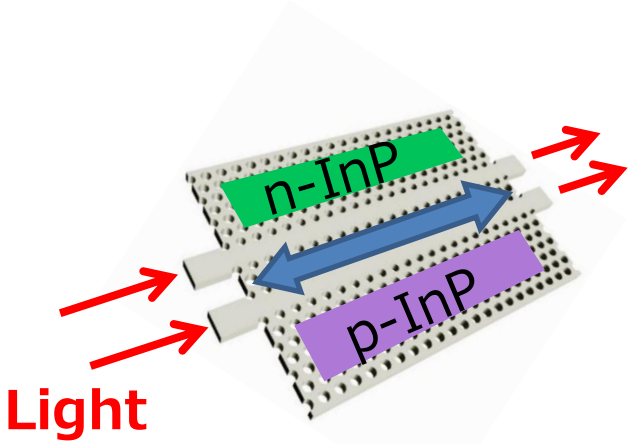
Integrated Optical Neural Networks Exploiting Light Speed Approximate Parallel Multipliers

Jun Shiomi[1], Tohru Ishihara[2], Hidetoshi Onodera[1],
Akihiko Shinya[3], Masaya Notomi[3]

[1]Graduate School of Informatics, Kyoto University, Japan
[2]Graduate School of Informatics, Nagoya University, Japan
[2]NTT Nanophotonics Center / NTT Basic Research Laboratories, Japan

Integrated Optical Logic Circuits Using Nanophotonics



Device length ~ 100 μm

E.g. Photonic crystal-based directional coupler

✓ Direction control on the order of the light wavelength

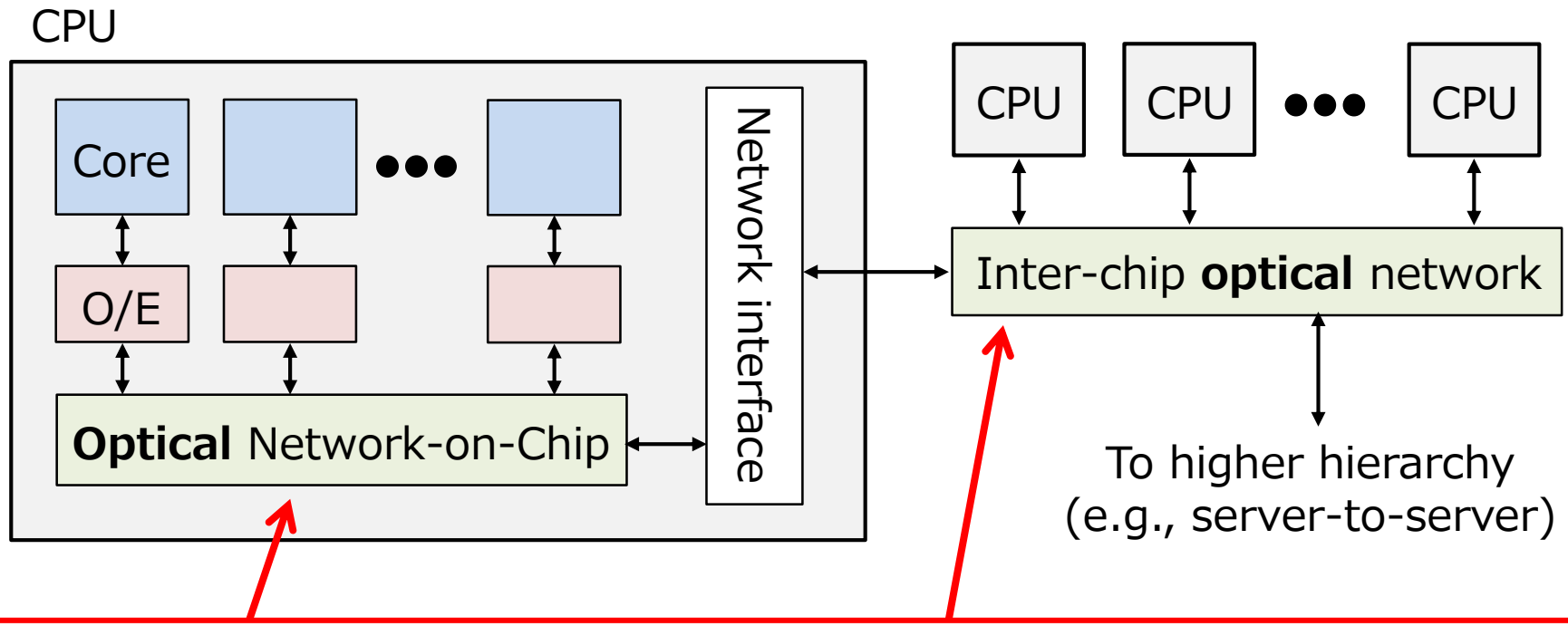
➡
- ☺ On-chip implementation
- ☺ **Ultra-low latency**

