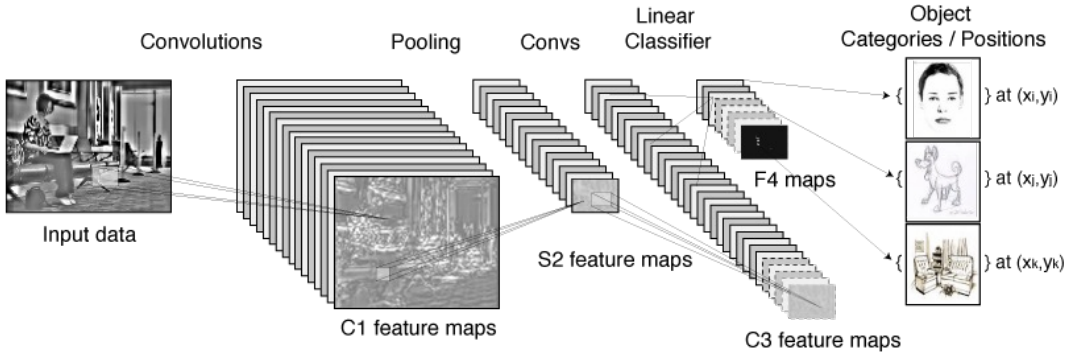
Research Overview of Energy-Efficient Multimedia Systems Group

Vivienne Sze

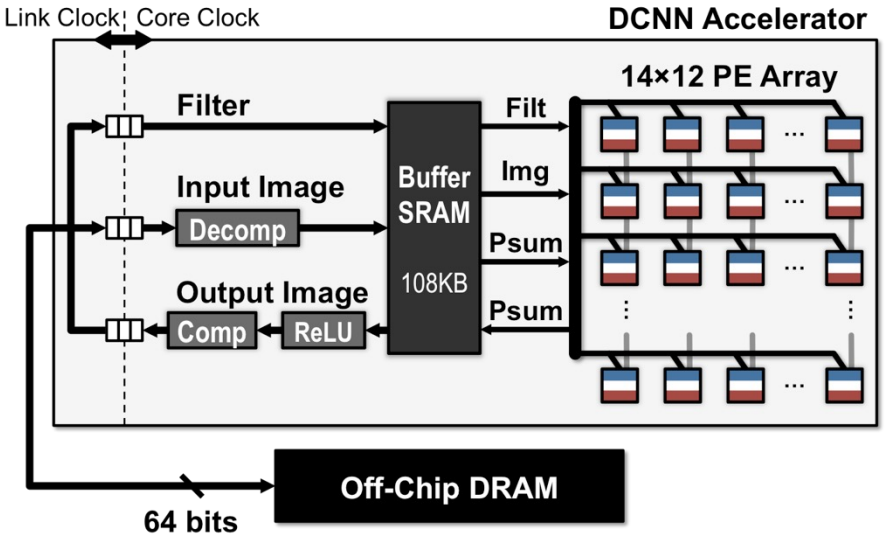
Efficient Computing with Cross-Layer Design

