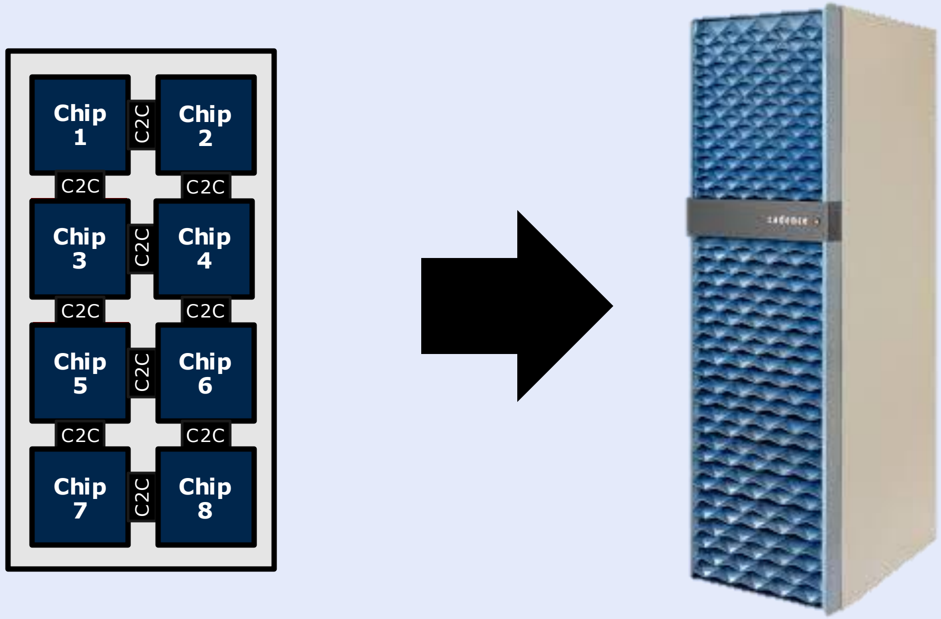
2.5x due to idling during messaging
Eliminated by Sparse C2C Use

16 chips, 288-layer 187 MB ResNet, 16 encoder 186 MB MobileBERT. Traditional parallel applied iso-hardware using approach in [Shao MICRO 19]. EDP relative to Dream Chip.

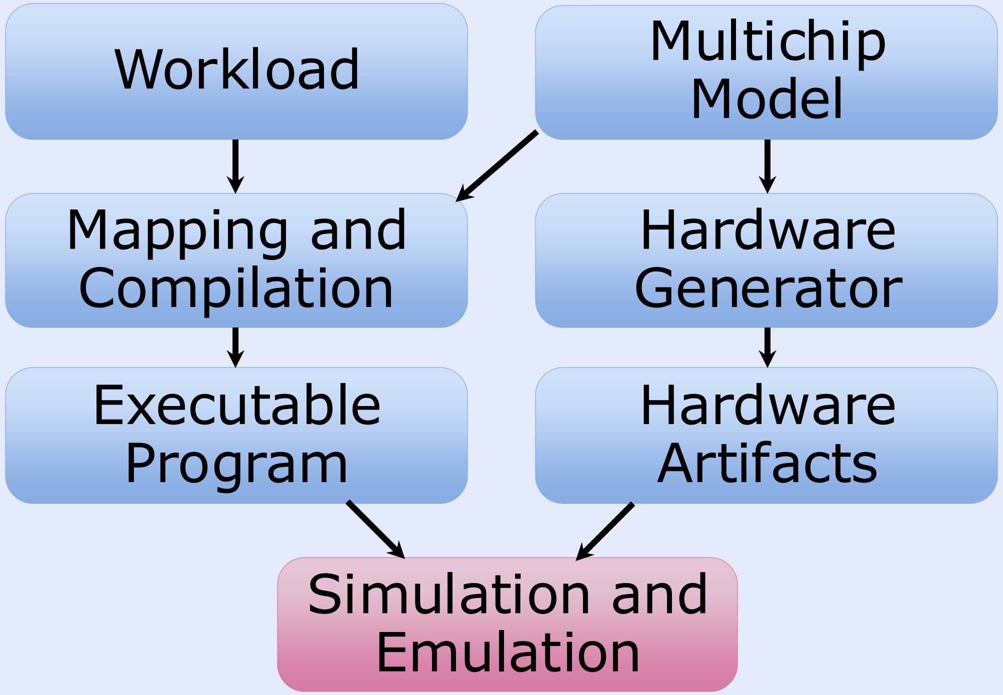
Illusion via **Emulation** (Beyond Hardware Demos)

Many thanks to Cadence!

