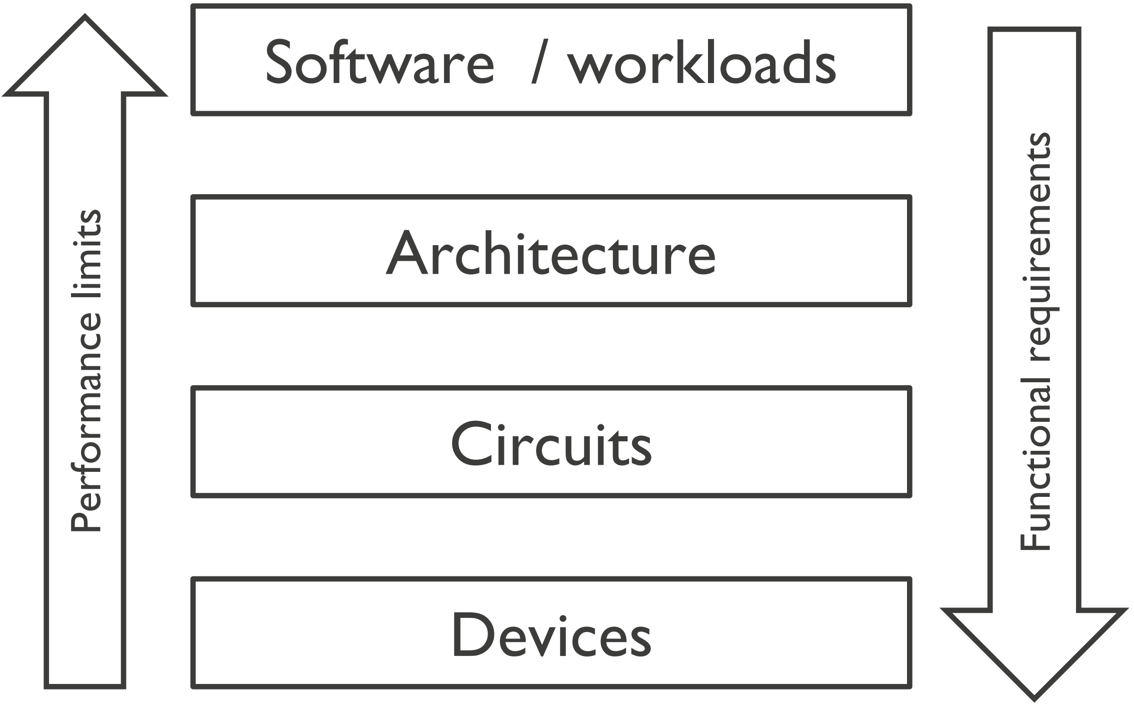
SYSTEM DRIVEN TECHNOLOGY OPTIMIZATION

Kaushik Roy, Edward G. Tiedemann Distinguished Professor of ECE
kaushik@purdue.edu
Arindam Mallik, IMEC
Anand Raghunathan, Silicon Valley Professor of ECE
Sumeet Gupta, Elmore Associate Professor of ECE

PURDUE UNIVERSITY®

Tools of the Trade



Performance limits

Software / workloads

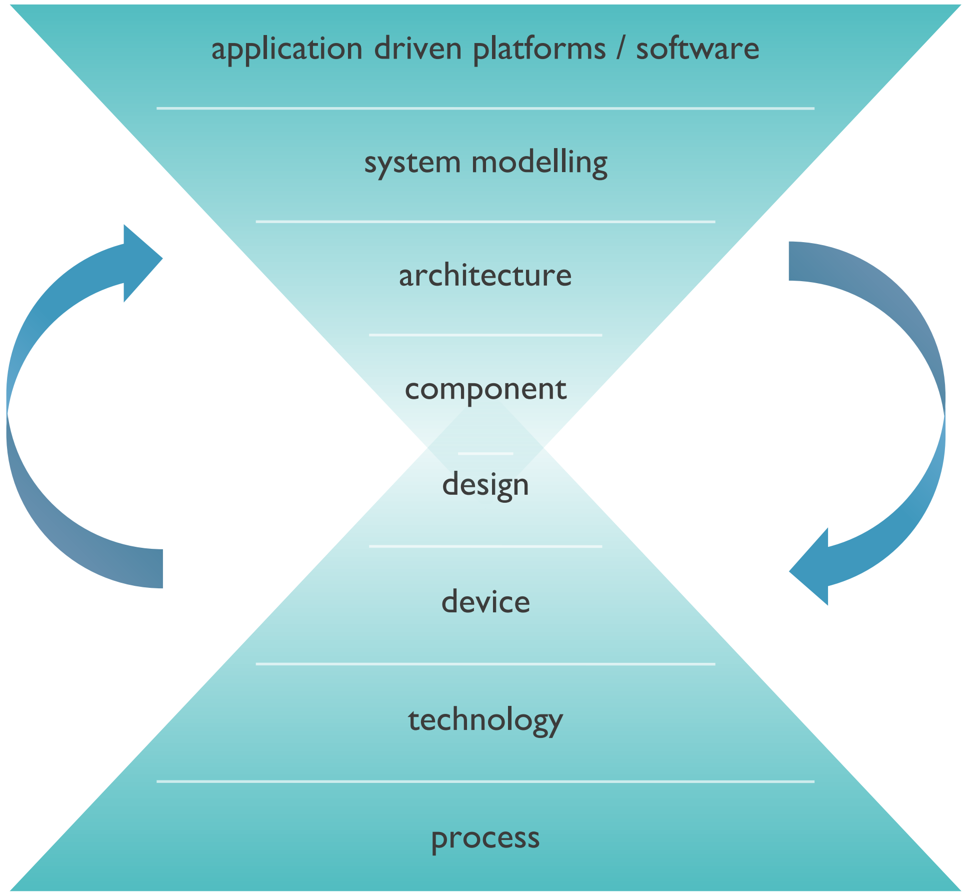
Architecture

Circuits

Devices

Functional requirements

Need for a full stack approach



application driven platforms / software

system modelling

architecture

component

design

device

technology

process

Technology impact might not be observable or significant in a specific architecture, but extremely important in another.

