Materials Engineering for AI Compute

Tony Chiang, Ph.D.
VP/CTO, New Markets and Alliances Group
Applied Materials

NICE Workshop, March 28, 2019

World's #1

semiconductor and display systems company

$17.3 billion
revenue

>12,500
patents

$2 billion
R&D spending

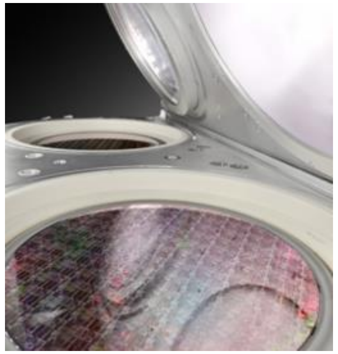
Applied Materials is the leader in materials engineering solutions used to produce virtually every new chip in the world

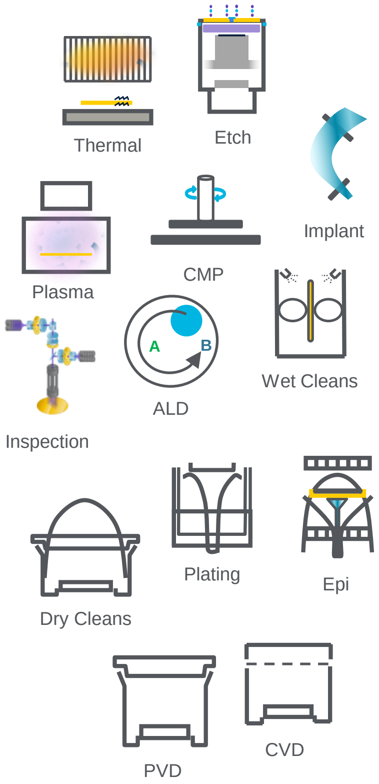
Data as of fiscal year end, October 28, 2018

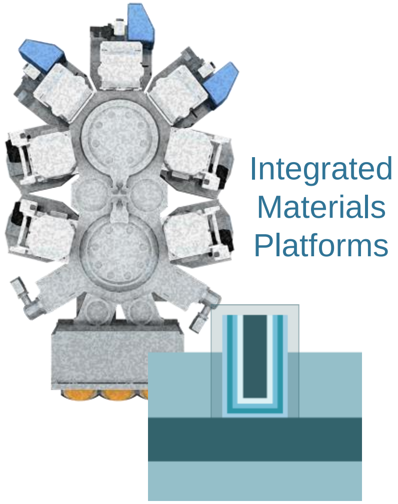
Applied's Materials Engineering

Enabling the Semiconductor Roadmap

50 years of materials innovation







Where we want to go next

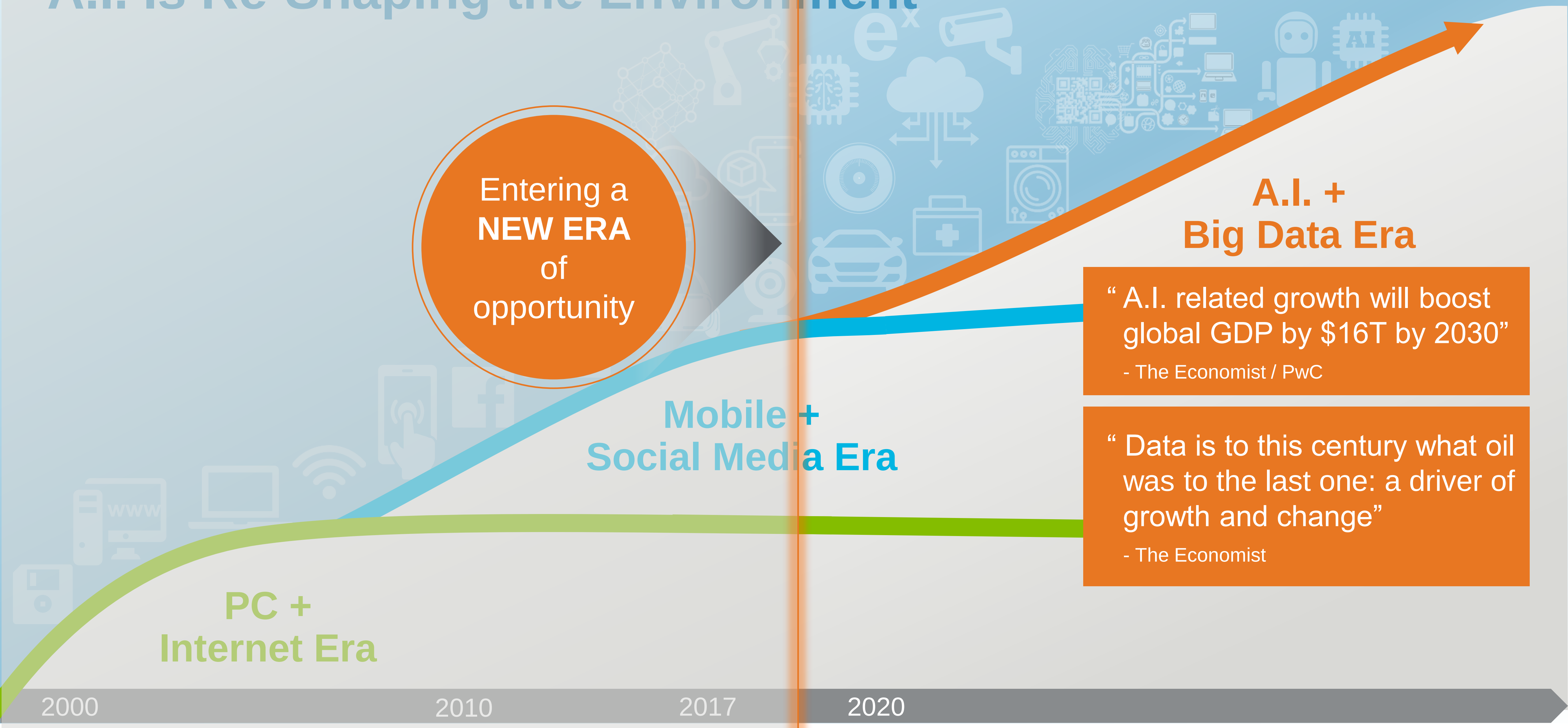
Software/Hardware Co-Optimization

Enable faster connection between materials innovation and AI performance

Semiconductor Processing Systems → Broad Portfolio of Technologies → Integrated Materials Solutions → Materials, Device, Circuits, Systems

APPLIED MATERIALS

A.I. Is Re-Shaping the Environment



" A.I. related growth will boost global GDP by $16T by 2030"