Parameterized system-scale emulation



Compiler-Based Flow



Workload

Multichip Model

Mapping and Compilation

Hardware Generator

Executable Program

Hardware Artifacts

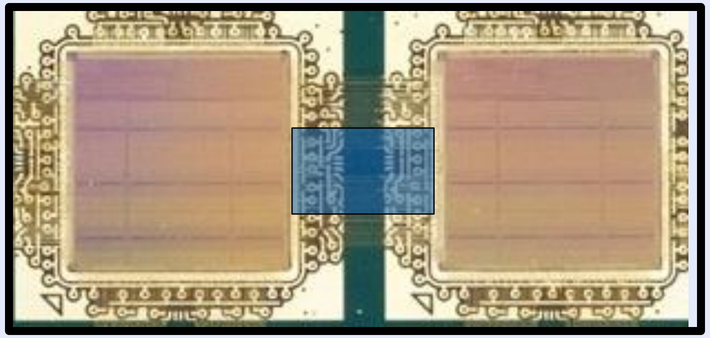
Simulation and Emulation

Cycle-accurate results in ***minutes not days***

Emulation Challenges for Illusion: Electrical Aspects

Chip-to-Chip Links

Realized parasitic vs. modeled differ

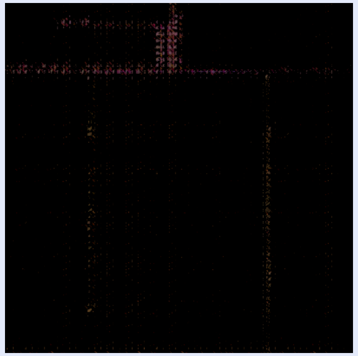


Energy 3.4× less

(Conservative model for timing closure)

Power Delivery Network (PDN)

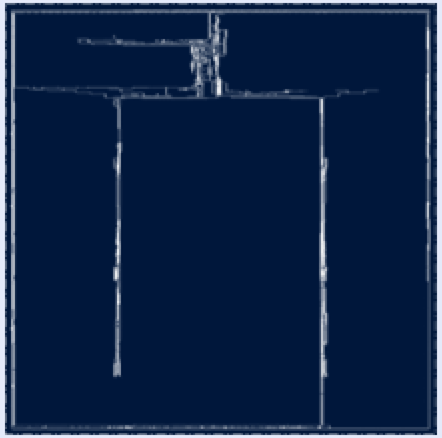
Power cycling charges & discharges PDN



120 uJ to wakeup & shutdown

Clock Tree

Nonzero off power



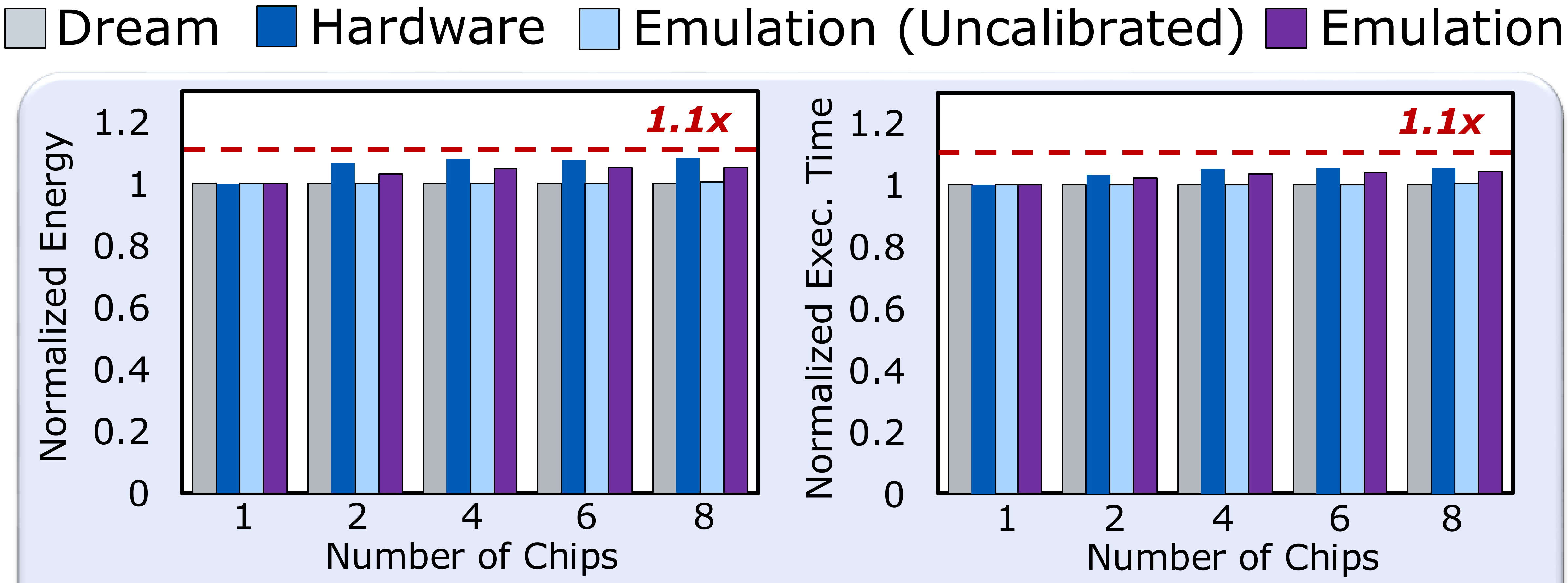
0.56 mW of always on power

Meticulous hardware calibration *essential*

(multi-chip MINOTAUR system in our case)