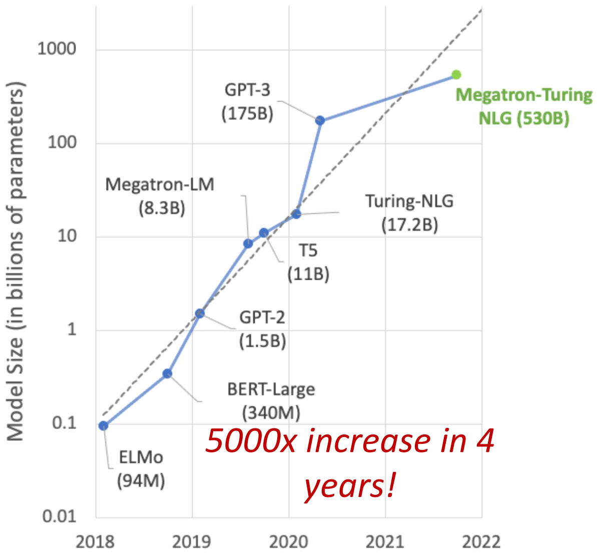
Technology: What, When, Where?

To translate technology improvements to ***workload*** performance, a HW-SW codesign methodology is needed.

Specific Workloads will Drive Demand for Computing and Devices

- AI will continue to set the pace in demand for compute efficiency

NLP model growth

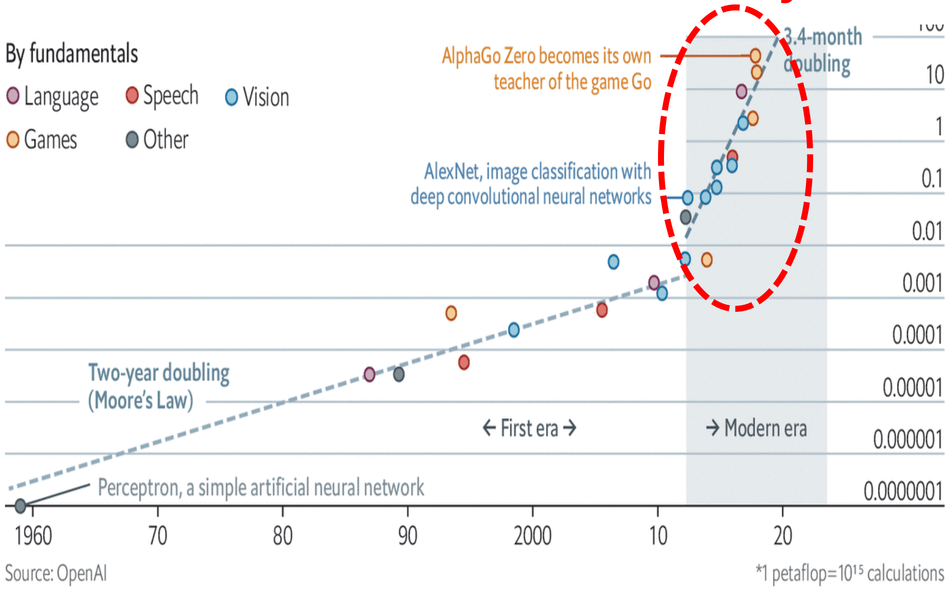


5000x increase in 4 years!

Source: NVIDIA

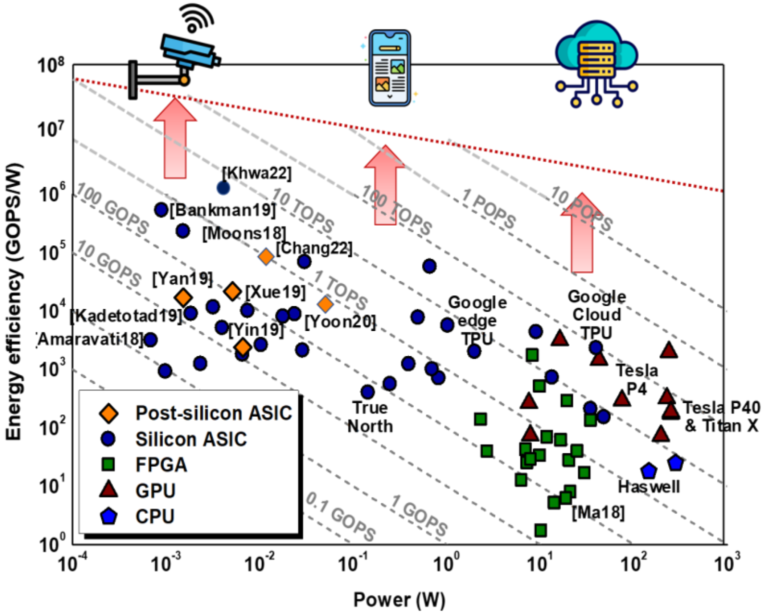
AI Training Compute Demands

10X/year!



