Research Article | Vol. 29, No. 9 / 26 April 2021 / Optics Express 13153
Optics EXPRESS

Optical and electrical programmable computing energy use comparison

CHRIS COLE*

II-VI Incorporated, 1389 Moffett Park Drive, Sunnyvale, CA 94089, USA
*chris.cole@ii-vi.com

Abstract: Optical computing has been proposed as a replacement for electrical computing to reduce energy use of math intensive programmable applications like machine learning. Objective energy use comparison requires that data transfer is separated from computing and made constant, with only computing variable. Three operations compared in this manner are multiplication, addition and inner product. For each, it is found that energy use is dominated by data transfer, and that computing energy use is a small fraction of the total. Switching to optical from electrical programmable computing does not reduce energy use.

1. Introduction

Electronic digital computers have been in use for over half a century [1]. They are ubiquitous because advances in silicon electronics steeply increase computing power while decreasing energy use. Together with light transmission over fiber optic cable, which uses significantly less energy than electrical transmission, this has enabled exponential growth in cloud computing. However, math intensive applications like machine learning are increasing energy use at a faster rate than can be decreased by advances in silicon electronics [2]. This has motivated a search for alternate means to lower energy use. One proposed approach is switching to analog from digital computing, with potential to reduce energy use at low bit precision [3,4]. More ambitious proposed approach is switching to optical from electrical computing, with promise of low energy use like in optical data transmission. This has attracted substantial research and development funding and generated enthusiastic publicity [5,6]. What has been missing is objective apples-to-apples comparison of optical and electrical programmable computing energy use. It is critical that such analysis is broadly discussed in a timely manner because investment in optical programable computing is rising. Energy use comparison of computing systems with limited or no programmability is not made in this paper, and the scope and value of their supported applications is not evaluated.

2. Computing models

Figure 1 shows a data transfer model of a computer optimized for math intensive operations. All elements are electrical. Computing is separated into two types, A and B. Type A is optimized for math operations like addition, multiplication and inner product [7]. Many such basic operations are executed to process a complex computing task. An example task is addition of values in a column. A more complex task is the inner product, which is the multiplication of adjacent values in two columns, followed by addition of all the resulting products. In machine learning applications inner product is executed multiple times to process the highest energy use computing task, which is matrix vector product.

Computing architectures used for energy comparison in this paper are formally Turing-complete, or informally generally programmable to process complex tasks like matrix vector product with changing coefficients, data and tensor size. Many optical computing publications report approaches with limited if any programmability. This greatly simplifies implementation, but also narrows the supported applications. All commercial computing systems are programmable,

#420027 https://doi.org/10.1364/OE.420027
Journal © 2021 Received 22 Jan 2021; revised 25 Mar 2021; accepted 26 Mar 2021; published 14 Apr 2021

Optics EXPRESS

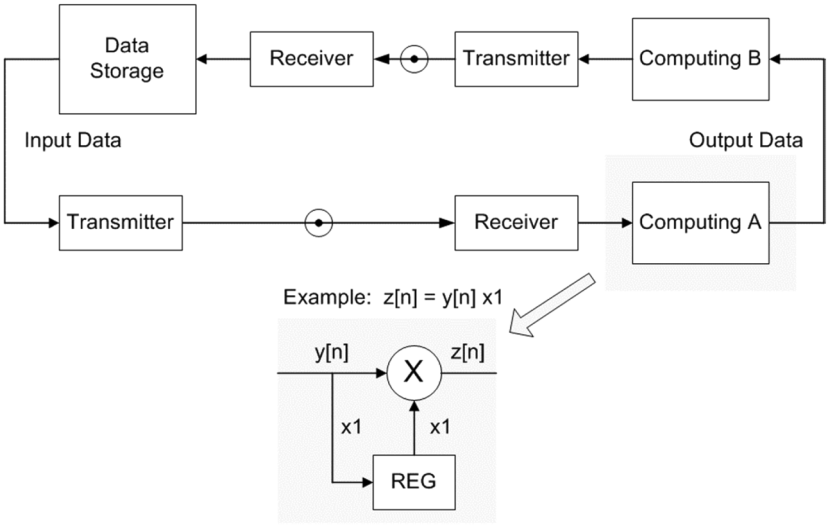

Fig. 1. Data transfer model of a computer optimized for math intensive operations. All elements are electrical. Input Data path flows left to right.

and the trend is towards greater complexity and flexibility in computing intensive applications like training of neural networks.