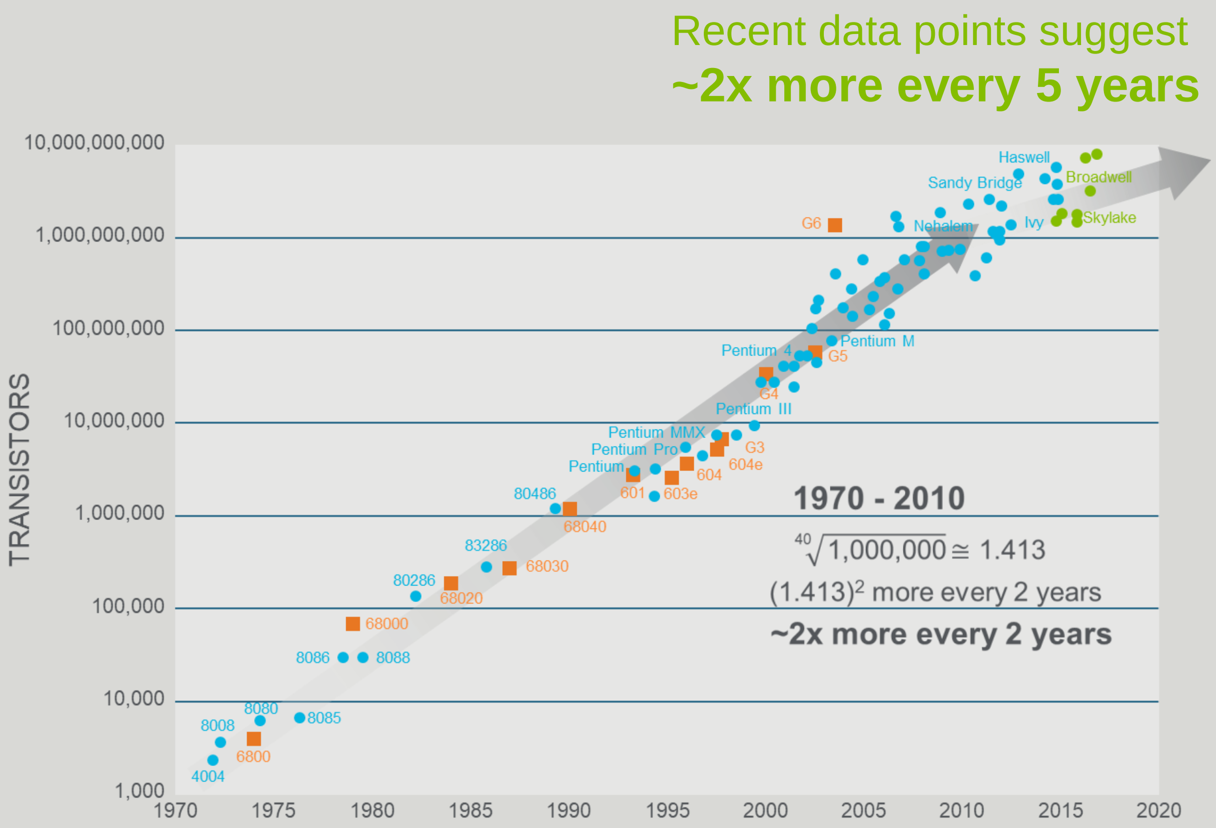
- The Economist / PwC

" Data is to this century what oil was to the last one: a driver of growth and change"

- The Economist

Performance Improvements Are Slowing…

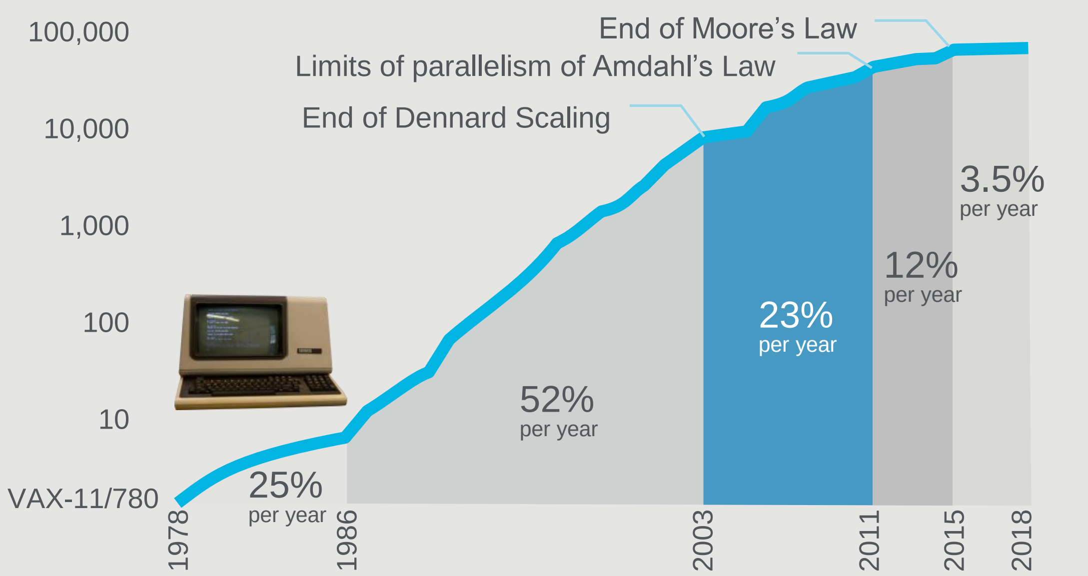
Moore's Law: Projection Held for 40 Years…



Classic 2D Feature Scaling Slowing

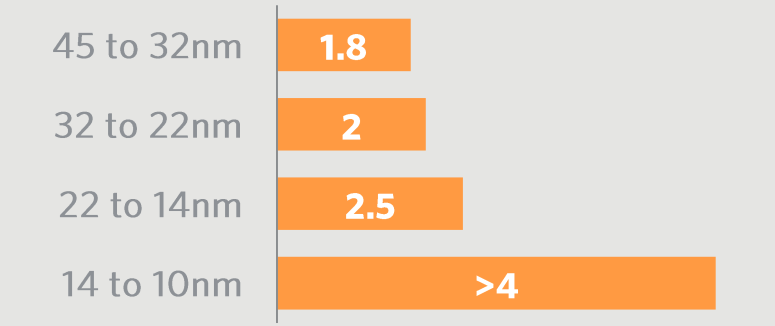
SOURCE: University of Wisconsin

PERFORMANCE IMPROVEMENTS OVER TIME
(VS. VAX-11/780)



SOURCE: Computer Architecture: A Quantitative Approach, Sixth Edition, John Hennessy and David Patterson, December 2017

