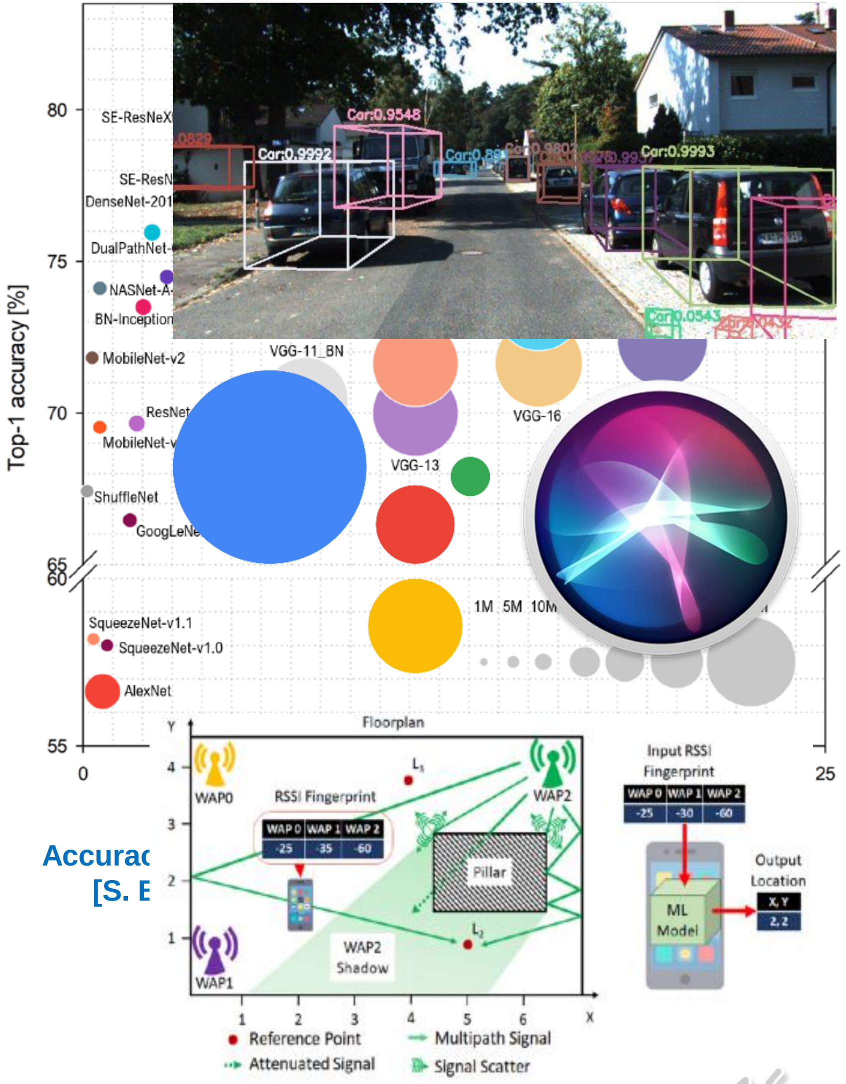
27th Asia and South Pacific Design Automation Conference (ASPDAC)
Jan 17-20, 2022

SONIC: A Sparse Neural Network Inference Accelerator with Silicon Photonics for Energy-Efficient Deep Learning

Febin P. Sunny, Mahdi Nikdast, Sudeep Pasricha
Department of Electrical and Computer Engineering
Colorado State University, Fort Collins, CO, USA
{febin.sunny, mahdi.nikdast, sudeep}@colostate.edu

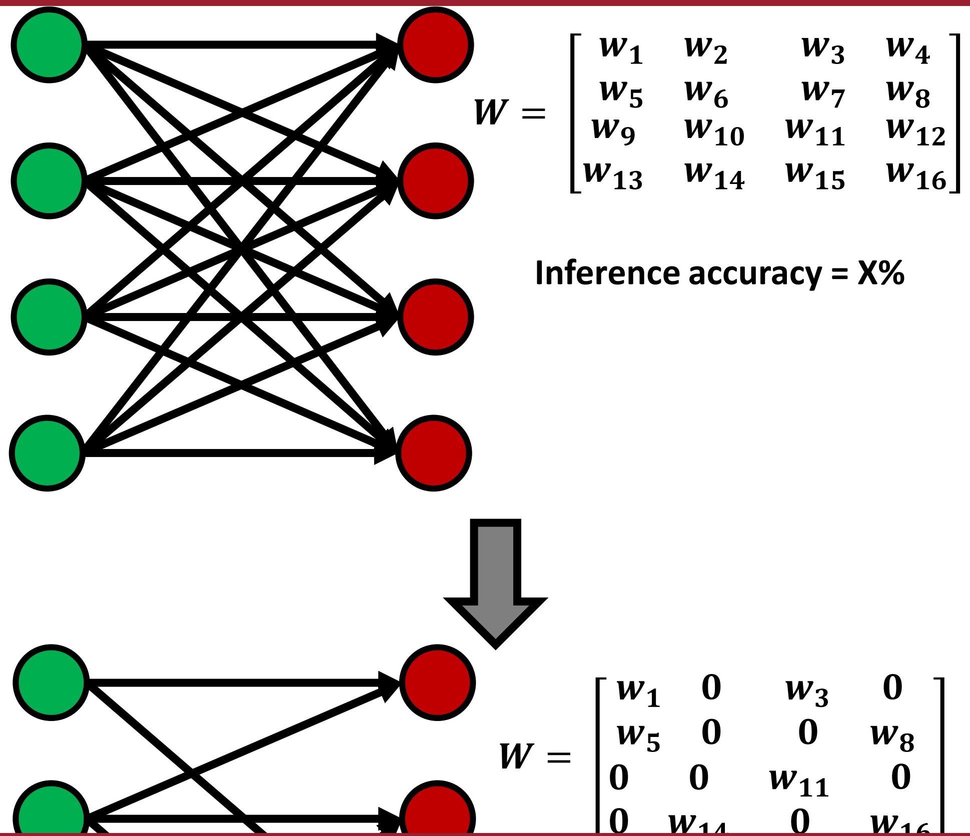
Emerging ML Applications

- **ML applications are becoming increasingly complex**
- **Some examples:**
  - **Object detection in autonomous vehicles**
    - **J. Dey, W. Taylor, and S. Pasricha, , "VESPA: Optimizing Heterogeneous Sensor Placement and Orientation for Autonomous Vehicles", *IEEE Consumer Electronics, Mar 2021.***
  - **Natural language processing**
    - **Google Assistant, Apple's SIRI**
  - **Deep learning models and optimizations for IoT applications**
    - **S. Tiku and S. Pasricha, "Overcoming Security Vulnerabilities in Deep Learning Based Indoor Localization on Mobile Devices", *ACM TECS, Jan 2020.***
- **Inference acceleration is becoming crucial**
  - **For energy- and resource-constrained platforms executing real-time embedded and IoT applications**
- **Domain-specific ML hardware accelerators are preferred**
  - **Provide energy and throughput benefits over GPUs and CPUs**



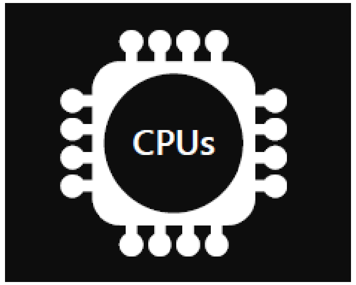
Sparsity in Neural Networks

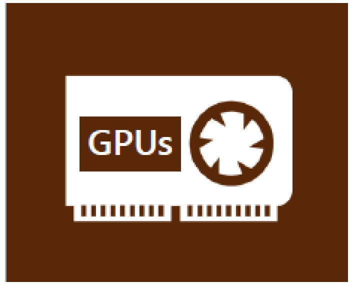
- **Sparsity in neural networks refers to presence of weight parameters which are zero valued**
  - **Allows for reducing number of neurons and synapses without impacting accuracy**
  - **An optimization method to reduce resource requirements, while maintaining good inference accuracy**
- **Sparse Neural Networks (SpNNs) should enable lower memory and computation requirement, BUT:**
  - **To obtain the reduced computation benefits from SpNNs, specialized hardware is necessary**
  - **Optimizations for Dense Neural Network (DNNs) does not make use of available sparsity in SpNNs**