Source: The Economist, OpenAI

AI Hardware Landscape

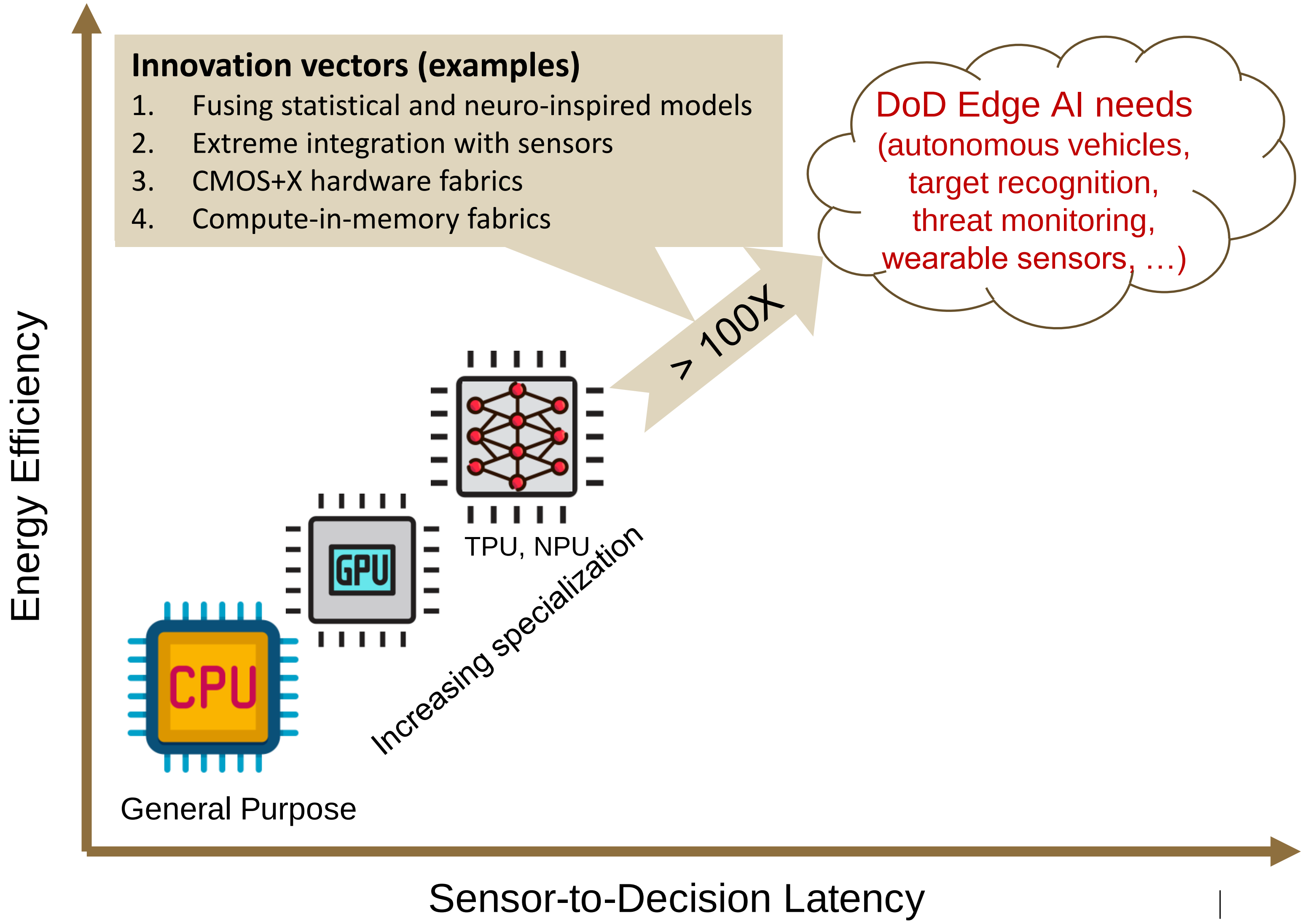


AI Hardware: Goals, Current Landscape

How’s it done today?

- AI is powered by GPUs, TPUs/NPUs
- Most of the improvements are from data-parallel architectures, specialization and high-bandwidth memory

Need for radically different approach: beyond von-Neuman based hardware



Innovation vectors (examples)

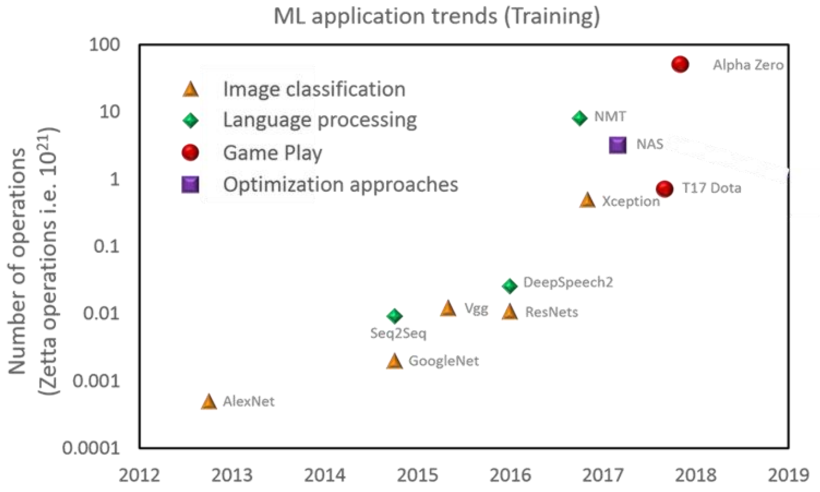
1. Fusing statistical and neuro-inspired models
2. Extreme integration with sensors
3. CMOS+X hardware fabrics
4. Compute-in-memory fabrics

DoD Edge AI needs (autonomous vehicles, target recognition, threat monitoring, wearable sensors, …)

Why co-design is pivotal for AI hardware?

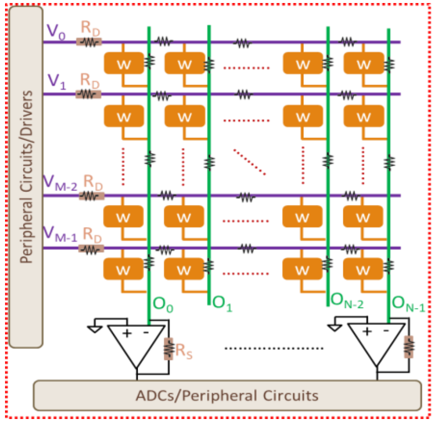
Data-intensive AI workloads

ML application trends (Training)



Enormous energy/performance overheads in traditional von-Neumann architectures

Solution: Computing-in-Memory (CiM) of Matrix Vector Multiplications



Additional design conflicts/complexities due to CiM

- Larger range of CiM output currents (than simple read) → More non-idealities
- Need for complex peripheral circuits

Aggravation of Technology Non-Idealities

- Wire resistivity/resistance
- Process variations

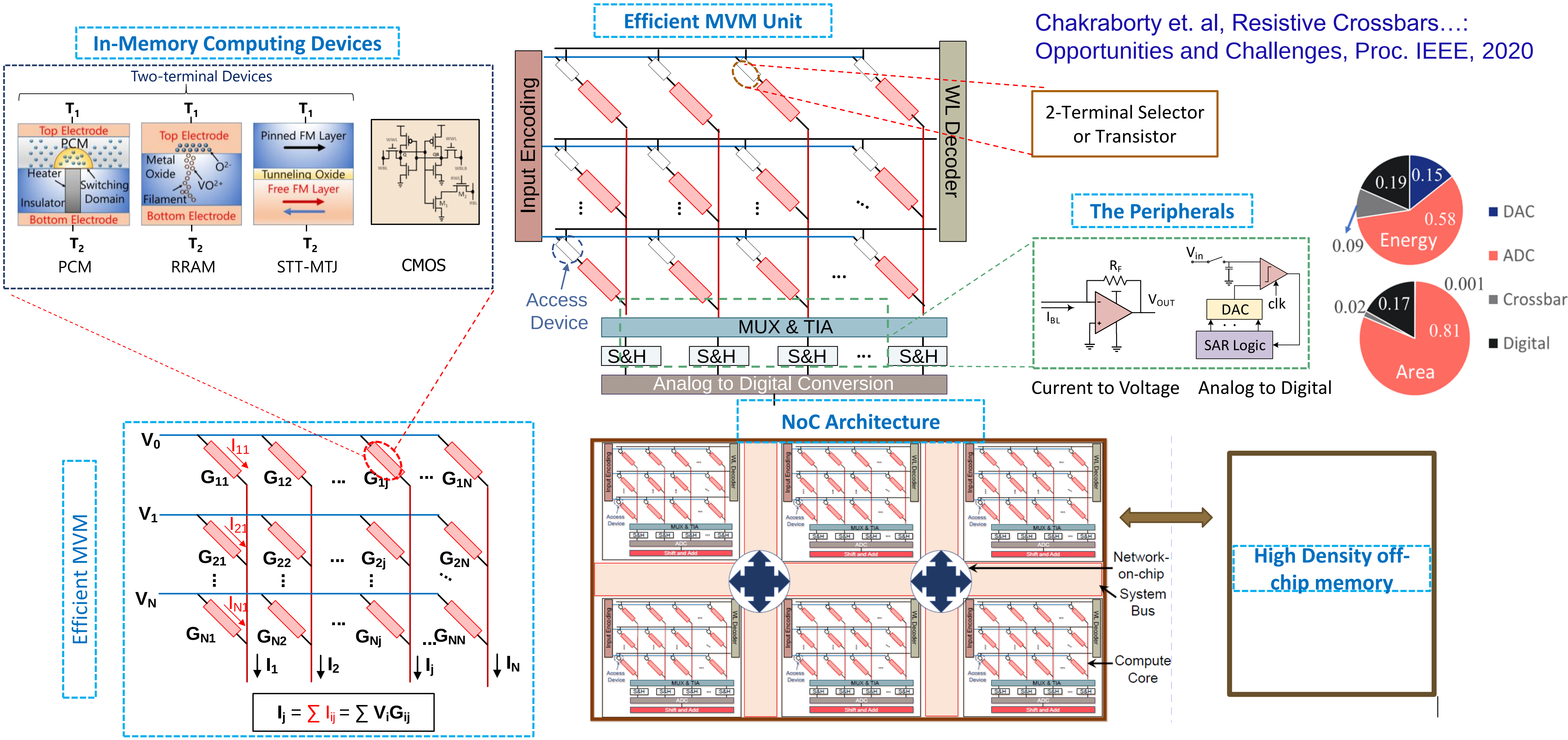
Technology scaling to support high data storage and energy/performance demands of AI hardware

