Optical and Electrical Computing Energy Use Comparison

MEC/DARPA Optical Computing Workshop

Electronics versus Optics at the System Level

2 April 2022

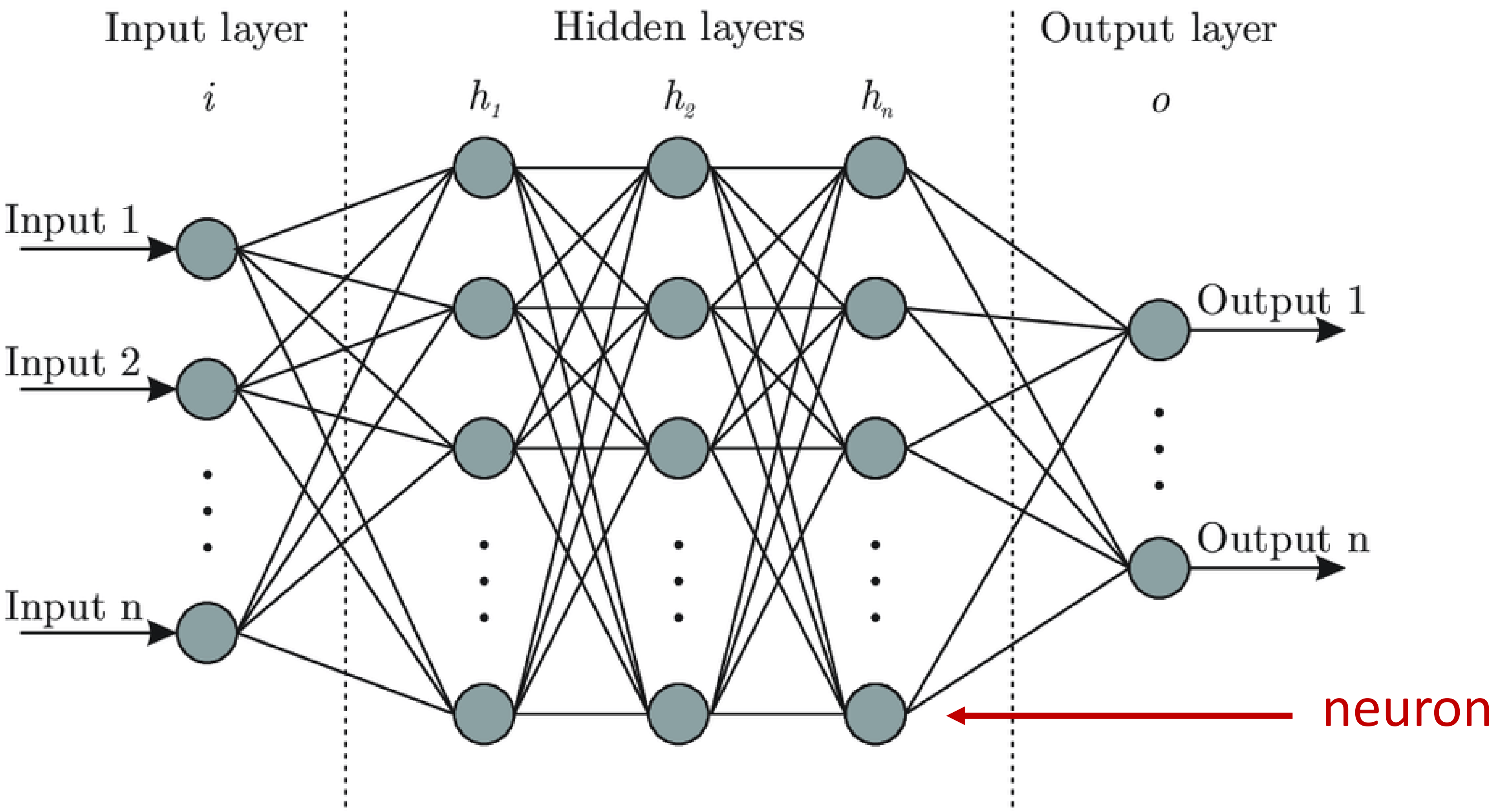
Chris Cole, II-VI Incorporated

Outline

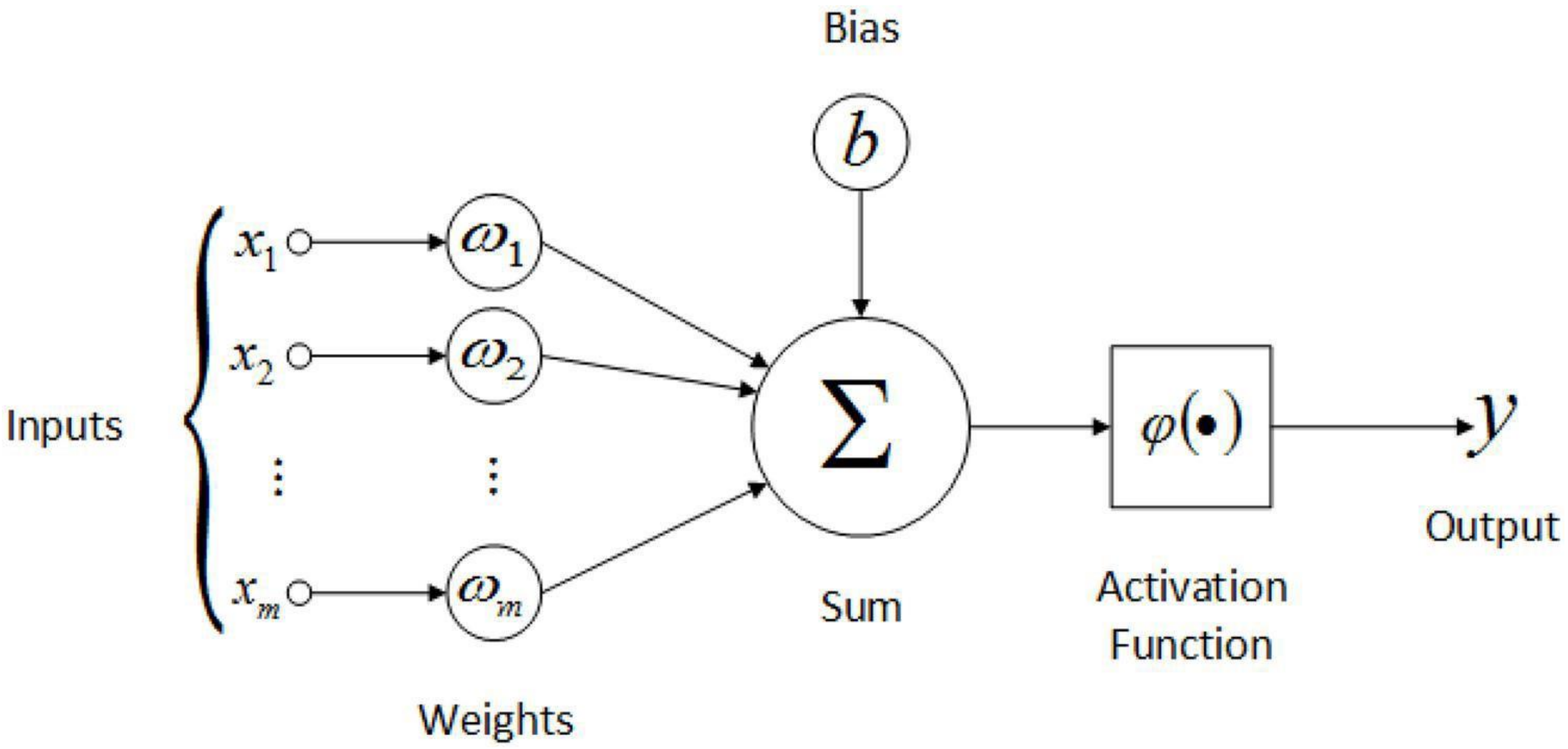
➢ Neural Networks Demystified

- Optical Computing Examples
- Addition
- Multiplication
- Convolution
- Discussion

Inference Neural Network



Neuron


Bias
b
Inputs
x_1
x_2
x_m
ω_1
ω_2
ω_m
Weights
Σ
Sum
$\varphi(\bullet)$
Activation Function
y
Output

Machine Learning Terms

Machine Learning Term		Computation	
Inference			
input	inputs	x	**x**
weight	weights	w	W
Applying weights		w x	W **x**
bias	biases	b	**b**
Activation		*f*(•)	
output	outputs	y	**y**
Neuron		y = *f*(**w**’ **x** + b)	
Layer (except input)		**y** = *f*(W **x** + **b**)	
Training			

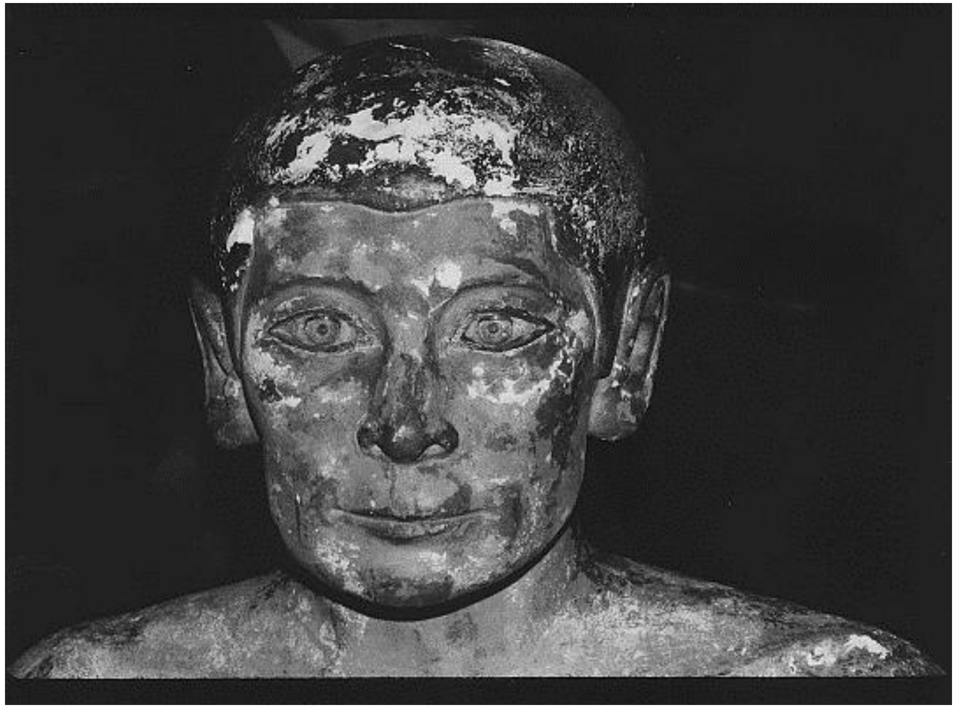
Machine Learning Terms and their Communication Equivalents