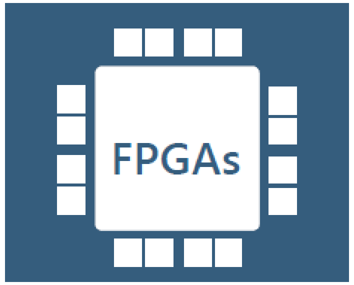


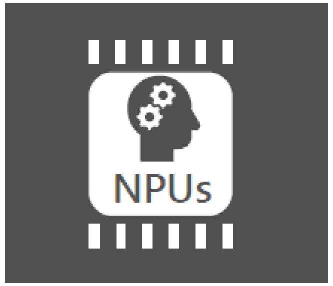
There is a need for specialized hardware to accelerate SpNNs while making use of the available sparsity

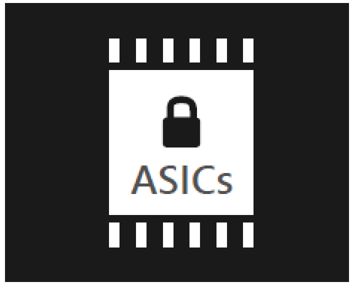
ML Acceleration Hardware



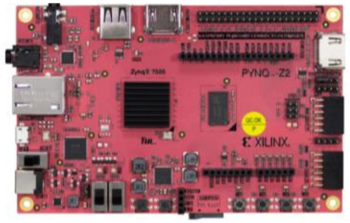


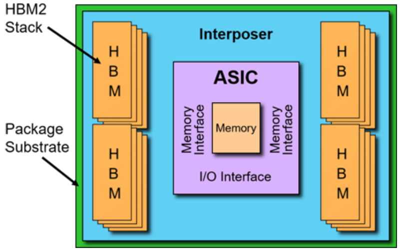






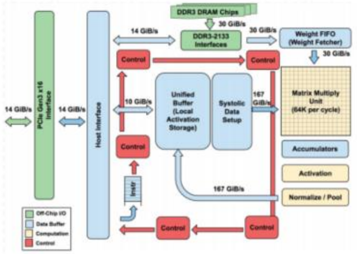
FLEXIBILITY ⟷ EFFICIENCY

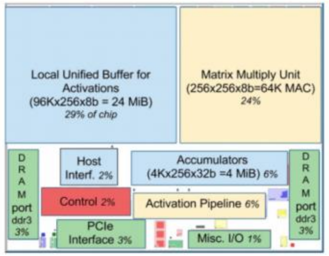




TU102 – TITAN RTX
18.6 BILLION TRANSISTORS

SM	72
CUDA CORES	4608
TENSOR CORES	576
RT CORES	72
GEOMETRY UNITS	36
TEXTURE UNITS	288
ROP UNITS	96
MEMORY	384-bit 7 GHz GDDR6
NVLINK CHANNELS	2



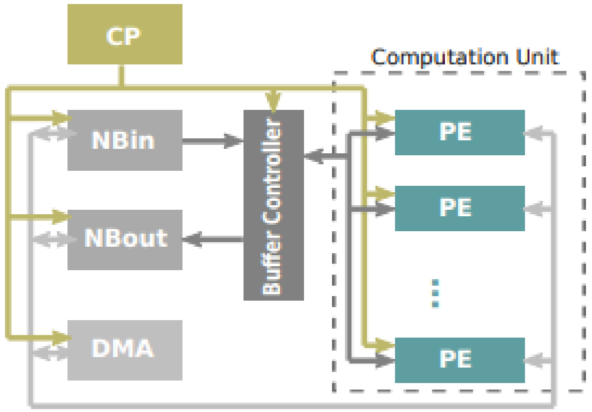


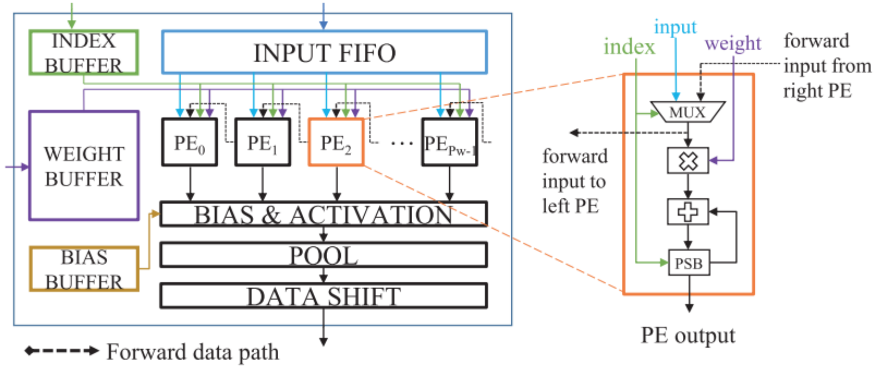
SpNN Accelerators

- **S. Zhang et al., in MICRO, 2016**
  - Electronic SpNN accelerator with custom instruction set architecture (ISA)
  - Specialized buffer controllers with indexing to keep track of sparse elements

- **W. You, C. Wu, in IEEE Access, vol. 9, Jan. 2021**
  - Software-hardware co-optimized reconfigurable sparse CNN accelerator
    - Proposed for FPGAs
  - Exploited both inter- and intra-output feature map parallelism
  - Kernel merging along with structured sparsity considered to improve overall efficiency

- **A. Aimar et al., in IEEE Trans. Neural Netw. Learn. Syst, vol. 30, no. 3, Mar. 2019**



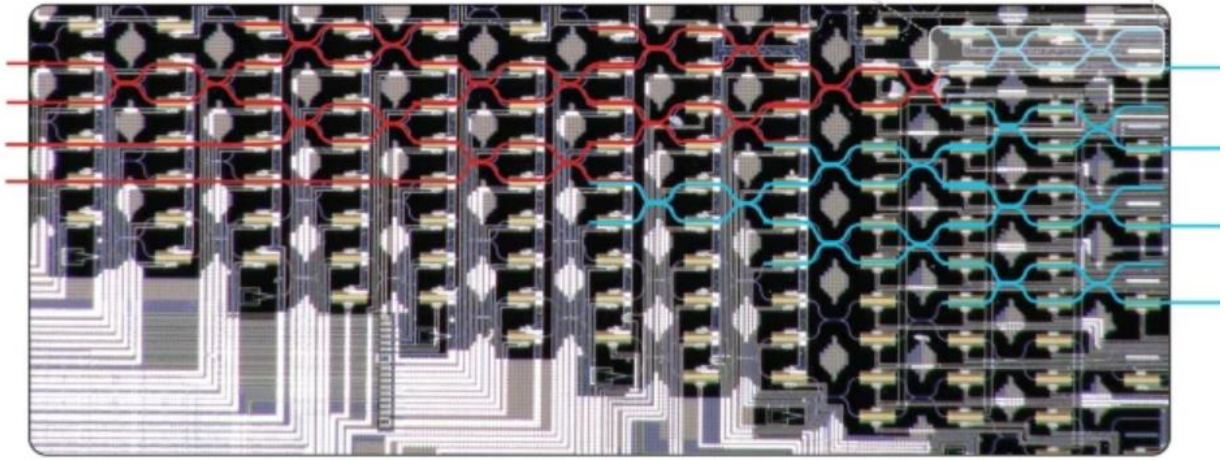


Nullhop CNN Accelerator

But electronic accelerators face the fundamental limitations due to slowdown of Dennard scaling

Silicon Photonics for ML Acceleration

- **Large fan-in and fan-out possible for linear algebra processing**
- **Energy efficient data transfer by using optical transceivers**
- **Low power or passive implementations of complex operations**
  - **Low power multiplication**
  - **Passive Fourier transforms**
- **Fast rate of operation**
  - **Theoretical limit 100 GHz (photodetection rate)**
  - **Lower latency than electronic processing**