CMOS logic gate	~ 10 ps
Nanophotonics-based	0.1 ps ~ 1 ps

Goal: Ultra-fast optical logic circuit design

Beyond Optical Communication

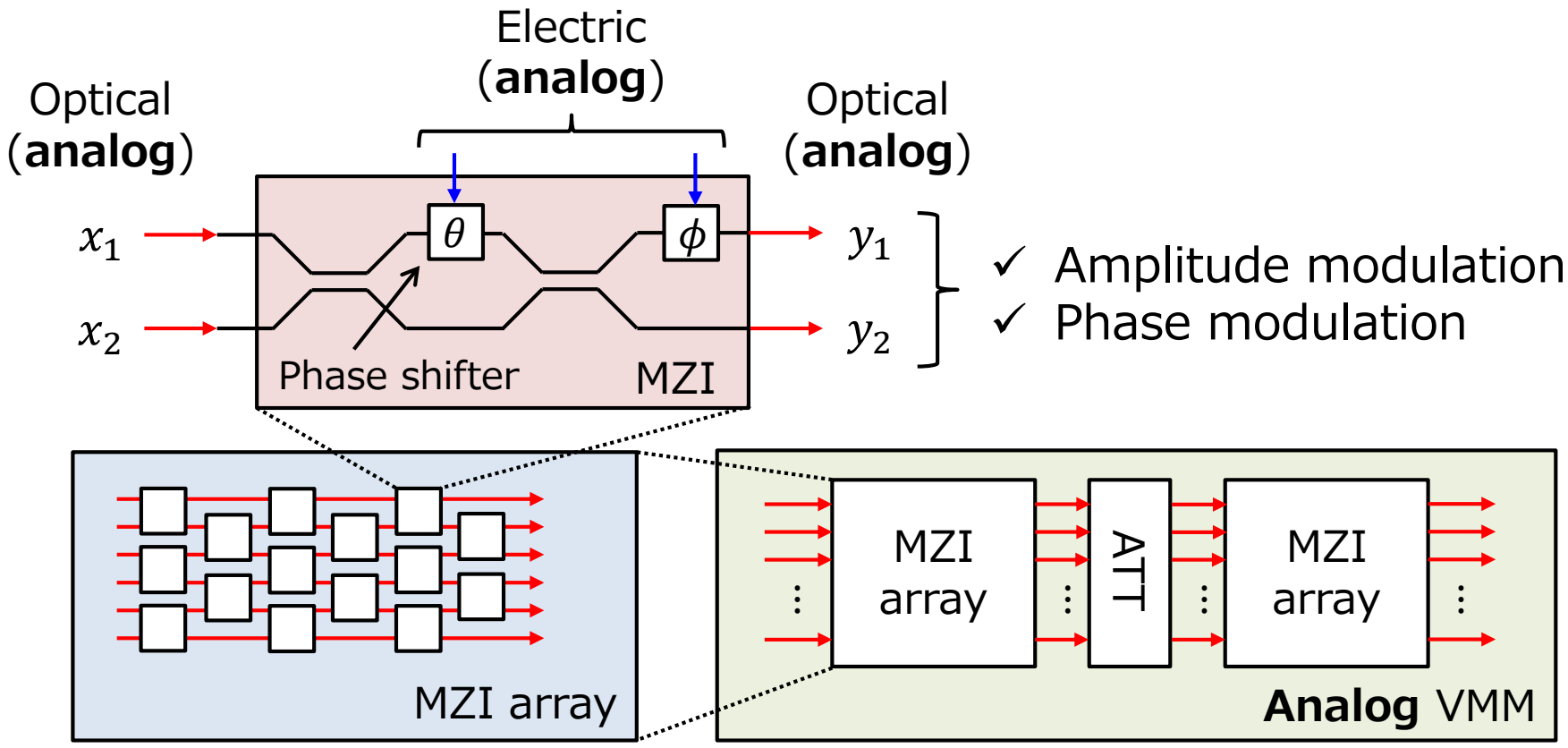
Existing technique's focus: on-chip optical communication



Goal: Add **functional unit** to boost up on-chip communication

E.g. Integrated Neural Network (NN) for packet filtering

Related Work: Vector Matrix Multiplication (VMM) Exploiting Mach–Zehnder Interferometer (MZI) [1]



Challenges in analog multiplication:

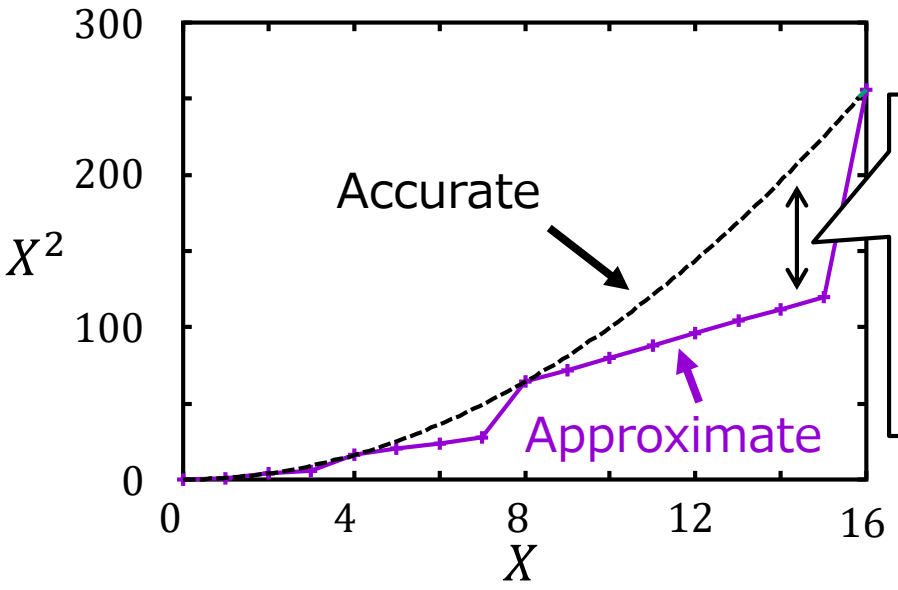
Environment-dependent error (e.g. noise & temperature)

[1] M. Reck et al., “Experimental realization of any discrete unitary operator”, Phys. Rev. Lett 73, 1994