Figure 1 Example shows type A Computing multiplying row of N values $y[n]$ by single cell value $x1$ stored in register REG, resulting in row of N values $z[n]$. Figure 2 shows the data sequencing.

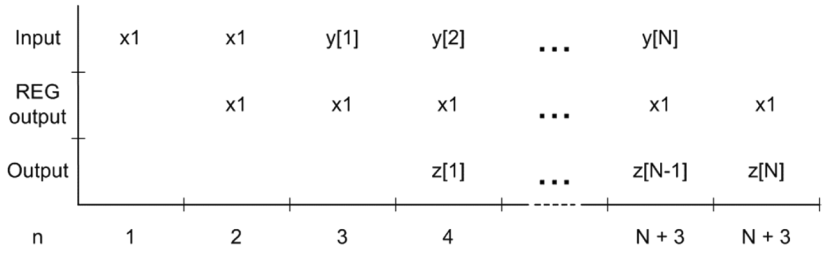

Fig. 2. Data sequencing for computing $z[n] = y[n]\ x1$.

Type B is all other computing like logical comparisons, decision making, and flow control. Computers require data and instructions to perform tasks, which the Fig. 1 computer model supports. Since energy use associated with instructions is negligible, they are not used in energy use comparisons. Despite math intensive nature of machine learning tasks, the total energy use is dominated by data transfer from and to Data Storage [8].

2.1. Electrical

Figure 3 shows a model of the same computer as in Fig. 1, except data is transferred by light instead of electricity [9]. Fiber optic cable, or other type of optical waveguide, is used for transmission. Data Storage and Computing remain electrical. Figure 3 Example shows details of the transmitter and receiver. Each value $y[n]$ modulates wavelength optical power P ($P\lambda$). A photo detector (PD), biased by voltage V, converts the optical power to signal current, which is

Optics EXPRESS

converted to electrical data $y[n]'$ by the trans-impedance amplifier (TIA) and analog to digital converter (ADC). Black and blue signify electrical and optical elements, respectively.

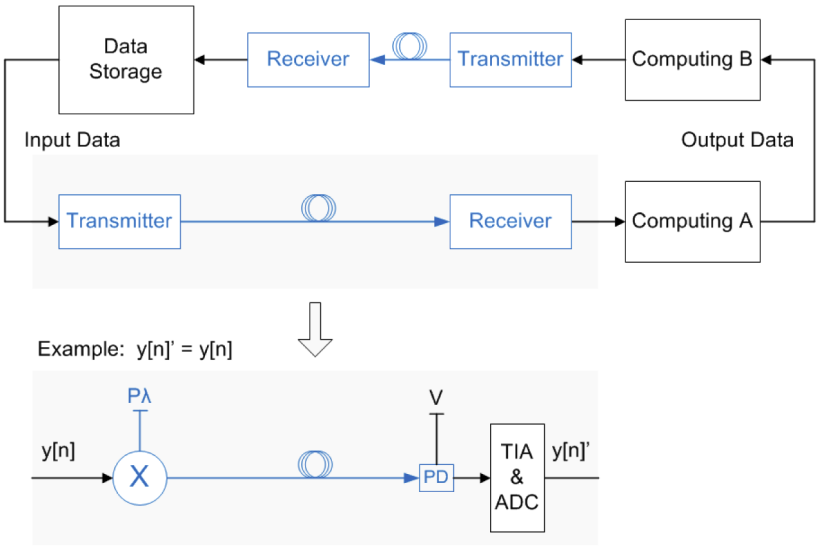

Fig. 3. Data transfer model of a computer using light for transmission. Black and blue signify electrical and optical elements, respectively. Input Data path flows left to right.