Algorithms

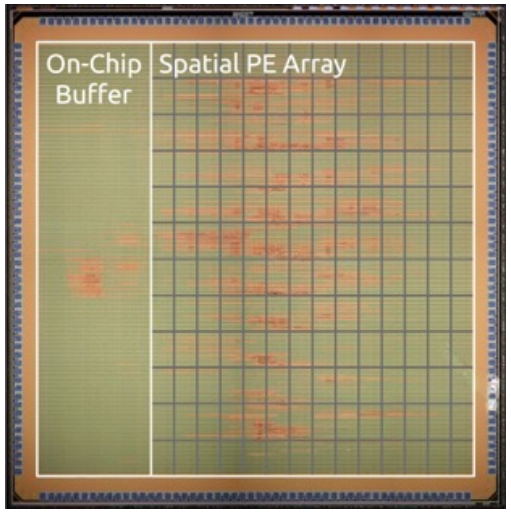


Systems

Architectures

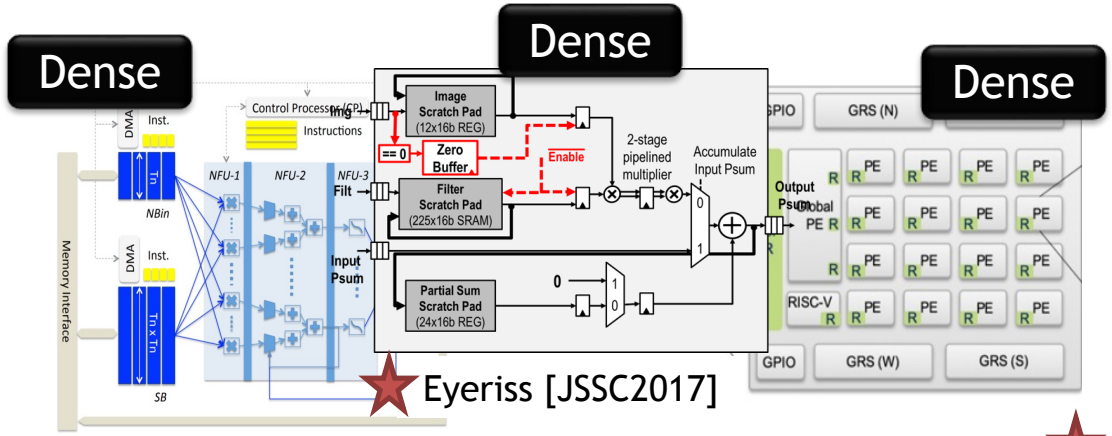


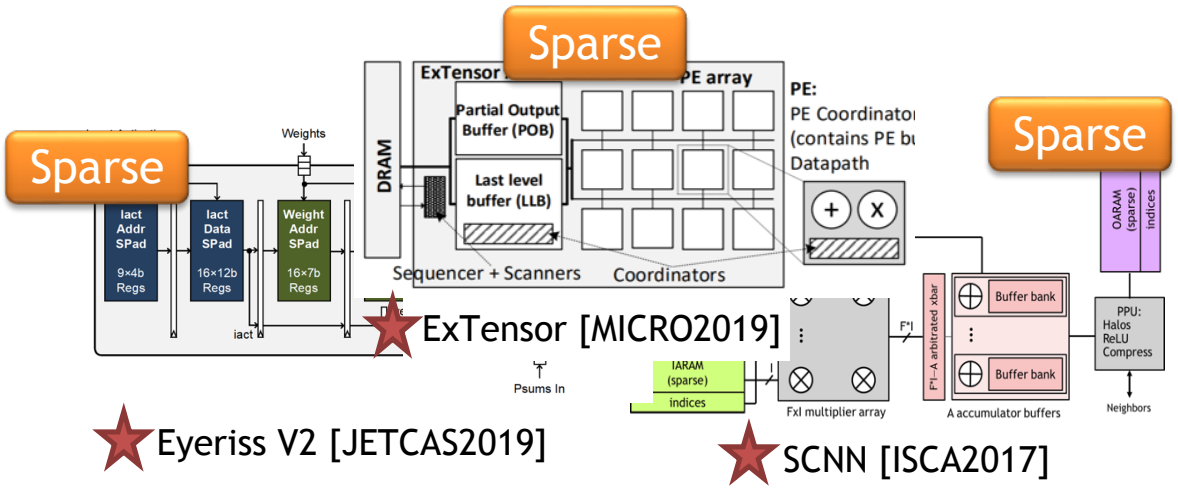
Circuits



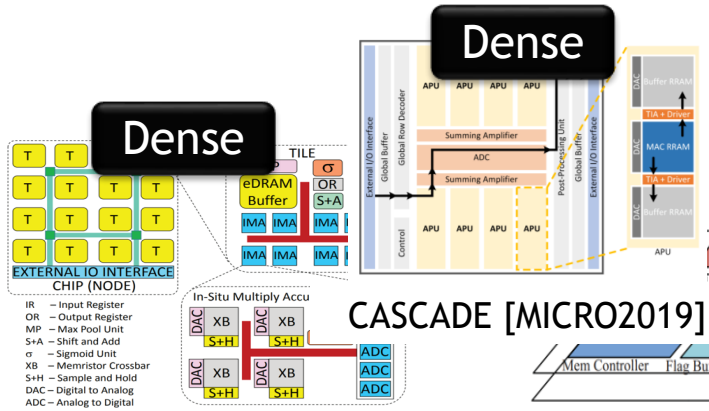
Design Space for DNN and Tensor Accelerators