Fully Digital Approximate Parallel Multiplier

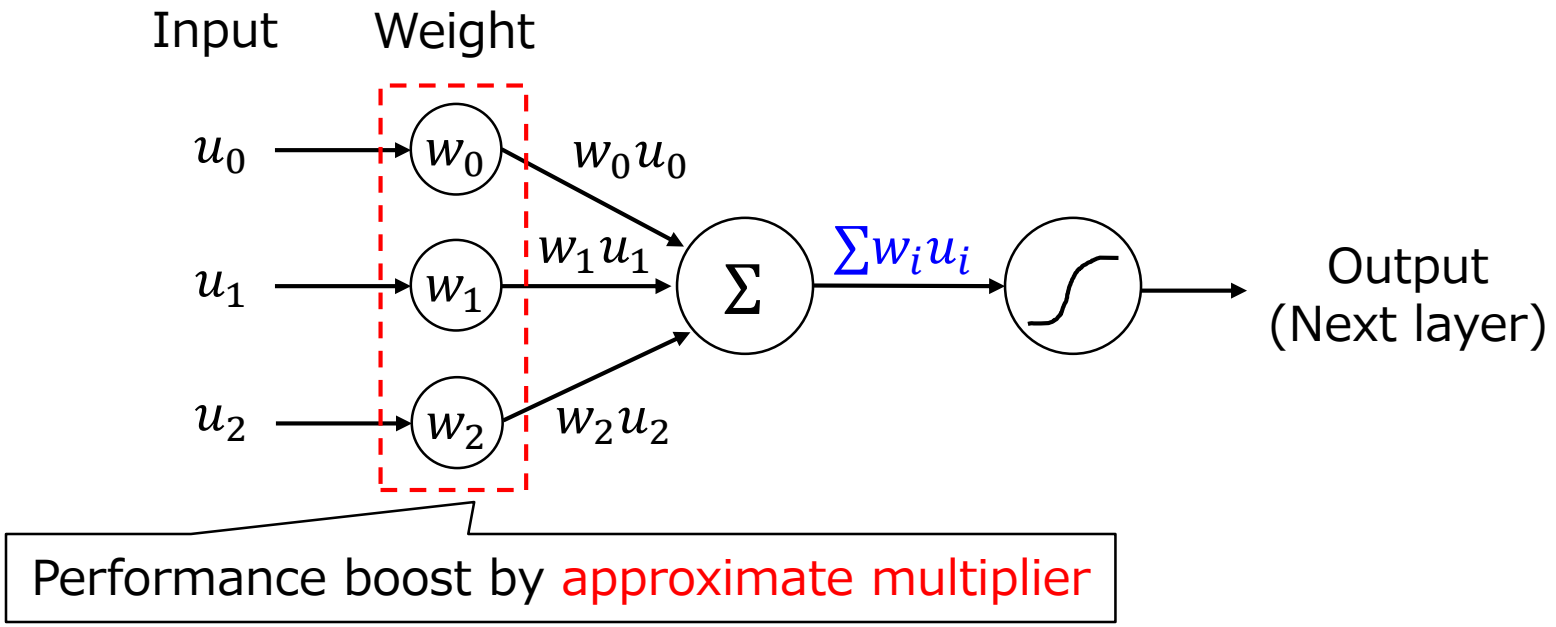
Accurate digital multiplier: larger delay than an analog multiplier

- ✓ This work: **Approximate structure** for performance boost
- ☹ Deterministic multiplication error



- ✓ Dependent on input value only
- ✓ Fixed error @ every calculation
  - ➢ Up to −50% error

Neural Network (NN) Exploiting Approximate Multiplier



- ✓ Pattern recognition: high resiliency to approximation errors
- ✓ Absorb the deterministic error by learning process

$$\boldsymbol{w_0}u_0 + \boldsymbol{w_1}u_1 + \boldsymbol{w_2}u_2$$

Weight vector optimization to cancel out deterministic error

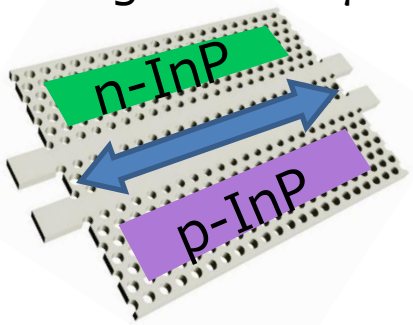
Outline

- Background
- Parallel Multiplier Using Nanophotonic Devices
- Approximate Parallel Multiplier for NN
- Conclusion

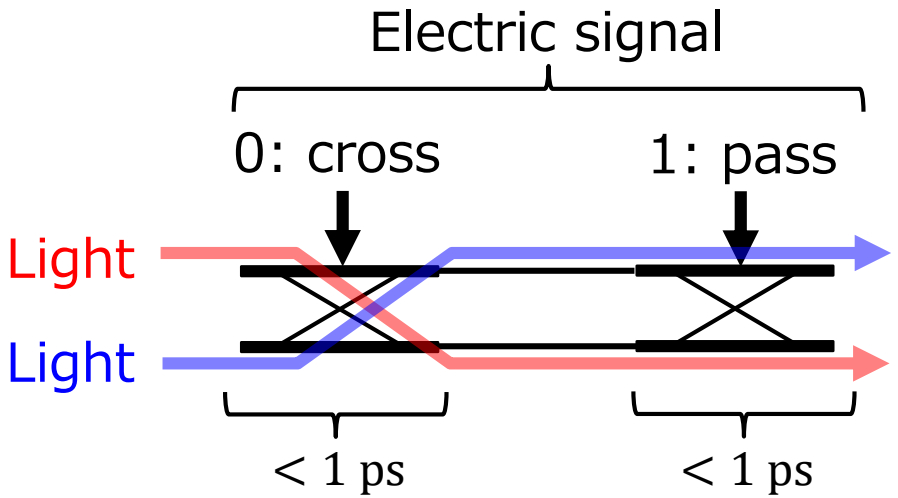
Photonic Crystal-Based Optical Pass-Gate (OPG)