Computational Errors, Energy Inefficiencies, Latency Increase !!

Need a technology-circuit-architecture-algorithm co-design solutions to tackle the increase in challenges

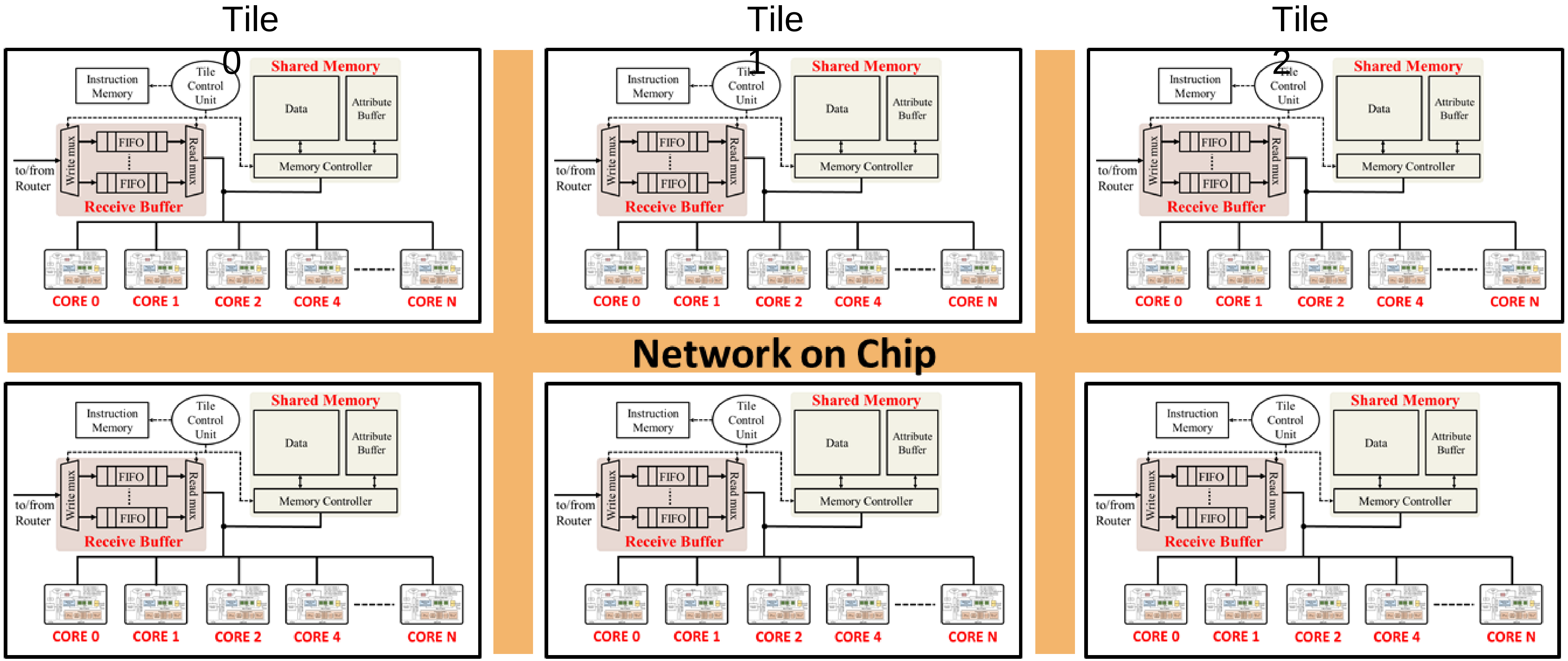
Major Compute Requirement: MMMs

Chakraborty et. al, Resistive Crossbars…: Opportunities and Challenges, Proc. IEEE, 2020