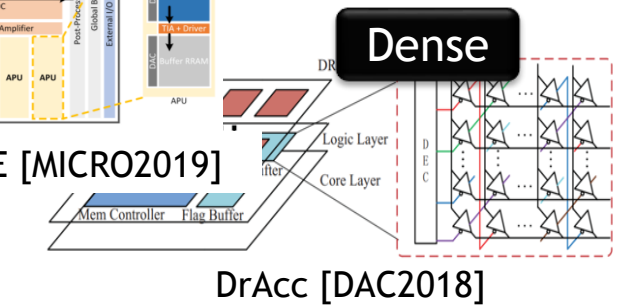
Digital-Compute Accelerators Design

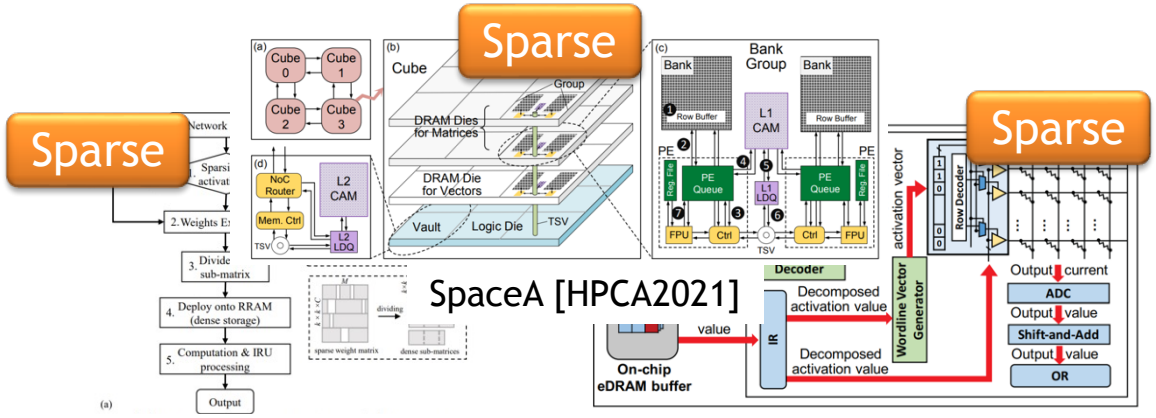




Analog-Compute (CiM) Accelerators Design

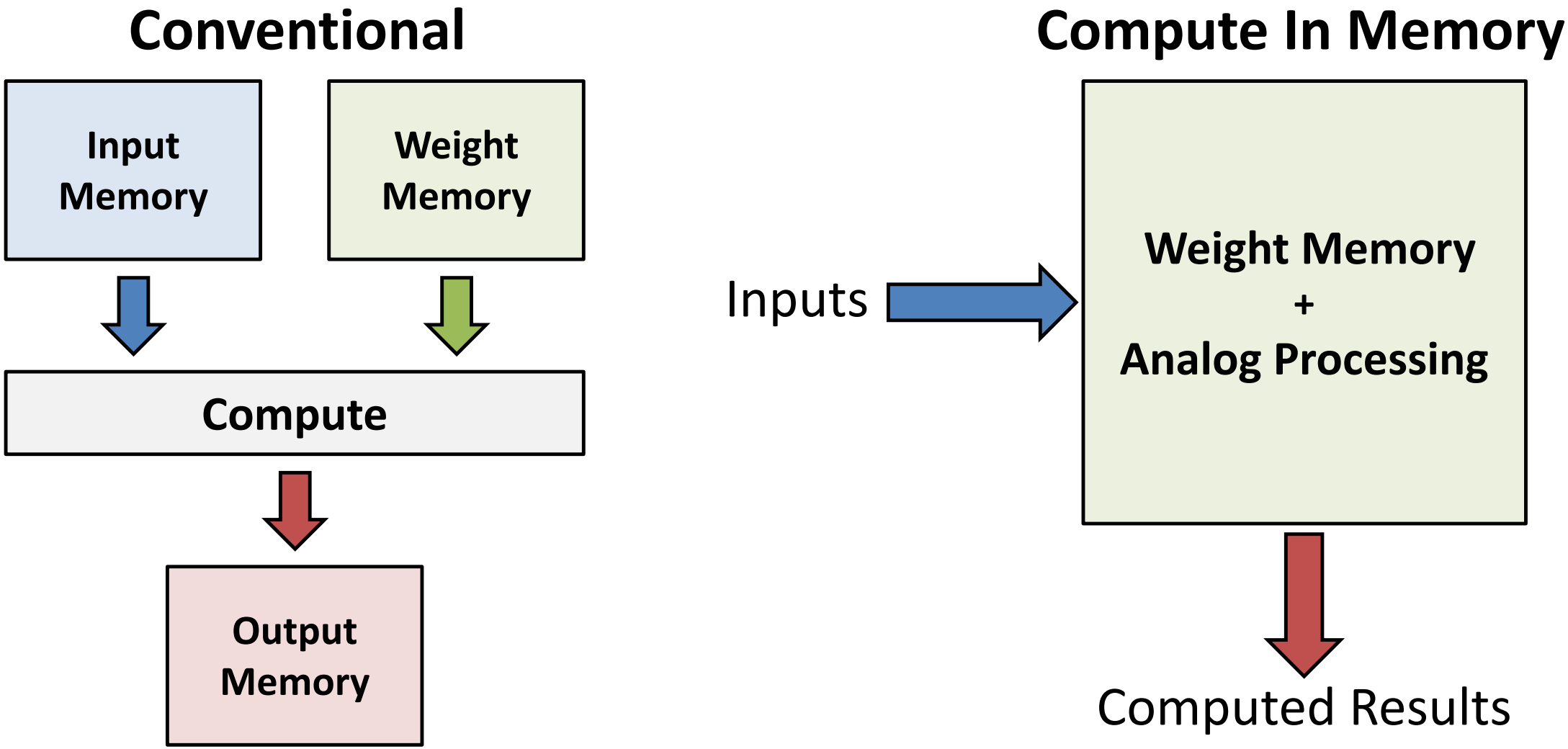






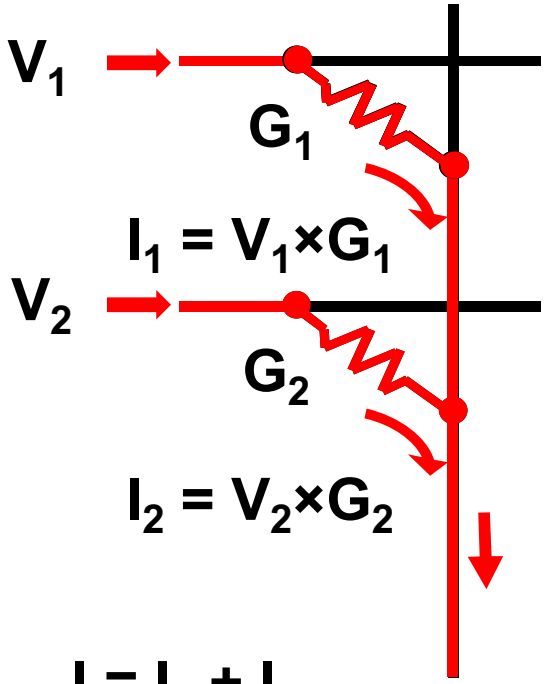
Works we contributed to

Compute In Memory (CiM) Accelerators



Compute In Memory

Activation is input voltage (V_i)
Weight is resistor conductance (G_i)



Psum is output current

$$I = I_1 + I_2 = V_1 \times G_1 + V_2 \times G_2$$

Image Source: [**Shafiee**, *ISCA* 2016]