Directional coupler

Length > 100 μm

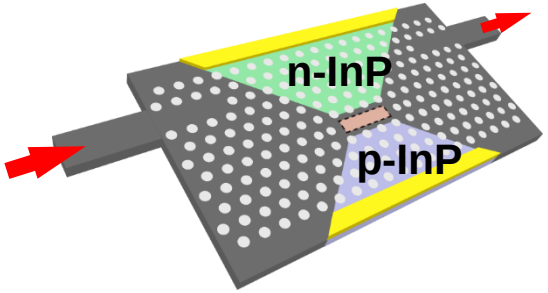


≡

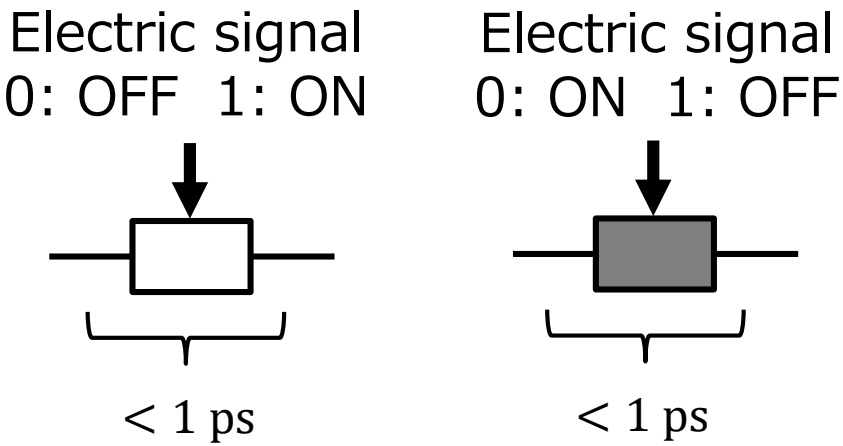


Modulator (switch)

Length ~1.3 μm

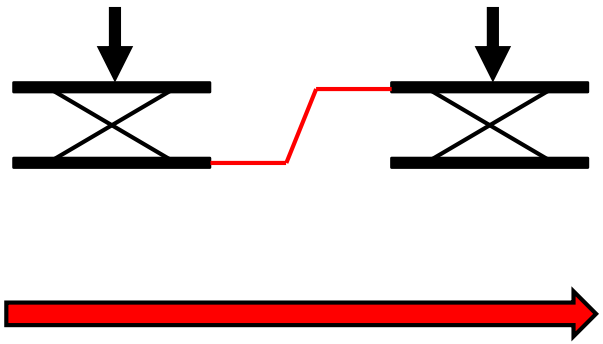


≡



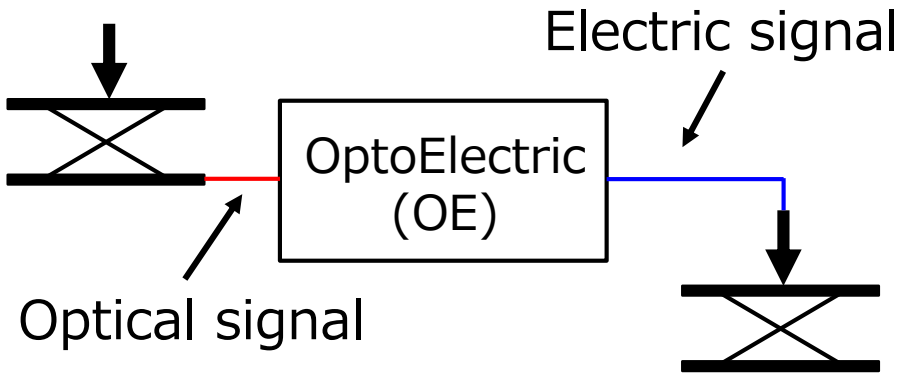
OptoElectric (OE) Conversion Delay

Serial connection



Light speed (~ **1 ps**/gate)