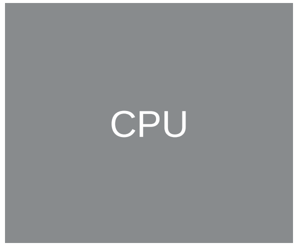
TIME BETWEEN LOGIC NODES IN YEARS



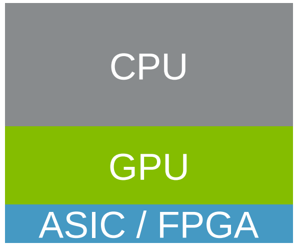
SOURCE: Bernstein

Data Explosion + Rise of A.I. = Heterogeneous Computing

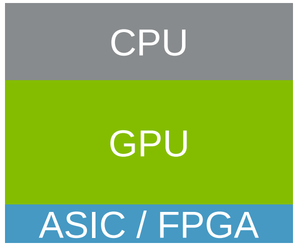
Rapidly Evolving Architectures



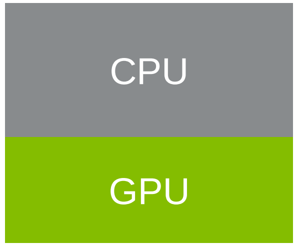
Normal Compiled / Managed Code
(Office, OS, Enterprise)



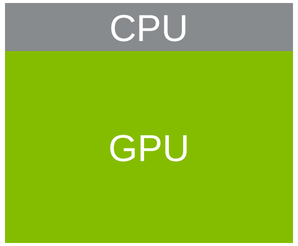
Imaging, Video Playback



AI Workloads:
Training, Inference, Analytics, etc.

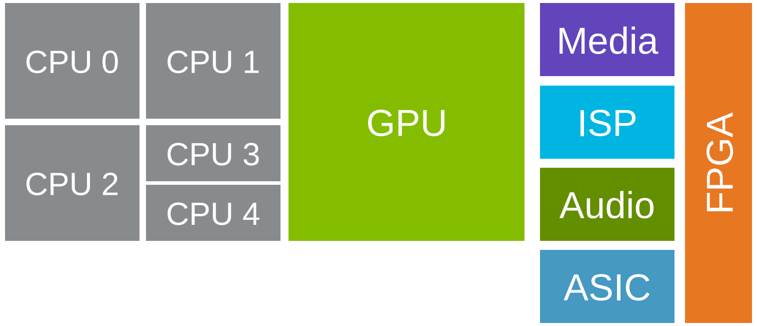


Client CC, Search, Audio, VOIP



Gaming, HPC, Highly Parallel Workloads

Heterogeneous Compute Era



Domain Specific Architectures:

System level optimization meeting performance, compute and area efficiency goals for targeted workloads

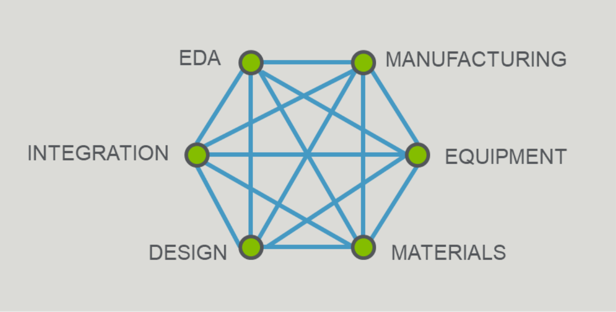
APPLIED MATERIALS

New Playbook Needed for Connectivity & Speed…

…to address Industry Challenges of Complexity ↑, Integration challenges ↑↑, Time to market ↑↑↑

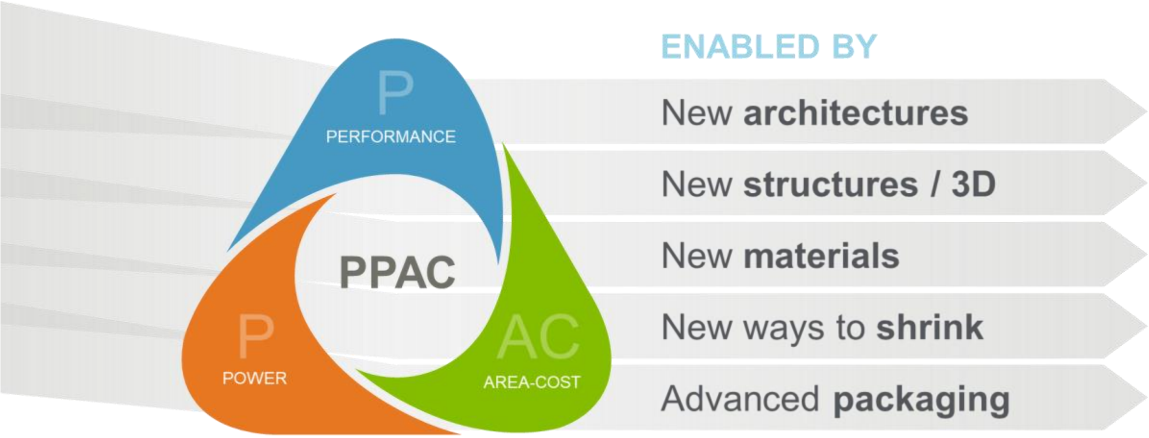
Serial mindset vs. **Connected** mindset

TODAY: **Serial** / compartmentalized interaction between key parts of eco-system



OPPORTUNITY: **Parallel** development to accelerate innovation

Connectivity to Accelerate Innovation



Foundation is Materials Engineering

APPLIED MATERIALS®

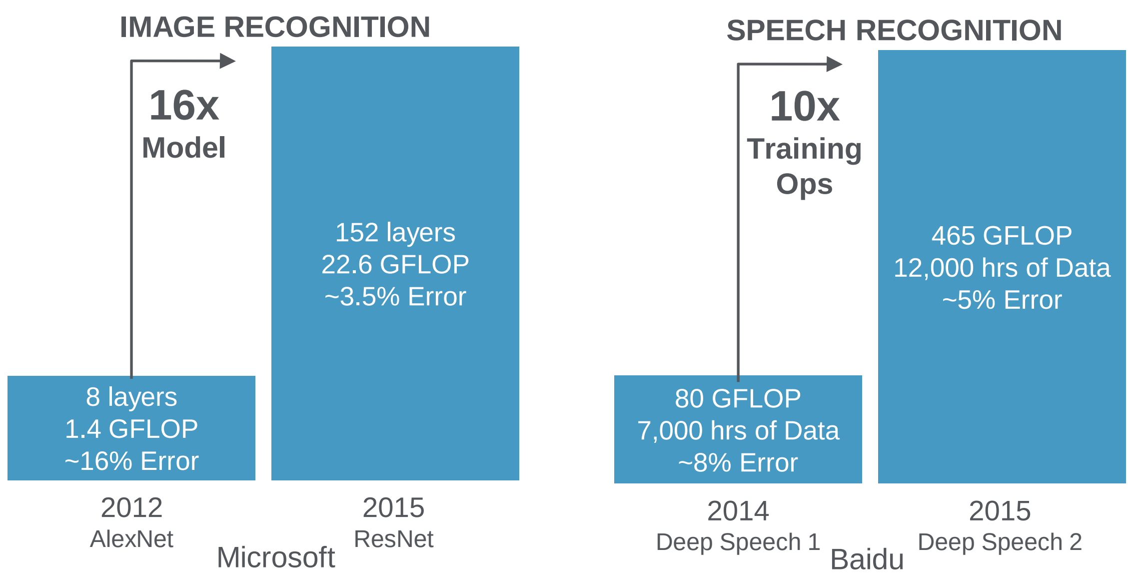
Key Challenges & Mitigation Strategies in AI