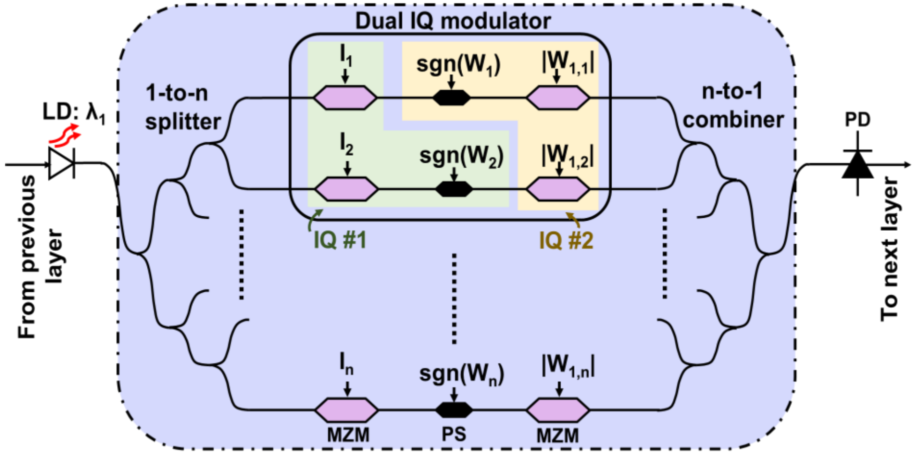
Silicon nano-photonic ANN accelerator prototype from Y. Shen et al., Nature Photonics, 2017

Silicon Photonic ANN accelerators have the potential to overcome the limitations electronic accelerators face due to Dennard scaling

Lightmatter ENVISE for general purpose AI inference acceleration

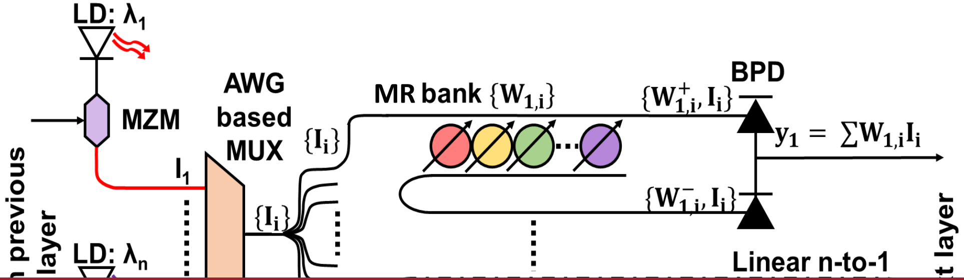
Computation Using Photonics

- **Option 1: Coherent computation**
  - Single wavelength
    - Weights represented using electrical field amplitude
  - Challenges
    - Scalability issues
    - Phase encoding noise
    - Phase error accumulation



LD: laser diode; PD: Photo diode; MZM: Mach Zehnder Modulator;
IQ modulation: In-phase and Quadrature modulation

- **Option 2: Noncoherent computation**
  - Multiple wavelengths used simultaneously
  - Phase-change in devices used to imprint weight/activation values on signal intensities
  - Advantages

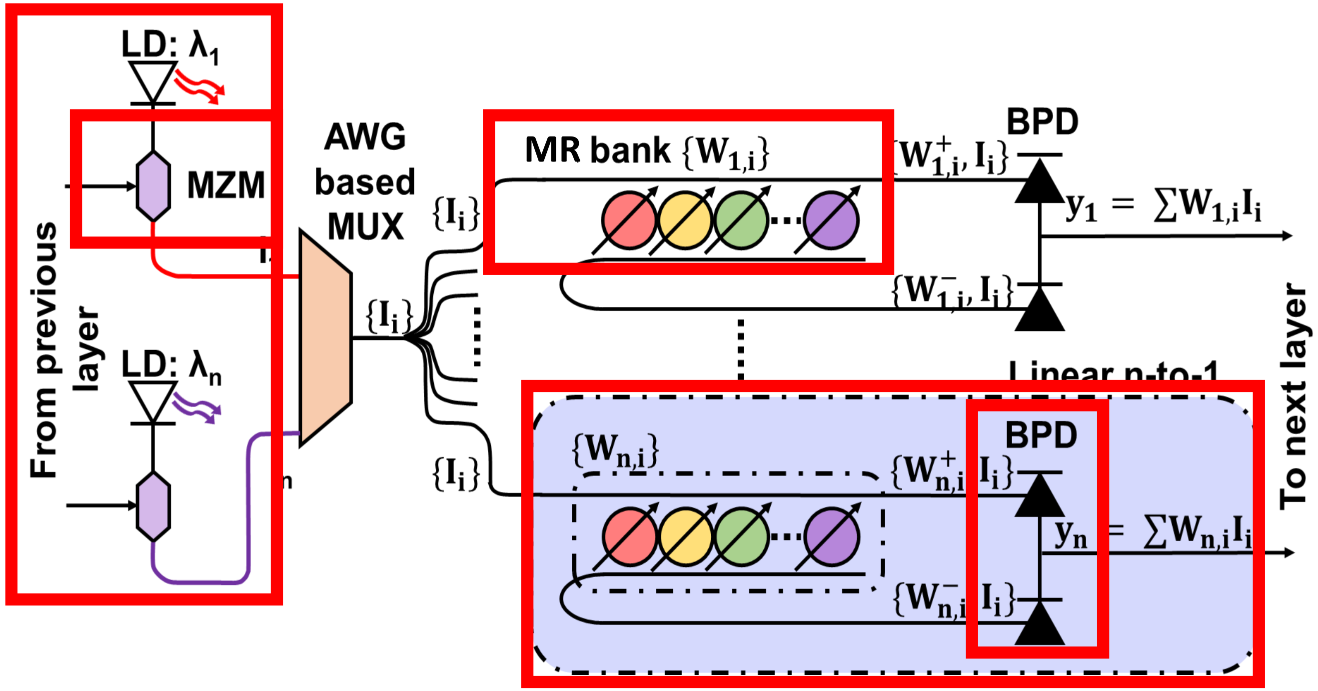


Non-coherent computation used in our work for smaller footprints and ability to handle larger number of parameters simultaneously

MUX: Multiplexer; MR: Microring Resonator

Colorado State University

Noncoherent Architectures



- **Utilizes the mature wavelength division multiplexing (WDM)/Dense WDM technology to represent large number of neurons**
- **Weight values are imprinted onto the wavelength amplitude**
- **Multiplication done by imprinting the activation value onto the signal**
- **Summation performed in photonic mac units using photodetectors**

BPD: Balanced PD; AWG: Arrayed Waveguide Grating; MUX: Multiplexer; MR: Microring Resonator

Our Contributions

- **SONIC: A novel, sparsity aware, photonic CNN accelerator**
  - **Utilizes a modular, vector granularity-aware structure to enable high throughput and energy-efficient execution**
  - **Utilizes sparsity-aware data compression and dataflow techniques for fully connected and convolution layers**
  - **Comprehensive comparison with state-of-the-art sparse electronic and dense photonic CNN accelerator platforms**

- **SONIC was compared against:**
  - **NVIDIA Tesla P100**
  - **RSNN [W. You, C. Wu, IEEE Access, vol. 9, Jan. 2021]**
  - **Null Hop [A. Aimar et al., IEEE Trans. Neural Netw. Learn. Syst, vol. 30, no. 3, Mar. 2019]**
  - **HolyLight [W. Liu et al., IEEE/ACM DATE, 2019]**
  - **LightBulb [F. Zokaee et al., IEEE/ACM DATE, 2020]**
  - **CrossLight [F. Sunny et al., IEEE/ACM DAC 2020]**

Photonic dense CNN accelerators

Model Sparsification and Weight Clustering

- We utilized a layer-wise sparsity-aware training approach for inducing sparsity
  - Layer-wise sparsity considered to avoid overly sparsifying layers
  - The weights in chosen layer are sorted by their absolute values and smallest magnitude weights are masked to zero until specified sparsity levels are reached
  - We opt for sparsity-aware training instead of post-training sparsification, as the latter approach can indiscriminately remove neurons
    - This can adversely affect inference accuracy

- We performed post-training quantization in the form of weight clustering
  - Utilized a centroid based weight clustering approach
  - If there are C centroids, and thus C clusters, the model will end up with C unique weights
    - Required DAC resolution reduced to $\log_2 C$
    - Reduced DAC resolution enables power and latency savings