Machine Learning Term		Computation		Communication Term	
Inference				Signal Processing	
input	inputs	x	$\mathbf{x}$	input	input vector
weight	weights	w	W	coefficient	coefficient matrix
Applying weights		$w\ x$	$W\ \mathbf{x}$	Filtering	
bias	biases	b	$\mathbf{b}$	threshold	threshold vector
Activation		$f(\bullet)$		Detection	
output	outputs	y	$\mathbf{y}$	output	output vector
Neuron		$y = f(\mathbf{w}'\ \mathbf{x} + b)$		MISO Receiver	
Layer (except input)		$\mathbf{y} = f(W\ \mathbf{x} + \mathbf{b})$		MIMO Receiver	
Training				Adaptation	

Outline

- Neural Networks Demystified
- ➢ Optical Computing Examples
- Addition
- Multiplication
- Convolution
- Discussion

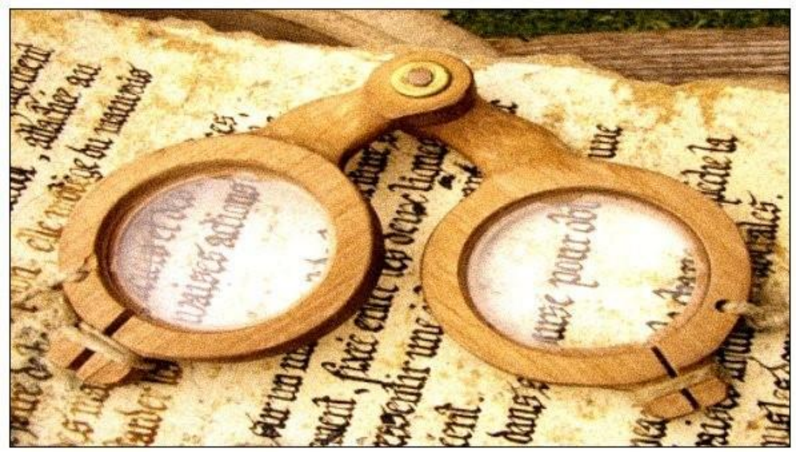
1st Optical Computing Example: 4600-year-old Egyptian Lens



- The "eyes" appear to follow the observer as they move about the statue
- On display at Louvre Museum, Paris

Widely Used Optical Computing Example: Eyeglasses

- Two lenses in a wooden frame, Italy, 1280’s
- Lens processing is 2-D spatial filtering or 2-D convolution, i.e., inference
- A hypothetical electronic lens processes 24-bit RGB 512x512 pixel image at 120 frames/sec
  - → ~25 trillion 8-bit Multiply-Accumulates/sec
- Zero incremental energy

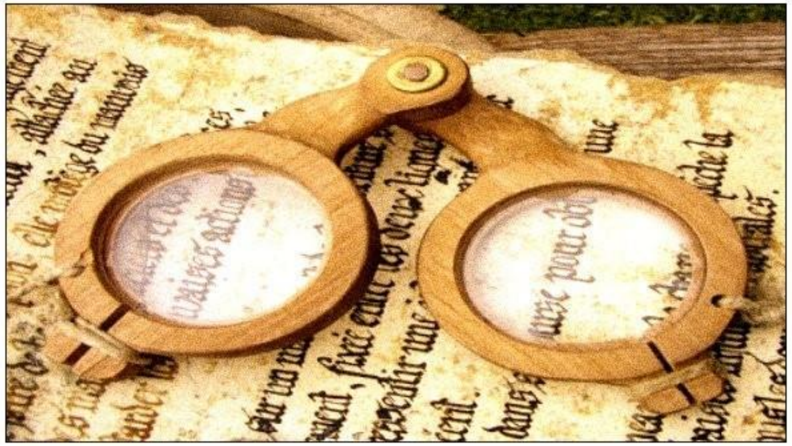


Widely Used Optical Computing Example: Eyeglasses

- Two lenses in a wooden frame, Italy, 1280’s
- Lens processing is 2-D spatial filtering or 2-D convolution, i.e., inference
- A hypothetical electronic lens processes 24-bit RGB 512x512 pixel image at 120 frames/sec
  → ~25 trillion 8-bit Multiply-Accumulates/sec
- Zero incremental energy

- Problem: fixed focus (fixed coefficients)
- Solution: Ben Franklin bi-focal eyeglasses
  → 1 bit of programmability

- How do we train eyeglasses?



Training Eyeglasses

Telecom Optical Computing Example: DCF

- DCF (Dispersion Compensation Fiber) used in every Telecom link in the ’90s
- Passive, complex signal processing
- Zero incremental energy use (ignoring amplification for loss)
- Fixed compensation; requires a custom length spool for every link
- The only optical compensation approach despite extensive R&D into alternatives

Telecom Optical Computing Example: DCF

- DCF (Dispersion Compensation Fiber) used in every Telecom link in the ’90s
- Passive, complex signal processing
- Zero incremental energy use (ignoring amplification for loss)
- Fixed compensation; requires a custom length spool for every link
- The only optical compensation approach despite extensive R&D into alternatives
- Coherent DSP CMOS ASIC with adaptive equalization introduced 20 years ago
- Over time, completely replaced DCF and all other optical compensation techniques

Telecom Optical Computing Example: DCF

- DCF (Dispersion Compensation Fiber) used in every Telecom link in the ’90s
- Passive, complex signal processing
- Zero incremental energy use (ignoring amplification for loss)
- Fixed compensation; requires a custom length spool for every link
- The only optical compensation approach despite extensive R&D into alternatives
- Coherent DSP CMOS ASIC with adaptive equalization introduced 20 years ago
- Over time, completely replaced DCF and all other optical compensation techniques
- Same thing as happened to all successful analog computing approaches; there were all replaced by digital computing

Outline

- Neural Networks Demystified
- Optical Computing Examples

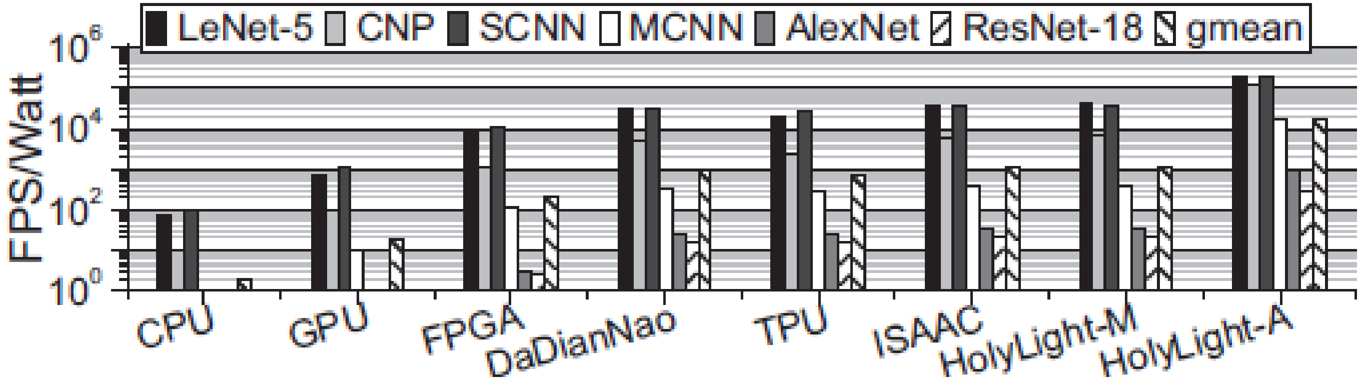
➢ **Addition**

- Multiplication
- Convolution
- Discussion

Optical Computing Example 1: Nanophotonic Accelerator

HolyLight (DATE Conf., Mar. 2019)

- micro-disk adders and shifters
- claimed **10x to 100x** lower energy compared to conventional GPUs and TPUs



- 42mW central-computing-block made up of 16x 16-bit 13 Gops/sec optical adders

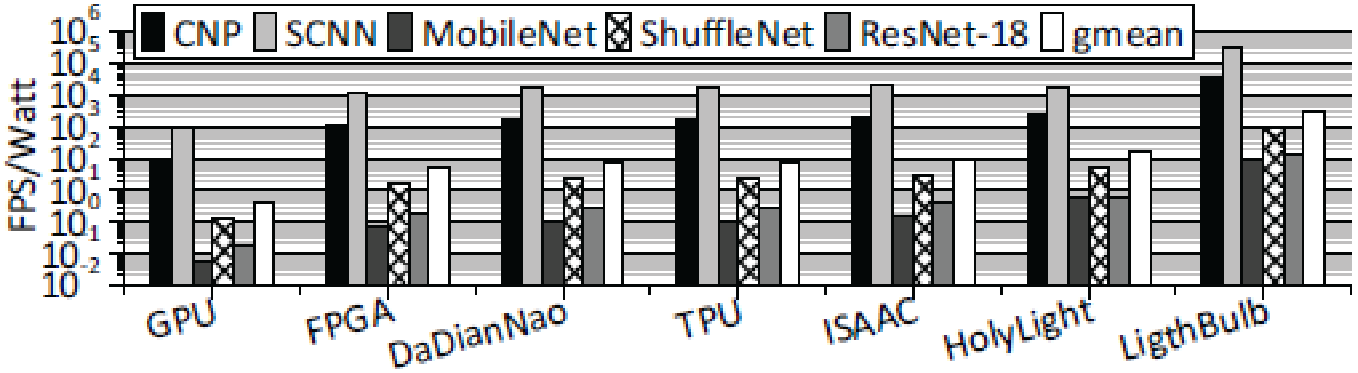
  → optical 13 Gops/sec adder energy use = **13 fJ/bit**

W. Liu, W. Liu, Y. Ye and Q. Lou, "HolyLight: A Nanophotonic Accelerator for Deep Learning in Data Centers," in Design, Automation & Test in Europe Conference & Exhibition (DATE), pp. 1483-1488, March 2019.

Optical Computing Example 2: Nanophotonic Accelerator

LightBulb (DATE Conf., Mar. 2020)

- HolyLight enhanced with photonic local storage registers
- claimed **20x to 600x** lower compared to conventional GPUs and TPUs



- 1060mW central-computing-block made up of 25x 16-bit 50 Gops/sec optical adders

  → optical 50 Gops/sec adder energy use = **53 fJ/bit**

F. Zokaee, Q. Lou, N. Youngblood and W. Liu, "LightBulb: A Photonic-Nonvolatile-Memory-based Accelerator for Binarized Convolutional Neural Networks," in Design, Automation & Test in Europe Conference & Exhibition (DATE), pp. 1438-1443, March 2020.

CMOS Adder Energy Use

CMOS node	Delay	Energy/op (max)	Input	Rate	Energy
nm	ps	fJ	bits/op	Gops/s	fJ/bit
7	40	50	16	25	2.9
7	30	40	16	33	2.5
				average	**2.7**

Q. Xie, X. Lin, S. Chen, M. Dousti and M. Pedram, "Performance Comparisons between 7nm FinFET and Conventional Bulk CMOS Standard Cell Libraries," *IEEE Transactions on Circuits and Systems II: Express Briefs,* vol. 62, no. 8, pp. 761-765, August 2015.

A. Vatanjou, E. Lte, T. Ytterdal and S. Aunet, "Ultra-low Voltage and Energy Efficient Adders in 28nm FDSOI Exploring Poly-biasing for Device Sizing," Microprocessors & Microsystems, vol. 56, no. C, pp. 92-100, February 2018.