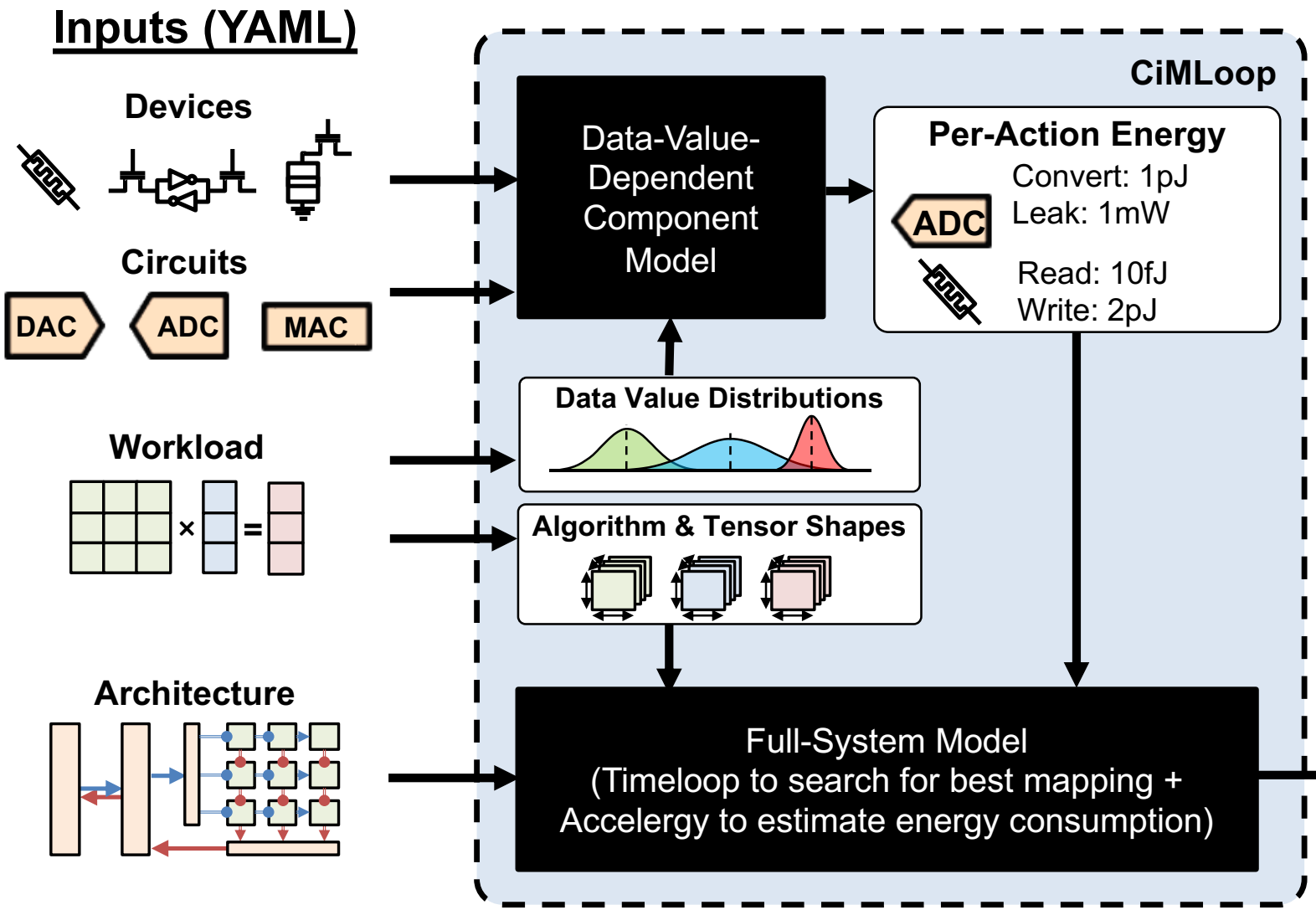
- Reduce data movement by **moving compute into memory**
- Compute MAC with memory storage element
- **Analog Compute**
  - Activations, weights and/or partial sums are encoded with analog voltage, current, or resistance
  - Increased sensitivity to circuit non-idealities
  - A/D and D/A circuits to interface with digital domain
- Leverage **emerging memory device technology**

CiM Research Spans Full Stack

- **Devices:** The components forming each memory cell (e.g., SRAM, DRAM, ReRAM, STT-RAM)
- **Circuits:** The components performing computation, analog/digital conversion, storage, data movement, and other actions
- **Architecture:** The organization of components into a larger system (e.g., the number of each component and how components are connected)
- **Workload:** The DNN to be processed, which we model as a series of extended-Einsum operations with tensors of varying shapes and values
- **Mapping:** The temporal and spatial scheduling of the workload onto the system

Need for modeling tool to enable apple-to-apple comparison and design space exploration → **CiMLoop**

CiMLoop: A Flexible, Accurate, and Fast CiM Modeling Tool



CiMLoop is built on the Timeloop [**Parashar**, *ISPASS* 2019] + Accelergy [**Wu**, *ICCAD* 2019] Infrastructure

Code available at https://github.com/mit-emze/cimloop

Outputs

Energy

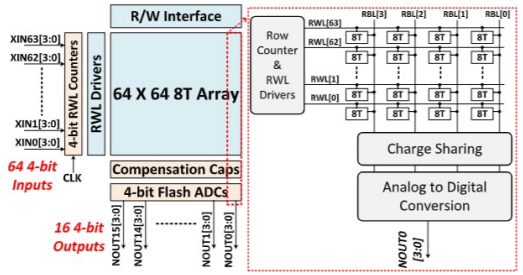
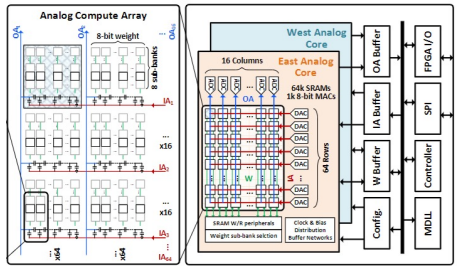
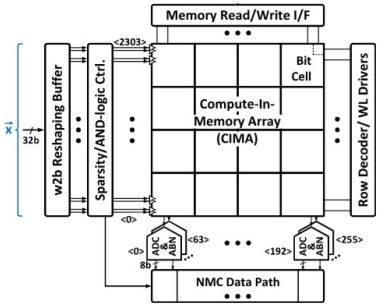
Area