Key Challenges

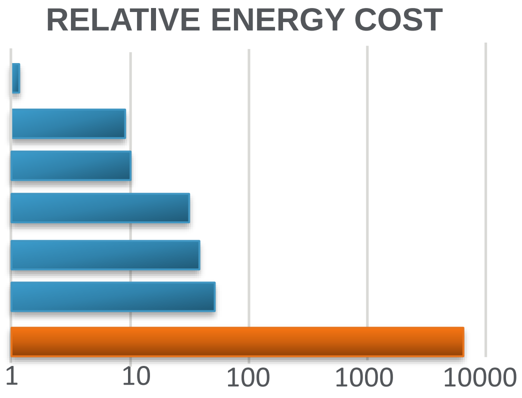
- Volumes of data and sizes of models are exploding
- Longer training times
- Larger model → more memory ref. → more energy
- Power use is not scalable -- energy efficiency issue

Mitigation Strategies

- Algorithm and Hardware Co-design
  - Customize/optimize per workload type
  - Minimize data moves
  - Move memory closer to computation
- New Devices
  - Integrate computation into the memory (analog compute)
  - New compute paradigms – quantum, synaptic



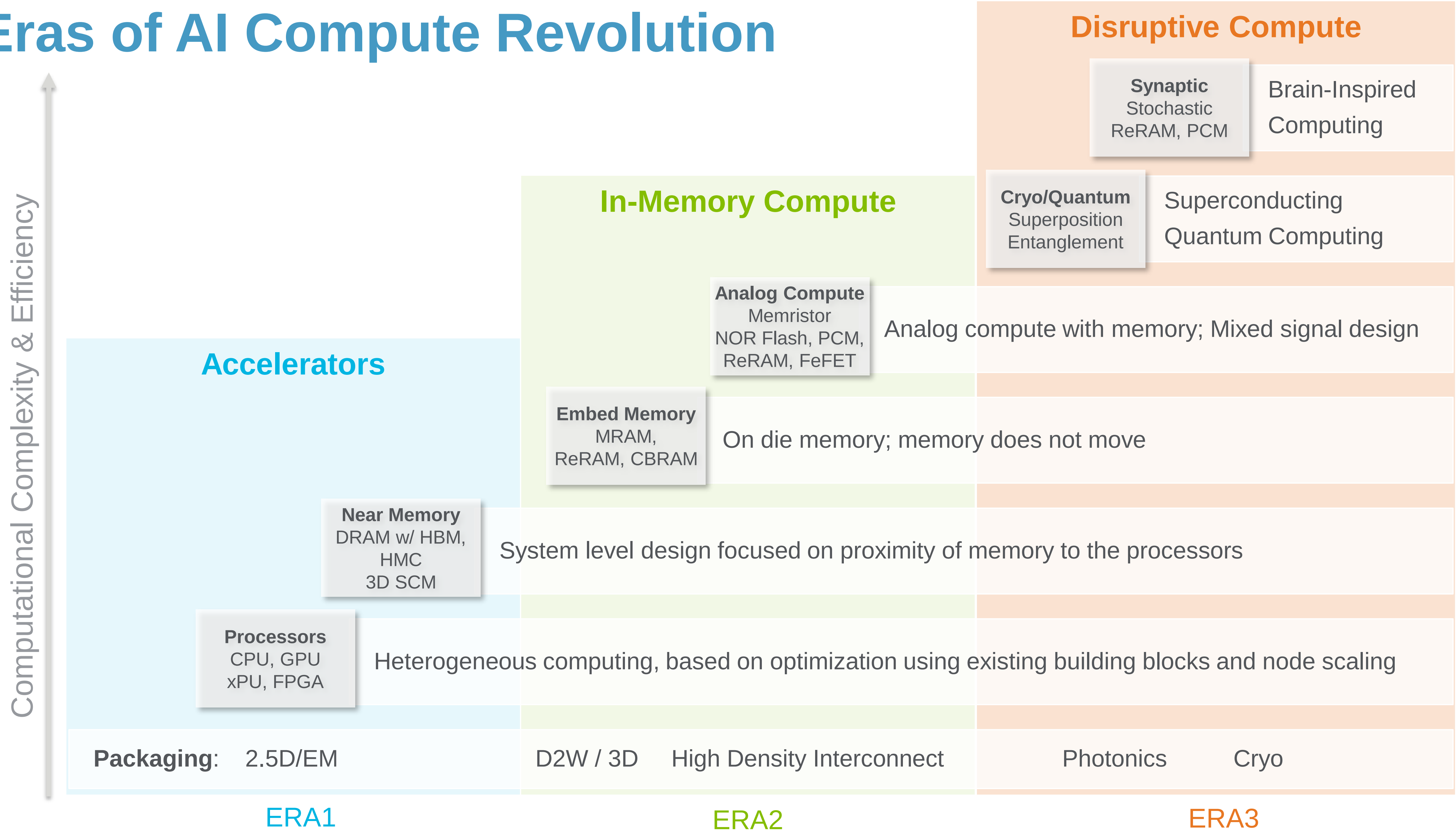
Operation	Energy (pJ)
32 bit int ADD	0.1
32 bit float ADD	0.9
32 bit Register File	1
32 bit int MULT	3.1
32 bit float MULT	3.7
32 bit SRAM Cache	5
32 bit DRAM Memory	**640**



Source: B. Dally (Chief Scientist Nvidia/Stanford), S. Han (Stanford), Efficient Methods and Hardware for Deep Learning (2017), NIPS 2016 Workshop on Efficient Methods for Deep Neural Networks (2016); V. Sze (MIT), Efficient Processing of Deep Neural Networks: A Tutorial and Survey (2017)