2.2. Optical

Figure 4 shows type A Computing performed optically, reusing Input Data path light. Type B Computing is difficult to perform optically and stays electrical. Figure 4 Example 1 shows an optical modulator, like in an optical transmitter, used for multiplication. Row of N values $y[n]$ are multiplied by single cell value $x1$, same as in Fig. 1 Example. Figure 4 Example 2 shows a WDM MUX (wavelength division multiplexer) used for addition. Each value x[m] modulates wavelength m optical power P ($P\lambda m$). Column of M values $x[m]$ is optically summed together by the WDM MUX combining all M wavelengths, resulting in single cell value $z1$.

Optical computing uses passive or active elements, for example lenses or optical modulators, respectively, or a combination of the two. Lenses rely entirely on the energy of the source, do not store data, are not programmable, and perform spatial filtering, i.e., two-dimensional convolution.

The usefulness of lenses has been recognized for over four millennia [10]. A hypothetical electronic lens projects 24-bit color, 120 frame per second, 512×512 image. This is ~25 trillion 8-bit multiply-add operations per second, and is the same math as used in machine learning. This level of processing at zero energy use is very compelling and the reason such approaches have so much research interest. Unfortunately, the lack of programmability restricts applicability. Neuromorphic computing is an example of cascaded optical computing with potential for good energy efficiency [11–14]. Like lenses, it's limited programmability restricts its use, for example to image pre-processing. As stated previously, this type of optical computing is not analyzed nor compared in this paper.

For decades, the holy grail in optical computing has been all-optical random-access memory (RAM) [15], to enable all-optical programmable computers. Except for niche applications in which low-density data storage is useful, for example FFT processing [16], no practical high-density all-optical RAM exists. All practical digital computers use electrical RAM, which requires conversion between optical and electrical, if light is used for data transfer and/or computing.

Optics EXPRESS

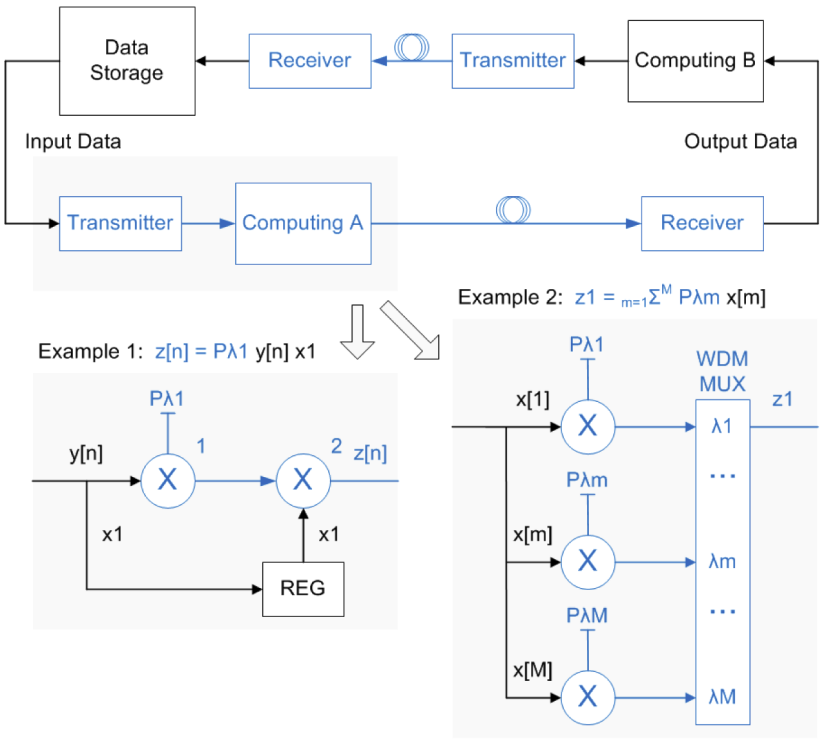

Fig. 4. Data transfer model of a computer using shared light for transmission and optical computing. Input Data path flows left to right.

