Cross-Layer Design of Deep Learning Accelerators with Silicon Photonics

OPTICS Workshop, April 15, 2021

Sudeep Pasricha

Walter Scott Jr. College of Engineering Professor

Director of Embedded, High Performance, and Intelligent Computing (EPIC) Lab

Colorado State University, Fort Collins, CO

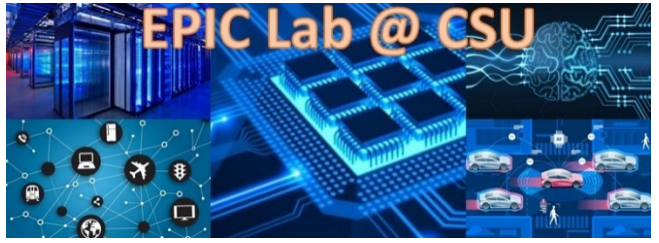
sudeep@colostate.edu

Colorado State University

Collaborative work with:

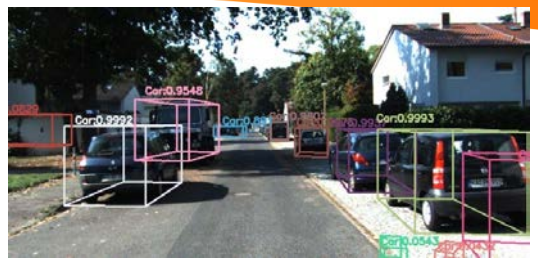
Prof. Mahdi Nikdast

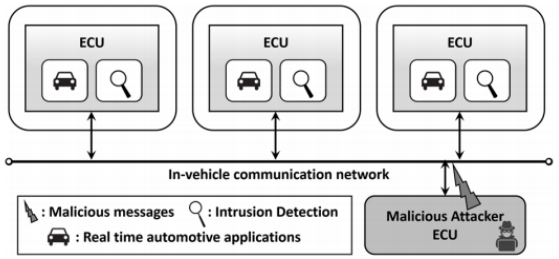
Febin Sunny, Asif Mirza (Ph.D. students)

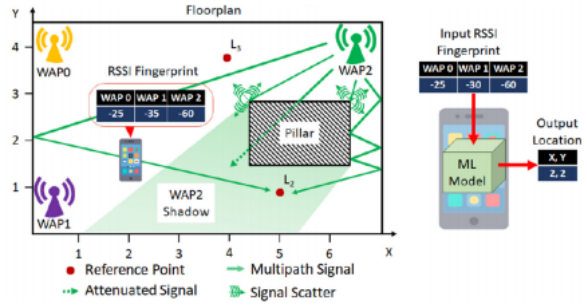


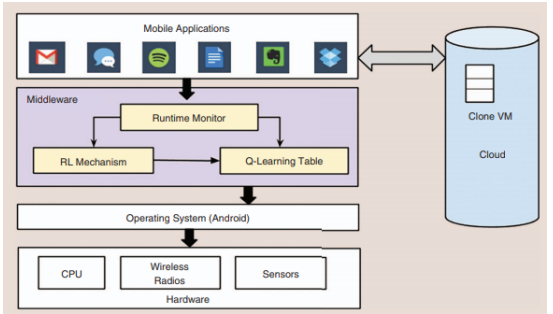
Emerging ML Applications

- **ML applications are becoming increasingly complex**
- **Some recent examples from our lab:**
  - 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.***
  - Unsupervised deep learning for network anomaly detection
    - V. K. Kukkala, S. V. Thiruloga, and S. Pasricha, "INDRA: Intrusion Detection using Recurrent Autoencoders in Automotive Embedded Systems", ***IEEE TCAD, Nov 2020.***
  - 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.***
  - Deep reinforcement learning for embedded mobile devices
    - A. Khune and S. Pasricha, "Mobile Network-Aware Middleware Framework for Energy Efficient Cloud Offloading of Smartphone Applications", ***IEEE Consumer Electronics, 2019.***
- **Inference acceleration is becoming crucial**
  - for energy- and resource-constrained platforms executing real-time embedded and IoT applications
- **Domain-specific ML hardware accelerators preferred**
  - provide many benefits over GPUs and CPUs