Illusion Demonstrated ***In Hardware and Emulation***

Illusion demonstrations agree within 5%



BERT Encoders on MINOTAUR (12 MB per chip): 16-chip workloads also emulated

BERT Scaled Deeper by the Number of Chips

Heuristics Essential to MIQP Scalability

128X Larger ResNet Still Tractable With MIQP

Model Size	Nodes	Edges	Chips	Variables	Constraints	Illusion EDP Overhead vs. Dream	Solve Time
128X	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Also explored:

Highly parallel models (64 branches)

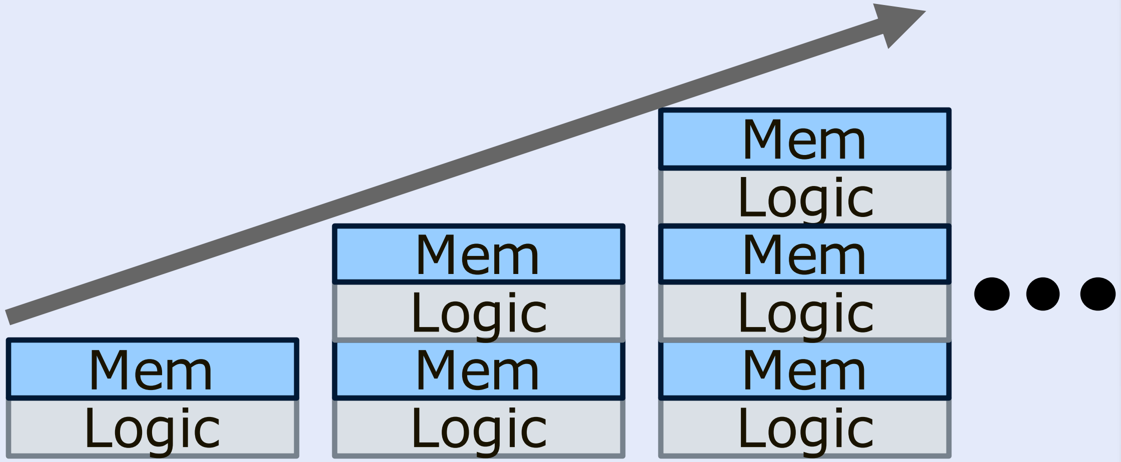
Fine-grained Transformer parallelism

Different chip and network configurations

…

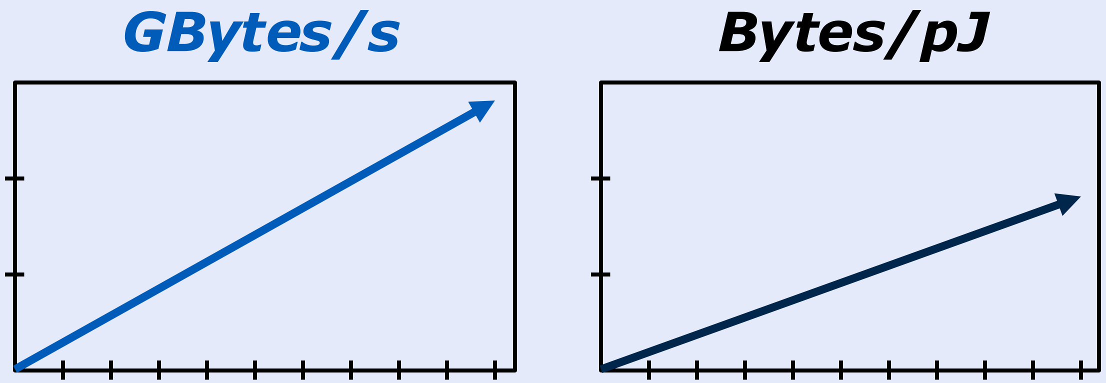
Illusion Scaleup: On-going

Increase dense 3D layers
Linear



Reduce *total messages*

Improve chip-to-chip links
Linear + one-time gains



Reduce *per-message cost*

Multiplicative effect

Maintain Illusion despite exponential workload growth over fixed time

[Radway IEDM 21] Linear improvements in electrical links through tight integration [das Sharma, Nature Electronics 24]

Conclusion

- ***N3XT 3D MOSAIC***
  - Overcome memory wall & miniaturization wall, successful lab-to-fab
  - Large system-level Energy Delay Product benefits for AI/ML
- **3D Thermal Scaffolding: high-power compute in 3D**
  - Co-design: thermal dielectric + 3D architecture + 3D physical design
- **Multi-chip Illusion: large AI/ML workloads**
  - Hardware results demo effectiveness, superior vs. traditional parallel