Dataflow Optimizations

- **Fully Connected (FC) layer**

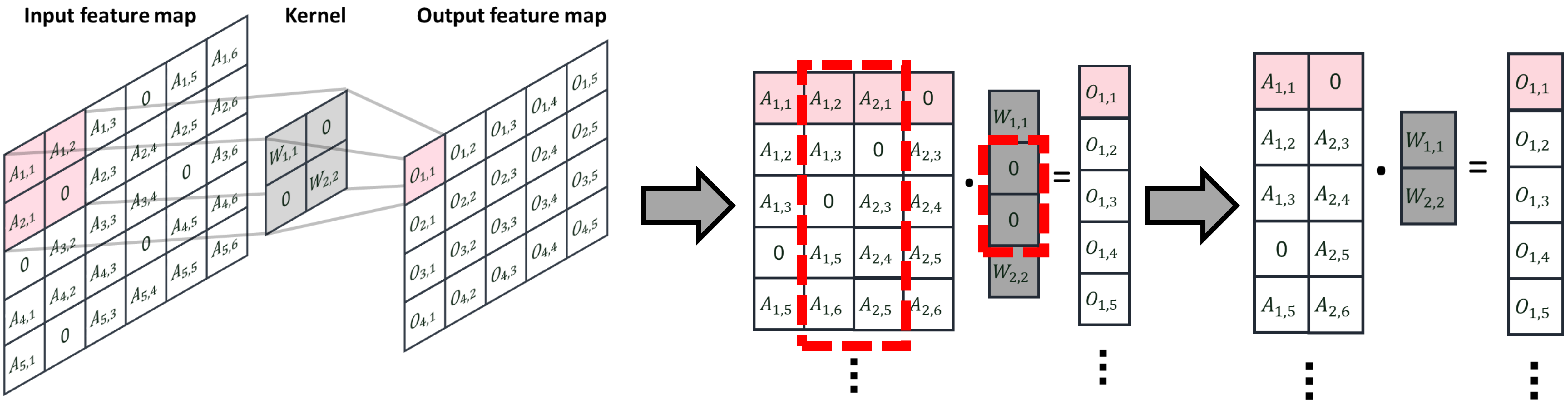
$$\underbrace{\begin{bmatrix} 0 & W_{1,2} & W_{1,3} & 0 & W_{1,5} & W_{1,6} \\ W_{2,1} & 0 & 0 & W_{2,4} & W_{2,5} & 0 \\ 0 & W_{3,2} & W_{3,3} & W_{3,4} & 0 & W_{3,6} \\ W_{4,1} & W_{4,2} & W_{4,3} & 0 & W_{4,5} & W_{4,6} \\ W_{5,1} & 0 & W_{5,3} & W_{5,4} & W_{5,5} & W_{5,6} \end{bmatrix}}_{\text{Weights}} \cdot \underbrace{\begin{bmatrix} A_1 \\ 0 \\ A_3 \\ 0 \end{bmatrix}}_{\text{Activations}} = \underbrace{\begin{bmatrix} O_1 \\ O_2 \\ O_3 \\ O_4 \\ O_5 \end{bmatrix}}_{\text{Output}}$$

$$\Rightarrow$$

$$\underbrace{\begin{bmatrix} 0 & W_{1,3} & W_{1,5} & W_{1,6} \\ W_{2,1} & 0 & W_{2,5} & 0 \\ 0 & W_{3,3} & 0 & W_{3,6} \\ W_{4,1} & W_{4,3} & W_{4,5} & W_{4,6} \\ W_{5,1} & W_{5,3} & W_{5,5} & W_{5,6} \end{bmatrix}}_{\text{Weights}} \cdot \underbrace{\begin{bmatrix} A_1 \\ A_3 \end{bmatrix}}_{\text{Activations}} = \underbrace{\begin{bmatrix} O_1 \\ O_2 \\ O_3 \\ O_4 \\ O_5 \end{bmatrix}}_{\text{Output}}$$

Dataflow Optimizations

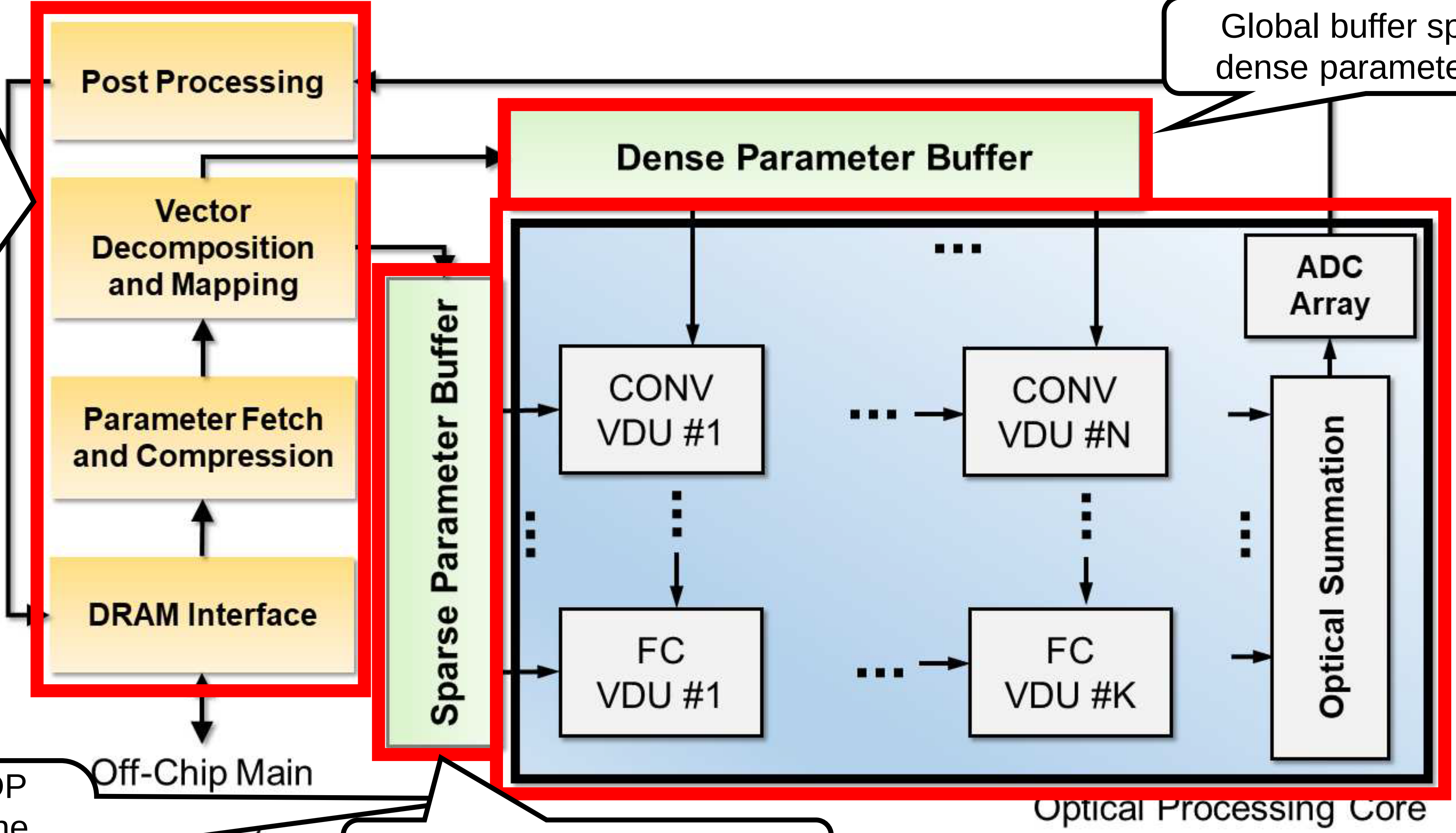
- **Convolution (Conv) layer**



SONIC Accelerator Overview

Electronic Control Unit:
(1) fetch model parameters from global memory;
(2) decomposing matrices to vectors;
(3) mapping vectors to the photonic accelerator;
(4) Applying non-linearities (post processing)