APPLIED MATERIALS

3 Eras of AI Compute Revolution



3 Eras of AI Compute Revolution

Computational Complexity & Efficiency

Disruptive Compute

Synaptic
Stochastic
ReRAM, PCM

Brain-Inspired Computing

Cryo/Quantum
Superposition
Entanglement

Superconducting Quantum Computing

In-Memory Compute

Analog Compute
Memristor
NOR Flash, PCM,
ReRAM, FeFET

Analog compute with memory; Mixed signal design

Embed Memory
MRAM,
ReRAM, CBRAM

C

Accelerators

Near Memory
DRAM w/ HBM,
HMC
3D SCM

System level design fo

Processors
CPU, GPU
xPU, FPGA

Heterogeneous computing, based

Packaging: 2.5D/EM D2W / 3D High D

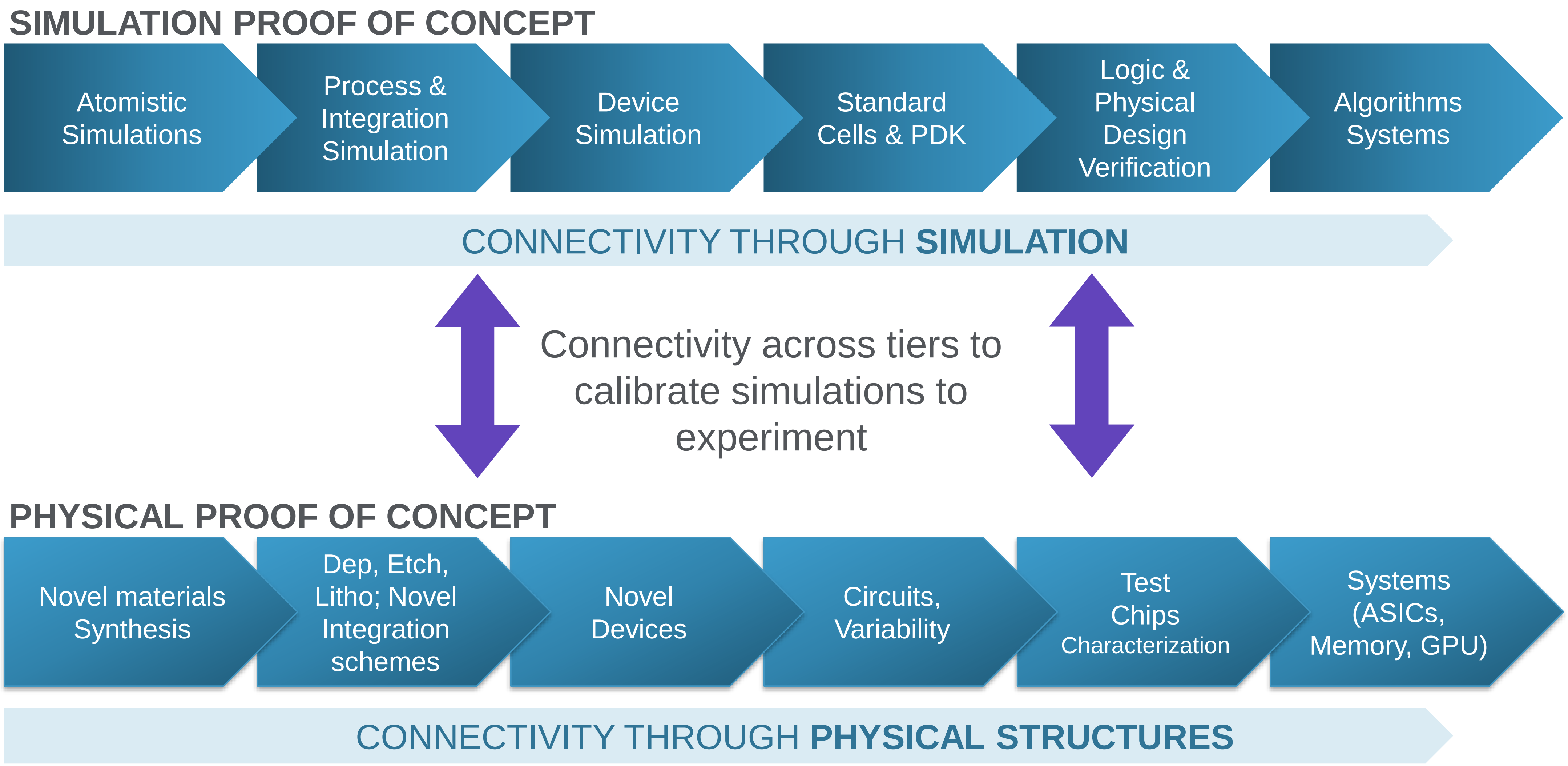
ERA1 ERA2 ERA3

Implications/Needs:

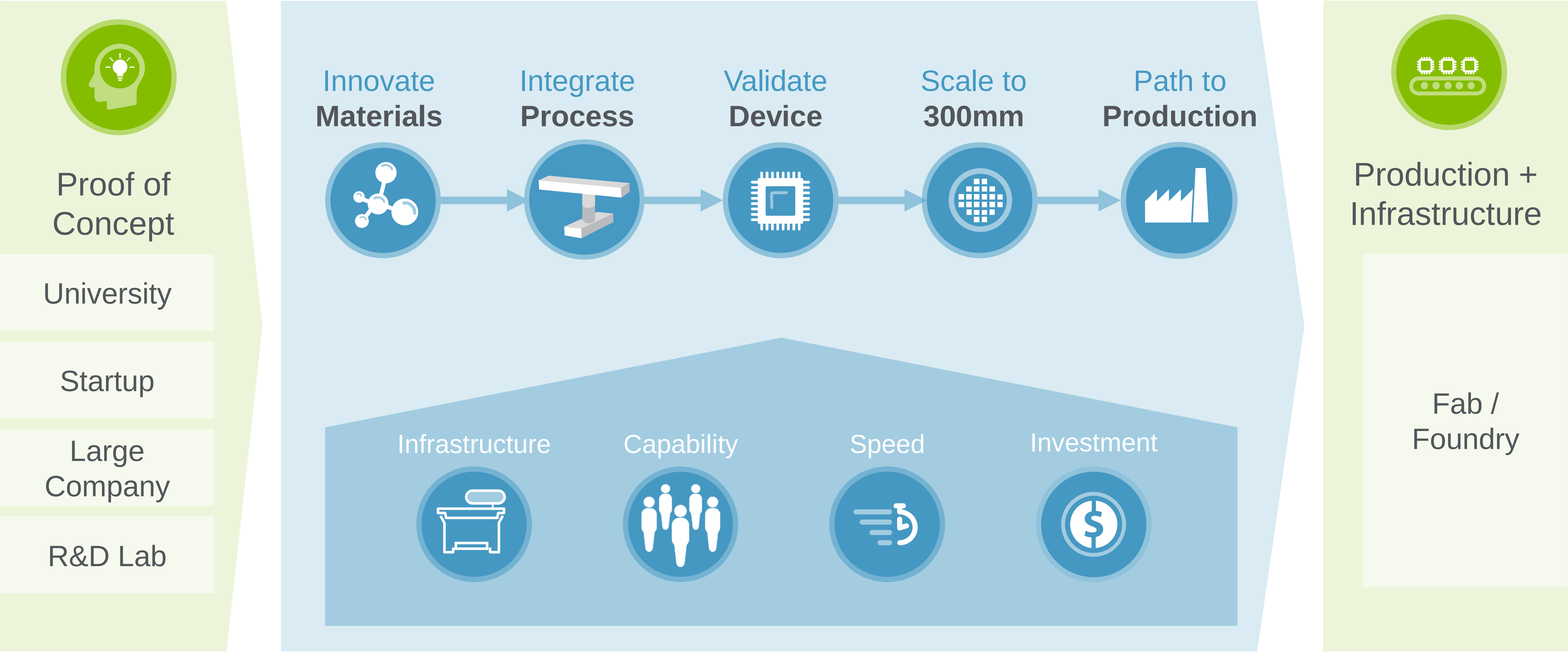
- Connectivity across the ecosystem
- Acceleration of learning cycles
- Co-optimization and customization of software & hardware, extending from materials to systems