2.3. *Comparison methodology*

Electrical computing energy use is well understood, but not relative to optical computing because it is difficult to separate data transfer from computing energy use. This leads to the first major problem in energy use comparisons. Optical computing proposals compare electrical data transfer and computing like in Fig. 1 to optical data transfer and computing like in Fig. 4. Even when energy use is dominated by data transfer, energy use advantages are attributed to optical computing [17–21].

Apples-to-apples energy use comparison must use optical data transfer for both and compare electrical computing as in Fig. 3 to optical computing as in Fig. 4. The implementation and operation of Data Storage, type B Computing, and Output Data paths must be identical so that their energy use is identical and does not affect energy use comparison. Only the Input Data paths are different. Then the Energy Total Electrical type A Computing Input Data path ($E_{Total-EtAComp}$) of Fig. 3 can be fairly compared to the Energy Total Optical type A Computing Input Data path ($E_{Total-OtAComp}$) of Fig. 4, both processing the same task.

$$E_{Total-OtAComp} = E_{TX-OtAComp} + E_{RX-OtAComp} + E_{Comp-OtAComp} \tag{1}$$

$$E_{Total-EtAComp} = E_{TX-EtAComp} + E_{RX-EtAComp} + E_{Comp-EtAComp} \tag{2}$$

$$E_{Total-EtAComp} \ ? \ E_{Total-OtAComp} \tag{3}$$

A characteristic of many optical computing implementations is inherent low precision, which leads to the second major problem in energy use comparisons. Optical computing is proposed for applications in which low precision maybe acceptable, for example certain neural networks.

Optics EXPRESS

Incomplete implementations are used to demonstrate feasibility. They are then compared to complete commercial systems implemented with Graphical Processing Units (GPUs) or Tensor Processing Units (TPUs), which support high precision floating point and high precision integer electrical computing [22–24].

Key requirement of apples-to-apples energy use comparison is having the same precision for data transfer to and from optical and electrical computing. This means if externally examined as black boxes, optical and electrical computing implementations cannot be differentiated. Precision is quantified by the signal-to-noise ratio (SNR). Fair comparison must have equal input data SNRs and equal output data SNRs. Three operations used in making this comparison are multiplication, addition and inner product, outlined in Table 1.

Table 1. Optical and electrical programmable computing energy use comparison outline.

Section	2. Computing model	3. Multiplication	4. Addition	5. Inner product
Task	all	vector scalar product	vector element sum	matrix vector product
Equation	all	$z[n] = y[n]\ x1$	$z1 = \sum_{m=1}^{M} x[m]$	$z[n] = Y[n,\ m]\ x[m]$
Electrical	Fig. 3 ∈ 2.1	Fig. 5 ∈ 3.1	Fig. 7 ∈ 4.1	Fig. 9 ∈ 5.1
Optical	Fig. 4 ∈ 2.2	Fig. 6 ∈ 3.2	Fig. 8 ∈ 4.2	Fig. 10 ∈ 5.2

3. Multiplication

N-element vector $y[n]$ is shown electrically and optically multiplied by scalar $x1$ in Figs. 5 and 6, respectively, resulting in N-element vector $z[n]$.

$$z[n] = y[n]\ x1 \tag{4}$$

Input K-bit data is read sequentially from Data Storage, multiplied to compute output data, and written sequentially to Data Storage.

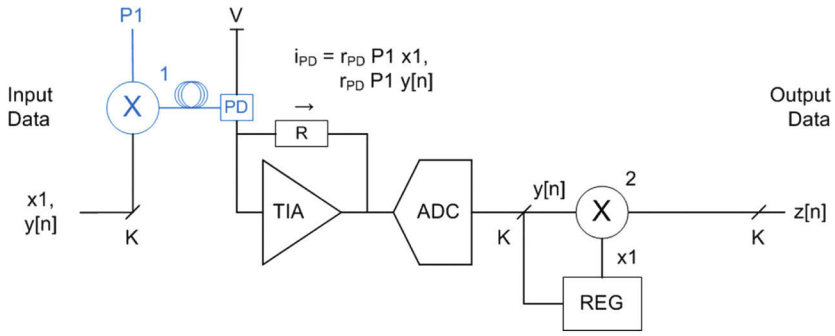

Fig. 5. Input Data path model with electrical multiplication of $z[n] = y[n]\ x1$.