Cascade connection



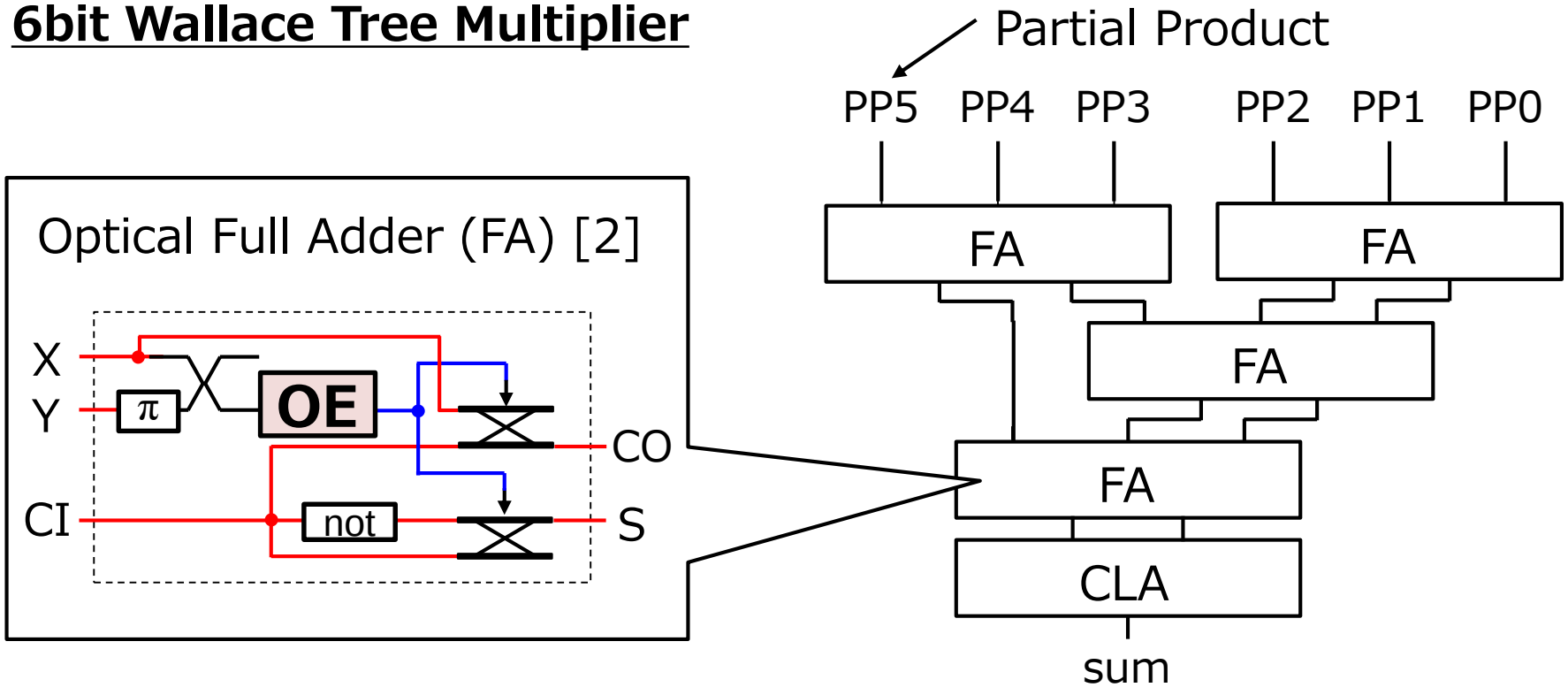
OE conversion (~ **25 ps**/OE)

✓ **Reducing OEs on a critical path** is a key to ultra fast operation

cf. CMOS ~ 10 ps/gate

Issues in Conventional Optical Parallel Multiplier

6bit Wallace Tree Multiplier



☹ Large #(OEs on critical path) for large bit width n

⇨ Unacceptable latency for large n

[2] T. Ishihara, et al., International SoC Design Conference, 2016

Outline

- Background
- Parallel Multiplier Using Nanophotonic Devices
- Approximate Parallel Multiplier for NN
- Conclusion

Log-quantized Approximate Multiplier

Weight → W × U ← Data

Key idea: $W \times U = W \times \mathrm{sign}\,(U) \times 2^{\log_2|U|} \quad (W, U \in \mathbb{R})$

$\log_2|U| \simeq \underline{\mathbf{floor}}(\log_2 U) := \tilde{u}$

Quantization to an integer

Example ($U = 42$)