A. Stillmaker and B. Baas, "Scaling equations for the accurate prediction of CMOS device performance," *Integration the VLSI journal,* vol. 58, pp. 74-81, February 2017.

Optical Computing Examples 1 & 2 Energy Use Comparison

Ex. 1 & 2 claim Optical Computing uses much less energy than CMOS:

- Ex. 1 claim: Nanophotonic accelerator is **1/10 to 1/100** of CMOS GPU/TPU
- Ex. 2 claim: Nanophotonic accelerator is **1/20 to 1/600** of CMOS GPU/TPU

Optical Computing Examples 1 & 2 Energy Use Comparison

Ex. 1 & 2 claim Optical Computing uses much less energy than CMOS:

- Ex. 1 claim: Nanophotonic accelerator is **1/10 to 1/100** of CMOS GPU/TPU
- Ex. 2 claim: Nanophotonic accelerator is **1/20 to 1/600** of CMOS GPU/TPU

Ex. 1 & 2 data show Optical Computing Blocks uses much more energy than CMOS:

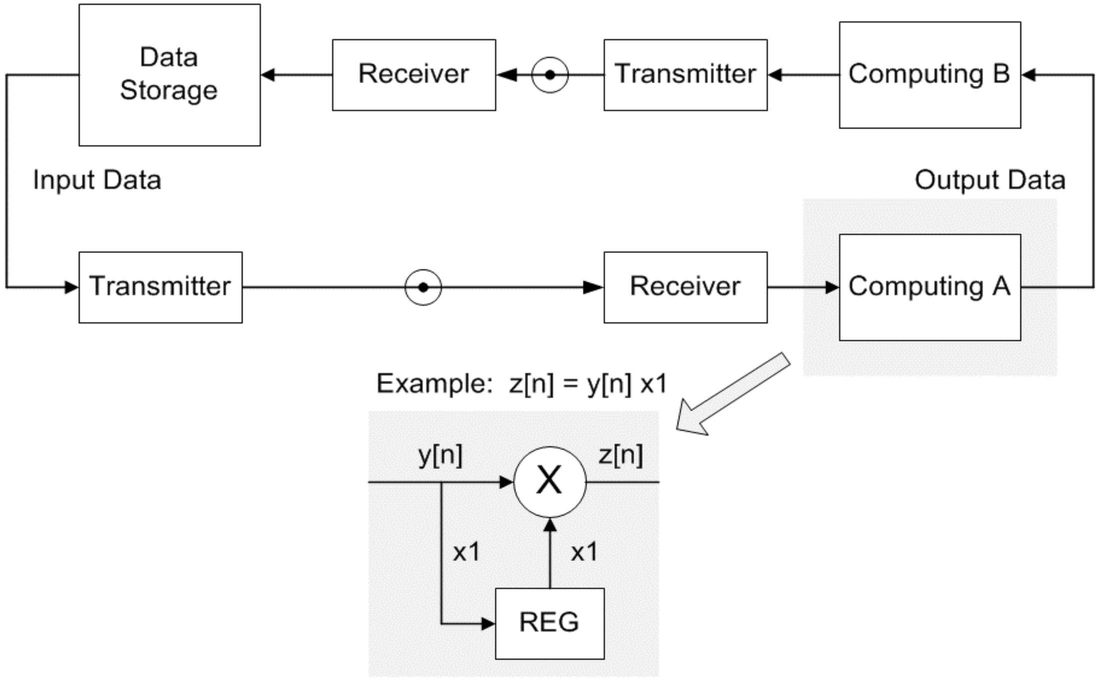
- Ex. 1 data: 13 Gops/sec Optical Block is **5x** of 30 Gops/sec CMOS (**13/2.7**)
- Ex. 2 data: 50 Gops/sec Optical Block is **20x** of 30 Gops/sec CMOS (**53/2.7**)

Computing Addition does not determine overall energy use

These photonic accelerator papers deserve credit for nicely separating assumptions, data, and claims, enabling independent and alternate conclusions.

