Optical and Electrical Computing Energy Use Comparison

Symposia: The Role of Machine Learning in Optical Systems and the Role of Optics in Machine Learning Systems

Session II: Photonics in Machine Learning and AI for Future Data Centers

June 9, 2021

Chris Cole, II-VI Incorporated

Approach taken by this Presentation is Described in an Anecdote

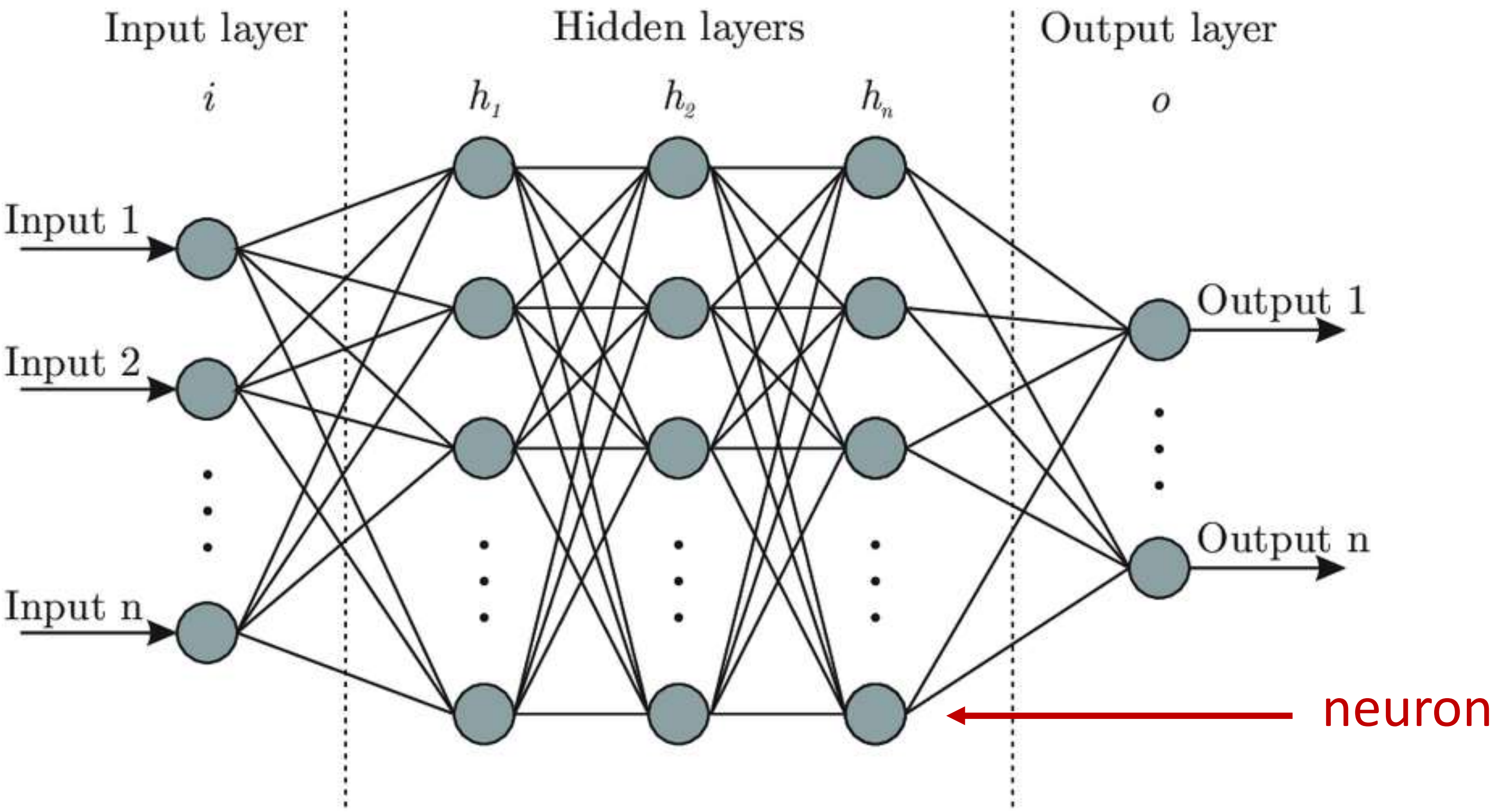
- An engineer and a mathematician are in a kitchen, and are asked to boil water
- They both fill a pan with water, put it on a stove, and boil the water
- Next, they are given a pan full of water, and are asked to boil it
- The engineer puts the pan on a stove, and boils the water
- The mathematician pours the water out of the pan, and stops …