Hardware Architecture

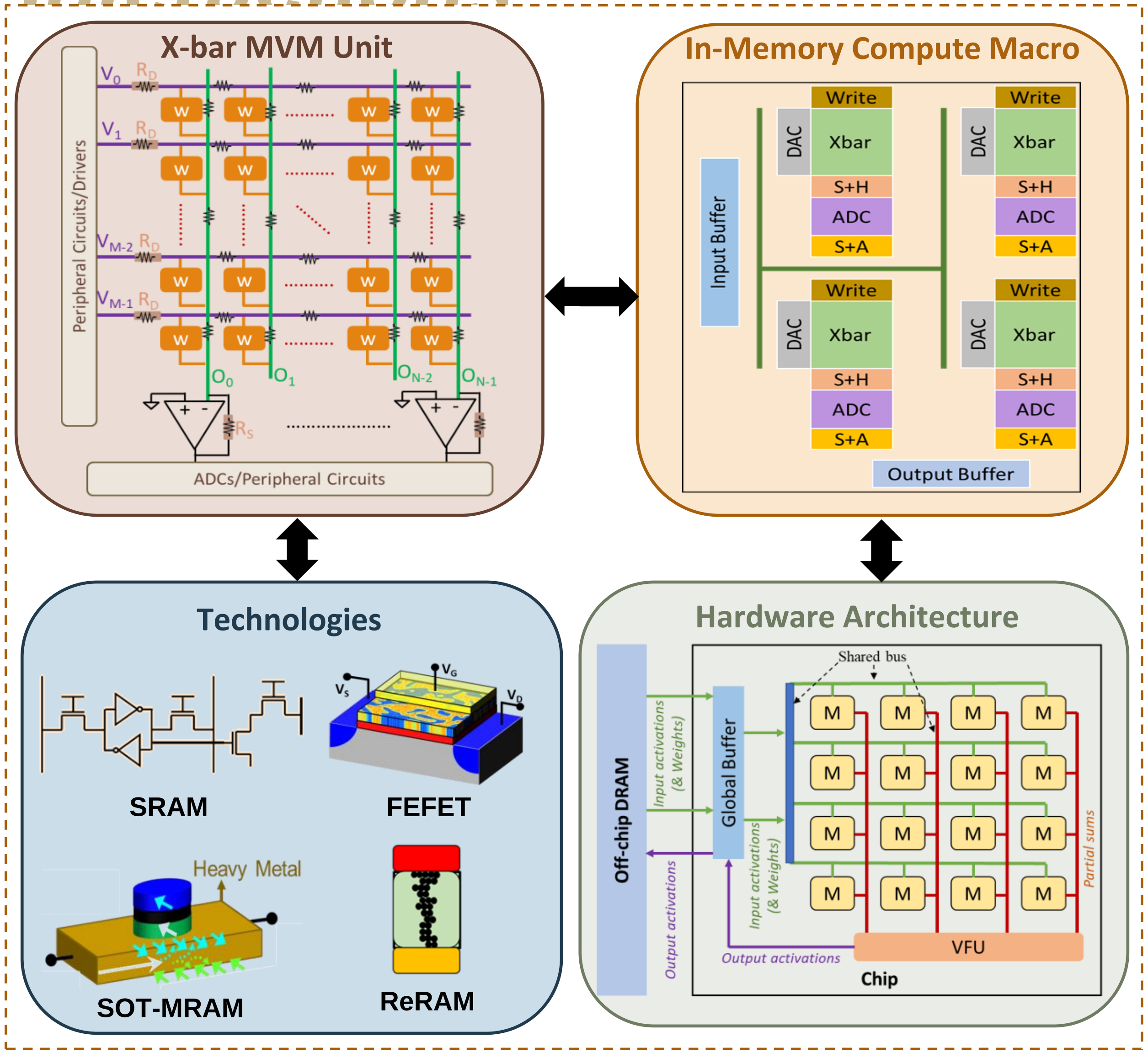
PUMA: A Programmable Ultra-efficient Memristor-based Accelerator for Machine Learning Inference

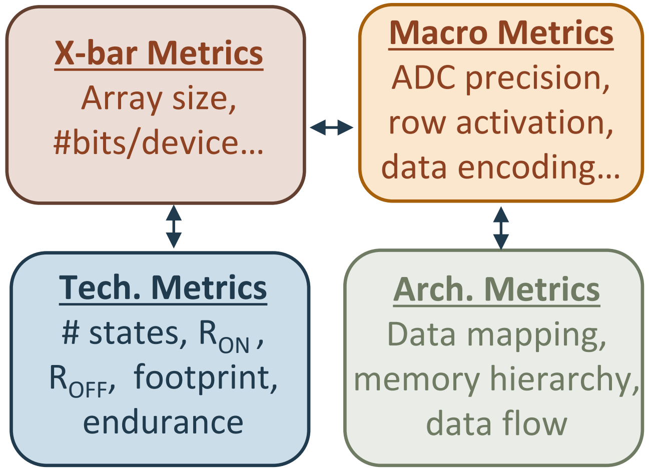


Massively parallel accelerator –> Amenable to Data-Level Parallelism -> Highly efficient ML inference

Ankit et. al, PUMA:…, ASPLOS 2019

AI Hardware Design: Intricate Cross-Layer Interactions





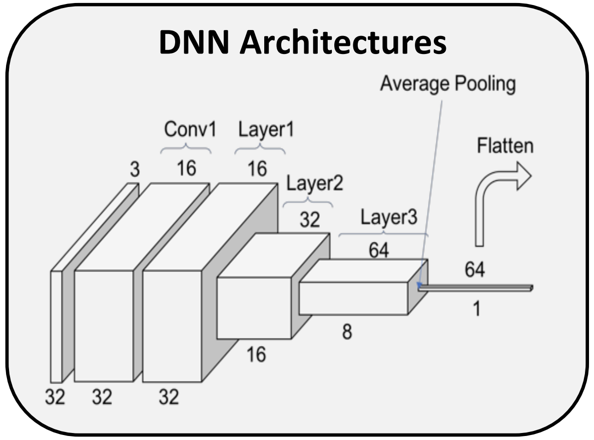
X-bar Metrics
Array size, #bits/device…

Macro Metrics
ADC precision, row activation, data encoding…

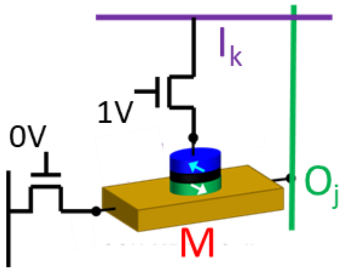
Tech. Metrics
# states, R_{ON}, R_{OFF}, footprint, endurance

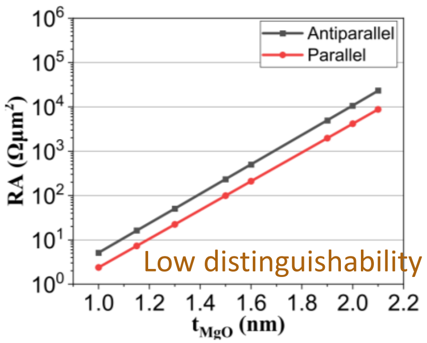
Arch. Metrics
Data mapping, memory hierarchy, data flow

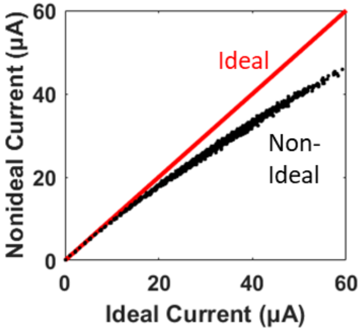
Need to consider the cross-layer optimization techniques to capture the effects of such interactions