$U =$ 0 0 1 0 1 0 1 0

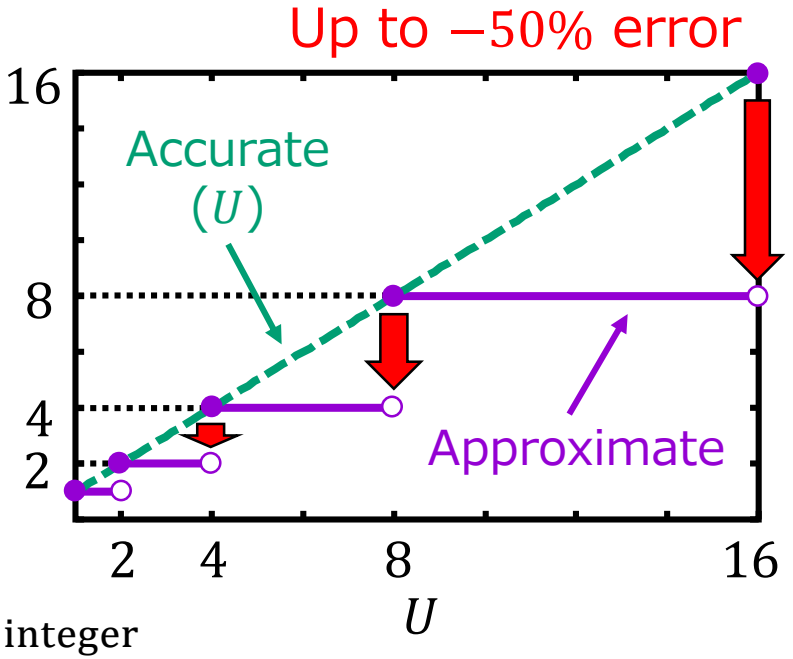
MSB LSB

Log-quantization

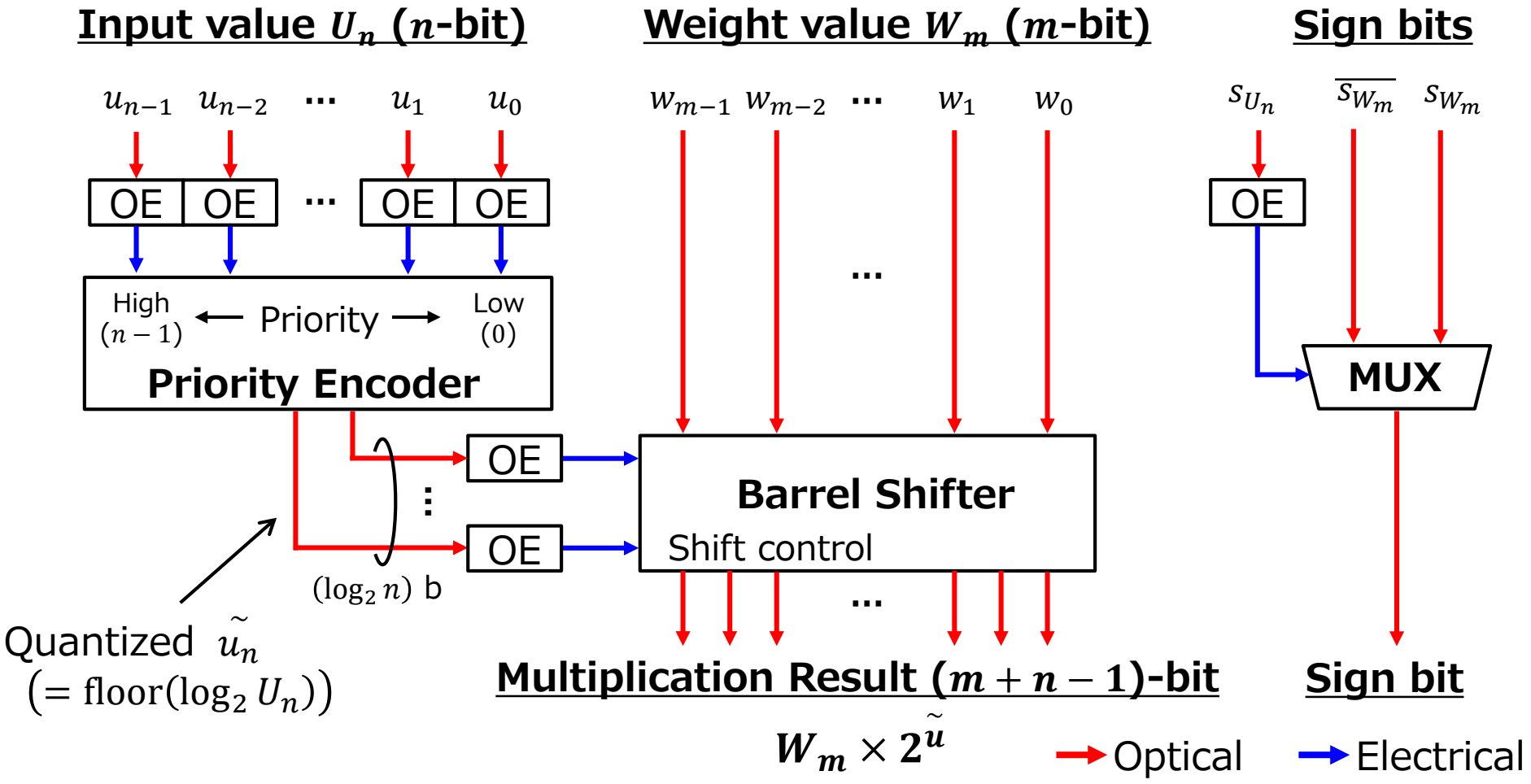
$2^{\tilde{u}} =$ 0 0 1 0 0 0 0 0

MSB LSB

Quantization to 2^{integer}