Throughput

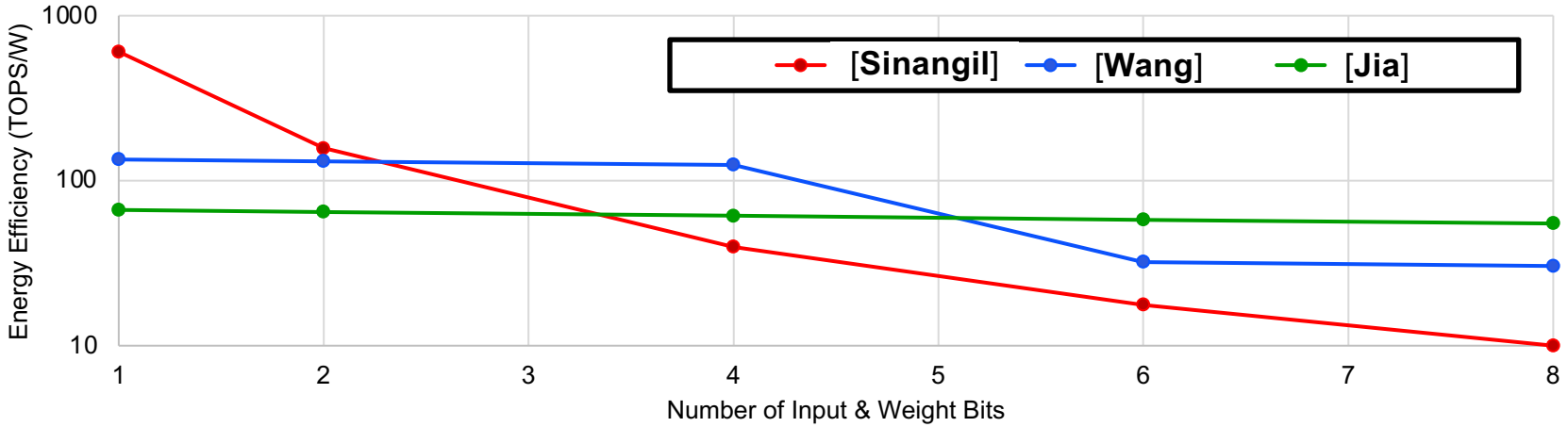
CiMLoop: A Flexible, Accurate, and Fast CiM Modeling Tool

- **Flexibility**
  - A flexible specification that lets users describe, model, and map workloads to both circuits and architecture
- **Accuracy**
  - A data-value-dependent energy model that captures the interaction between DNN operand values, data representations, and analog/digital values
  - ***Estimated values are within 8% of values reported for measured designs***
- **Speed**
  - A fast statistical model to enable for constant runtime w.r.t. number of components and amortizes overhead across mappings
  - ***Enables orders-of-magnitude speed up relative to other high-accuracy models***