Global buffer specific for dense parameter vectors

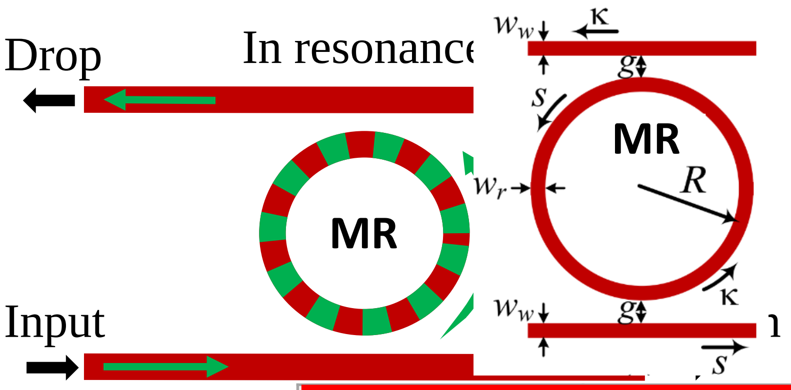


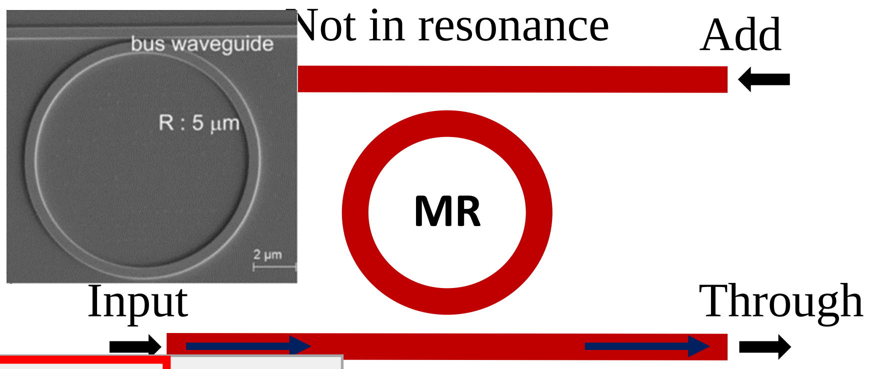
Layer-specific VDP unit array forms the core of SONIC's photonic accelerator substrate

Global buffer specific for sparse parameter vectors

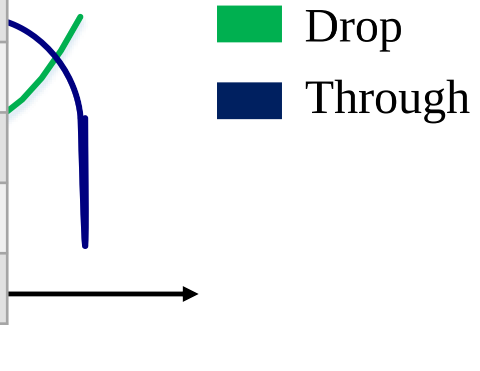
Microring Resonator (MR) Basics and Operations

- **MRs are prominently used in Noncoherent Silicon Photonic (SiPh) architectures**
- **Designed to be selective to a resonant wavelength (λ_{MR})**





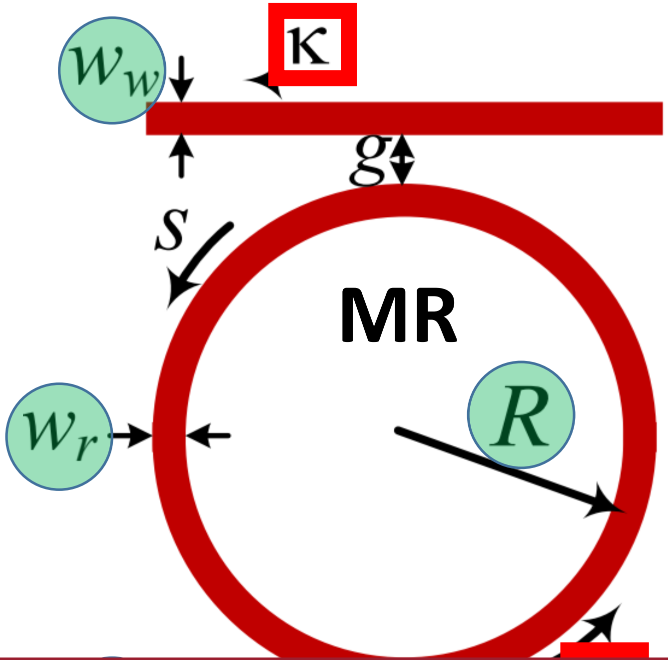
w_w	Input waveguide width
w_r	Ring waveguide width
g	Gap
s	Straight-through coupling
k	Cross-over coupling
R	Radius



Cross-Over Coupling (κ) in MRs

- **Determines the amount of optical power transferred between waveguides**
- **Cross-over coupling can be defined as a function of**
  - $k(\lambda_{MRR}, w_{w/r}, t) \propto f(n_{ew/er}(\lambda_{MRR}, w_{w/r}, t), g^{-1}, R)$

$n_{ew/er}$	Effective index of input/ring waveguide
g	Gap
R	Radius



A higher κ value is preferred for efficient operation of MRs

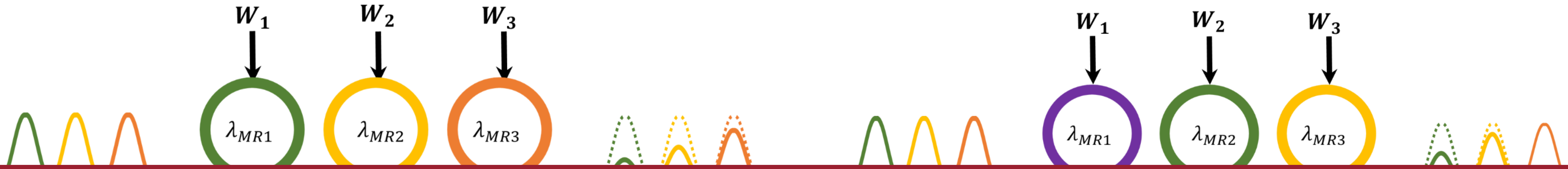
FPV Induced Resonant Wavelength Shift

n_{eff} is a function of ring waveguide thickness and width

$$\lambda_{MR} = \frac{\boxed{n_{eff}}\boxed{L}}{m}, \qquad m = 1,2,3\ldots$$

$$\boxed{L = 2\pi R}$$

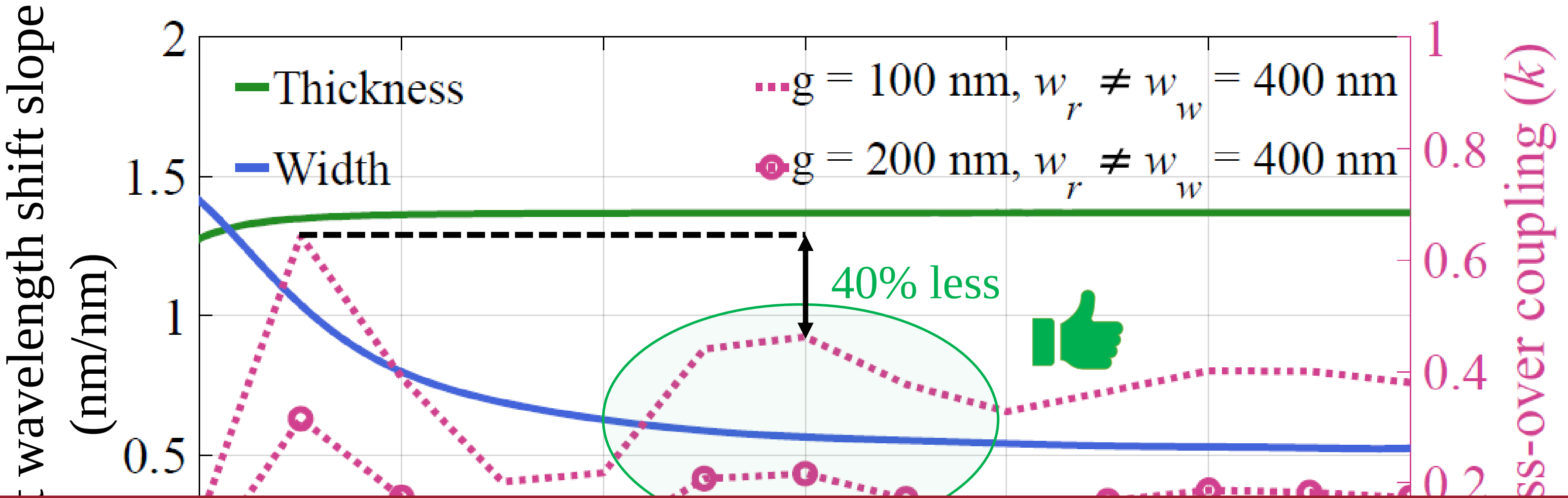
- **FPV can also affect the gap distance between waveguides**
- **Causes change in λ_{MR}, i.e. resonant wavelength shift ($\Delta\lambda_{MR}$)**