  … he has reduced the problem to one previously solved
- This presentation takes the mathematical approach
- It shows that in the fiber optic communication field:
  - optical computing is well understood
  - computing elements used in receivers are used in machine learning
- Once the audience appreciates this, they can easily draw their own conclusions

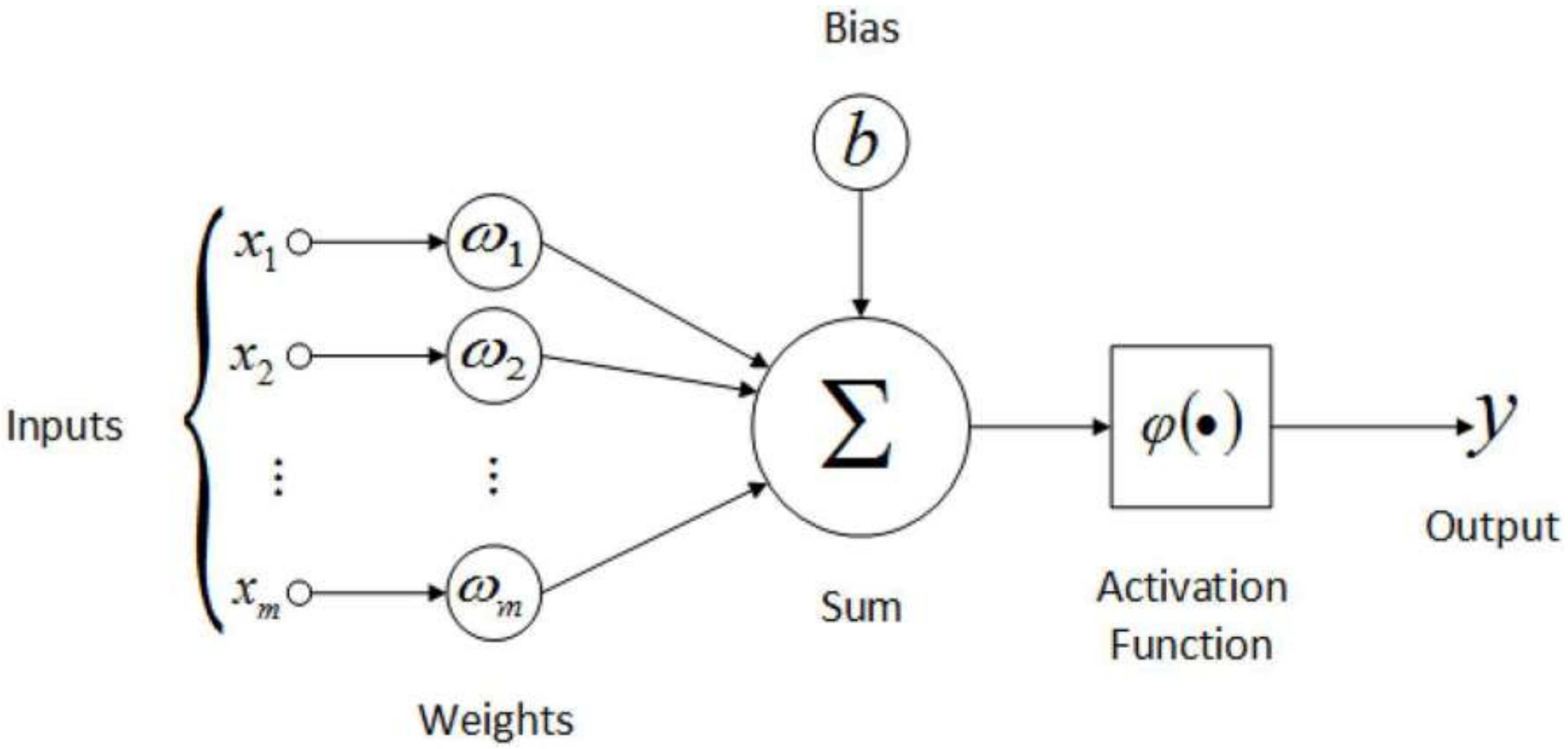
Attributes of a Field with Exclusivity and Cachet