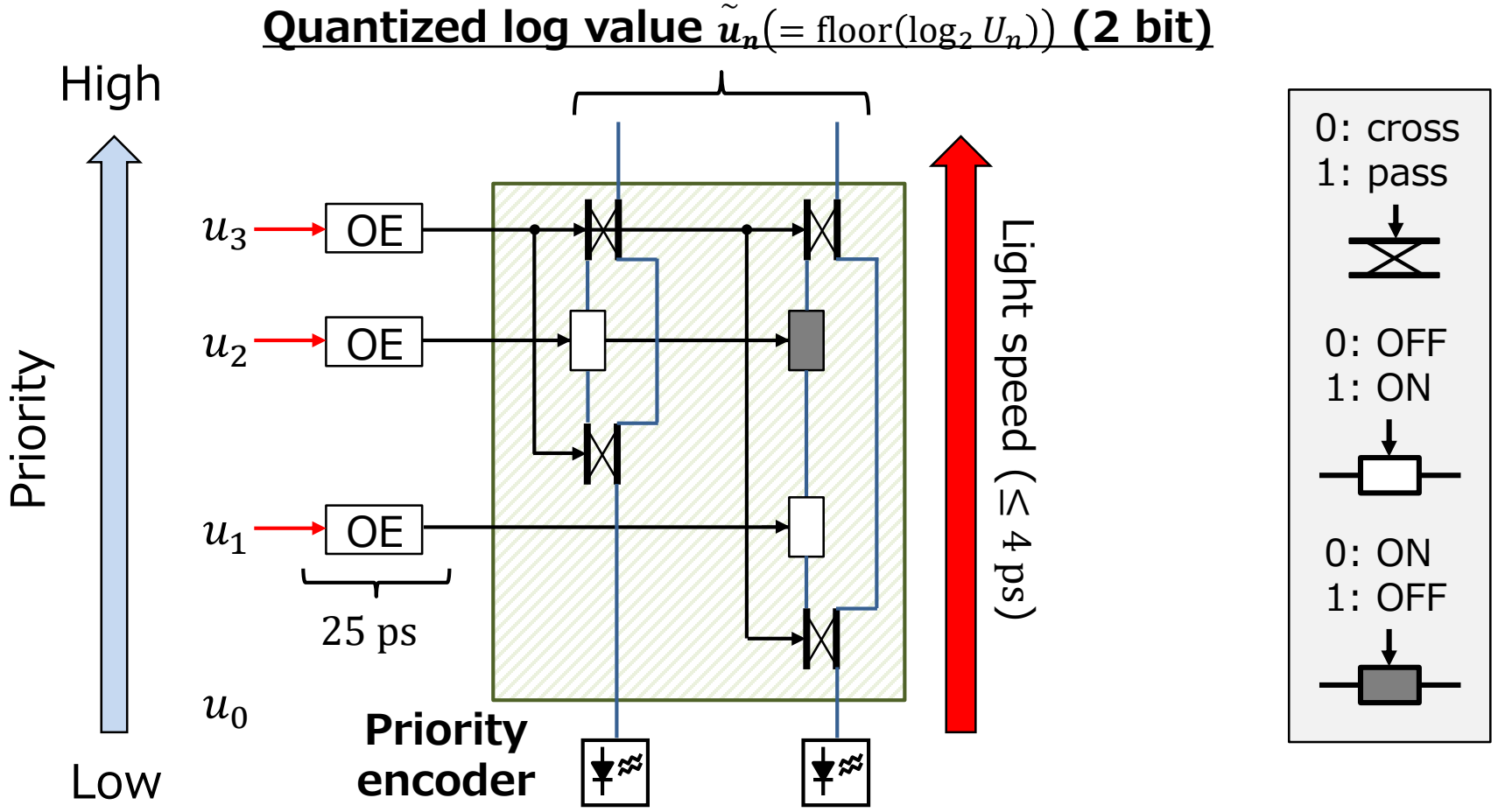


Overview of the Approximate Multiplier



✓ Only two OE converters on critical paths for any bit widths

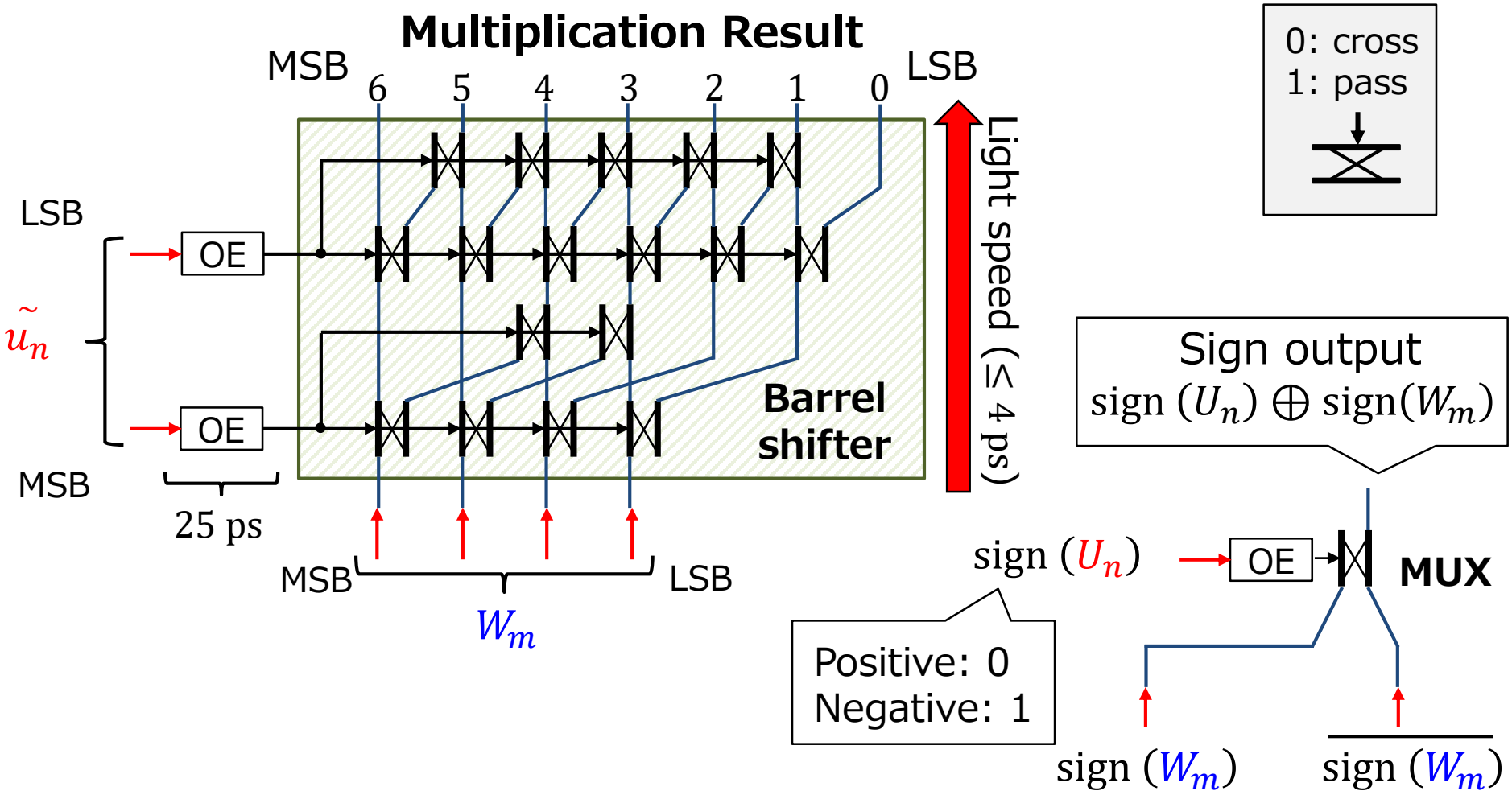
Optical Implementation of the Log-Quantizer ($n = 4$)