Data Transfer Compute Model Optimized for Math Operations

Electrical Computing w/ electrical DT

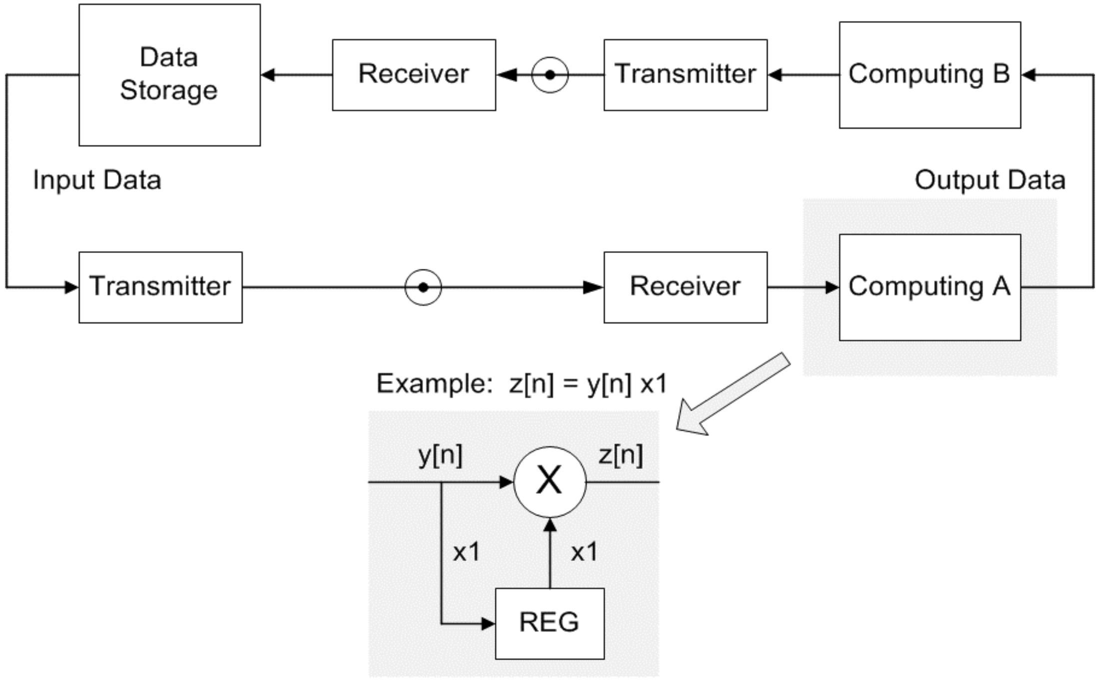


DT: Data Transfer

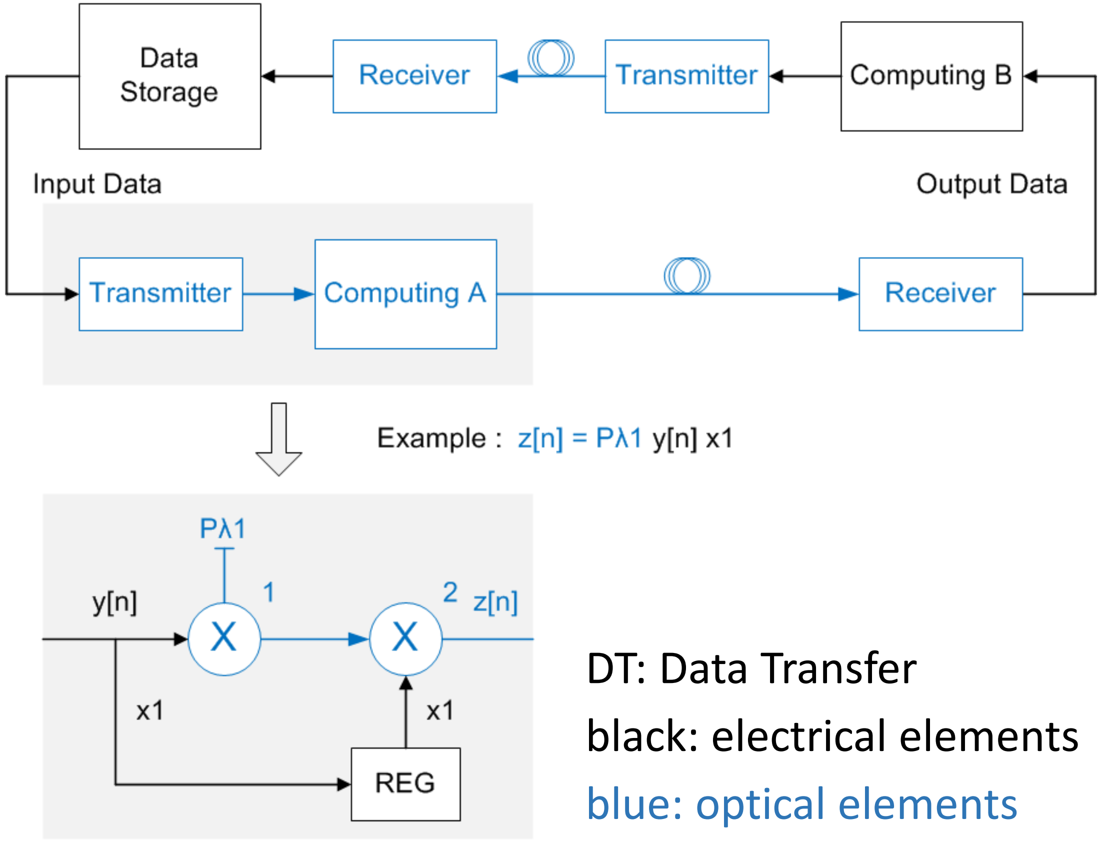
black: electrical elements

Apples-to-Oranges Energy Use Comparison Models

Electrical Computing w/ electrical DT



Optical Computing w/ optical DT



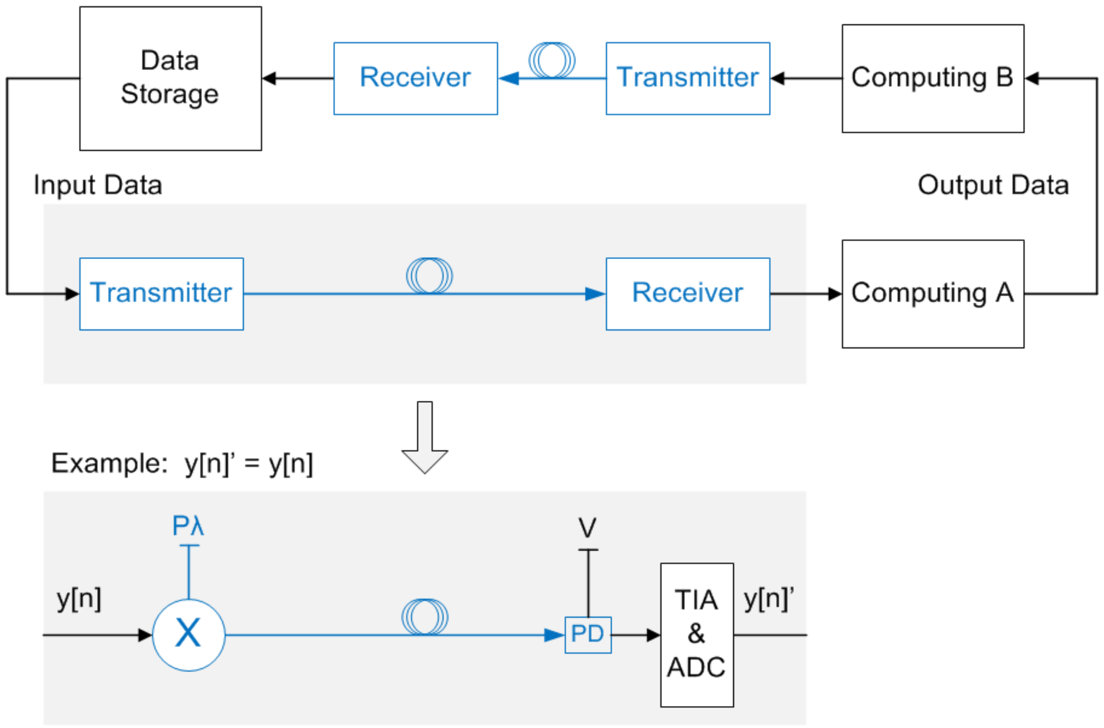
DT: Data Transfer

black: electrical elements

blue: optical elements

Data Transfer Compute Model Optimized for Math Operations

Electrical Computing w/ optical DT



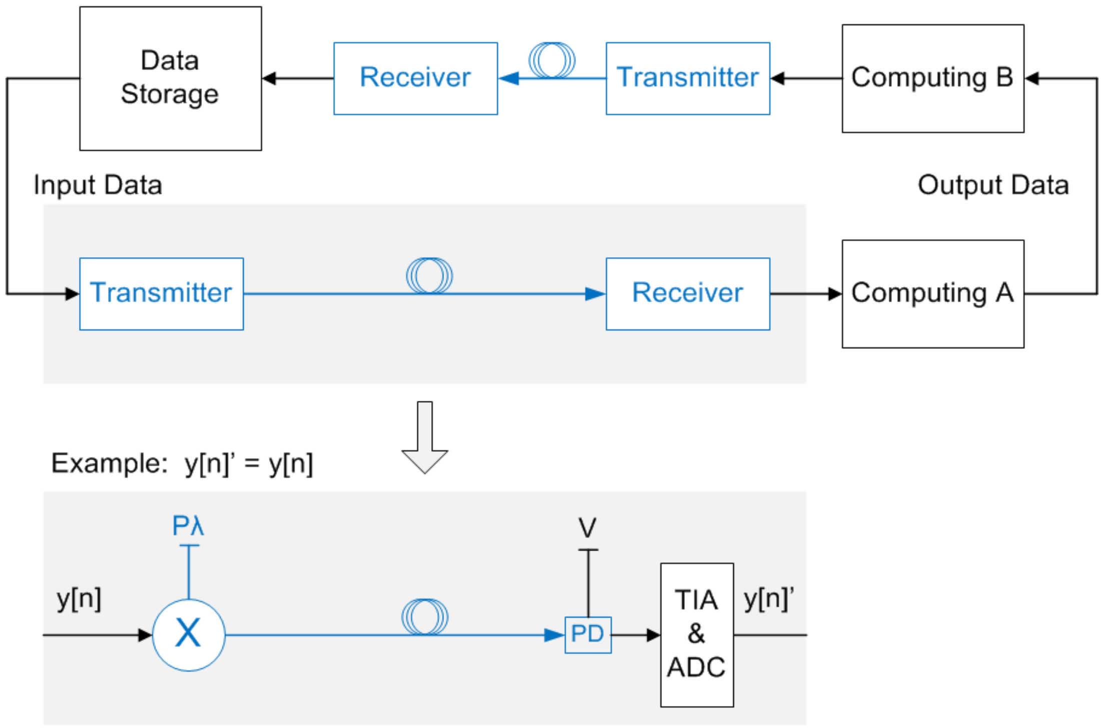
DT: Data Transfer

black: electrical elements

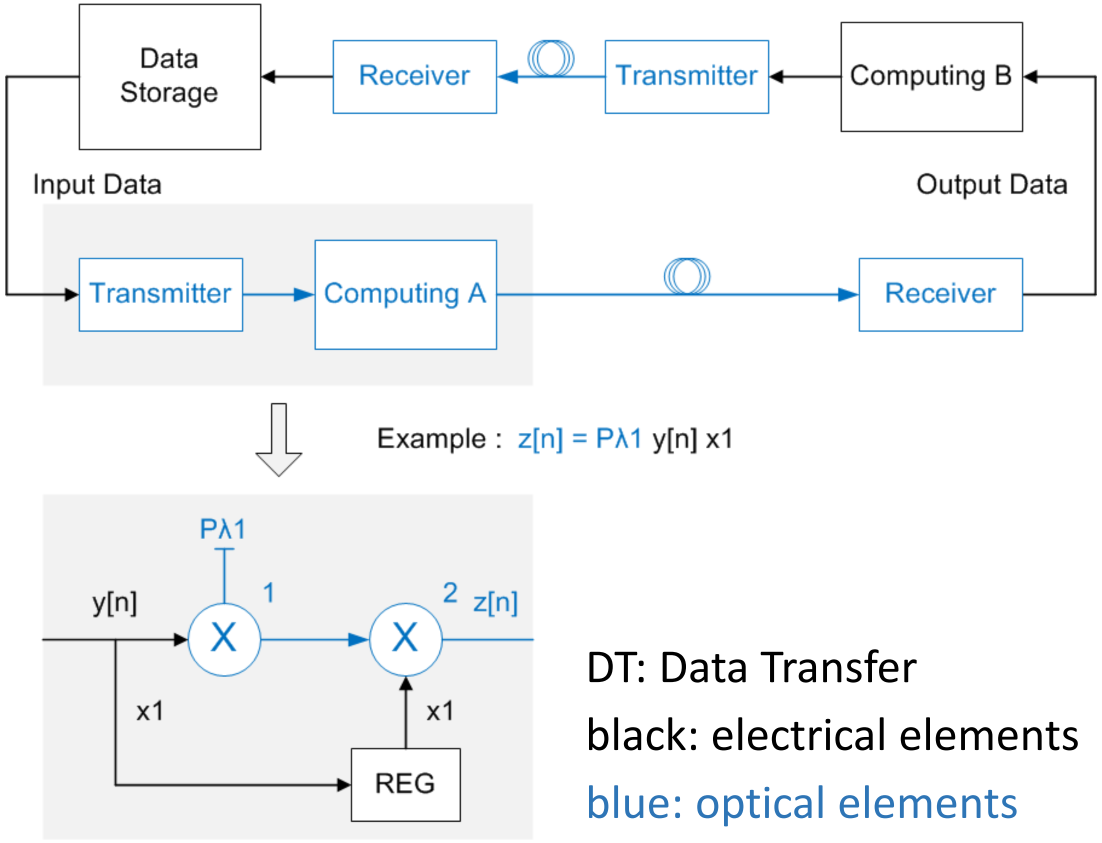
blue: optical elements

Apples-to-Apples Energy Use Comparison Models

Electrical Computing w/ optical DT



Optical Computing w/ optical DT

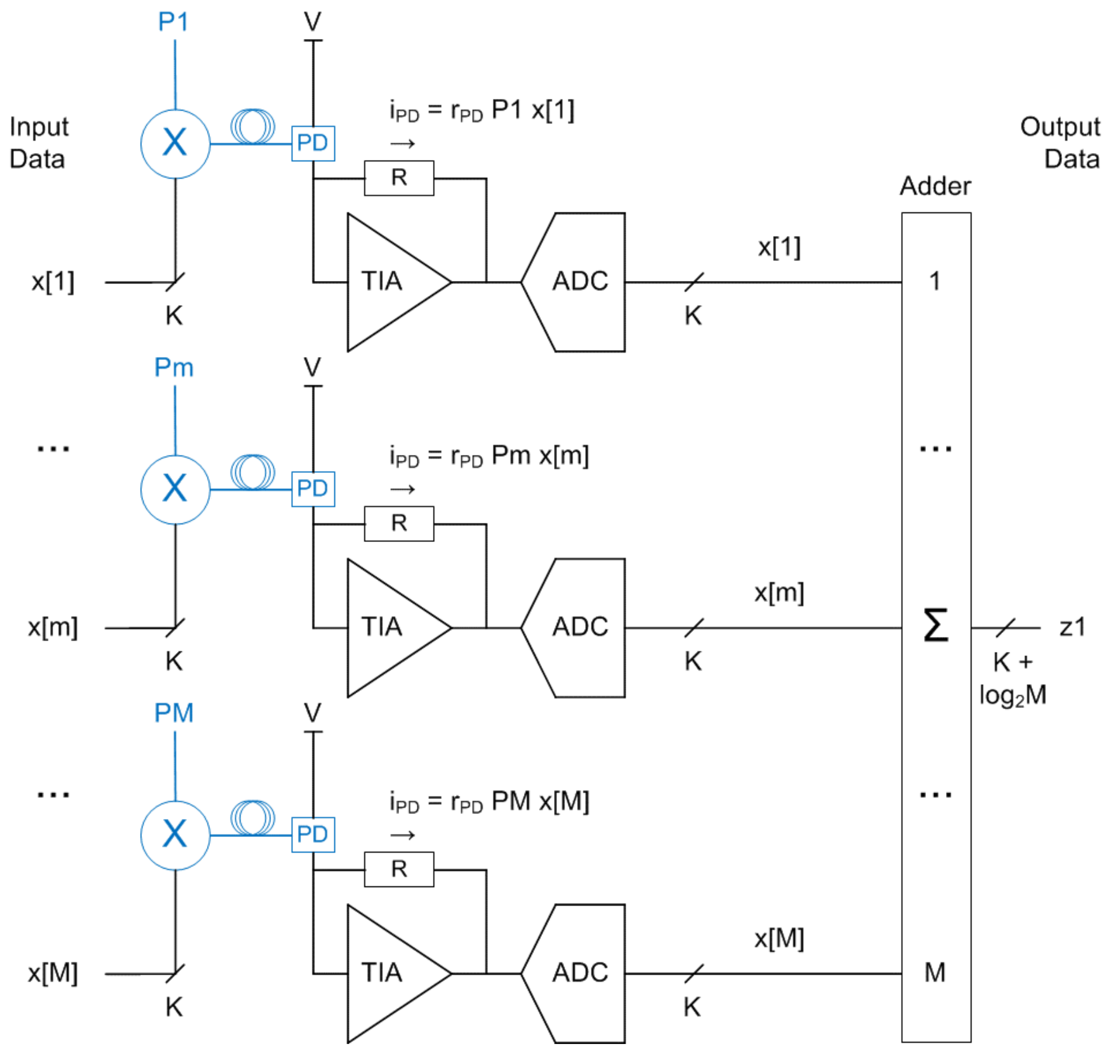


DT: Data Transfer

black: electrical elements

blue: optical elements

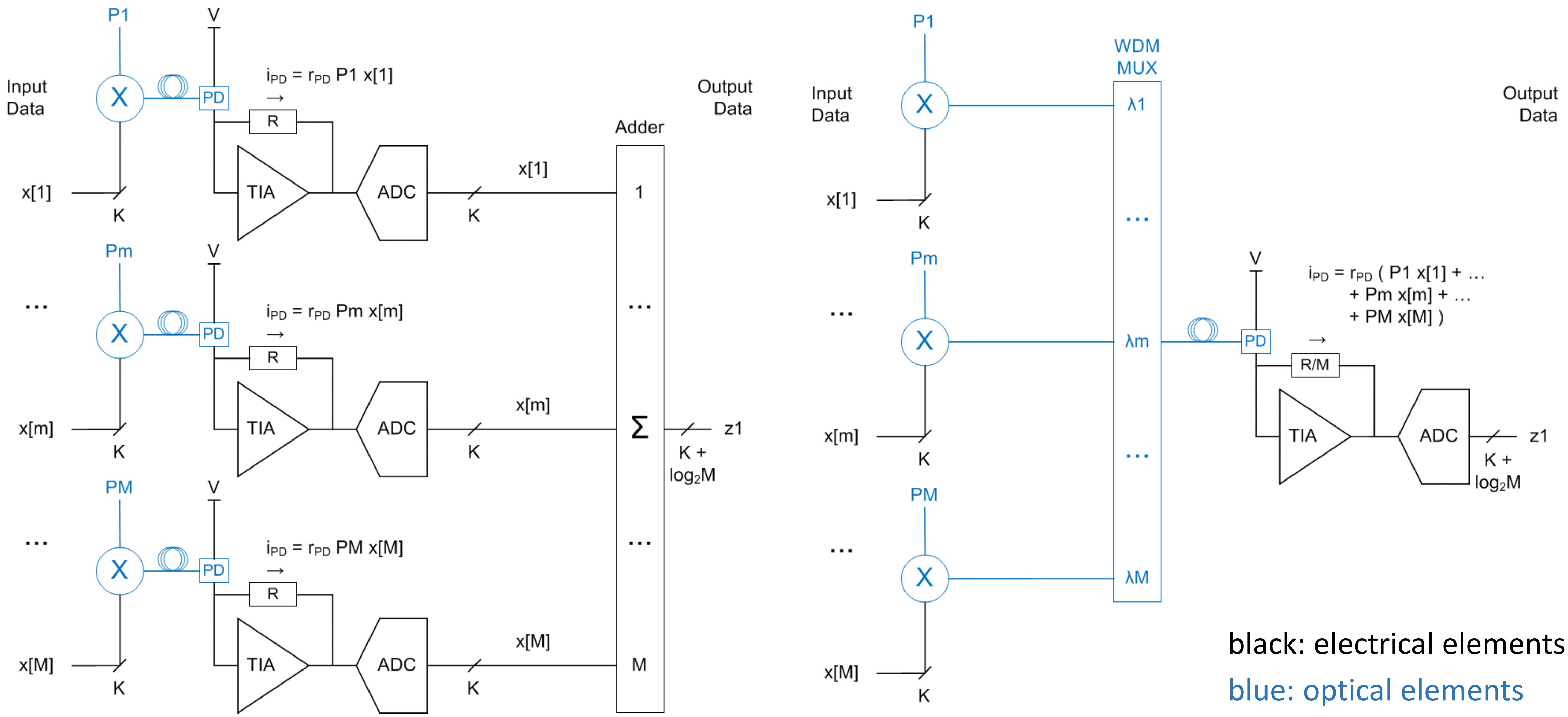
Electrical Parallel Addition (column sum) Model



black: electrical elements

blue: optical elements

Electrical and Optical Parallel Addition (column sum) Models



black: electrical elements

blue: optical elements

Energy Use of High-speed CMOS ADCs

Output	Rate	CMOS node	Effective bits	Energy	Reference
Bits	GS/s	nm	ENOB	**fJ/bit**	
6	24	28	4.5	**210**	[32]
6	3.3	28	5.4	310	[33]
8	10	65	6.4	800	[34]
8	1	28	7.3	350	[35]
8	28	7	5.0	355	[36]

References from C. Cole, "Optical and electrical programmable computing energy use comparison," Optics Express, Vol. 29, Issue 9, pp. 13153-13170, 2021.