- Unfamiliar terms
- Hushed rituals
- Hidden truths
- Red robes

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

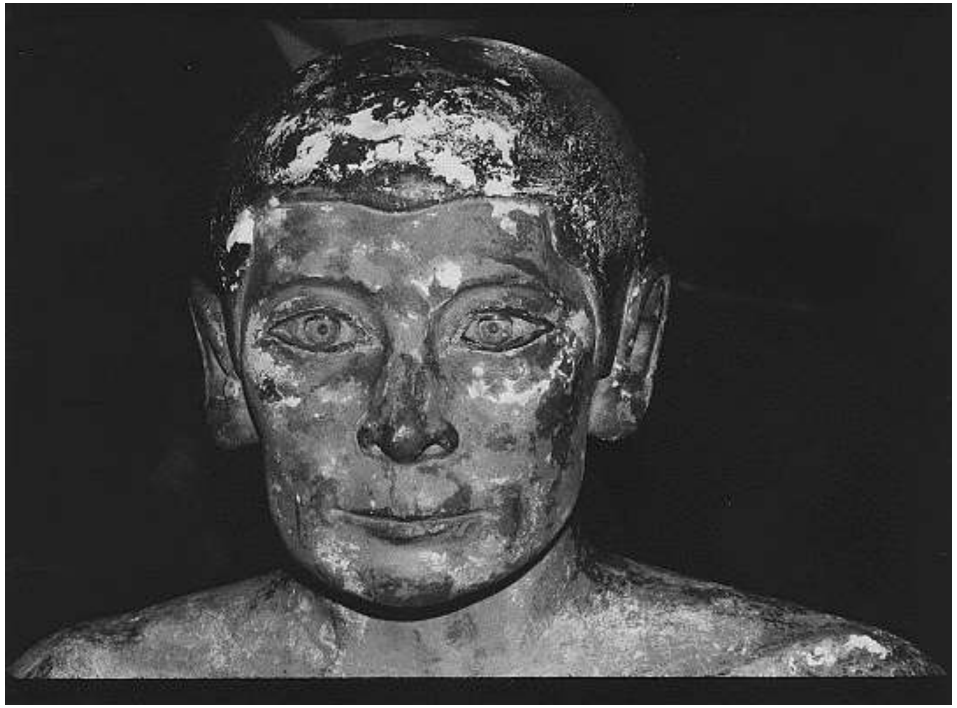
Machine Learning Terms

Machine Learning Term		Computing	
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		Computing		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		$f()$		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	

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



- 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
- 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

Telecom Optical Computing Example: DCF

- Passive, complex signal processing
- Zero incremental energy (ignoring amplification if required because of path loss)
- Problem is fixed compensation, which requires a custom length spool for every link
- Replaced by Coherent DSP CMOS ASIC which adaptively equalizes all links
- Since then, there has been no serious proposal to replace Coherent DSP CMOS with DCF, despite the latter's zero energy use
- Coherent DSP also eliminated all other optical and electronic analog compensation techniques

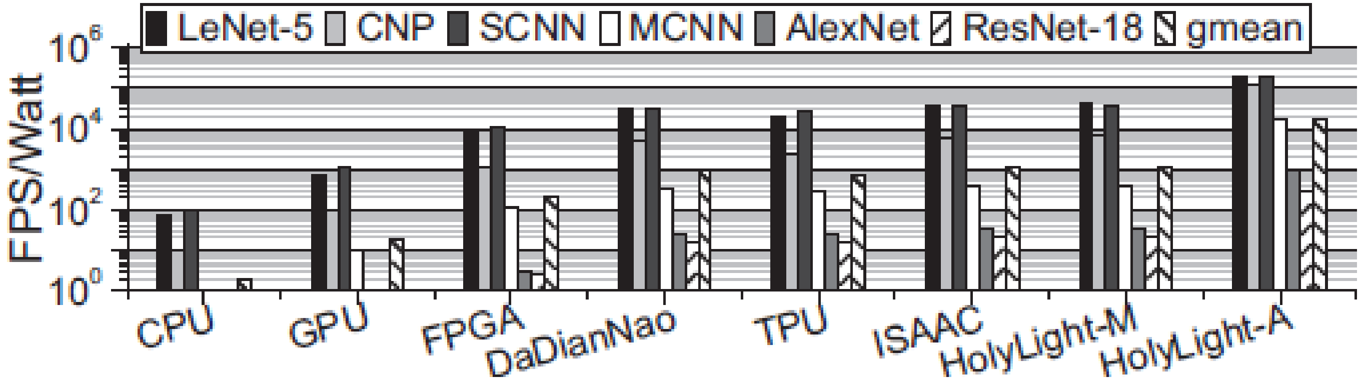
Optical Datacom Analog Computing Example: 10GBASE-LRM

- Problem was 10Gb/s serial transmission over legacy MMF
- 1st generation ASIC
  - 10GBaud analog signal path
  - Programmable analog FFE/DFE receiver
  - DSP based coefficient adaptation
  - Transceivers from all optics suppliers deployed in 2005
- 2nd generation ASIC
  - ADC and all DSP CMOS
  - Took over 10G direct attach copper cable market
- Since 1st gen. LRM, there has been no successful optical or analog computing implementation of a complex equalizer (adaptive filter) for optical links. All are DSP.