Example: Apples-to-Apples Comparison

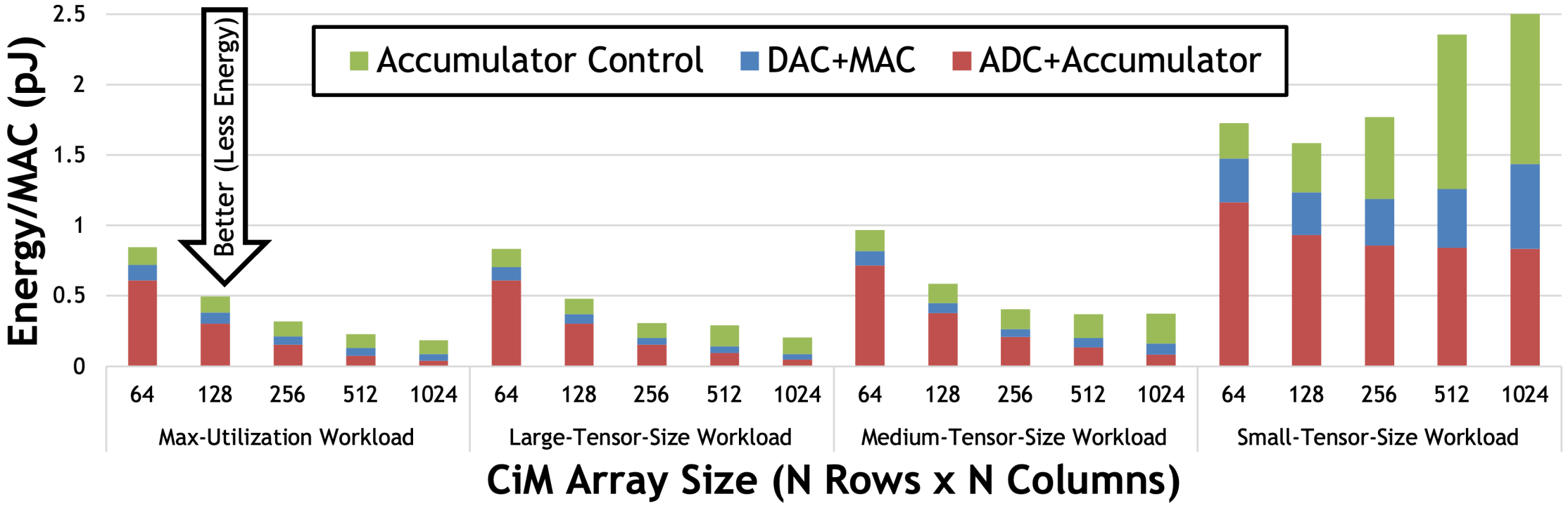
	Macro	Macro	Macro
			
	[**Sinangil**, *JSSC* 2021]	[**Wang**, *VLSI* 2022]	[**Jia**, *JSSC* 2020]
Technology Node	7nm	22nm	65nm
ADC Type	4b Flash	8b SAR	8b SAR
Memory Device	6T SRAM	8T SRAM + Capacitor	6T SRAM

Compare Designs:
Same technology, ADC, device for all macros



[**Andrulis**, *ISPASS* 2024]