APPLIED MATERIALS®

Materials to Systems



Accelerating the Path to Productization: Lab to Fab



Models are becoming Deeper and Larger

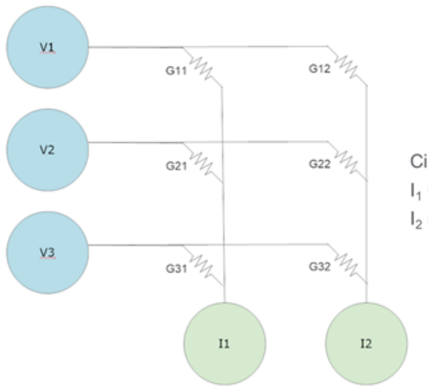
Common NN's

Neural Network	Type	# Weights	# MACs
AlexNet	CNN	61M	724M
GoogLeNet	CNN	7M	1.43G
VGG-16	CNN	138M	15.5G
ResNet50	CNN	25.5M	3.9G
RestNet152	CNN	60M	11.3G
MLP0	MLP	20M	
MLP1	MLP	5M	
LSTM0	LSTM (RNN)	52M	
LSTM1	LSTM (RNN)	34M	
CNN0	CNN	8M	
CNN1	CNN	100M	

95% of TPU Workload

Analog Multiply-Accumulate

X-bar array



Circuit representation of "weighted sum"

$I_1 = G_{11}*V_1 + G_{21}*V_2 + G_{31}*V_3$

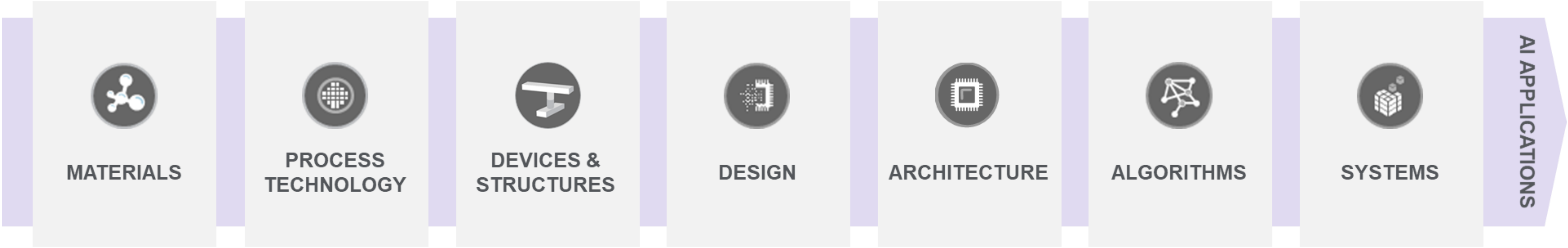
$I_2 = G_{12}*V_1 + G_{22}*V_2 + G_{32}*V_3$

Vector Matrix Multiplication performed by sensing current; weights stored as cell conductances

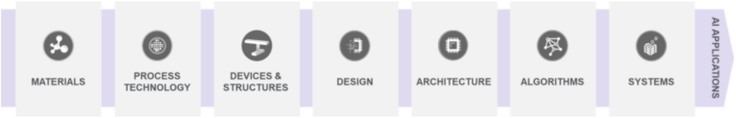
How to store weights on die?
How to make MACs more efficient?

Sources: V. Sze (MIT), Efficient Processing of Deep Neural Networks: A Tutorial and Survey (2017),
Canziani, Alfredo; Paszke, Adam; Culurciello, Eugenio; An Analysis of Deep Neural Network Models for Practical Applications; (2016),
N. P. Jouppi et. al, In-Datacenter Perf. Analysis of a Tensor Processing Unit, 44th International Symposium on Computer Architecture (ISCA), Toronto, Canada, June 24-28, 2017

Materials to Systems Approach is Needed



Materials Innovation Drives Performance of AI Devices



Performance of Different In-Memory Compute Elements

Based on MNIST data set

	Metric	Device type	Digital synapse	Potential Analog Synapses				
				ReRAM	ReRAM	ReRAM	PCRAM	FeRAM
			6-bit SRAM	Ag:a-Si	AlOx/HfO2	TaOx/HfOx	GST	HZO
Properties of materials system	# of conductance states	#	--	97	40	128	100-120	32
	Nonlinearity (weight up/down)	ratio	--	2.4 / -4.9	1.9 / -0.6	0.04 / 0.2	0.1 / 2.4	1.6 / 1.8
	RON	kΩ	--	26,000	17	86	5	500
	ON/OFF ratio	ratio	--	12.5	4.4	10	20	~1,300
	Weight increase pulse	V/μs	--	3.2 / 300	0.9 / 100	1.6 / 0.05	0.7 / 6	2.17 / 50
	Weight decrease pulse	V/μs	--	-2.8 / 300	-1 / 100	1.5 / 0.05	3 / 0.125	-1.62 / 50
	Cycle-to-cycle variation (σ)	%	--	3.5%	5%	~3.5%	1.5%	<1%
Power Performance Area (PPA)	Area	μm^2	10,311	1,072	3,657	1,513	7,233	1,194
	Latency (optimized)	sec	0.5217	64,200	4,440	10	413	480
	Energy (optimized)	mJ	2.2	15	146	0.81	1,340	0.21
	Inference Latency	msec	29.2	0.24	0.20	0.20	0.20	0.20
	Inference Energy	μJ	26.1	2.4	5.0	3.1	6.5	2.7
ML Algorithm	Online learning accuracy	%	~94%	~73%	~41%	~73%	~87%	~90%