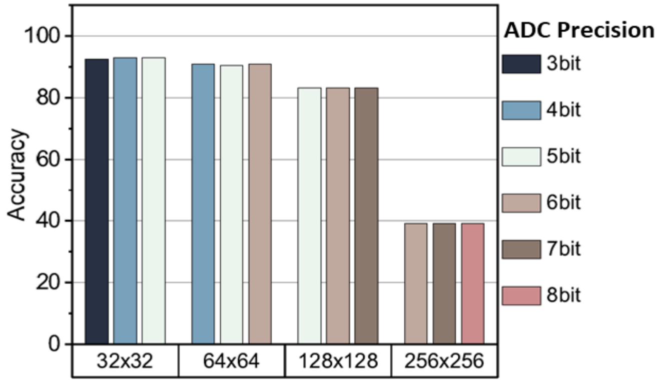
Example: Spin-based Deep Neural Networks



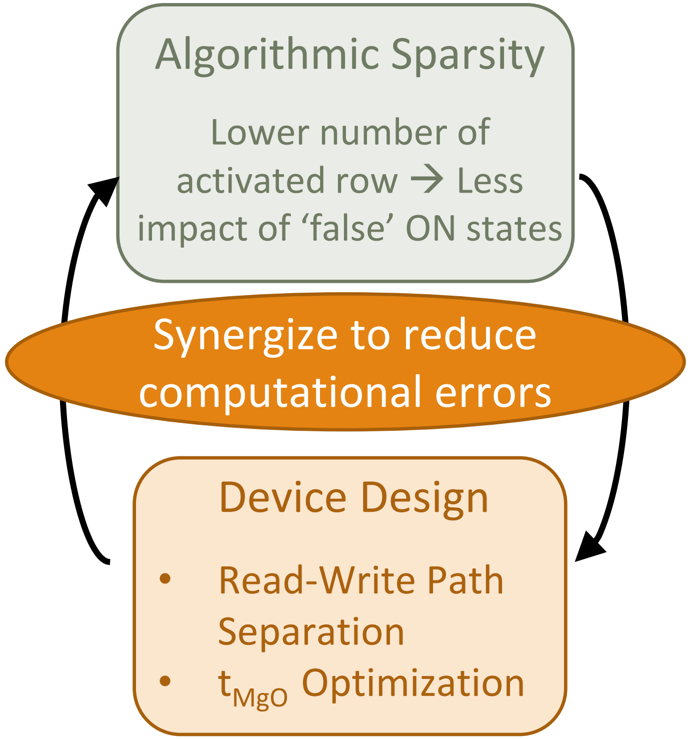




$$NF = \frac{I_{Ideal} - I_{Non-Ideal}}{I_{Ideal}}$$



Algorithm-hardware co-design to enhance system accuracy



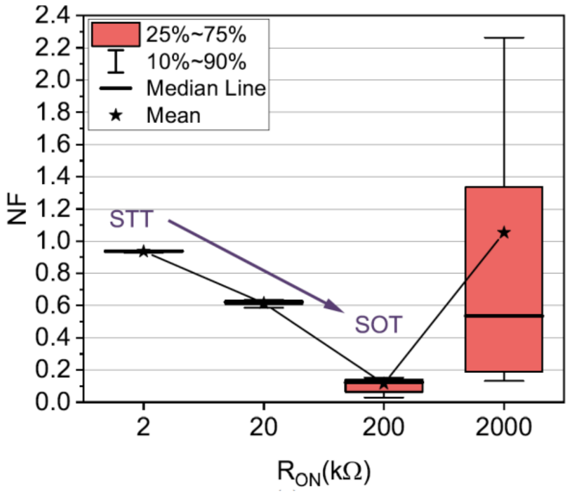
Algorithmic Sparsity

Lower number of activated row → Less impact of ‘false’ ON states

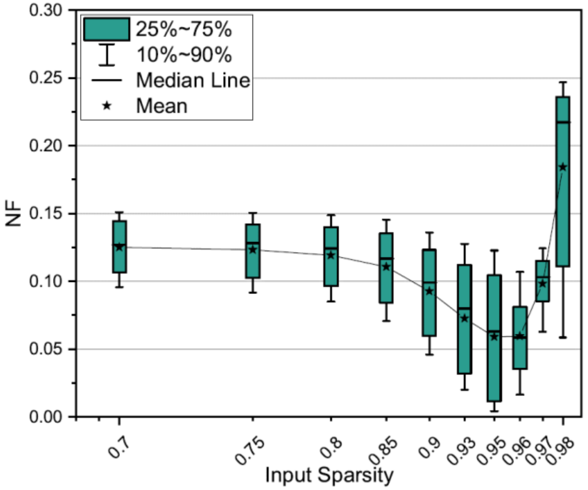
Synergize to reduce computational errors