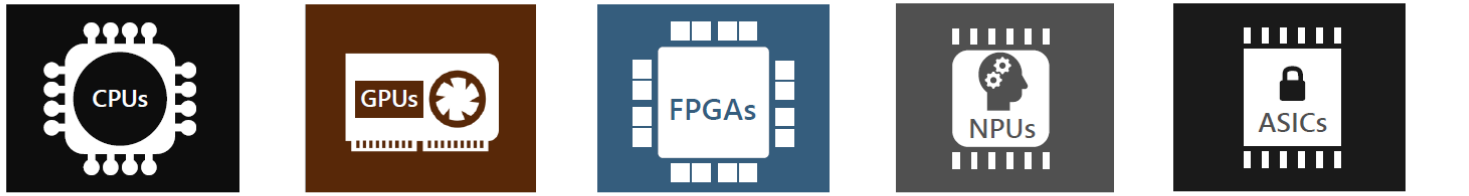


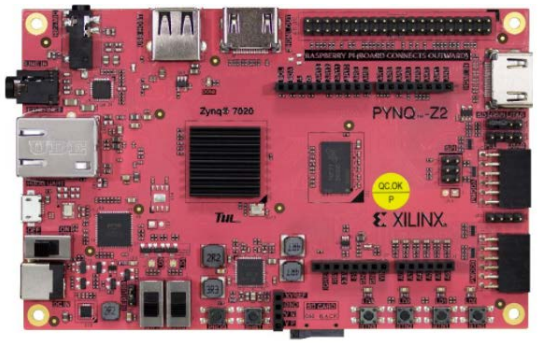


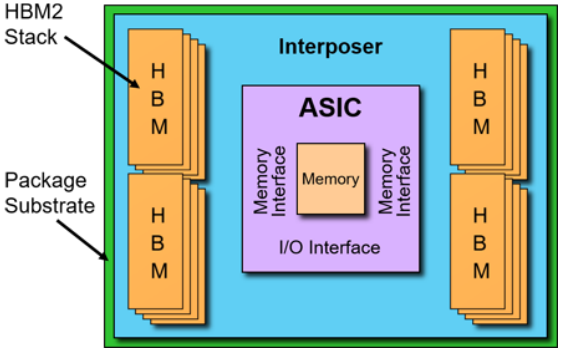


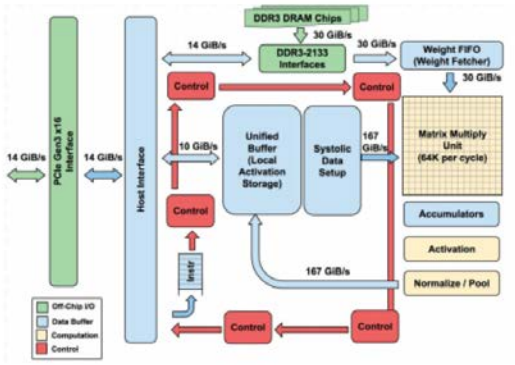


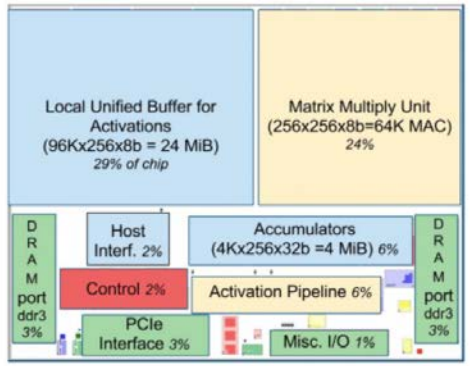
Hardware Accelerators for ML





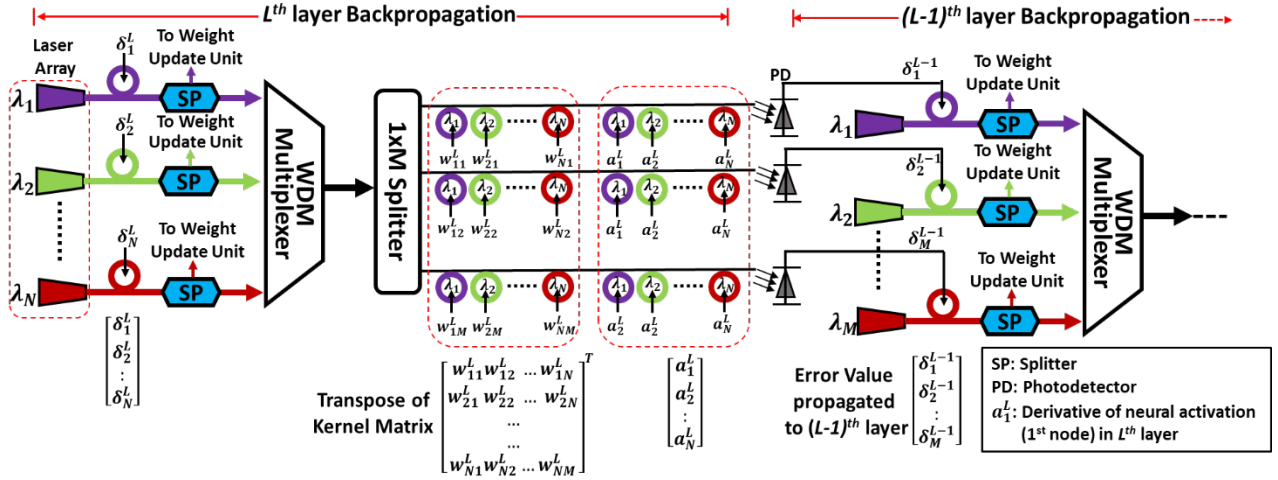


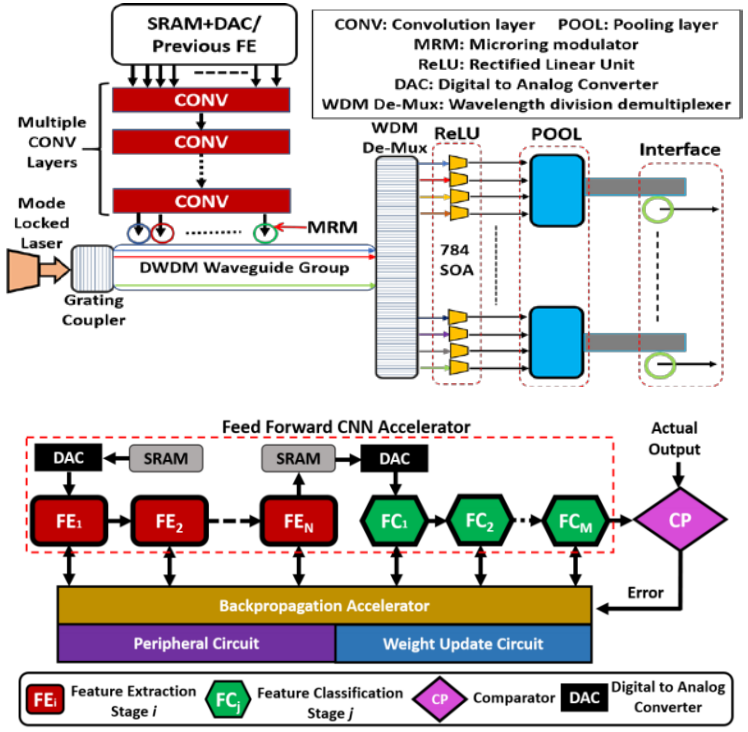


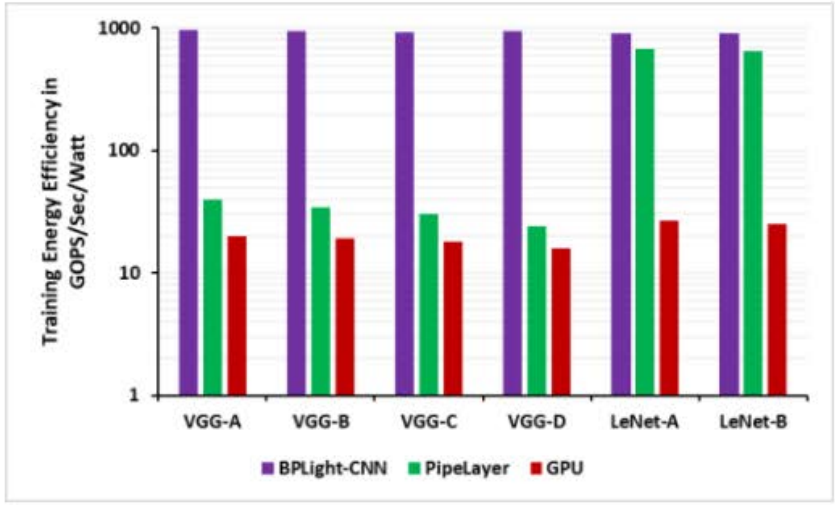


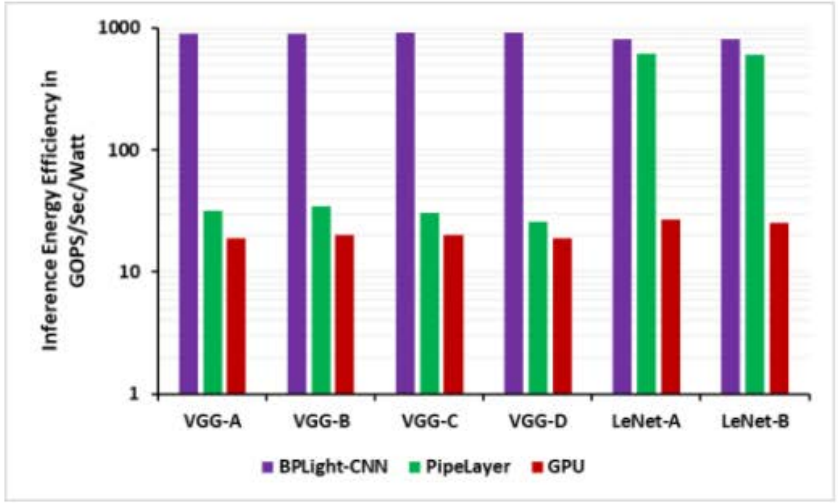
Case for Silicon Photonics in ML Accelerators

- **Example: Photonic-Memristor AI Training/Inference Accelerator**
  - D. Dang, S. V. R. Chittamuru, S. Pasricha, R. Mahapatra, D. Sahoo, "BPLight-CNN: A Photonics-based Backpropagation Accelerator for Deep Learning", ***ACM JETC, 2021.***



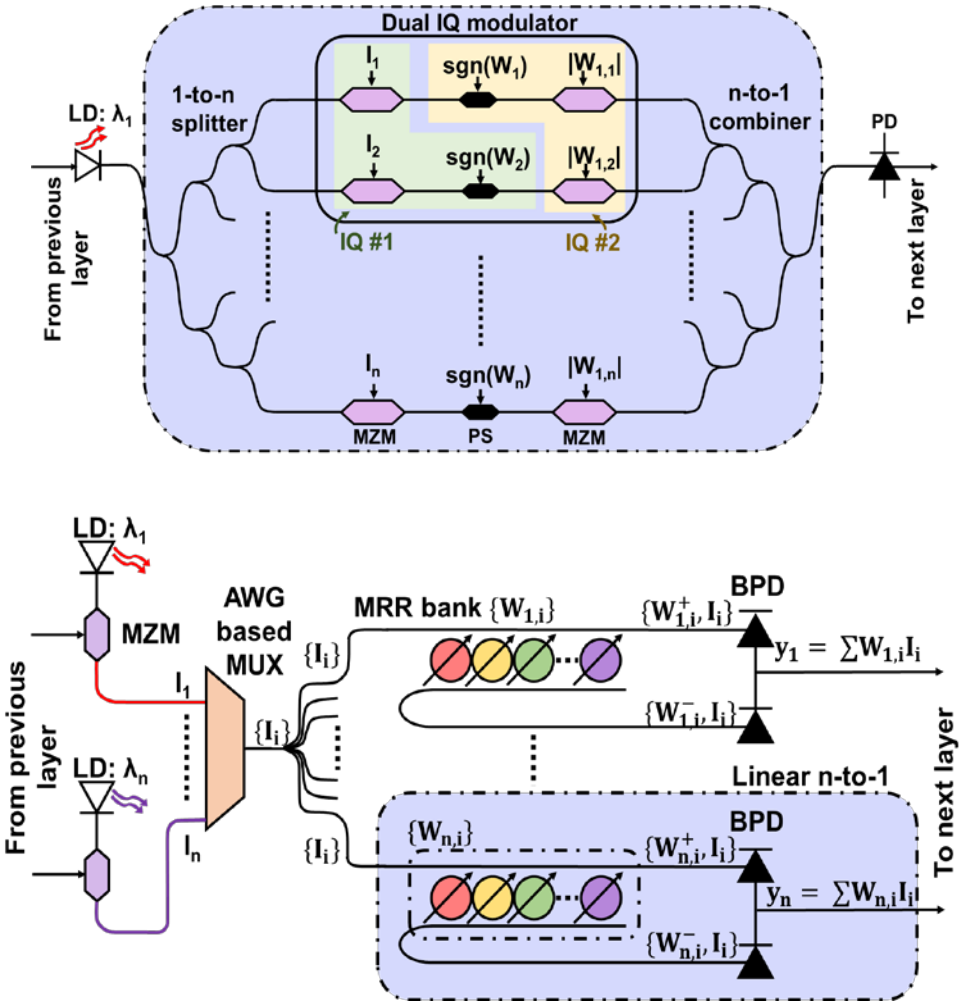






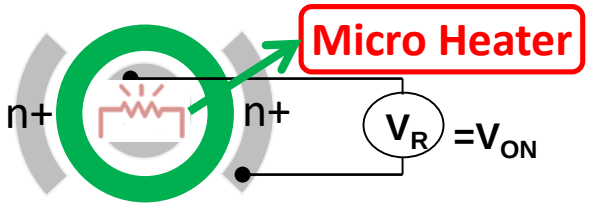
Computing and Communication with Photonics