✓ Only one OE converter on a critical path for any n

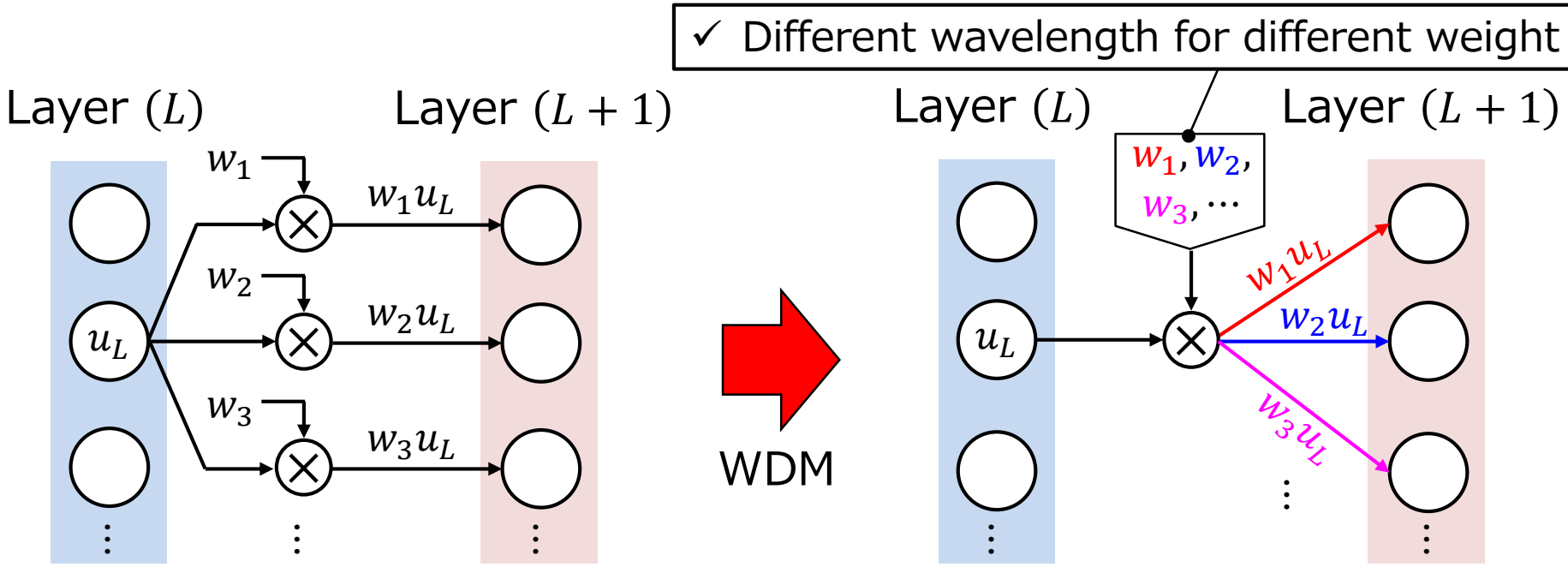
Barrel Shifter-based Multiplier ($n = 4$)

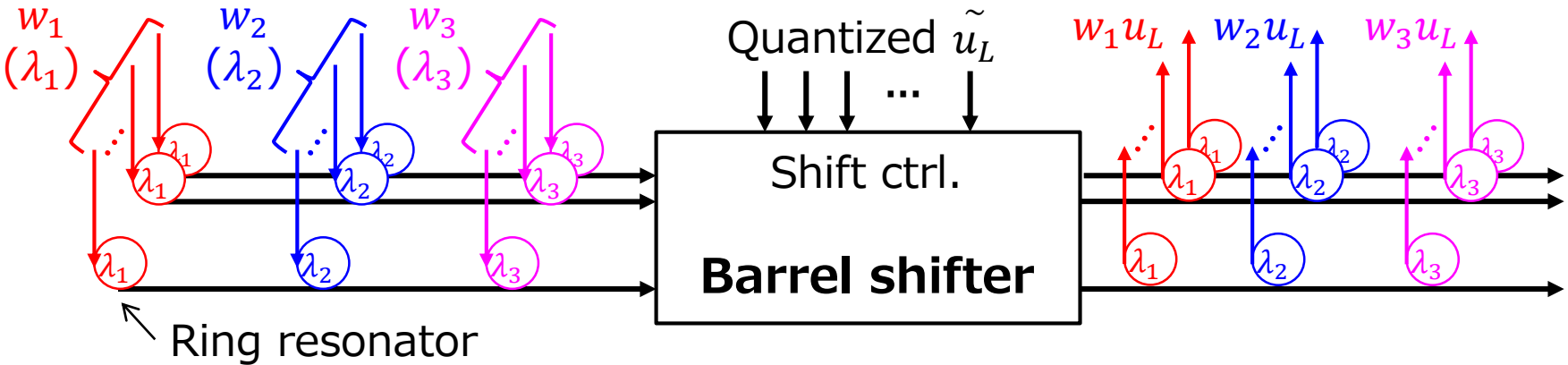
$W_m \times 2^{\tilde{u}_n}$



✓ Only one OE converter on a critical path for any n

Area Reduction of Optical NN Circuits Exploiting Wavelength Division Multiplexing (WDM)





Conclusion

- Parallel multiplier for neural networks
  - Performance boost by approximate structure
  - Priority encoder & barrel shifter
  - Area reduction exploiting WDM
- Future work
  - Detailed analysis
  - Detailed implementation

Acknowledgement

This work is partly supported by CREST (Core Research for Evolutional science and Technology) of JST (Japan Science and Technology Corporation)