Reference

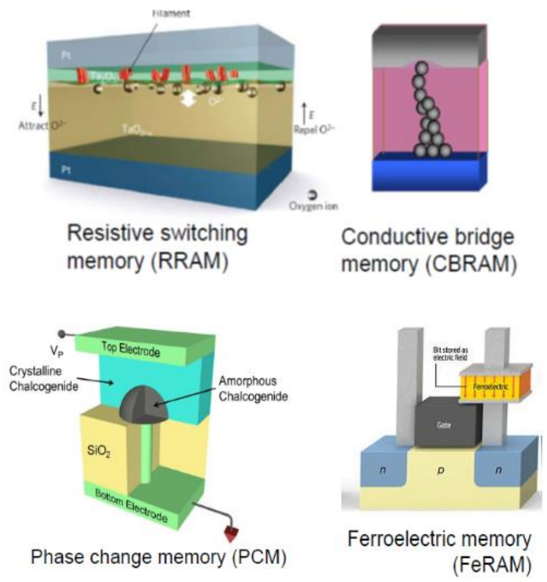
Adapted from S. Yu, ASU/Georgia Tech

Different materials systems

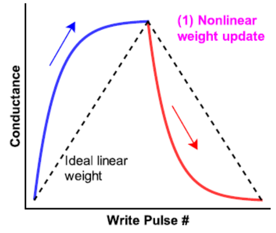
Properties developed by materials engineering

Resulting Power, Performance, Area

AI Model Accuracy

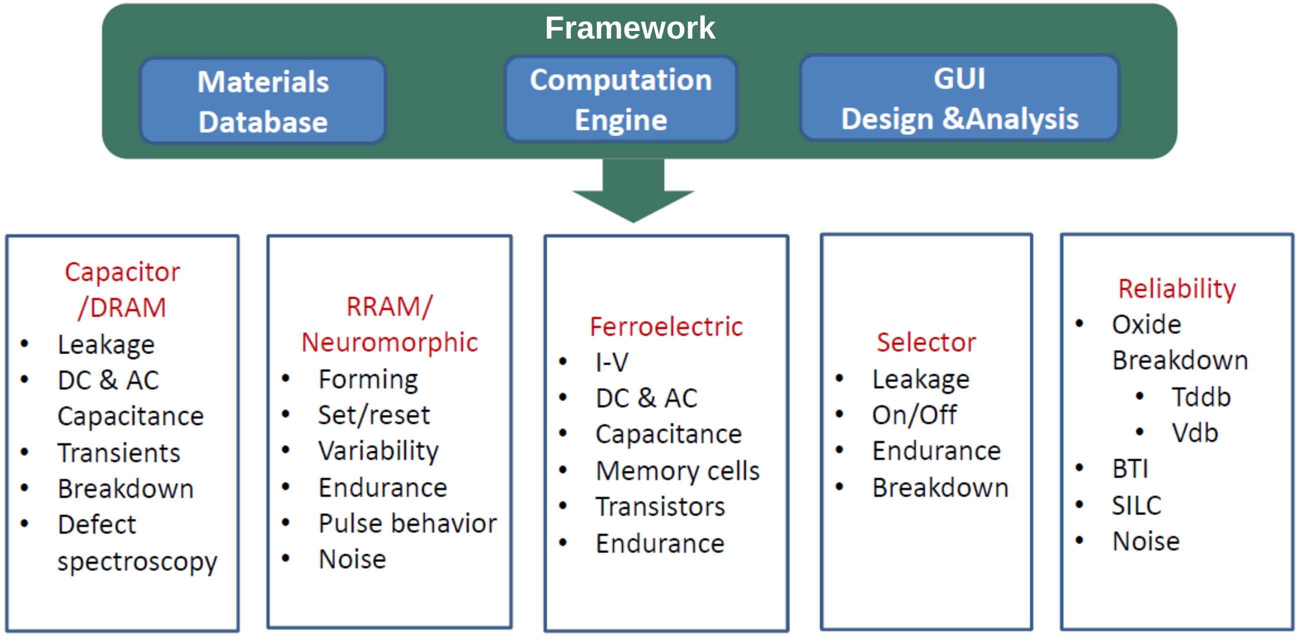


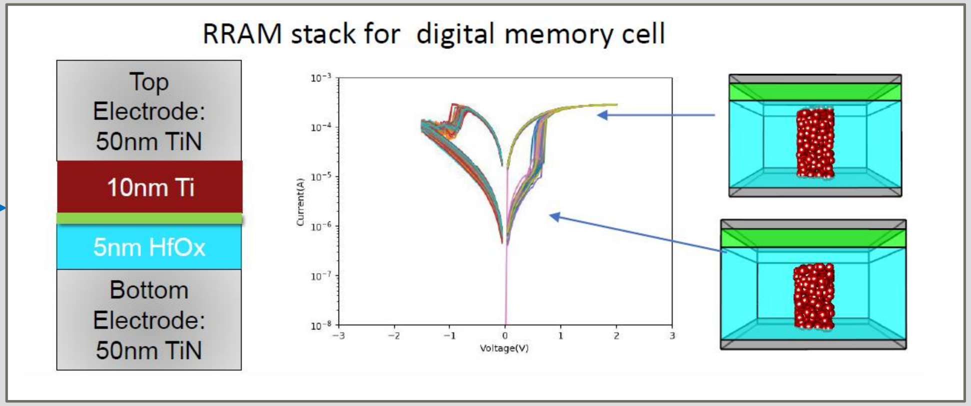
Images: Ielmini, "In-memory computing with emerging memory devices", Politecnico di Milano

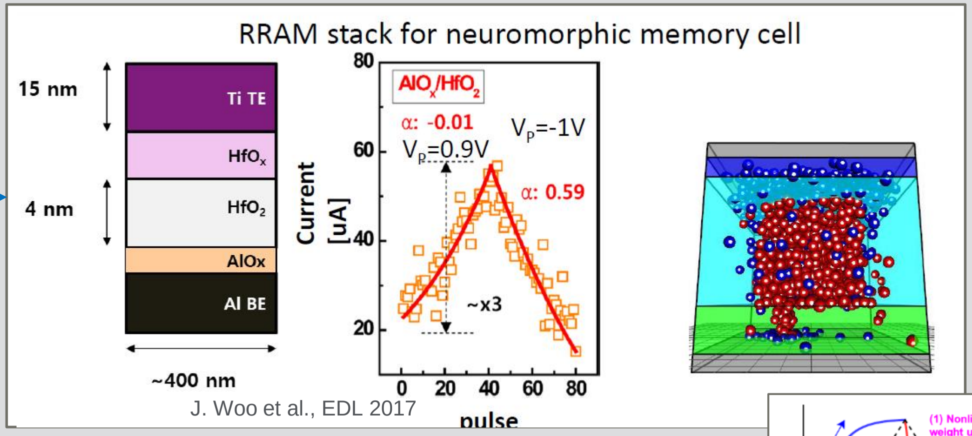


Many different device types & mechanisms: Need to leverage intrinsic physics for AI compute