3.1. Electrical

Figure 5 expands the Fig. 3 Input Data path. It shows one modulator single transmitter converting electrical digital data to optical signal power by multiplying continuous wave optical power of one wavelength by electrical signal representation of the data. Transmitter a_M and other optical path losses are not included.

Figure 5 shows a single receiver, comprised of one PD, TIA, and ADC. The PD converts optical signal power to signal current i_{PD}. The TIA and ADC convert the signal current to

Optics EXPRESS

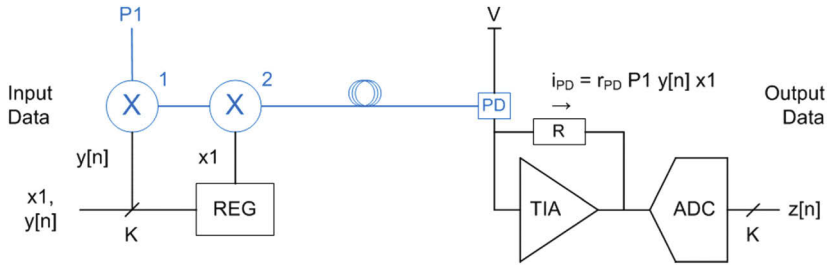

Fig. 6. Input Data path model with optical multiplication of $z[n] = y[n]\ x1$.

electrical digital data. The feedback resistor R is adjusted to map the TIA output into the full ADC preamplifier input range ($V_{ADC-input-range}$).

$$R = \frac{V_{ADC-input-range}}{max(i_{PD})} \tag{5}$$

Performance of high-speed CMOS TIAs and ADCs is listed in Tables 2 and 3, respectively. The cited technology is advanced but practical and implementable today. Future technology is not included. Exciting research, like novel PD waveguide geometry and material system [25] and photonic crystal platform modulator and receiver [26], is critical to future energy use breakthrough reductions. However, it is not considered. The criteria for including a technology in energy use comparisons in this paper is credible demonstration for today's designs.

Table 2. Energy of five high-speed CMOS TIAs, not including output buffers.

Bit Rate	Symbol Rate	Bandwidth	CMOS node	Noise	Energy	Reference
Gb/s	GBaud	GHz	nm	pA/√Hz	fJ/bit	
15	15	10	90	70	150	[27]
53	53	27	28	14	600	[28]
54	54	28	28	42	550	[29]
106	53	27	16	17	580	[30]
112	56	45	28	22	540	[31]

Table 3. Energy of five 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]

Figure 2 shows the data sequencing. During cycle 1 scalar x1 is transmitted, received and stored in REG. During cycle 2 scalar x1 path through multiplier 2 settles. During cycles 3 to N+2, N vector y[n] values are transmitted and electrically multiplied by scalar $x1$. Vector $z[n]$ values $z[1]$, $z[2]$, ... $z[N]$ are output on cycles 4, 5, ... N+3, respectively.

Electrical multiplier with two K-bit inputs has a 2K-bit output product. For zero mean, independent processes, multiplier 2 output SNR ($SNR_{mult-output}$) is simply half the multiplier 2

Optics EXPRESS

input SNR ($SNR_{mult-input}$). SNR is a ratio of powers, or variances (var).

$$SNR_{mult-input} = \frac{var(x)}{var(N_x)} = \frac{var(y)}{var(N_y)} \tag{6}$$

$$\begin{aligned} SNR_{mult-output} &= \frac{var((x + N_x)(y + N_y))}{var(xN_y + yN_x + N_xN_y)} \\ &= \frac{(SNR_{mult-input})^2}{(2SNR_{mult-input} + 1)} \\ &\approx \frac{SNR_{mult-\mathrm{input}}}{2} \end{aligned} \tag{7}$$

Because the multiplier 2 output SNR is half the multiplier 2 input SNR, which is represented by K-bit data, it is not necessary to use 2K-bits for multiplier 2 output. Using K-bits preserves the multiplier 2 output SNR.