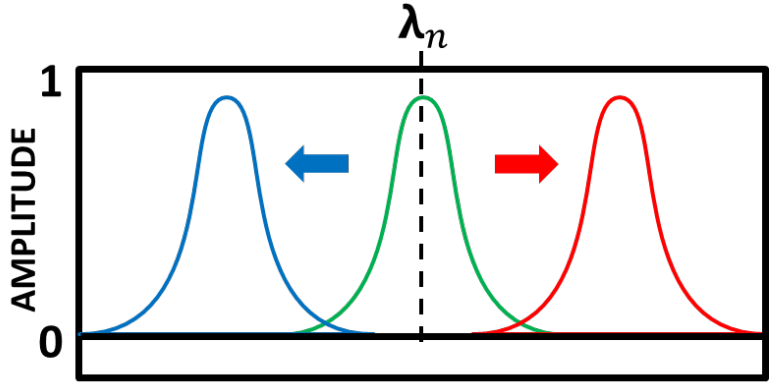
- **Coherent computation**
  - Single wavelength; weights represented using electrical field amplitude
  - Challenges:
    - Scalability issues
    - Phase encoding noise
    - Phase error accumulation
- **Noncoherent computation**
  - Phase-change in devices used to imprint weight/activation values on signal intensities of multiple λ

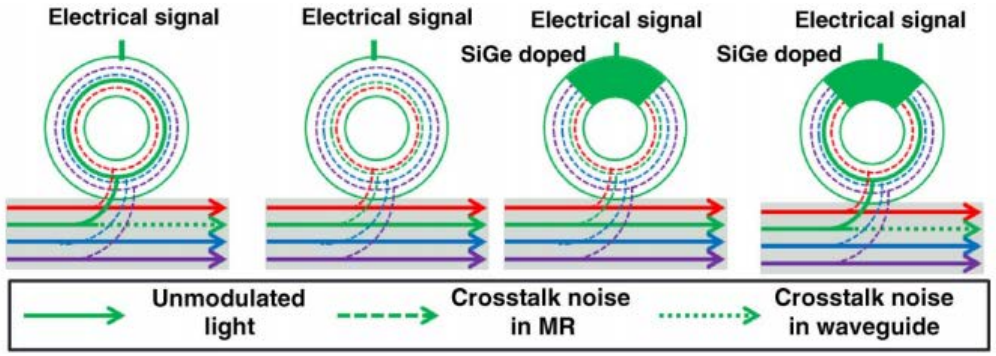


Challenges in Noncoherent Accelerators

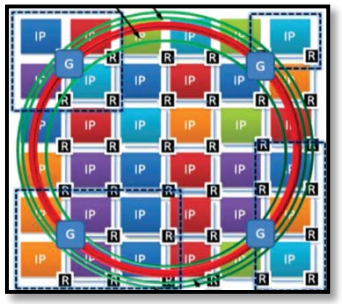
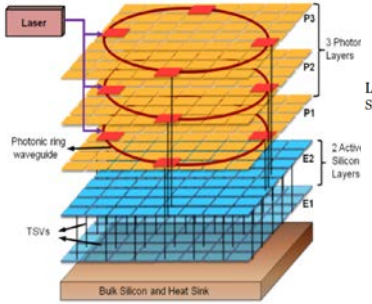
- **Large latencies induced by thermo-optic tuning (μs scale)**
  - Thermo-optic tuning preferred for its large tuning range (~15 nm)
- **MR resonance shifts $\Delta\lambda_{MR}$**
  - Fabrication process variation induced
  - Thermal variation induced
- **Thermal crosstalk in MRs**
  - limits the achievable resolution

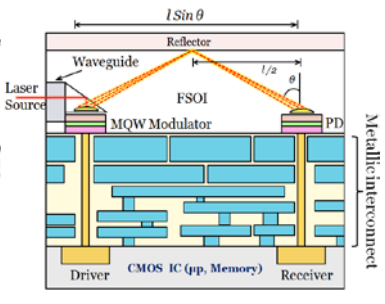




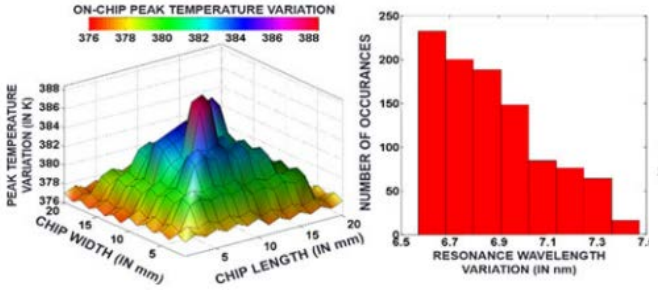


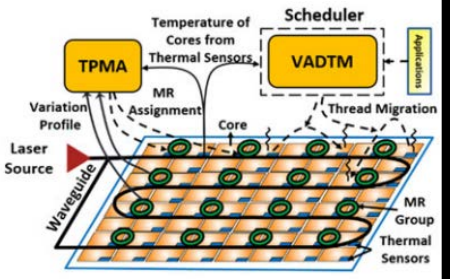
Potential of Cross-Layer Design

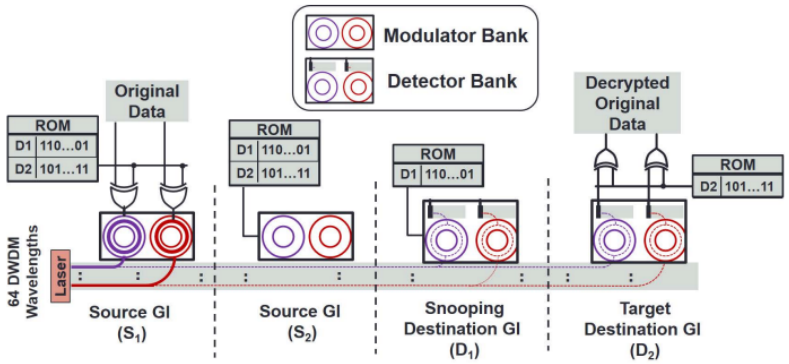


Network architectures: [TPDS'18], [JETC'17], [GLSVLSI'15 (Best Paper)], [TECS'14], [ASPDAC'11], [ISQED'10 (Best Paper)], [ASPDAC'08]

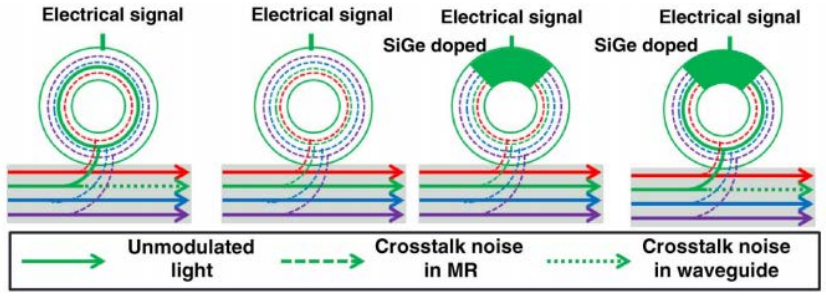




Thermal optimization: [ICCAD'17], [ASPDAC'17], [VLSID'16]



Security mechanisms: [TCAD'20], [NOCS'18], [DAC'18]

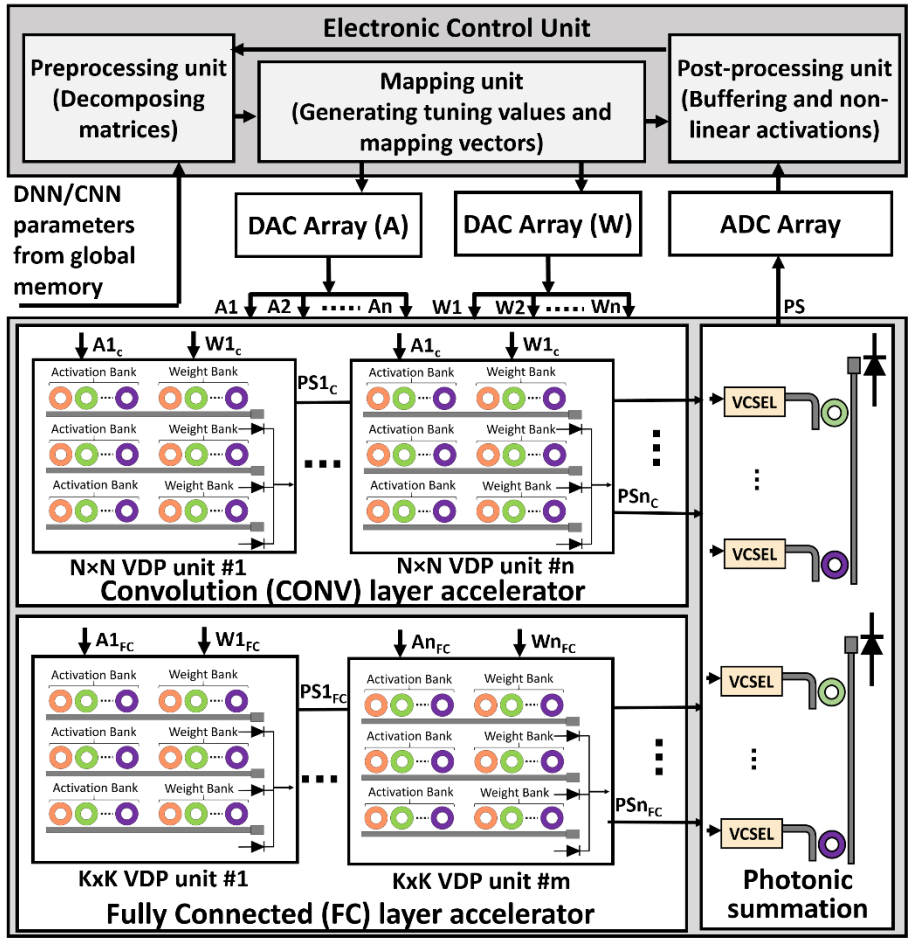


Crosstalk reduction: [TMSCS'18], [NOCS'17], [ISQED'16 (Best Paper Cand.)], [CODES+ISSS'16], [D&T'15]