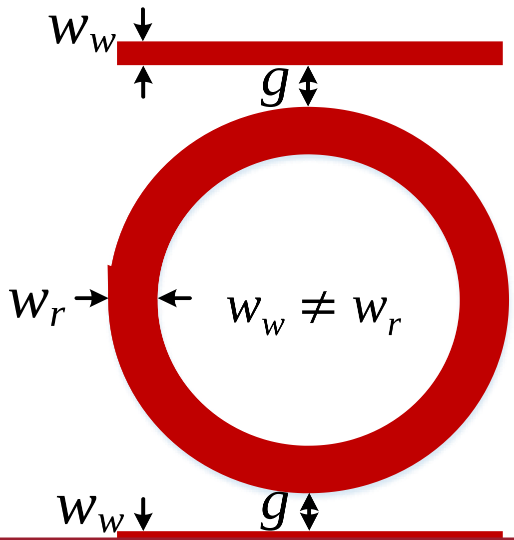


FPV induced $\Delta\lambda_{MR}$ causes computation errors in optical computing

MR Device Engineering

- **Increasing width of input and MR waveguides can help reduce FPV sensitivity**
  - **Causes drop in κ for conventional MR designs $(W_w = W_r)$**
- **We utilize unconventional MR designs $(W_w \neq W_r)$ for obtaining FPV resilience while obtaining upto 40% better κ than conventional MR designs**





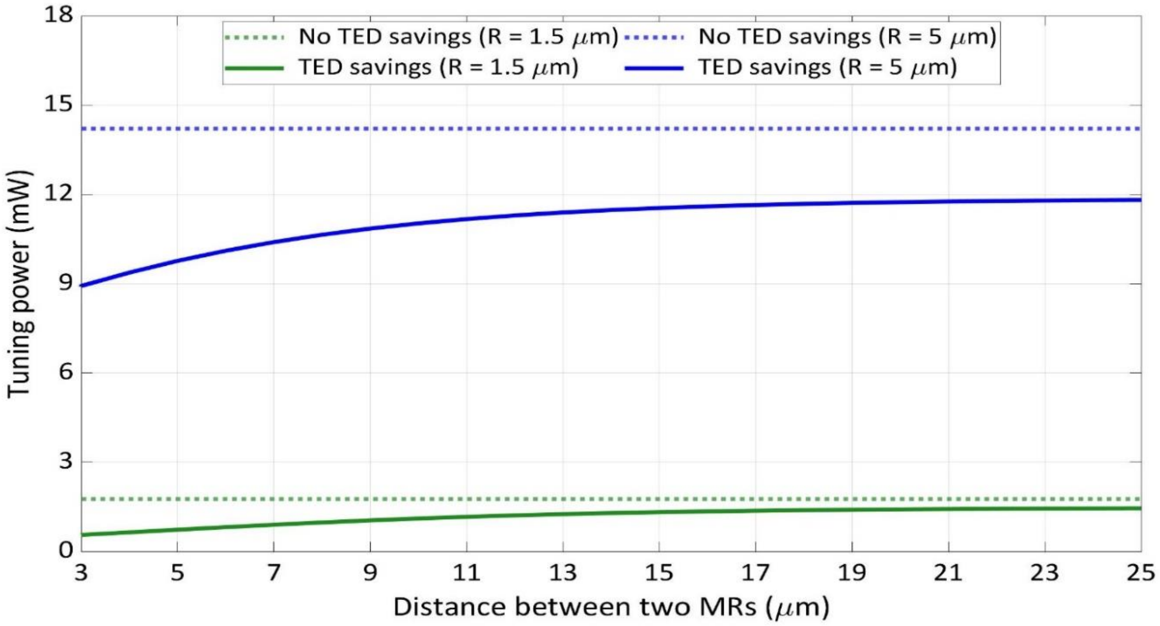
Our meticulous MR design space exploration gives the benefits of FPV resilience while having better κ values

g Gap

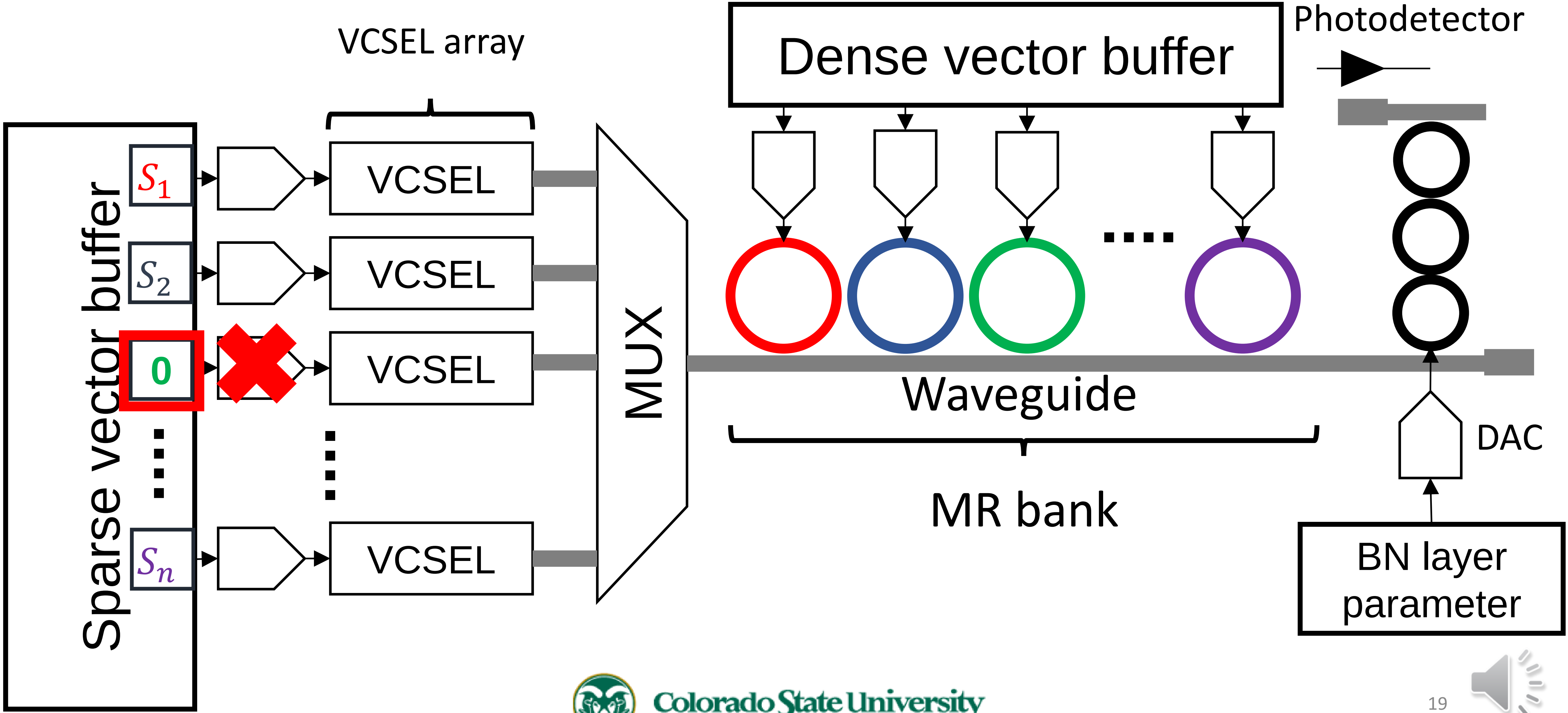
Colorado State University

Robust Tuning Approach

- **Hybrid Electro-Optic (EO) + Thermo-Optic (TO) tuning for reduced latencies**
- **EO tuning for speed and lower energy consumption (Range < 1.5 nm)**
  - **Used to imprint weights and activations on wavelengths**
- **Thermal Eigen mode Decomposition (TED) based TO tuning**
  - **To collectively tune all MRs in a VDP and cancel thermal crosstalk**
  - **Reduces the effective area and laser power over conventional approach**