Electrical & Optical Addition Energy Use Comparison

- 30 Gops/sec 16-bit 7nm CMOS adder is **3/210 = 1/70** of 28 Gops/sec 8-bit 7nm CMOS ADC
- Energy use of CMOS Adder compared to ADC is insignificant, and can be ignored
- To increase ADC effective resolution by √M bits (same increase in SNR as from summing the output of M ADCs) requires M times the energy (theoretical)
- Energy use of M K-bit ADCs **equals** energy use of one (K + $\log_2$M)-bit ADC

<u>Computing Addition optically instead of electrically does not save energy</u>

Electrical & Optical Addition Energy Use Comparison

- 30 Gops/sec 16-bit 7nm CMOS adder is **3/210 = 1/70** of 28 Gops/sec 8-bit 7nm CMOS ADC
- Energy use of CMOS Adder compared to ADC is insignificant, and can be ignored
- To increase ADC effective resolution by √M bits (same increase in SNR as from summing the output of M ADCs) requires M times the energy (theoretical *)
- Energy use of M K-bit ADCs **equals** energy use of one (K + $\log_2 M$)-bit ADC

<u>Computing Addition optically instead of electrically does not save energy</u>

* ADC changes required to increase resolution by 1-bit, with constant power supply voltage:

→ 4x lower ADC noise

→ 4x higher ADC signal capacitor(s) C

→ 4x higher gm to maintain constant gm/C (bandwidth)

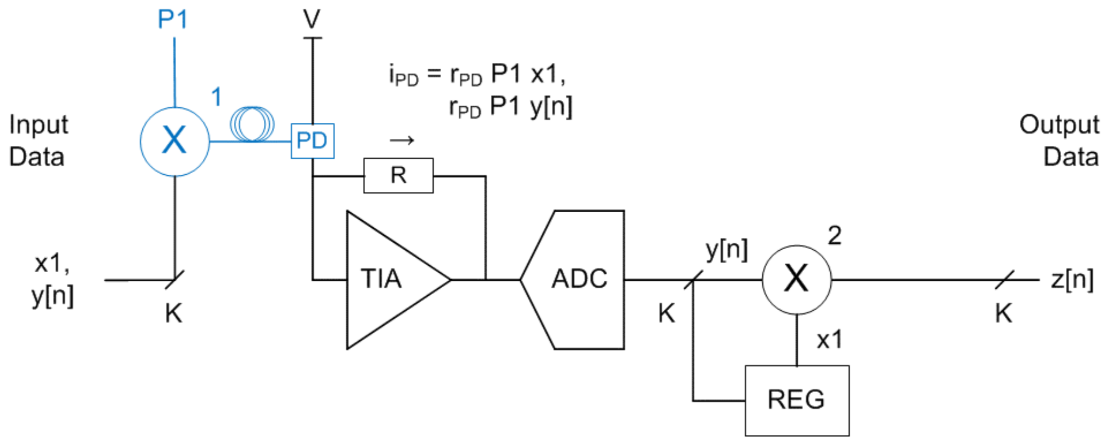
→ 4x higher i_{drain}

→ 4x higher ADC energy use

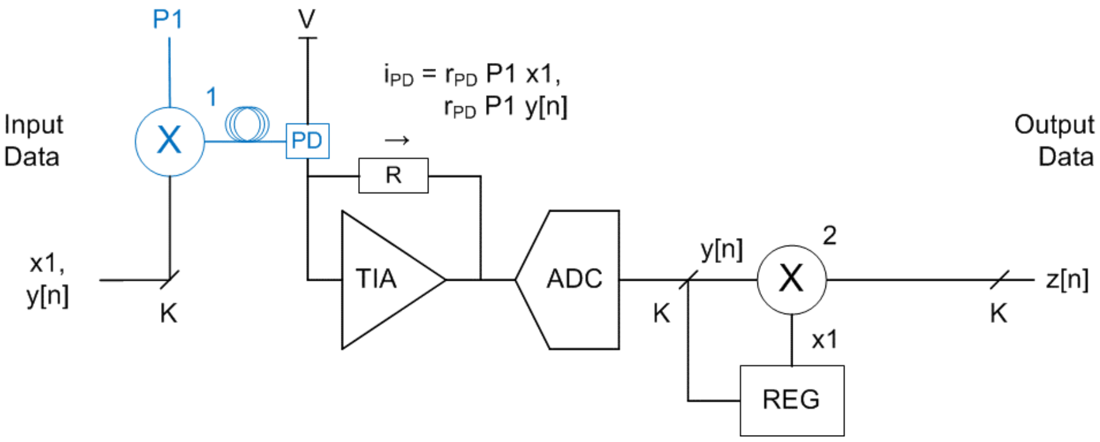
Outline

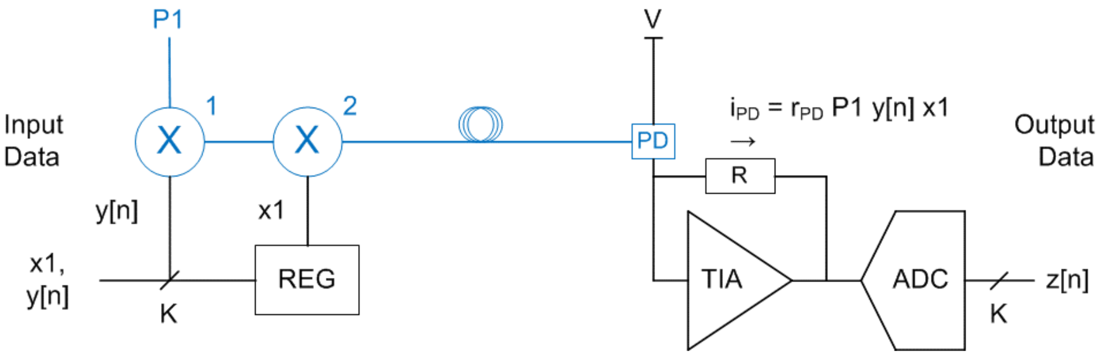
- Neural Networks in Demystified
- Optical Computing Examples
- Addition
- ➢ **Multiplication**
- Convolution
- Discussion