Performance of electrical 16-bit multipliers in 45 and 7nm CMOS process nodes is listed in Table 4. The 7nm values are estimated from the 45nm values using 10x energy scaling factor [37]. Energy use of 8-bit multipliers is ideally quarter of these values. Energy use is much less than of the TIAs and ADCs listed in Tables 2 and 3, respectively, and is negligible in calculating total energy use of the Input Data path. This is the same as in electrical computing where electrical data transfer dominates energy use [8].

$$E_{RX-EtAComp-Mult} \approx E_{RX-EtAComp-Mult} + E_{Comp-EtAComp-Mult} \tag{8}$$

As CMOS scales down in feature size, power drops for analog circuits like TIAs and ADCs and digital circuits like multipliers. However, a serious challenge facing CMOS analog design is that power decrease is plateauing with finer CMOS nodes. In contrast, power of CMOS digital circuits is steadily decreasing. Over time, the energy use advantage of CMOS digital circuits will continue to increase over analog circuits.

Table 4. CMOS 16-bit multiplier delay and energy use.

CMOS node nm	Delay ps	Energy/op (max) fJ	Input bits/op	Rate Gops/s	Energy fJ/bit	Reference
45	1440	2955	16	0.7	185	[38]
45 → 7	58	296	16	17.5	19	[37,38]
45	1010	3100	16	1	194	[39]
45 → 7	40	310	16	25	19	[37,39]

3.2. Optical

Figure 6 expands the Fig. 4 Input Data path. It shows two modulator single transmitter converting electrical digital data to optical signal power by multiplying continuous wave optical power of one wavelength by electrical signal representation of the data. Transmitter a_M^2 and other optical path losses are not included. For K=1, 0 and 1 binary levels are used, which implements digital optical multiplication.

Figure 2 shows the data sequencing. During cycle 1 scalar $x1$ is stored in REG. During cycle 2 scalar $x1$ path settles through modulator 2. During cycles 3 to N+2, modulator 1 transmits N vector $y[n]$ values, which are multiplied by scalar $x1$ in modulator 2. Vector $z[n]$ values $z[1]$, $z[2]$, ... $z[N]$ are output on cycles 4, 5, ... N+3, respectively. Modulator 2 operates at a

Optics EXPRESS

much lower rate than modulator 1, and its energy use is negligible in comparison.

$$E_{TX-OtAComp-Mult} \approx E_{TX-OtAComp-Mult} + E_{Comp-OtAComp-Mult} \tag{9}$$

Figure 6 shows a single receiver, comprised of one PD, TIA and ADC. The PD converts optical signal power to signal current i_{PD}. The TIA and ADC convert the signal current to electrical digital data. TIA output is adjusted to map into the full ADC preamplifier input range.

Equal Output Data SNRs in Figs. 5 and 6 requires equal TIA output SNRs ($SNR_{out-TIA}$).

$$SNR_{out-TIA-Fig.6} = SNR_{out-TIA-Fig.5} \tag{10}$$

Ideally, TIA noise is dominated by thermal noise of the feedback resistor R. The TIA resistors in Figs. 5 and 6 have the same value which results in equal output SNR.

In Fig. 6, the PD signal current (i_{PD}) is the input optical power times PD responsivity (r_{PD}).

$$i_{PD} = r_{PD}\ P1\ y[n]\ x1 \tag{11}$$

The average (avg) of the PD signal current is the product of the averages of component terms.

$$avg(i_{PD}) = r_{PD}\ P1\ avg(y[n])\ avg(x1) \tag{12}$$

The scalar x1 can have a value anywhere in the full range. However, the TIA must support the maximum (max) value of $x1$, resulting in same TIA signal current as in Fig. 5.

$$max(i_{PD}) = r_{PD}\ P1\ max(x1) \tag{13}$$

For canonical configurations, TIA energy use scales with the maximum total input current $max(i_{PD})$ [40]. Since the TIAs in Figs. 5 and 6 support the same maximum current $max(i_{PD})$, they use the same energy.