S. Pasricha, M. Nikdast, "A Survey of Silicon Photonics for Energy Efficient Manycore Computing" *IEEE Design and Test, Aug 2020.*

CrossLight

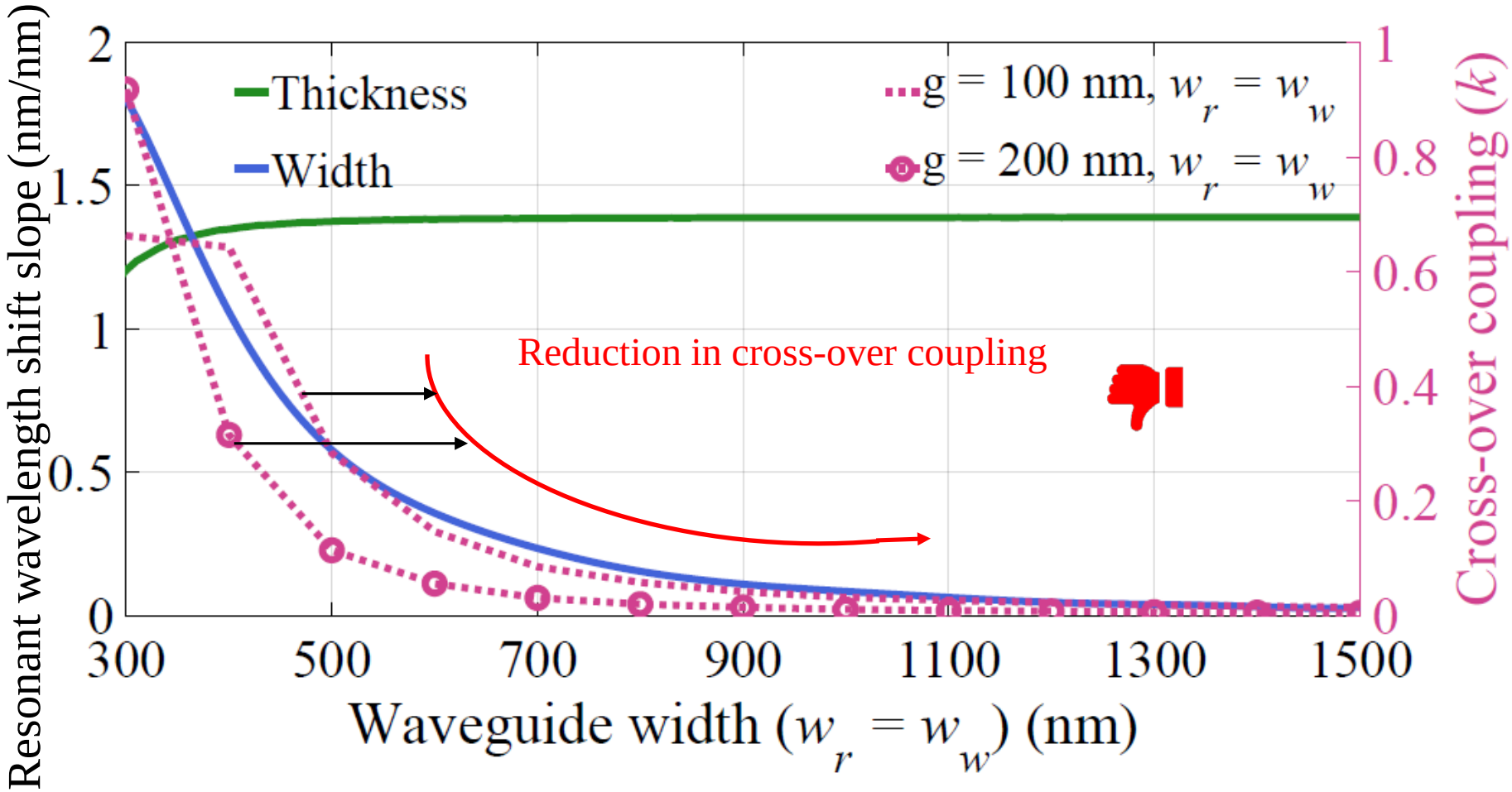
- **A Cross-Layer Optimized SiPh Neural Network Accelerator**
  - Device-level optimizations
    - Improved SiPh device designs for FPV resilience
  - Circuit-level optimizations
    - An enhanced tuning circuit to support large thermal-induced resonance shifts and high-speed, low-loss device tuning
    - Consideration of thermal crosstalk mitigation methods to improve the weight resolution achievable by our architecture
  - Architecture-level optimizations
    - Improved wavelength reuse
    - Smart matrix decomposition for scalable mapping
    - Both optimizations increase throughput and energy-efficiency

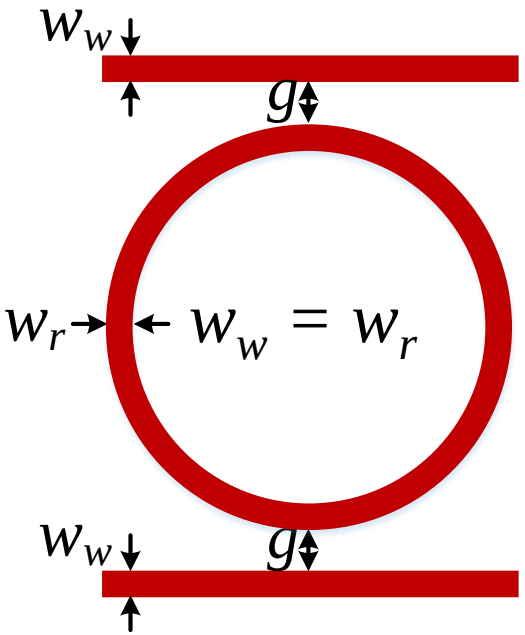


F. Sunny, A. Mirza, M. Nikdast, S. Pasricha, "CrossLight: A Cross-Layer Optimized Silicon Photonic Neural Network Accelerator", ***to appear, IEEE/ACM Design Automation Conference (DAC), 2021***

FPV Resilient MR Devices

- **Increase in FPV tolerance with increasing widths, when $w_w = w_r$**
  - at the cost of low cross-over coupling