Electrical Vector Multiplication Model

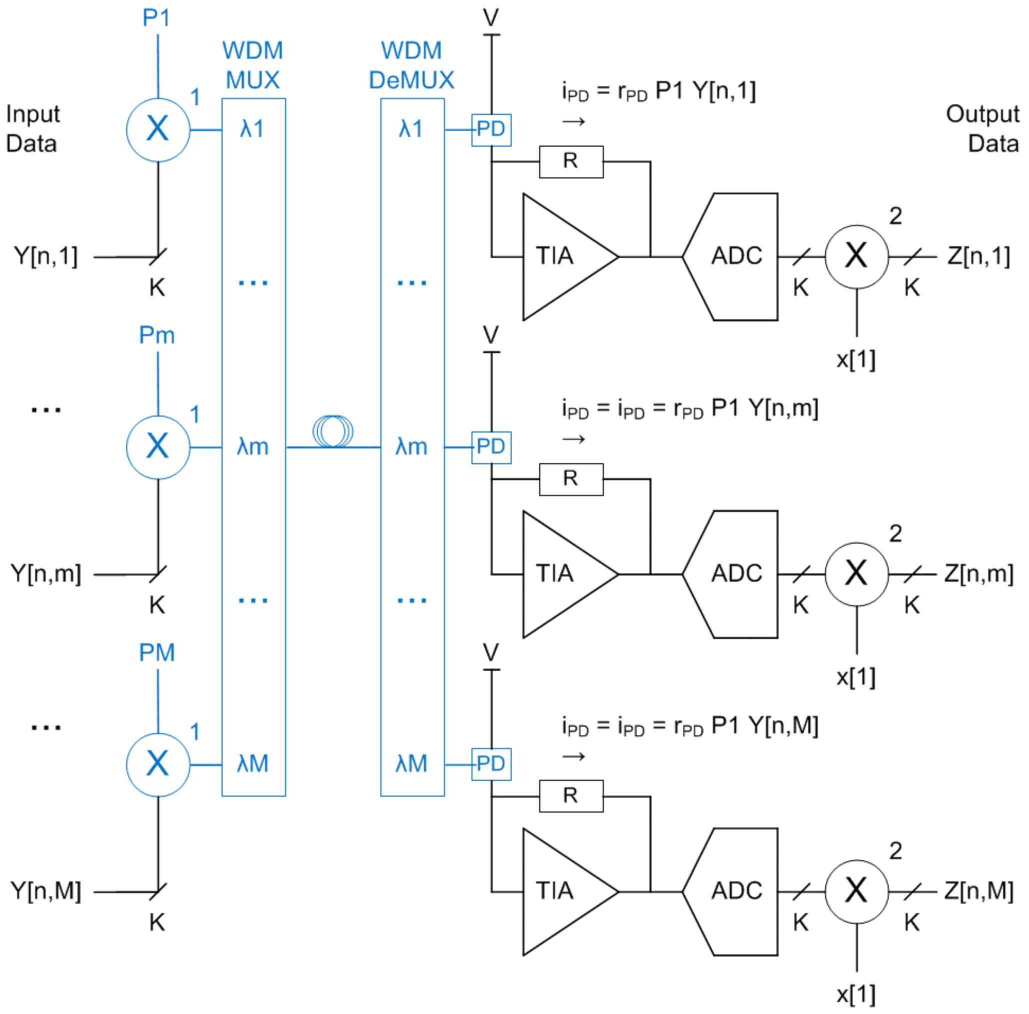


Electrical and Optical Vector Multiplication Models

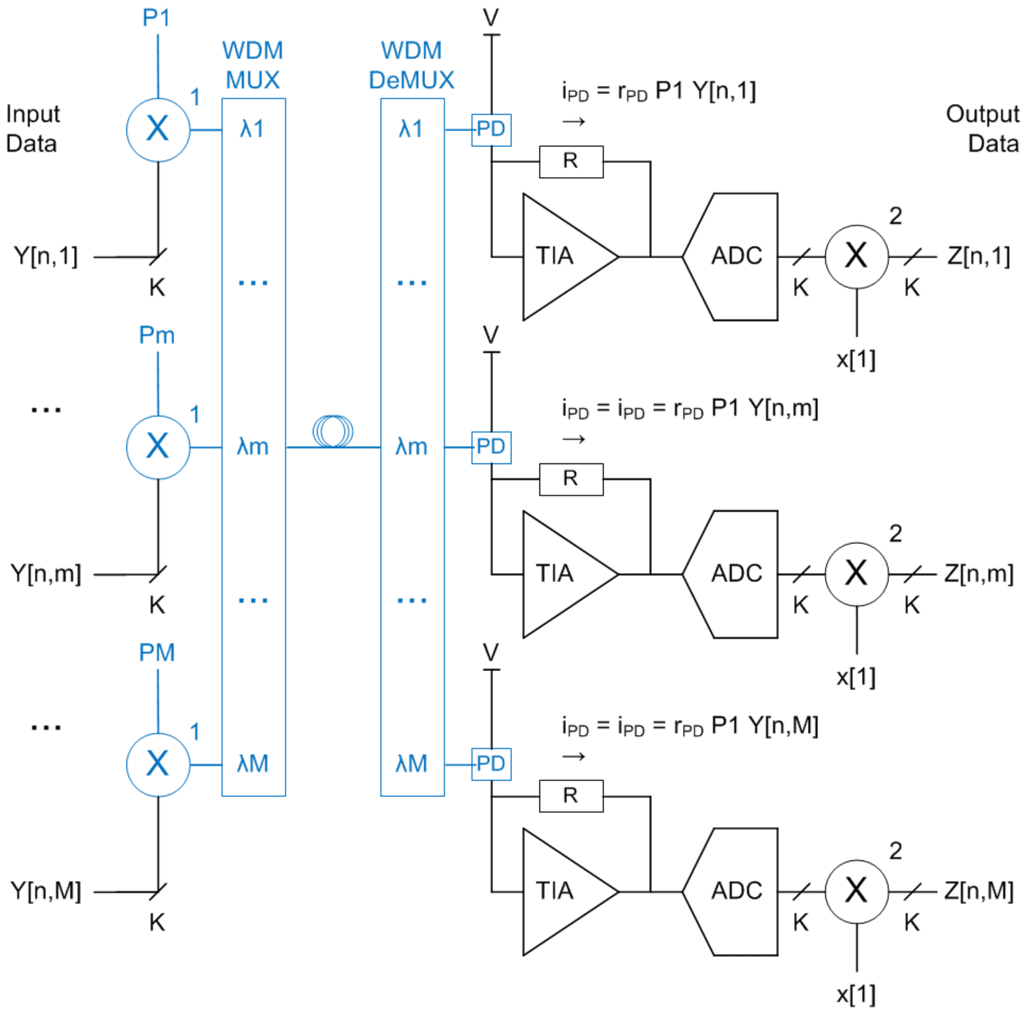


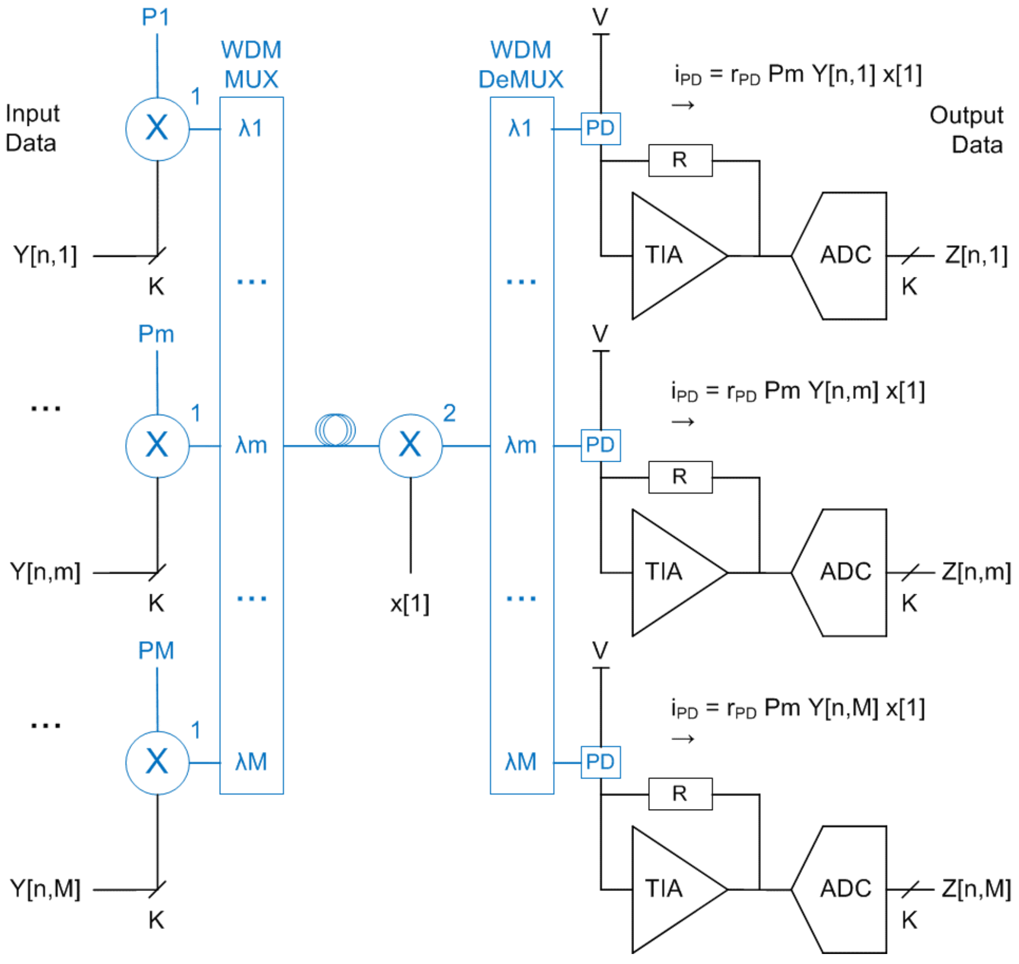


Electrical Parallel Vector Multiplication Model



Electrical and Optical Parallel Vector Multiplication Models





Energy Use of CMOS 16-bit Multipliers

CMOS node	Delay	Energy/op (max)	Input	Rate	Energy
nm	ps	fJ	bits/op	Gops/s	fJ/bit
7	58	296	16	17.5	19
7	40	310	16	25	19

Q. Xie, X. Lin, S. Chen, M. Dousti and M. Pedram, "Performance Comparisons between 7nm FinFET and Conventional Bulk CMOS Standard Cell Libraries," *IEEE Transactions on Circuits and Systems II: Express Briefs,* vol. 62, no. 8, pp. 761-765, August 2015.

D. Baran, M. Aktan, and V. Oklobdzija, “Energy Efficient Implementation of Parallel CMOS Multipliers with Improved Compressors,” in ACM/IEEE International Symposium on Low-Power Electronics and Design (ISLPED), pp. 147–152, August 2010.

A. Stillmaker and B. Baas, "Scaling equations for the accurate prediction of CMOS device performance," *Integration the VLSI journal,* vol. 58, pp. 74-81, February 2017.