The K-bit ADC in Fig. 6 preserves the modulator 2 output SNR, as per the multiplier SNR analysis in Section 3.1. Therefore, the ADCs in Figs. 5 and 6 use the same energy.

3.3. Summary

Transmitters in Figs. 5 and 6 have the same input SNR, operate at the same modulation rate and optical power, and their modulators use the same energy, resulting in same energy use.

$$E_{TX-EtAComp-Mult} = E_{TX-OtAComp-Mult} \tag{14}$$

Receivers in Figs. 5 and 6 have the same output SNR, operate at the same data sampling rate, and their ADCs and TIAs use the same energy, resulting in same energy use.

$$E_{RX-EtAComp-Mult} = E_{RX-OtAComp-Mult} \tag{15}$$

Therefore, optical and electrical Input Data paths computing vector scalar product employing multiplication operation use the same total energy.

$$E_{Total-OtAComp-Mult} = E_{TX-OtAComp-Mult} + E_{RX-OtAComp-Mult} \tag{16}$$

$$\begin{aligned} E_{Total-EtAComp-Mult} &= E_{TX-EtAComp-Mult} + E_{RX-EtAComp-Mult} \\ &= E_{Total-OtAComp-Mult} \end{aligned} \tag{17}$$

4. Addition

M elements of vector $x[m]$ are shown electrically and optically added together in Figs. 7 and 8, respectively, resulting in scalar $z1$.

$$z1 = \sum_{m=1}^{M} x[m] \tag{18}$$

Input K-bit data is read in-parallel from Data Storage, added to compute output data, and written to Data Storage. WDM implementation constraints limit the number of elements M.

Optics EXPRESS

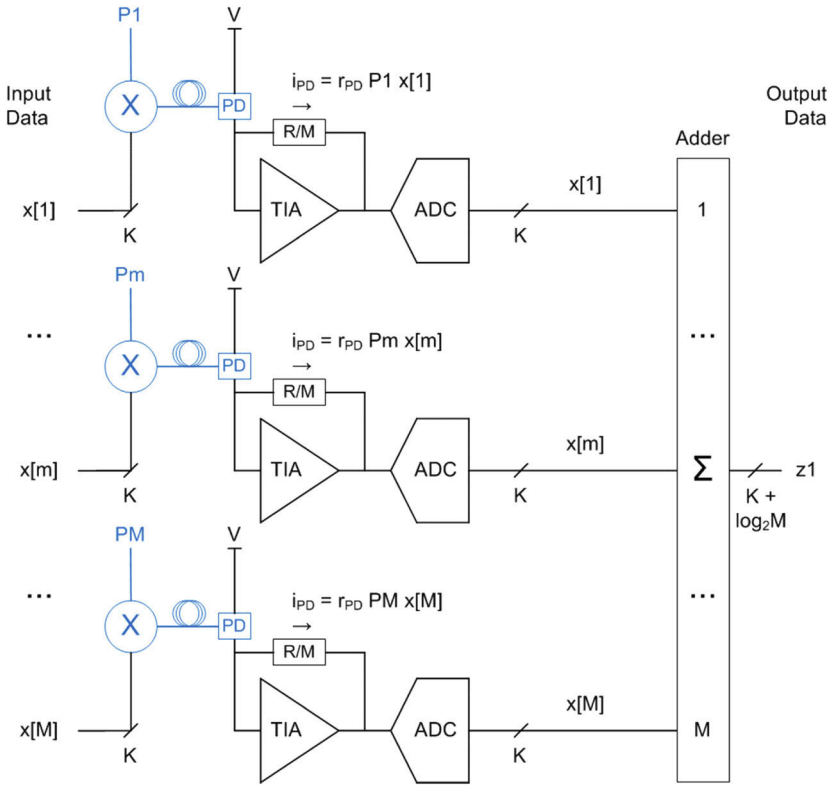

Fig. 7. Input Data path model with electrical addition of $z1 = \sum_{m=1}^{M} x[m]$.