Device Design

- Read-Write Path Separation
- t_{MgO} Optimization



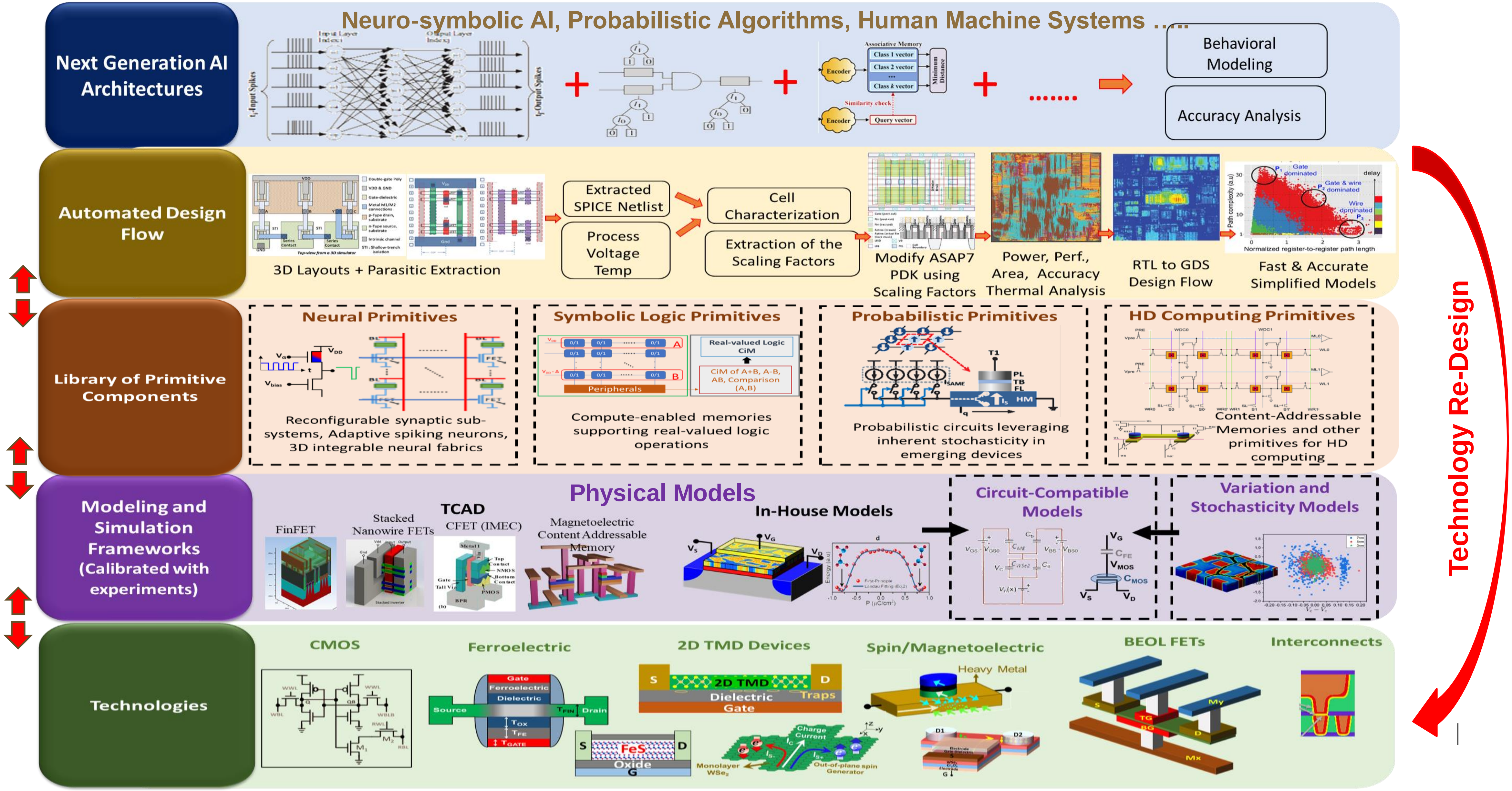
R_{ON} optimization to reduce NF



Algorithm design to reduce NF

T. Sharma, K. Roy *et al*, ISLPED 2021

Cross-layer flow for System-Technology Co-Design



Next Generation AI Architectures

Neuro-symbolic AI, Probabilistic Algorithms, Human Machine Systems …..

Behavioral Modeling

Accuracy Analysis

Automated Design Flow

3D Layouts + Parasitic Extraction

Extracted SPICE Netlist

Process Voltage Temp

Cell Characterization

Extraction of the Scaling Factors

Modify ASAP7 PDK using Scaling Factors

Power, Perf., Area, Accuracy Thermal Analysis

RTL to GDS Design Flow

Fast & Accurate Simplified Models

Library of Primitive Components

Neural Primitives

Reconfigurable synaptic sub-systems, Adaptive spiking neurons, 3D integrable neural fabrics

Symbolic Logic Primitives

Compute-enabled memories supporting real-valued logic operations

Probabilistic Primitives

Probabilistic circuits leveraging inherent stochasticity in emerging devices

HD Computing Primitives

Content-Addressable Memories and other primitives for HD computing

Modeling and Simulation Frameworks (Calibrated with experiments)

Physical Models

FinFET

Stacked Nanowire FETs

TCAD

CFET (IMEC)