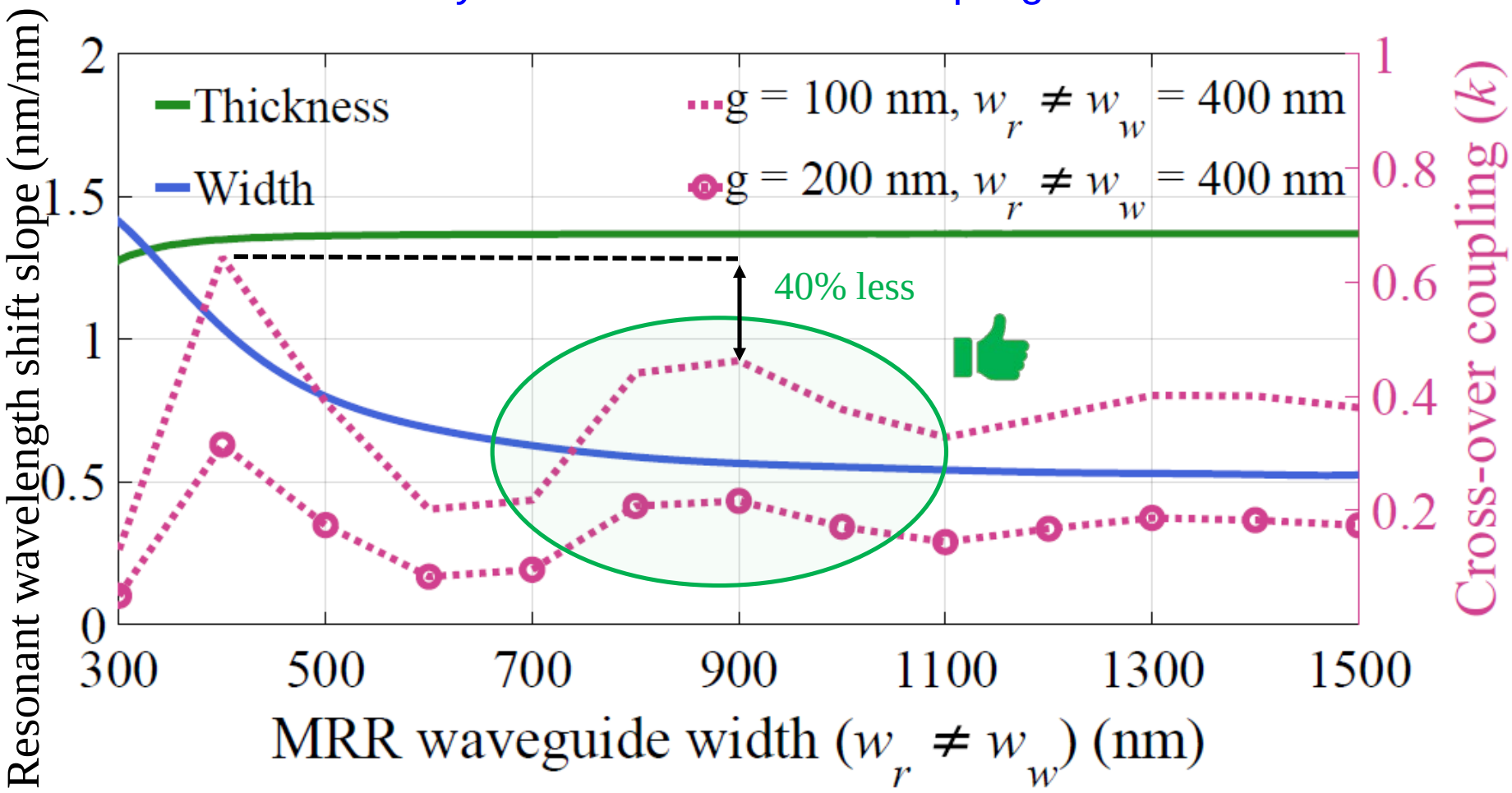


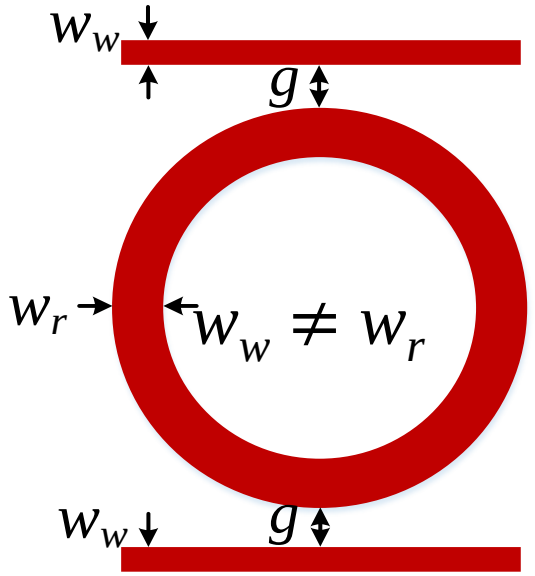


w_w	Input waveguide width
w_r	Ring waveguide width
g	Gap

MR Device Engineering

- **Increase in FPV tolerance when $w_w \neq w_r$**
  - With a relatively smaller reduction in coupling!



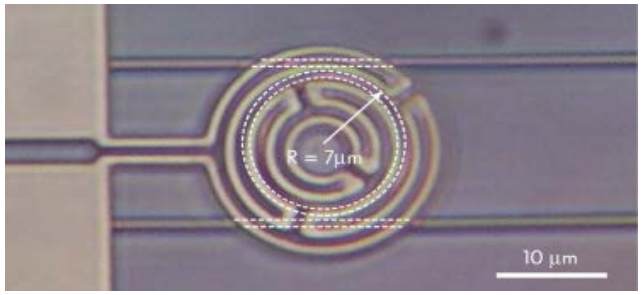


w_w	Input waveguide width
w_r	Ring waveguide width
g	Gap

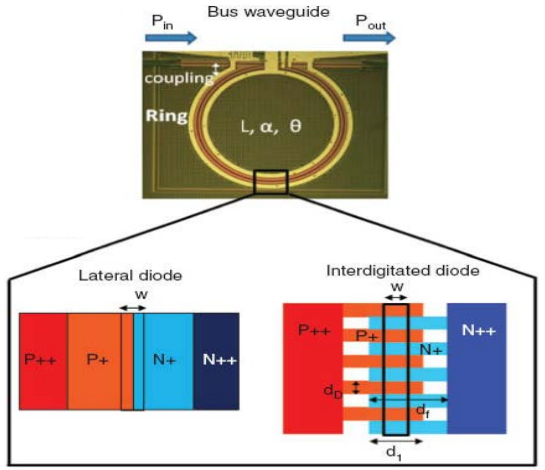
A. Mirza, F. Sunny, S. Pasricha, and M. Nikdast, "Silicon Photonic Microring Resonators: Design Optimization under Fabrication Non-uniformity", ***IEEE/ACM DATE, 2020.***

Tuning Circuits

- **FP/thermal variations compensated using tuning circuits**
  - Thermo-optic (TO) tuning
    - Uses in-built heaters to change effective index
    - Large latencies (μs scale)
    - Induces thermal crosstalk; mitigation approaches:
      - **MRs placed far apart (125 µm to 200 µm)**
      - **Increases area, waveguide length, laser power**
  - Electro-optic (EO) tuning for small variations
    - Carrier injection to impact effective index
    - Faster and more energy-efficient than TO tuning
    - But much smaller range than TO tuning