Electrical & Optical Multiplication Energy Use Comparison

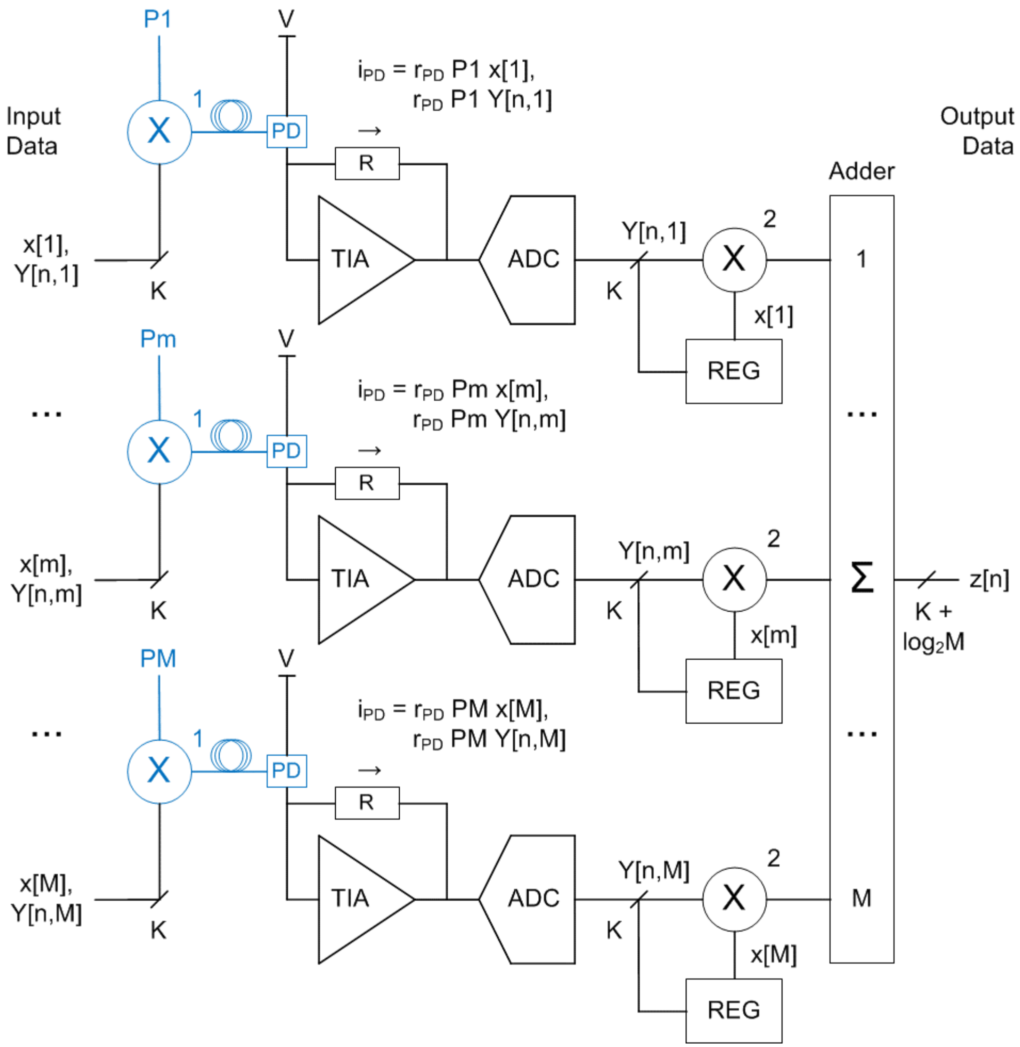
- 25 Gops/sec 16-bit 7nm CMOS multiplier is **19/210 = 1/11** of 28 Gops/sec 8-bit 7nm CMOS ADC
- Energy use of CMOS Multiplier compared to CMOS ADC is insignificant, and can be ignored
- Electrical and Optical Multiplication models have the same ADC energy use

<u>Computing Multiplication optically instead of electrically does not save energy</u>

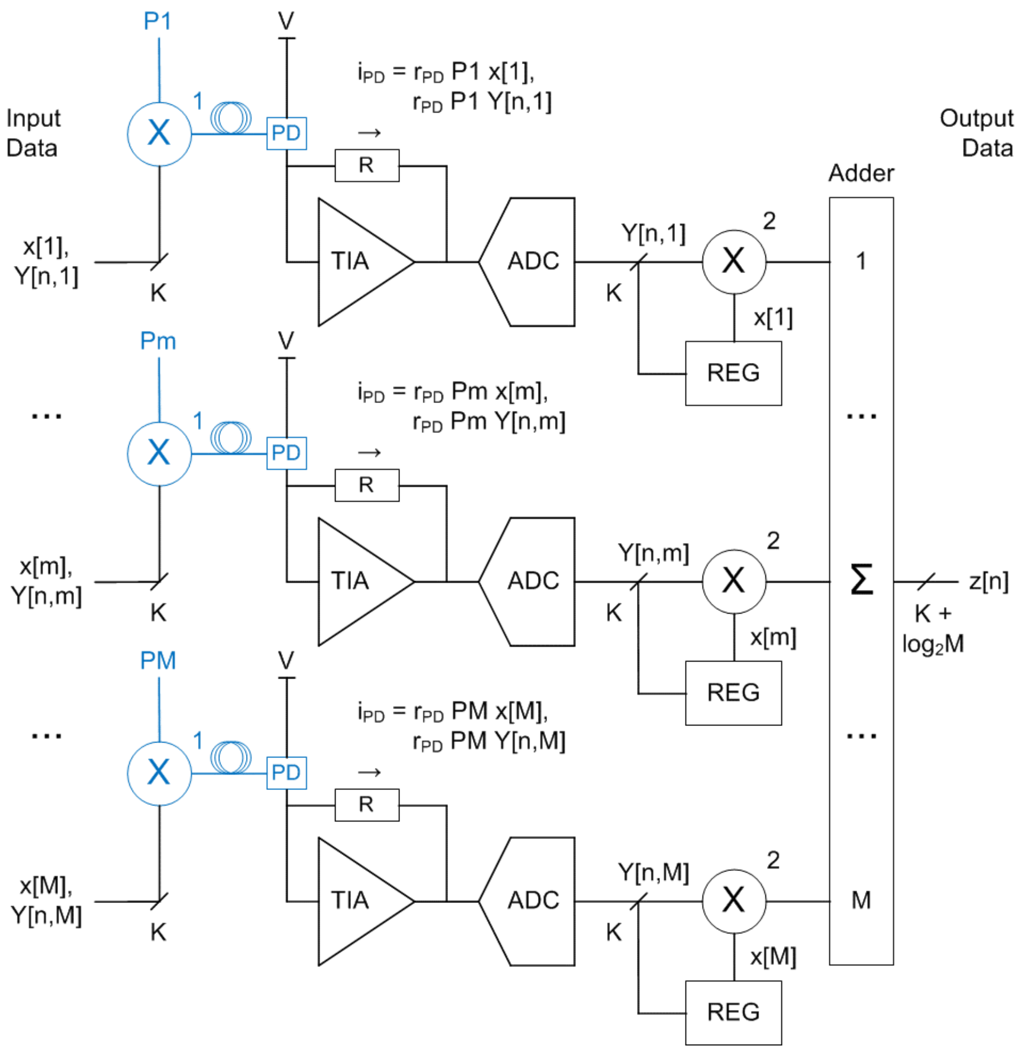
Outline

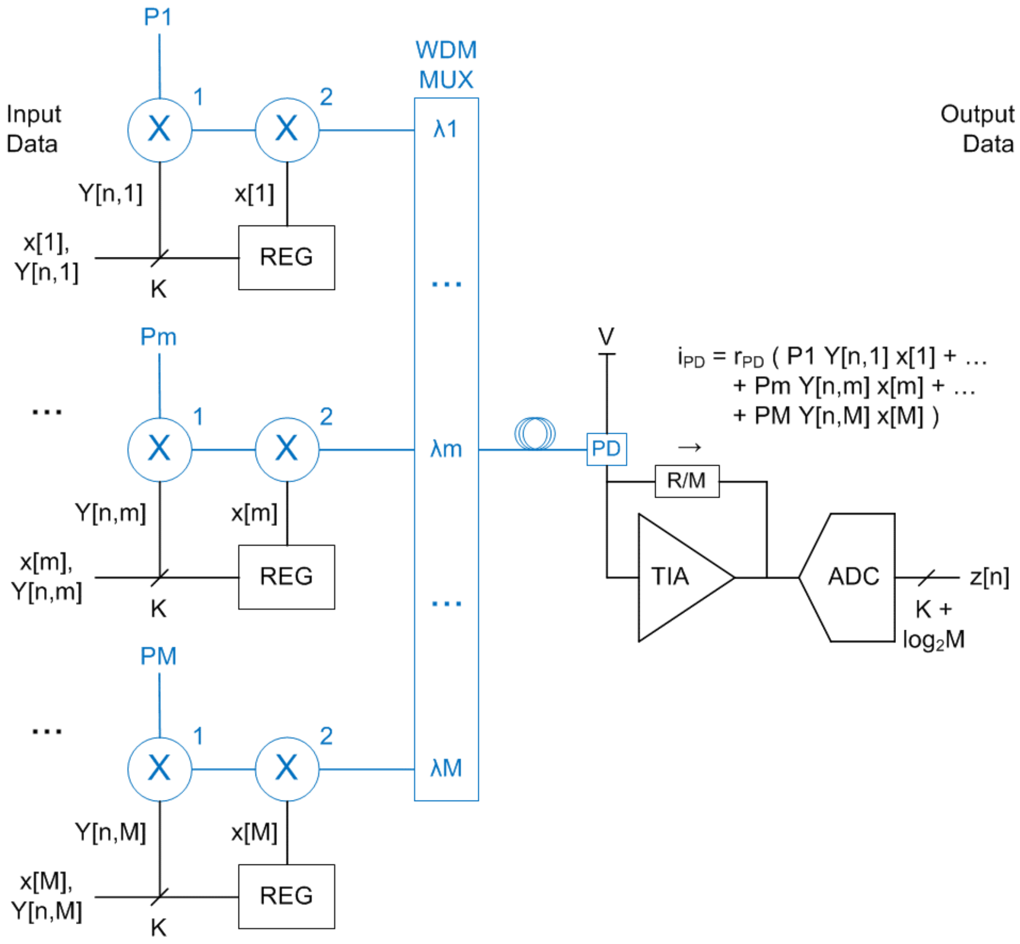
- Neural Networks Demystified
- Optical Computing Examples
- Addition
- Multiplication
- ➢ Convolution
- Discussion