Magnetoelectric Content Addressable Memory

In-House Models

Circuit-Compatible Models

Variation and Stochasticity Models

Technologies

CMOS

Ferroelectric

2D TMD Devices

Spin/Magnetoelectric

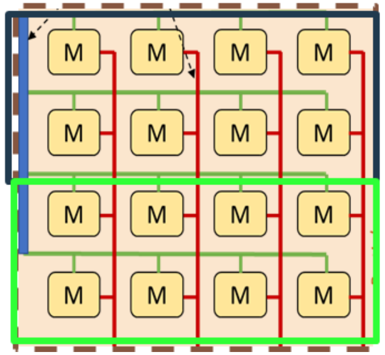
BEOL FETs

Interconnects

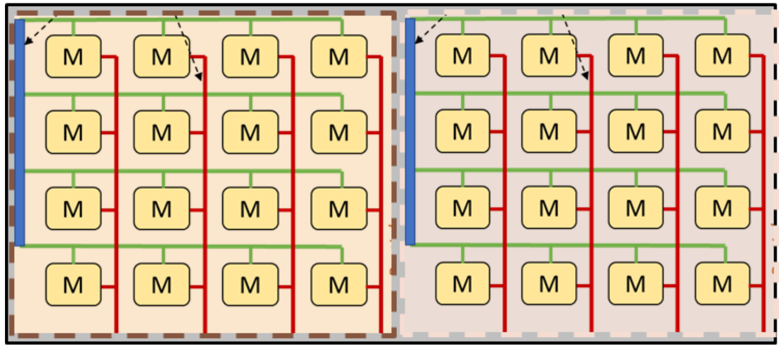
Technology Re-Design

Technology Exploration/Selection

Area required for fully mapping the workload



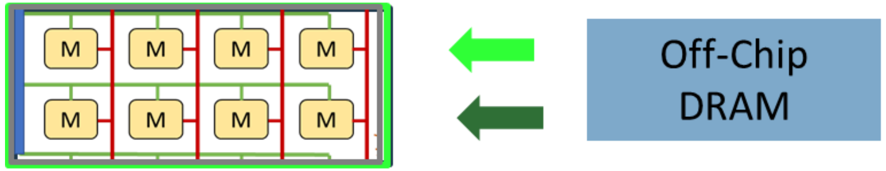
Large area budget



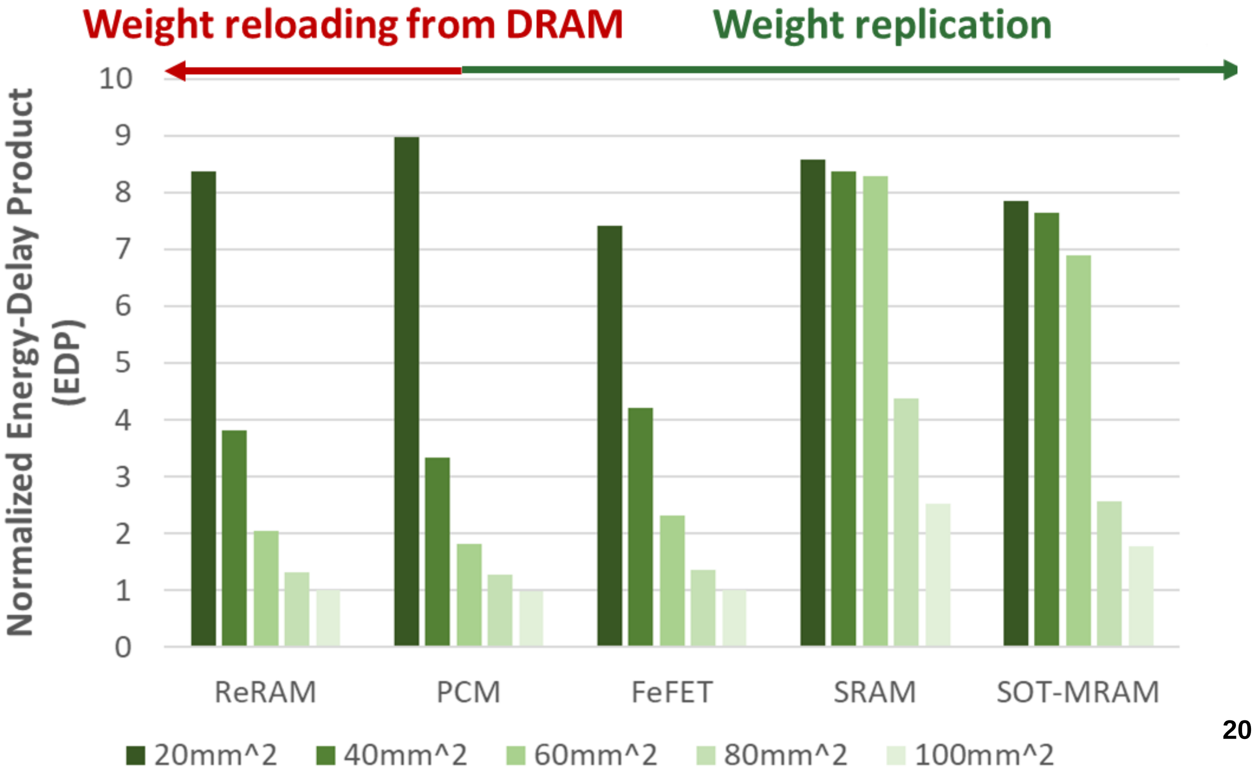
Weight replication (Increases parallelism)

He, Roy et al, ICCAD 2022

Small area budget



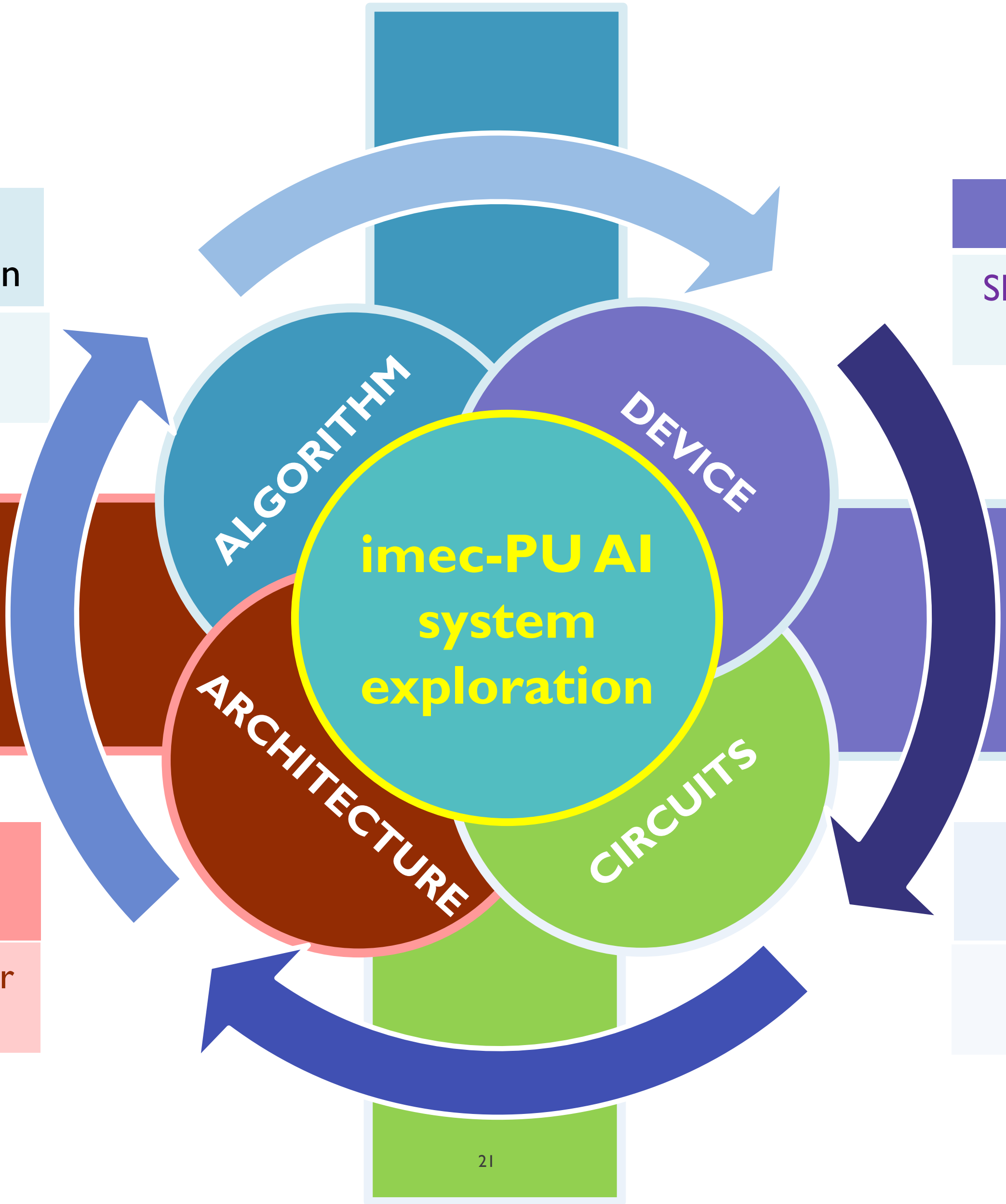
Weight reloading (Needs DRAM access)



HW-SW Codesign for future AI compute systems

Application and algorithm optimization
Low-precision neural network exploration

LLM,
neuromorphic exploratory algorithm



Use/drive emerging technologies

SRAM, STT/SOT-MRAM, Flash, FeFET, IGZO, PCM

System architecture exploration
Digital hardware acceleration

CiM accelerator component, middleware for accelerator integration in the system

Compute-in-Memory (CiM)
macro using emerging memories

1T1R, 2T1R, 2T1C, 2T0C cells for CiM macro

THANK YOU!

PURDUE
UNIVERSITY®