Optical Computing Example 1: Nanophotonic Accelerator

HolyLight (DATE Conf., Mar. 2019)

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



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

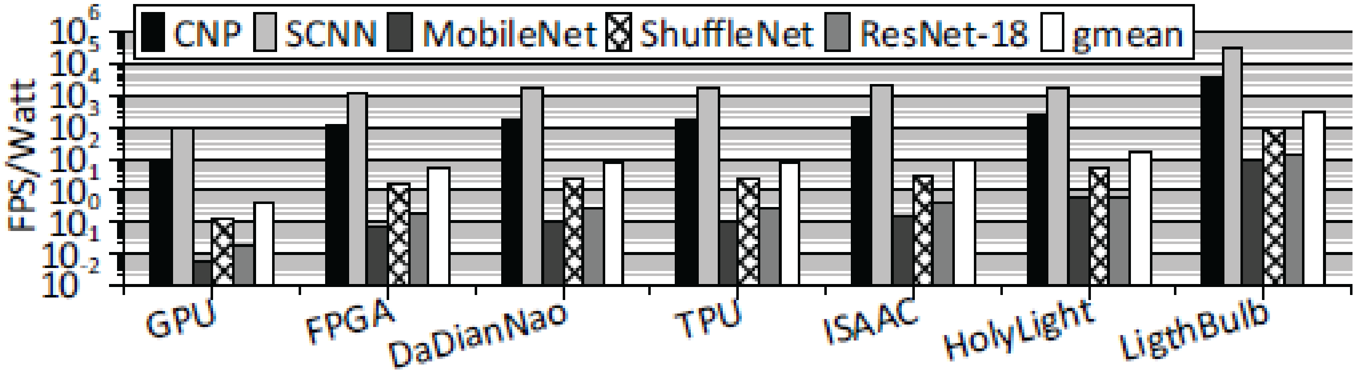
  → optical 13 Gops/sec adder energy use = 13 fJ/bit

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 energy use than 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

→ CMOS 30 Gops/sec adder energy use = 2.7 fJ/bit

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 and CMOS Adders Energy Use Comparison

- 13 fJ/bit optical 13 Gops/sec ≈ **5x** 2.7 fJ/bit CMOS 30 Gops/sec
- 53 fJ/bit optical 50 Gops/sec ≈ **20x** 2.7 fJ/bit CMOS 30 Gops/sec

- Optical chip claims: 10x to 600x lower than CMOS GPUs or TPUs
- Optical adder measurements: 5x to 20x higher than CMOS adders
- These results do support the conclusion that optical computing reduces energy use

Note

- These accelerator papers are excellent publications in the optical computing field.
- This is because their assumptions, descriptions and measurements are detailed, and separated from conclusions so that readers can draw their own.
- That's why they can be used in this presentation.

Optical Datacom Filter CMOS Computing Example: Fast FFE

- FFE processing is convolution, i.e., same processing as in a neuron
- Used in high volume 56 Gb/sec per lane and higher rate PHYs (CDRs)
- ADC + all DSP CMOS (w/ 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
- So why do all optical receiver FFEs use CMOS DSP, and none use optical computing?

CMOS Multiplier and FFE MAC Energy Use

Type	CMOS node	Delay	Energy/op (max)	Input	Rate	Energy
	nm	ps	fJ	bits/op	Gops/s	fJ/bit
Multiplier	7	58	296	16	17.5	19
	7	40	310	16	25	19
FFE MAC	7	58	367	16	17.5	23
	7	11	159	8	90	20

CMOS 90 Gops/sec MAC energy use = 20 fJ/bit

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

- Example CMOS MAC 90 Gops/sec 8-bits: 20 fJ/bit
- Much lower energy use than amplification to compensate for loss through an optical modulator (multiplier)
- Programmability, testability, repeatability and manufacturability matter for commercial products
- If 20 fJ/bit CMOS MAC energy use is too high, wait a few years:
  - core digital logic power is dropping with each node shrink
  - in contrast, analog circuit power is plateauing with node shrinks

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 Observations using Mathematical Approach

- Optical computing makes sense as fixed or limited programmability optical pre-processing before optical signal to digital data conversion, like eyeglasses.
- The trend in machine learning applications is towards greater scale, complexity and programmability; increases must be orders of magnitude to be meaningful.
- Optical computing precision, complexity and size scale poorly.
- Many optical computing proposals compare limited-use low-precision optical, to general-purpose high-precision electrical, like TPU or GPU, i.e., apples to oranges.
- Total optical or electrical programmable computer energy use is dominated by data movement to and from memory; computing is a negligible in comparison.
- Many optical computing proposals compare optical computing and optical data movement, to electrical computing and electrical data movement, and incorrectly attribute low energy use to just computing even though it’s insignificant.

Optical and Electrical Computing Energy Use Comparison

Thank you