Vector Dot Product Unit (VDU) Design



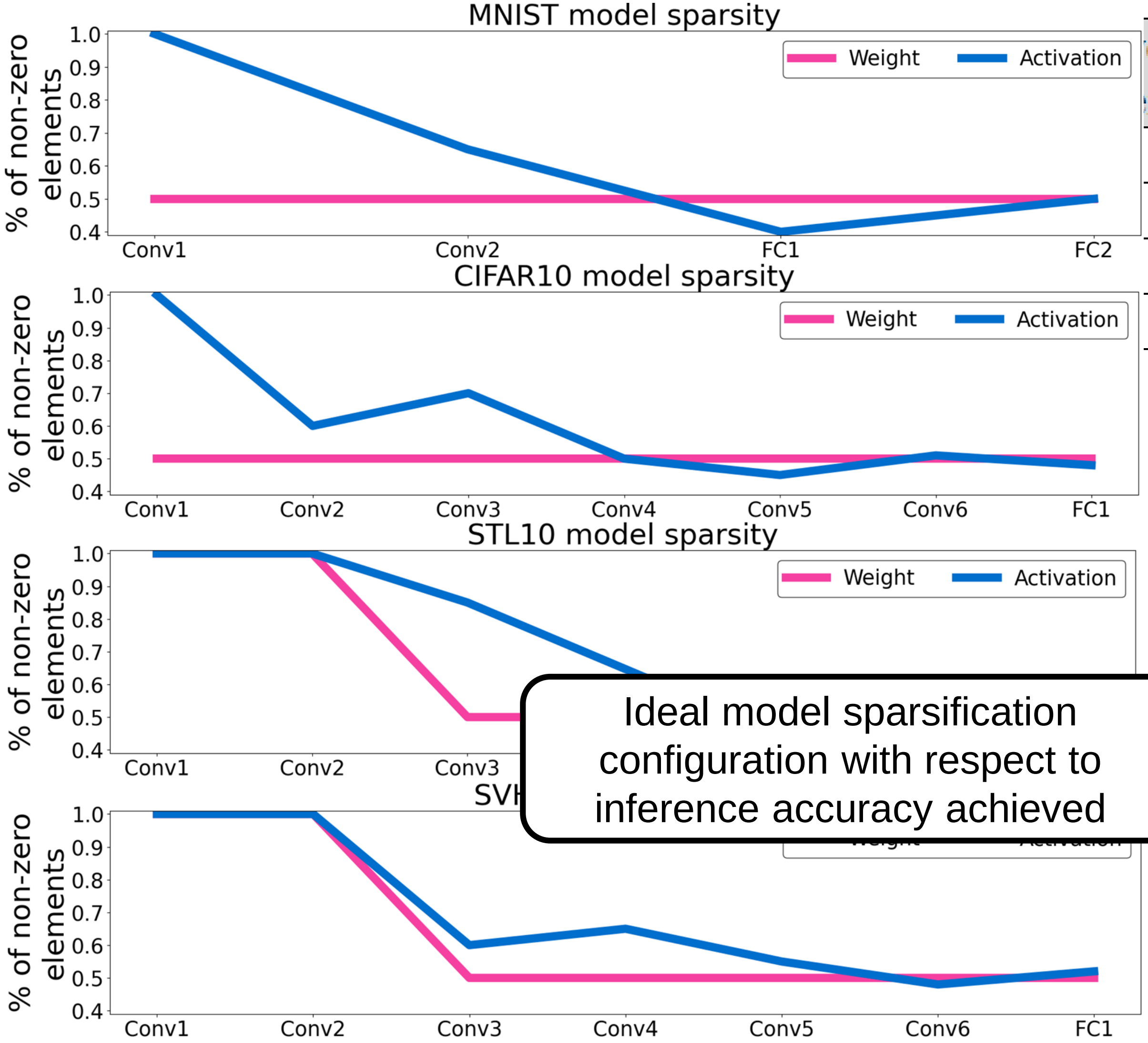
Colorado State University

Experiment Setup

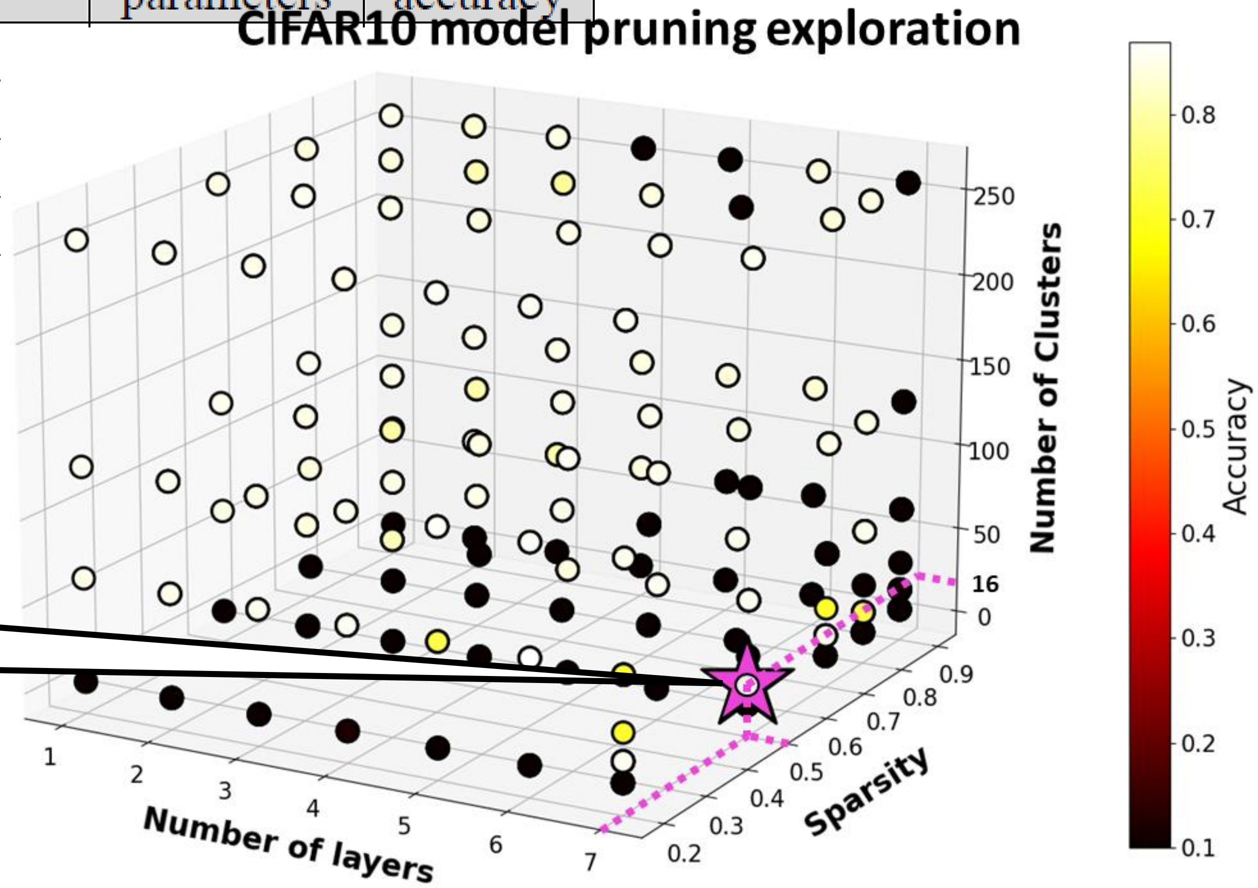
Datasets	Conv layers	FC layers	No. of parameters	Baseline accuracy
MNIST	2	2	1,498,730	93.2%
CIFAR10	6	1	552,874	86.05%
STL10	6	1	77,787,738	74.6%
SVHN	4	3	552,362	94.6%

Devices	**Latency**	**Power**
EO Tuning [A. Stefan et al., IEEE JLT, 2016]	20 ns	4 μW/nm
TO Tuning [P. Pintus et al., L&P reviews, 2019]	4 μs	27.5 mW/FSR
VCSEL [R. Ini et al., CICC, 2021]	0.07 ns	1.3 mW
Photodetector [B. Wang et al., IEEE JLT, 2020]	5.8 ps	2.8 mW
16-bit DAC [B. Wu et al., IEEE J. Solid-state circuits, 2016]	0.33 ns	40 mW
6-bit DAC [C. M. Yang et al., IEEE TCAS, 2021]	0.25 ns	3 mW
ADC [J. Shen et al., IEEE J. Solid-state circuits, 2018]	14 ns	62 mW