Electrical Matrix Vector Product Computation Model

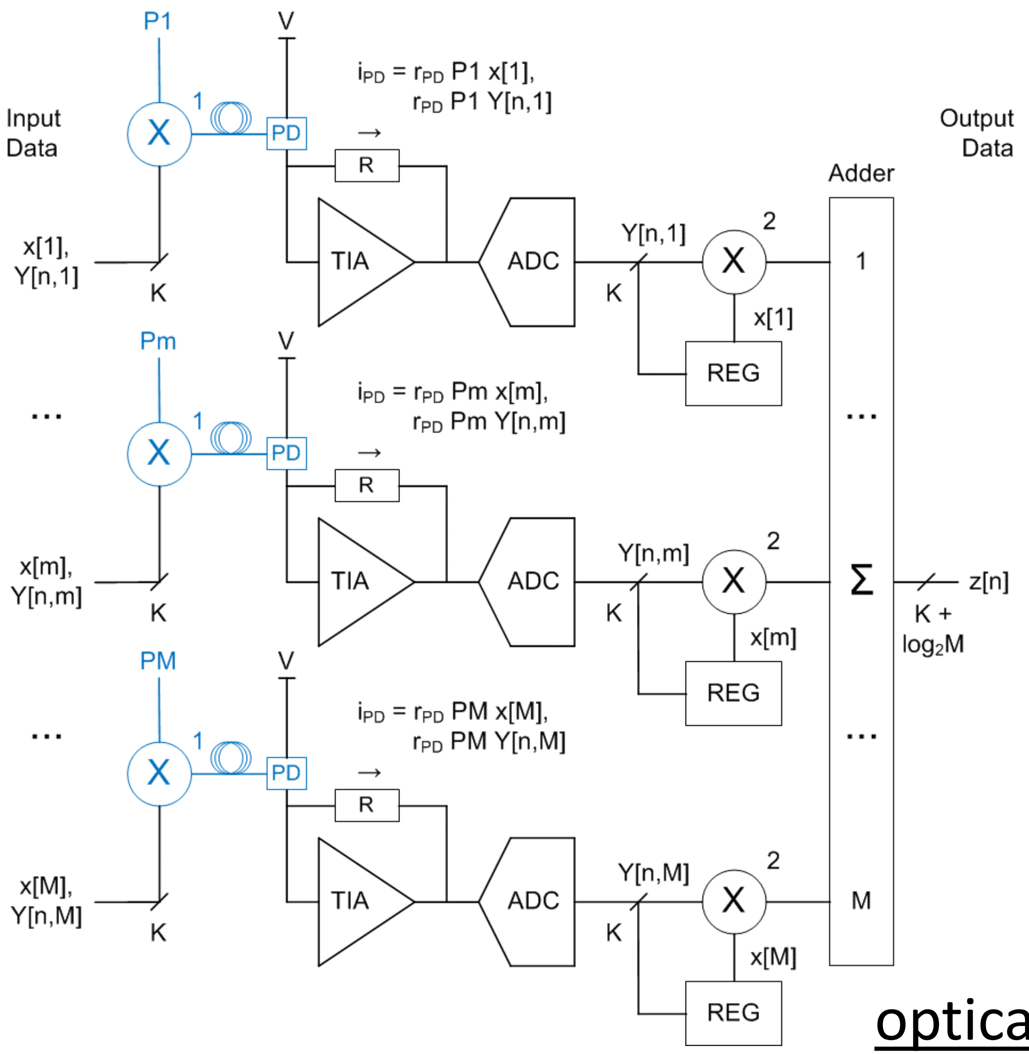


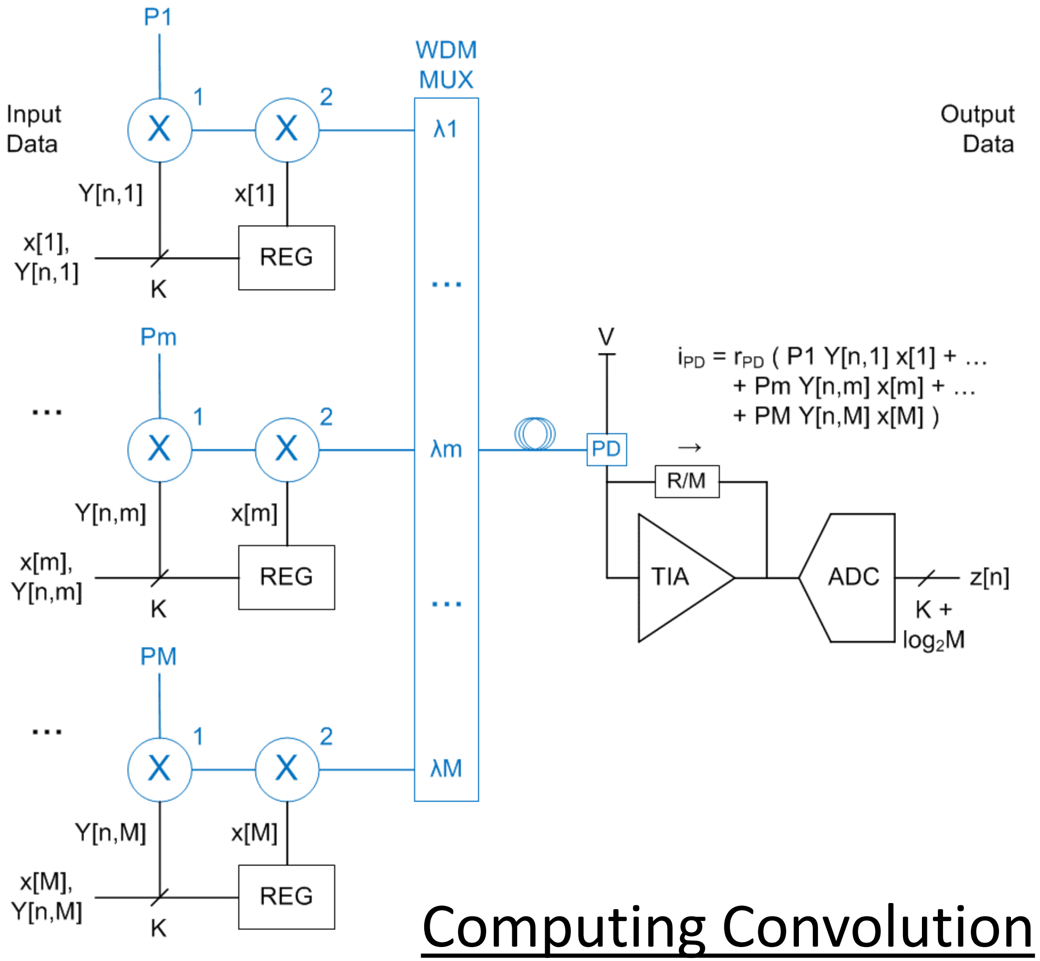
Electrical & Optical Matrix Vector Product Computation Models





Electrical & Optical Matrix Vector Product Computation Models





Computing Convolution

optically instead of electrically does not save energy

Optical Datacom Filter CMOS Computing Example: Fast FFE

- FFE (Feed Forward Equalizer) processing is convolution
- Same processing as applying weights in a neuron (inference)
- Used in high volume 56 Gb/sec and 112 Gb/sec per lane PHY (CDR) optical receivers
- Architecture: ADC + CMOS DSP with only CTLE analog pre-compensation
- Optical receiver FFE is the perfect problem for optical computing:
  - high bit rate
  - low precision
  - low number of coefficients
  - digital to optical & optical to digital conversion already in place
  - zero incremental energy use
- Yet all optical receivers use CMOS DSP FFEs, and none use optical computing

CMOS Multiplier and FFE MAC Energy Use

MAC Type	CMOS node	Delay	Energy/op (max)	Input	Rate	Energy
	nm	ps	fJ	bits/op	Gops/s	fJ/bit
Adder & Multiplier	7	58	367	16	17.5	23
FFE	7	11	159	8	90	**20**

Q. Xie, X. Lin, S. Chen, M. Dousti and M. Pedram, "Performance Comparisons between 7nm FinFET and Conventional Bulk CMOS Standard Cell Libraries," *IEEE Transactions on Circuits and Systems II: Express Briefs,* vol. 62, no. 8, pp. 761-765, August 2015.

C. Menolfi, M. Braendli, P. Francese, T. Morf, A. Cevrero, M. Kossel, L. Kull, D. Luu, I. Ozkaya and T. Toifl, "A 112Gb/s 2.6pJ/b 8-tap FFE PAM-4 SST TX in 14nm CMOS," in IEEE International Solid-State Circuits Conference Digest of Technical Papers, pp. 104-105, February 2018.

A. Stillmaker and B. Baas, "Scaling equations for the accurate prediction of CMOS device performance," *Integration the VLSI journal,* vol. 58, pp. 74-81, February 2017.

Why All Optical Receivers use CMOS DSP FFEs

- 7nm CMOS 90 Gops/sec 8-bit MAC is **20/210 = 1/10** of 28 Gops/sec 8-bit 7nm CMOS ADC
- Programmability, testability, repeatability and manufacturability are critical in commercial products
- If **20 fJ/bit** 7nm CMOS MAC energy use is too high, wait a few years:
  - core digital logic energy use is dropping with each node shrink
  - 3nm CMOS will be < **8 fJ/bit**
  - in contrast, I/O and analog circuit energy use is plateauing with node shrinks

<u>Computing Convolution optically instead of electrically does not save energy</u>

A. Stillmaker and B. Baas, "Scaling equations for the accurate prediction of CMOS device performance," *Integration the VLSI journal,* vol. 58, pp. 74-81, February 2017.

Outline

- Neural Networks Demystified
- Optical Computing Examples
- Addition
- Multiplication
- Convolution
- ➢ Discussion

Optical Pre-processing

- Optical computing is great as optical pre-processing before O-to-E conversion (ex. eyeglasses)
  - Digital camera front-end
  - LIDAR
  - FSO Beamformer
- Optical pre-processing is highly domain specific
- General purpose optical computing solutions are not usable
- Note that all electronic systems that use ADCs have some form of analog pre-processing (ex. anti-aliasing filters, AGC)

ML in the Datacenter

- The trend in machine learning applications is towards greater scale, complexity and programmability
- Model size increases are orders of magnitude to be meaningful.
- Optical computing precision and complexity scale poorly making them not useful in datacenters
- Web2.0 datacenter operators are not interested in low-resolution ML computing
    - TPUs and GPUs have 8-bit integer modes; these are rarely used

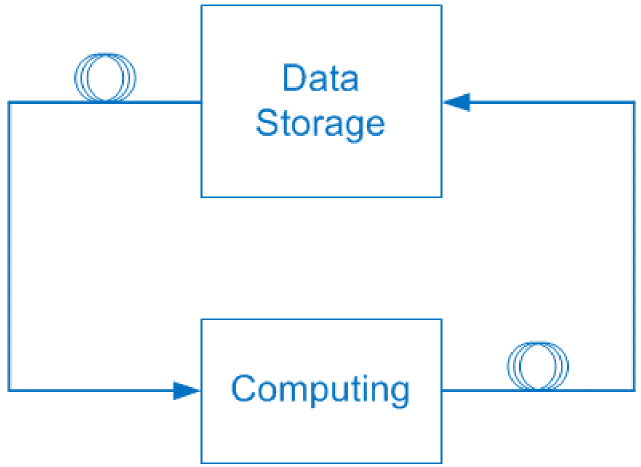
Optical Computing Summary

- Optical computing proposals compare low-precision optical, to general-purpose high-precision electrical, like TPU or GPU, i.e., apples-to-oranges.
- Optical computing proposals compare optical computing using optical data transfer, to electrical computing using electrical data transfer (like in a TPU or GPU), and incorrectly attribute low energy use to computing even though it's insignificant
- Total optical or electrical programmable computing energy use is dominated by data transfer to and from memory; computing is negligible in comparison

Optical Computing Summary

- Optical computing proposals compare low-precision optical, to general-purpose high-precision electrical, like TPU or GPU, i.e., apples-to-oranges.
- Optical computing proposals compare optical computing using optical data transfer, to electrical computing using electrical data transfer (like in a TPU or GPU), and incorrectly attribute low energy use to computing even though it's insignificant
- Total optical or electrical programmable computing energy use is dominated by data transfer to and from memory; computing is negligible in comparison
- General purpose optical computing offers no advantages
- However, it has huge implementation disadvantages which is why it is not used commercially

There is Hope for Optical Computing



blue: optical elements

We just need to invent competitive optical memory.

Optical and Electrical Computing Energy Use Comparison

Thank you