F. Gan et al., in *Photonics in Switching,* Aug. 2007

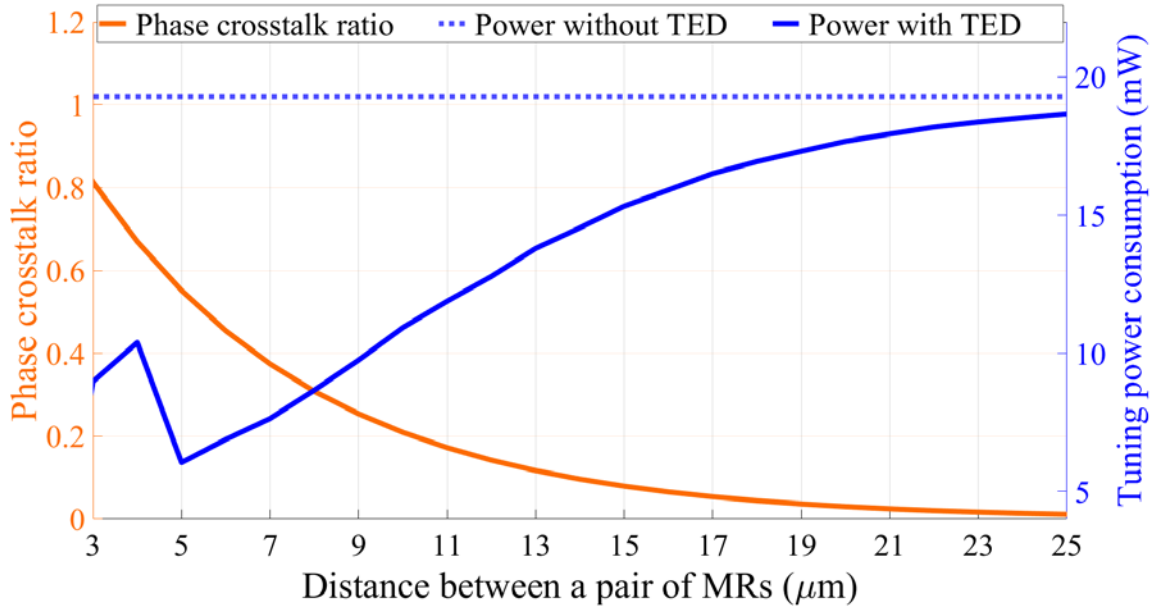


K. Padmaraju et al., in *Nanophotonics, vol.3, no. 4-5, Sept. 2013*

Tuning Circuit Optimization

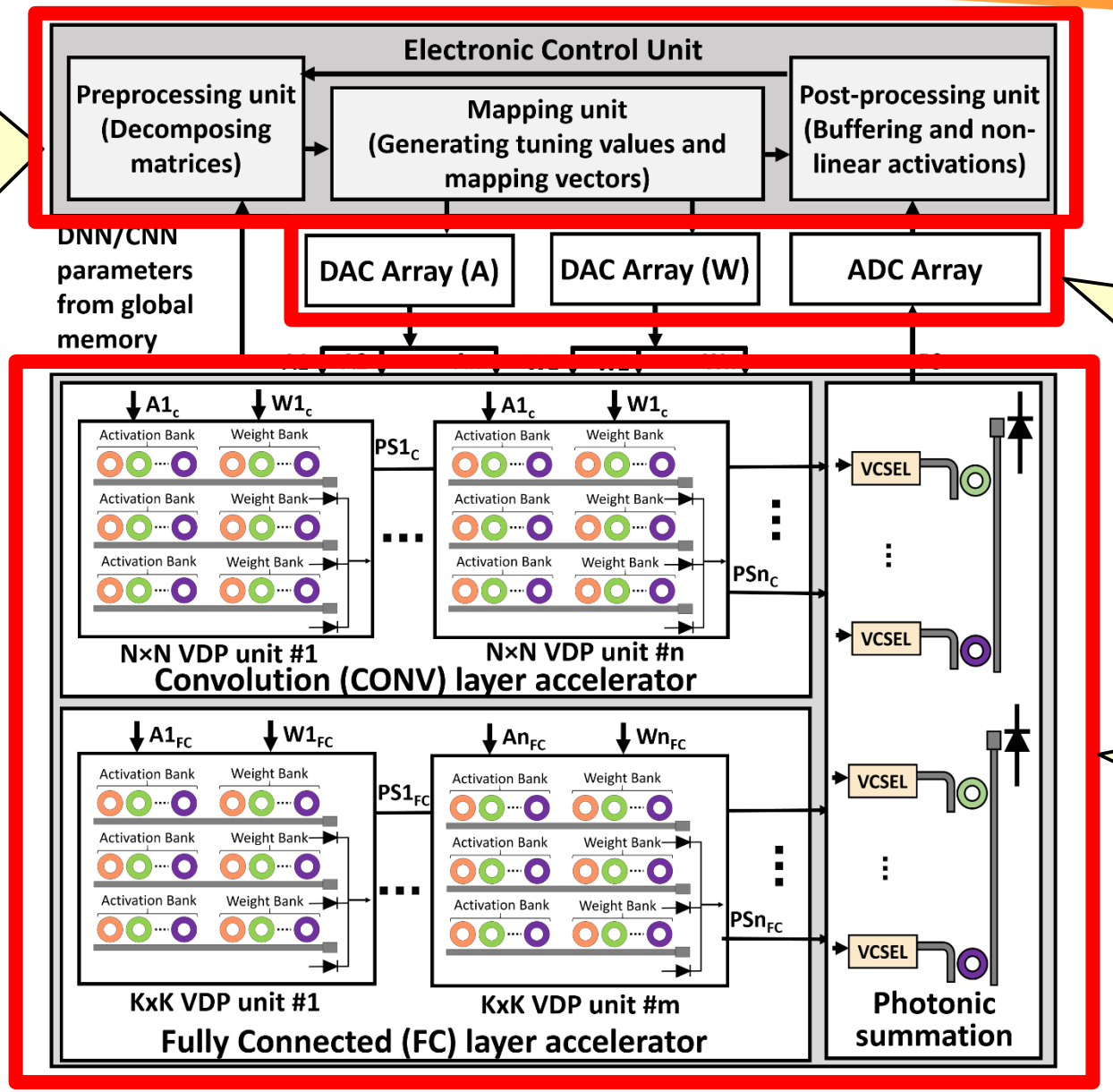
- **Hybrid EO + TO tuning for reduced latencies**
  - EO tuning for speed and lower energy consumption (Range < 1.5 nm)
    - used to imprint weights and activations on wavelengths
  - Thermal Eigenmode Decomposition (TED) based TO tuning
    - to collectively tune all MRs in an MR bank and cancel thermal crosstalk
    - reduces the effective area, waveguide length, laser power over conventional approach

- **Explored ideal layout for MRs**
  - Optimal MR radius: 5 μm
  - Optimal inter-MR distance: 5 μm



CrossLight Architecture

Electronic Control Unit for:
(1) fetching model parameters from global memory + decomposing matrices to vectors;
(2) mapping vectors to the photonic accelerators;
(3) Applying non-linearities

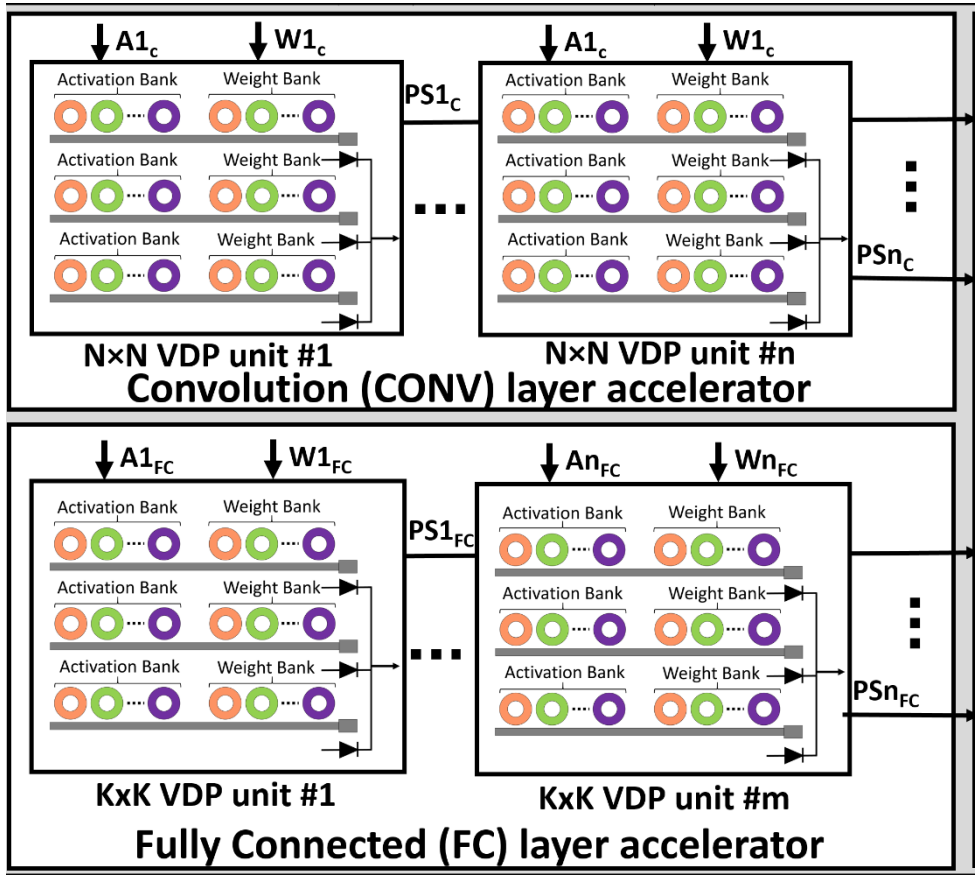


DAC units used to imprint values on to tuning circuits and ADC array to convert analog sum values to digital domain

Photonic domain for high throughput, energy efficient MAC operation

Photonic Computation

- **CrossLight has separate implementations for CONV and FC layer acceleration**
  - Vastly different order of vector dot product (VDP) required to implement each layer
    - CONV: n VDP units, supporting N×N dot product
    - FC: m VDP units, supporting K×K dot product
    - n > m
    - K > N
- **To reduce laser power we reuse the unique lasers needed per VDP**
  - Further dividing N or K into smaller values across waveguides
    - From analysis: N or K can be maximum of 15