Model Sparsification and Clustering Results



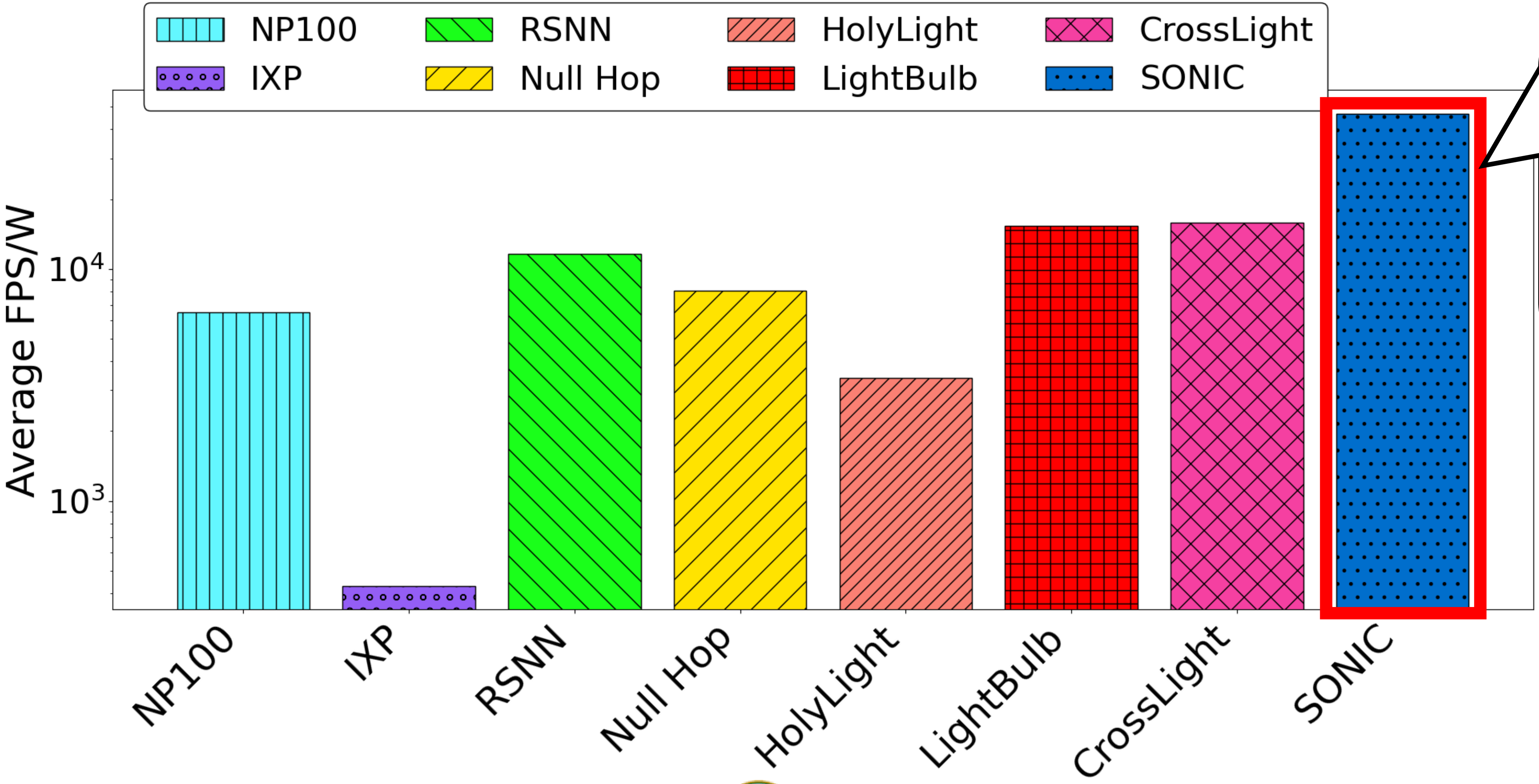
eight rs	No. of parameters	Final accuracy



Ideal model sparsification configuration with respect to inference accuracy achieved

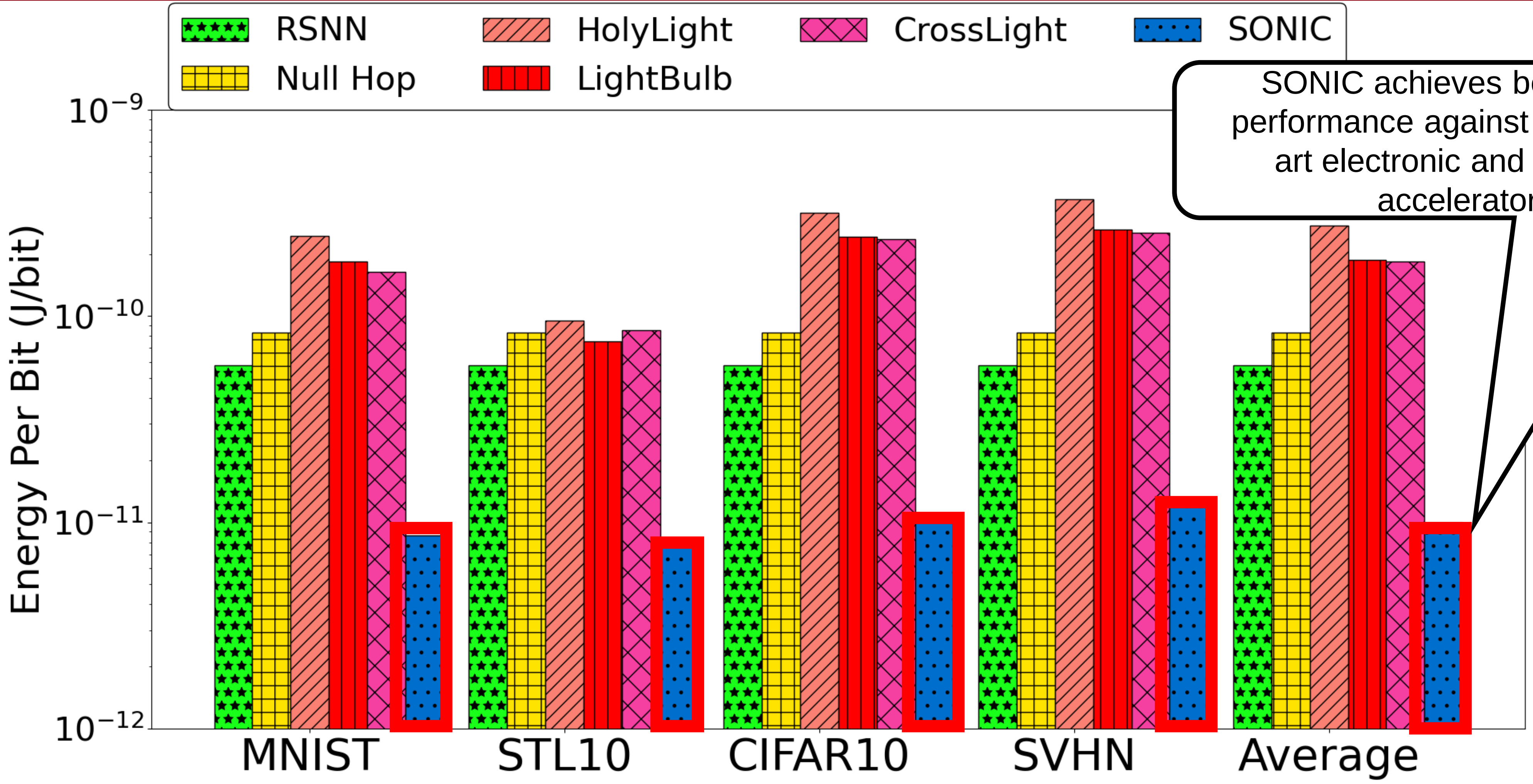
Colorado State University

Comparison Against Other Accelerators



SONIC achieves better power efficiency with respect to throughput due to the sparsity aware hardware and software optimizations utilized

Comparison Against Other Accelerators



SONIC achieves better EPB performance against state of the art electronic and photonic accelerators

Colorado State University

Conclusions

- This work presented a novel non-coherent photonic SpNN accelerator, SONIC, that integrates several hardware and software optimizations
- SONIC exhibits better performance in terms of power efficiency and EPB against state of the art electronic and photonic accelerators
  - Up to 5.8× better power efficiency, and 8.4× lower EPB than state-of-the-art electronic SpNN accelerators
  - Up to 13.8× better power efficiency and 27.6× lower EPB than state-of-the-art dense photonic neural network accelerators
- These results demonstrate the promising low-energy and low-latency inference acceleration capabilities of our SONIC architecture

Thank You

Questions?

- **Febin P. Sunny | febin.sunny@colostate.edu**
- **LinkedIn: https://www.linkedin.com/in/febin-sunny-86802389/**
- **Website: https://sites.google.com/view/febinpsunny**

Colorado State University