Example: Design Space Exploration

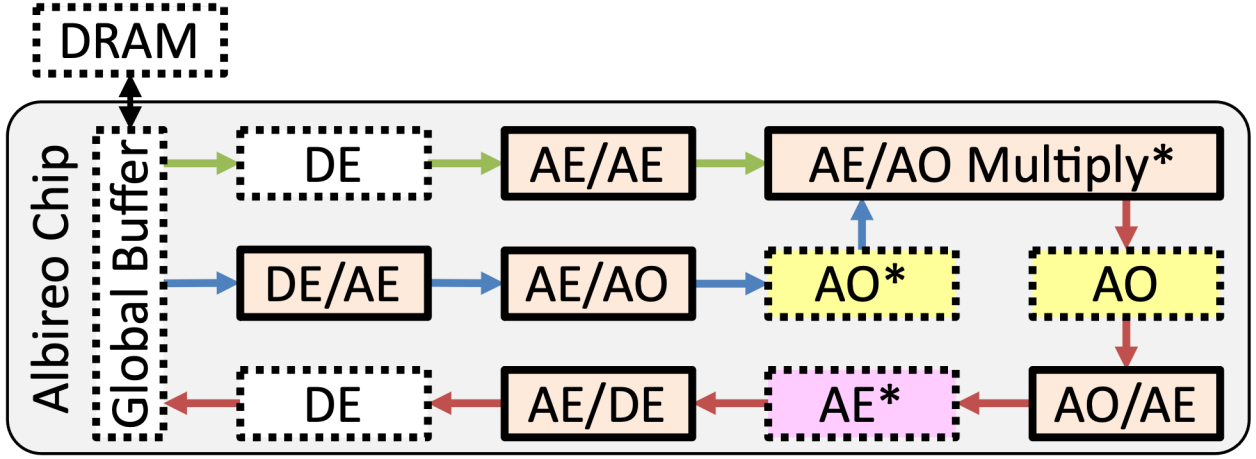


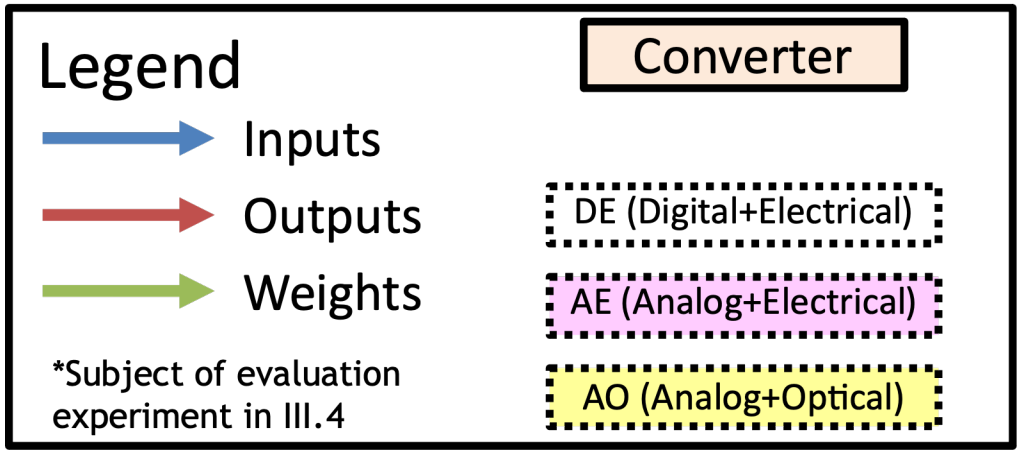
Explore array size (architecture) and DNN shapes (workload)

[**Andrulis**, *ISPASS* 2024]

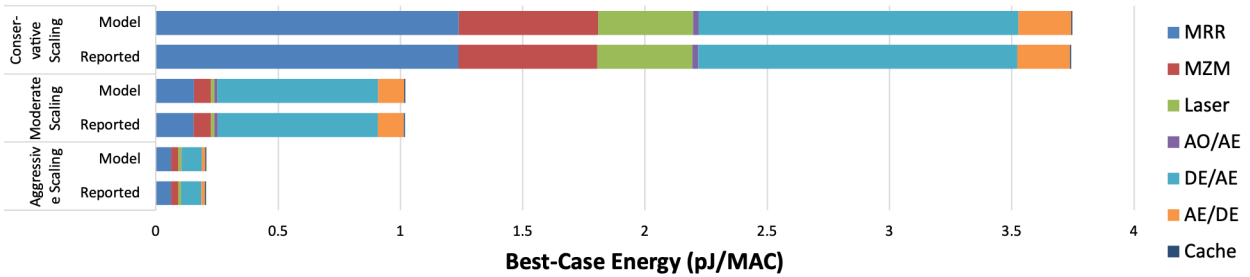
In collaboration with Joel Emer

CiMLoop for Photonic Accelerator Modeling

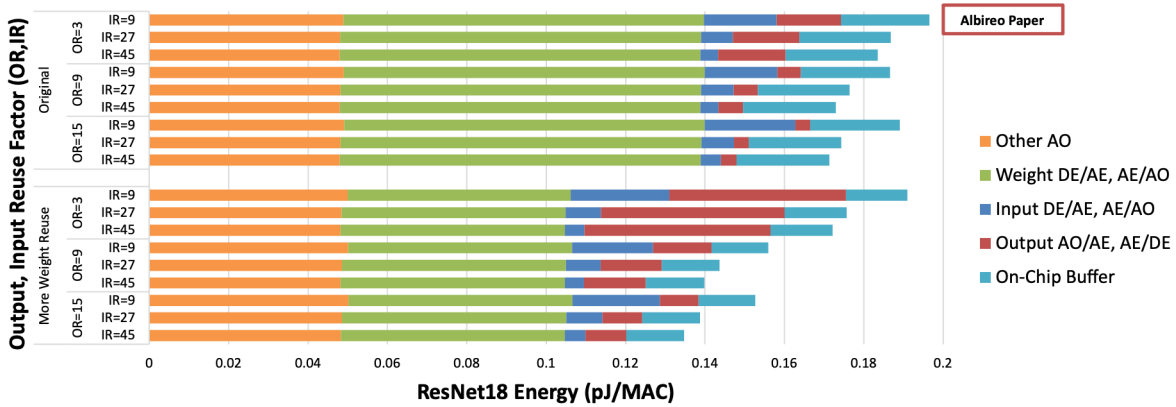




Validation



Design Space Exploration



Many similarities in design of CiM and Photonic accelerators → Can model with **CiMLoop!**