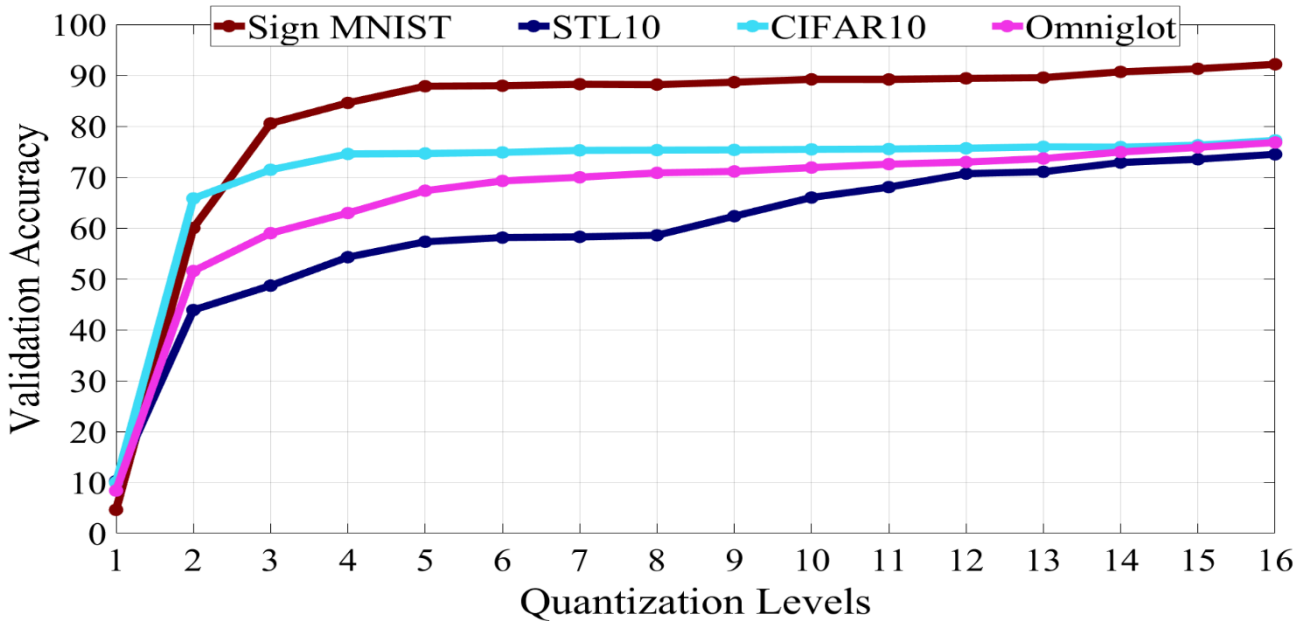


Experimental Setup

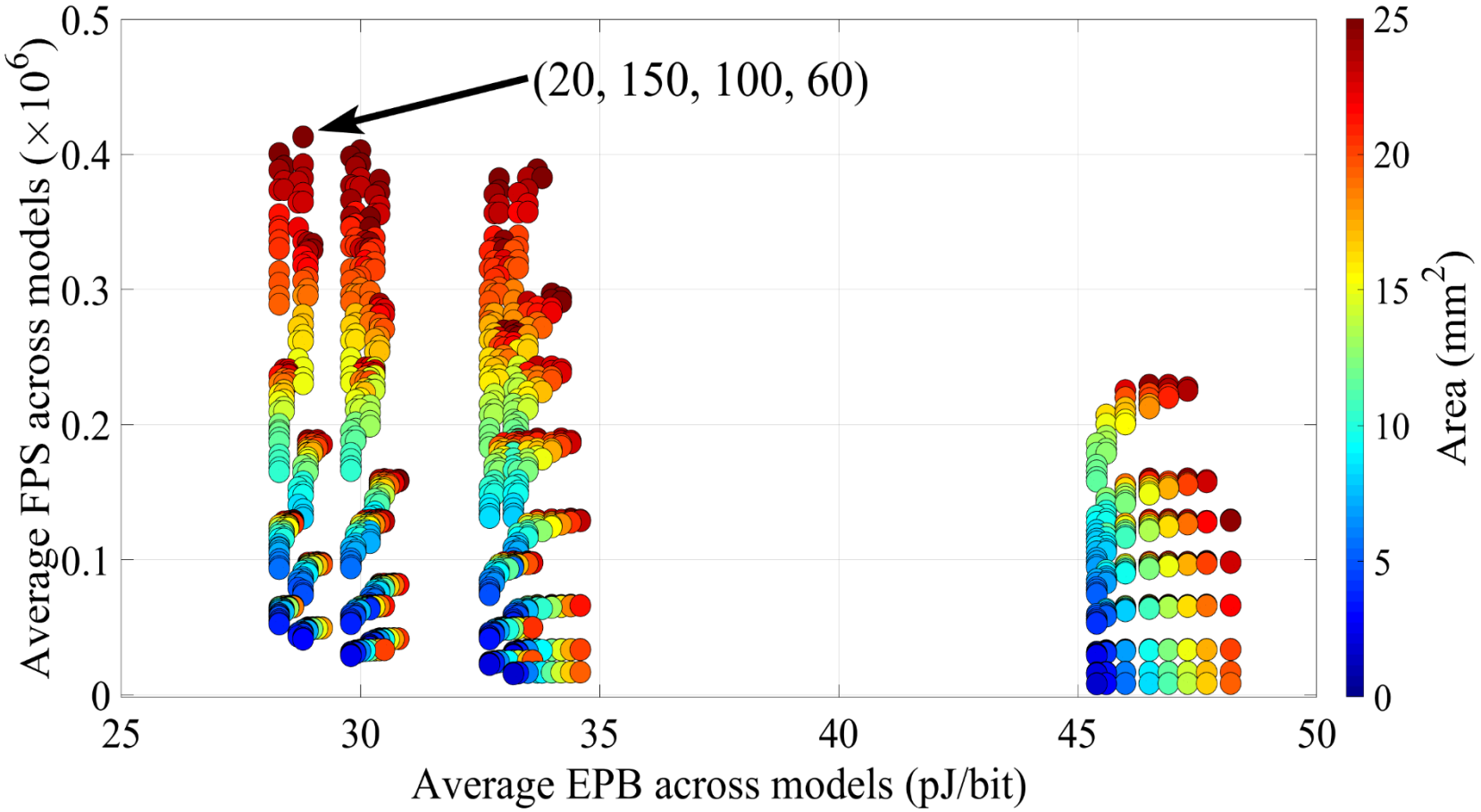
- **Crosslight able to achieve 16-bit resolution**
  - Sufficient for various NN models

Table I: Models and datasets considered for evaluation

Model no.	CONV layers	FC layers	Parameters	Datasets
1	2	2	60,074	Sign MNIST
2	4	2	890,410	CIFAR10
3	7	2	3,204,080	STL10
4	8	4	38,951,745	Omniglot



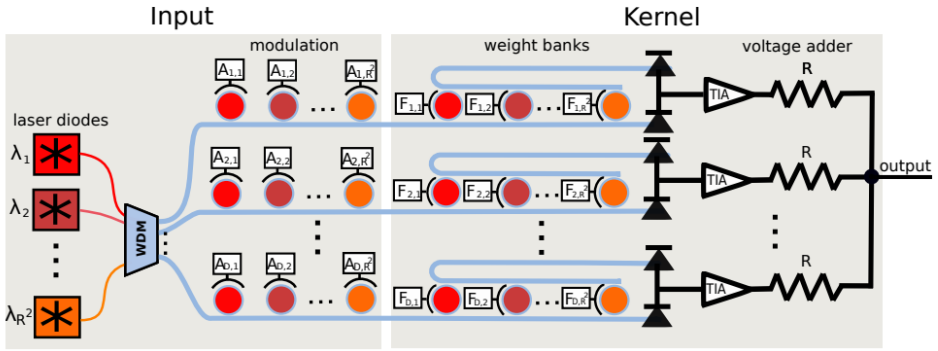
CrossLight Architectural Exploration

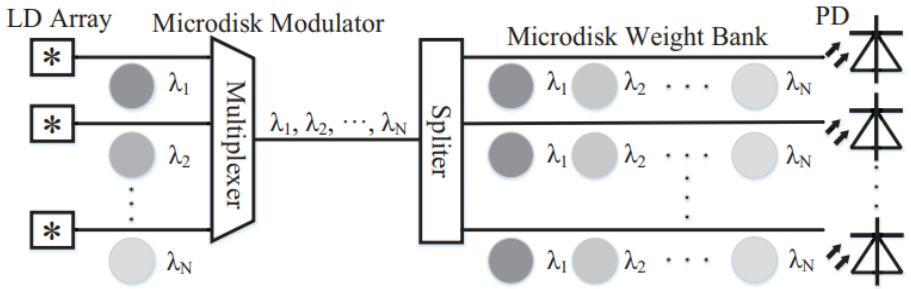


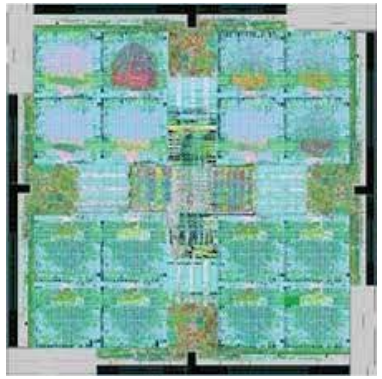
- **Optimal (N,K,n,m) configuration was found from analysis as: (20,150,100,60)**
  - n CONV VDP units, supporting N×N dot product; m FC VDP units, supporting K×K dot product
  - Optimal => best FPS/EPB

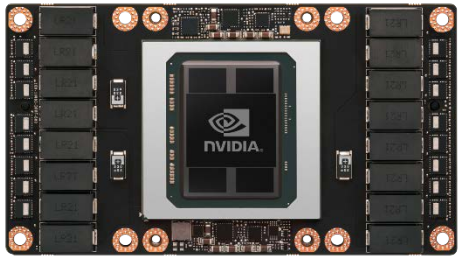
Comparison with Other Accelerators

- **DEAP-CNN**
  - [V. Bangari et al., IEEE JQE, 2020]
    - Implements a photonic CNN accelerator
    - Uses multiple, connected photonic CONV units
    - MR based architecture
- **HolyLight**
  - [W. Liu et al., IEEE/ACM DATE, 2019]
    - Microdisks considered instead of MRs
      - **For lower area and power consumption**
    - On-chip photonic microdisk based MACs
    - Photonic summation also used
- **Electronic-domain accelerators**
  - DaDianNao, Null Hop, and EdgeTPU
  - GPU: Nvidia Tesla P100
  - CPUs: Intel Xeon Platinum 9282 (IXP9282), and AMD Threadripper 3970x (AMD-TR)