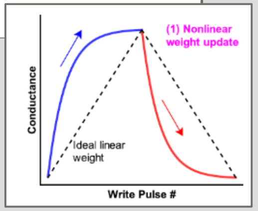
Device Physics to Cell Behavior







J. Woo et al., EDL 2017



Understand and Exploit Cell Physics

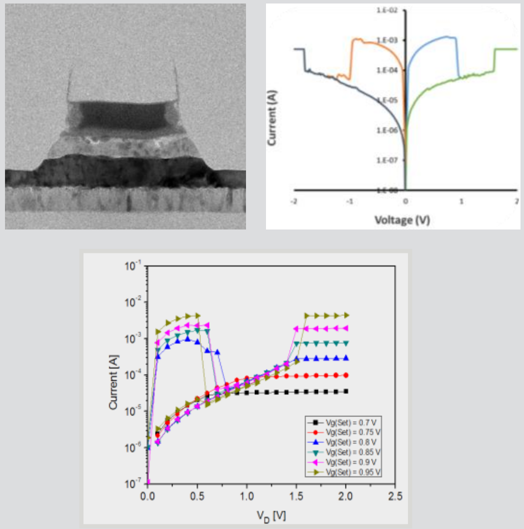
Engineer Cell Stack Based on Understanding

Inputs: Materials/properties, film thicknesses, layering scheme, etc.

Outputs: Forming behavior, potentiation/depression behaviors, variability, etc.

APPLIED MATERIALS

Connectivity Through Partnerships



Applied Materials team selected by **DARPA** to develop advanced technology for AI

Applied is working with **Arm** and **Symetrix** to develop a new neuromorphic switch based on CeRAM memory

Announced July 24th 2018

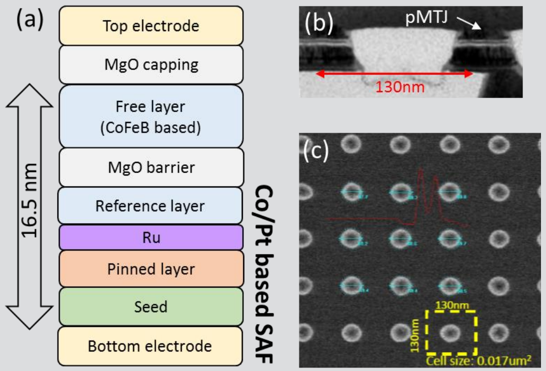
Source: SUNY Poly

ESD and SUNY Announce New Research Partnership with Applied Materials

New Applied Materials R&D Center to Help Customers Overcome Moore's Law Challenges

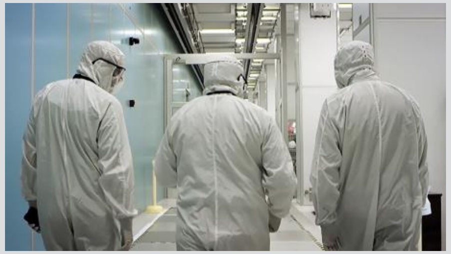
Applied Ventures and Empire State Development Aim to Accelerate Innovation in Upstate New York

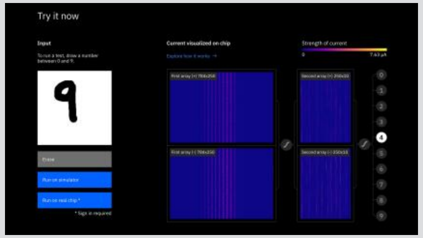
Announced Nov 15th 2018



Spin Memory Teams with Applied Materials to Produce a Comprehensive Embedded MRAM Solution

Announced Nov 11th 2018





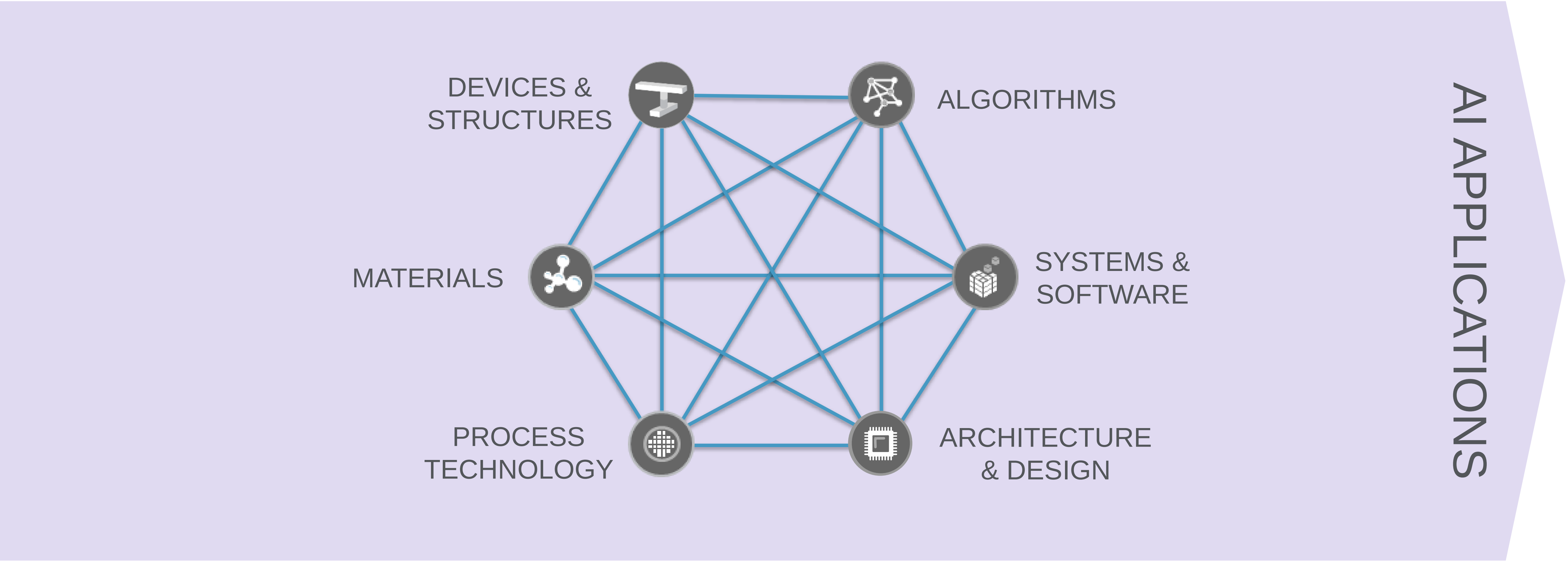
Source: IBM

IBM Launches Research Collaboration Center to Drive Next-Generation AI Hardware

Partnerships with leading semiconductor equipment companies Applied Materials… are crucial to the successful introduction of disruptive materials and devices to fuel our AI hardware roadmap.

Announced Feb 7th 2019

APPLIED MATERIALS


DEVICES &
STRUCTURES
ALGORITHMS
MATERIALS
SYSTEMS &
SOFTWARE
PROCESS
TECHNOLOGY
ARCHITECTURE
& DESIGN
AI APPLICATIONS

APPLIED MATERIALS®

make possible