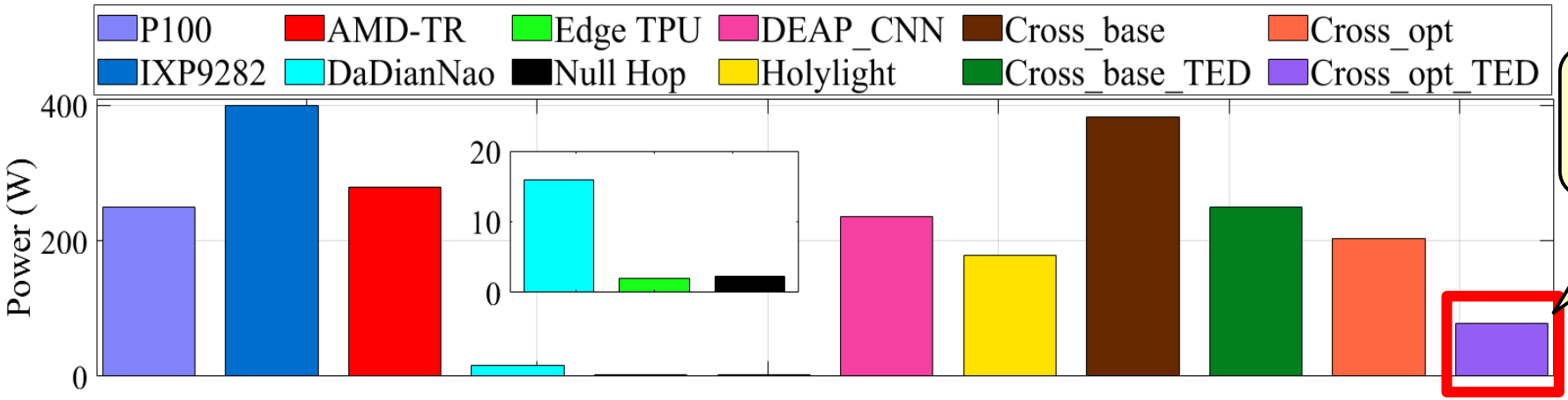




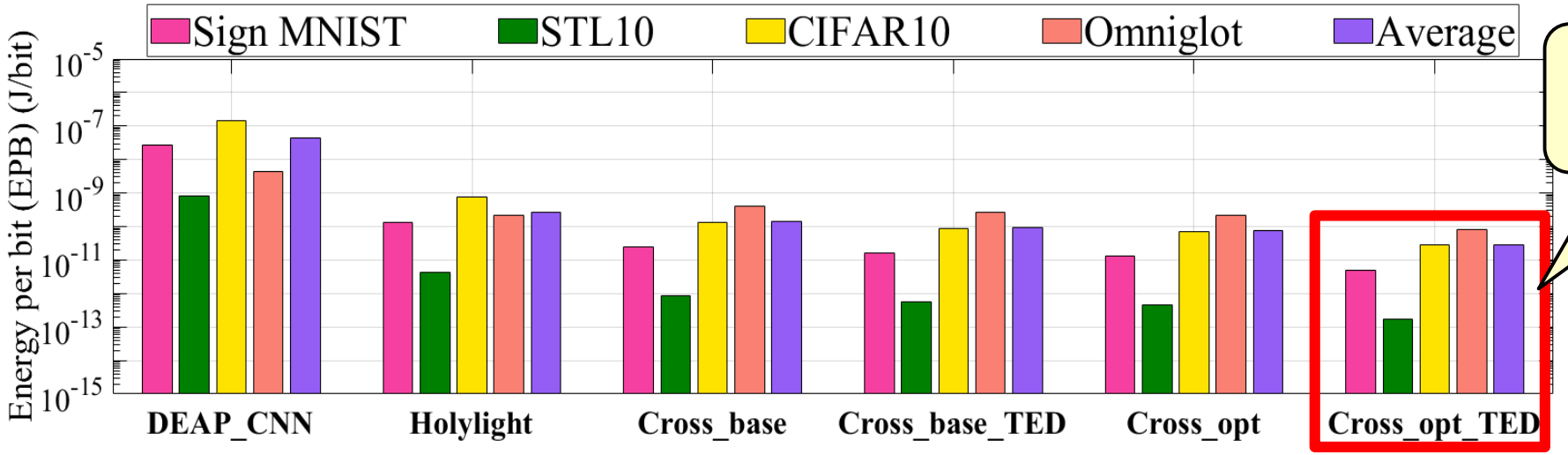




ML Accelerator Comparison

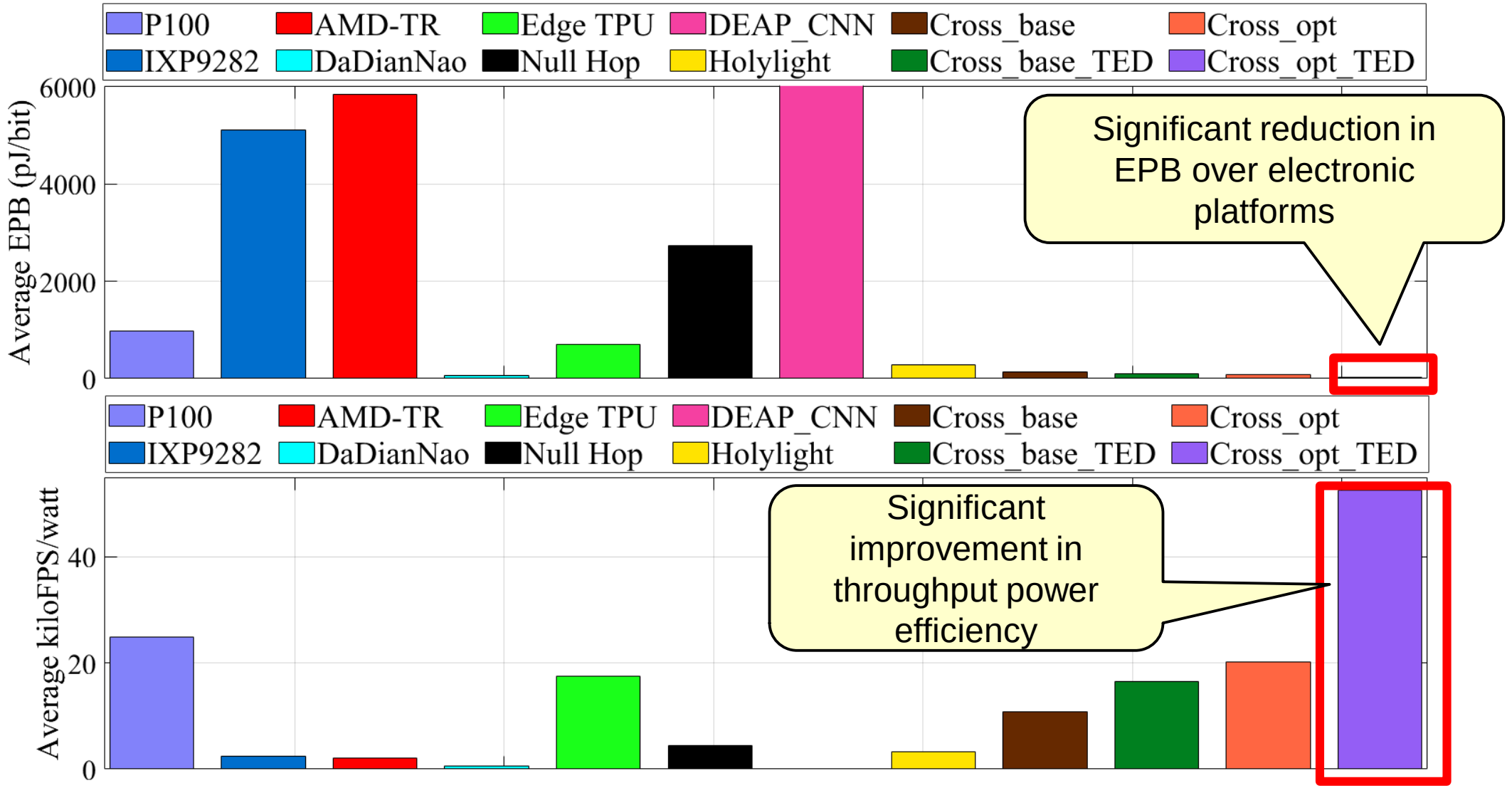


75 W power consumption



9.5x lower than HolyLight

ML Accelerator Comparison



Conclusion

- **CrossLight utilizes silicon photonic device-level optimizations along with tuning circuit and architecture level optimizations**
  - These optimization result in FPV and thermal crosstalk resilience, and lower laser and tuning power consumption
- **CrossLight is able to show improvements in FPS/Watt and EPB**
  - 9.5x better EPB, 15.9x better FPS/Watt vs. HolyLight [W. Liu et al., DATE, 2019]
- **These results showcase the effectiveness of cross-layer optimization efforts in realizing photonic NN accelerators**

Thank you!

Sudeep Pasricha (sudeep@colostate.edu)

Super-smart collaborators

Mahdi Nikdast

Febin Sunny

Asif Mirza

Generous sponsor acknowledgement