4.1. Electrical

Figure 7 expands the Fig. 3 Input Data path, except shows addition computation instead multiplication computation. It shows M transmitters converting electrical digital data to M optical signals by multiplying continuous wave optical power of each of the M wavelengths by electrical signal representations of the data. Transmitter a_M and other optical path losses are not included. Also not included is optional WDM MUX and DEMUX (demultiplexer), each with a_W path loss, combining and separating the M wavelengths into and from one fiber, if a single fiber link is desired.

Figure 7 shows M receivers, each comprised of one PD, TIA and ADC. Each PD converts optical signal power to signal current i_{PD}. Each TIA and ADC convert the signal current to electrical digital data. TIA output is adjusted to map into the full ADC preamplifier input range. Performance of high-speed CMOS TIAs and ADCs is listed in Tables 2 and 3, respectively.

The electrical adder with M K-bit inputs has a $(K + \log_2 M)$-bit output sum. An implementation of the M-input adder in Fig. 7 is a summing tree using M-1 two-input adders, or approximately one two-input adder per ADC. The simplest implementation reuses latches of each two-input adder resulting in $\log_2 M$ pipeline stages. A synthesis tool optimizing the M-input adder design, using adder precision, clock rate, and process parameters as constraint variables, generates fewer pipeline stages. Pipeline delay is not important in this application.

Because the signal sums coherently, while the noise sums in power, the adder output SNR ($SNR_{out-adder}$) increases 10dB per M decade.

$$SNR_{out-adder} = SNR_{in-adder} + 10\log_{10} M \tag{19}$$

Optics EXPRESS

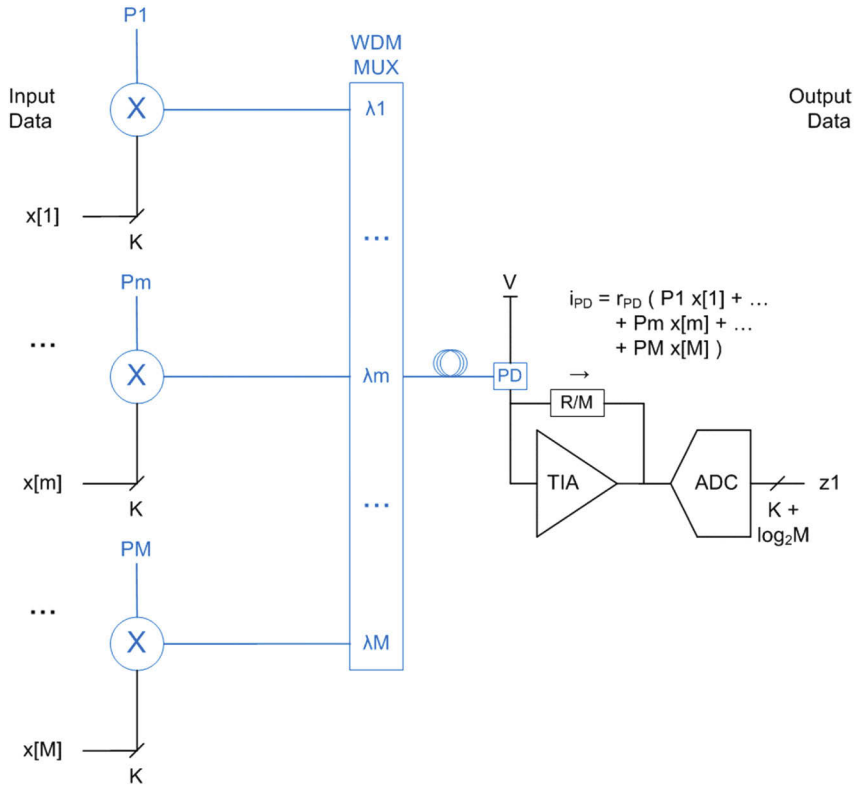

Fig. 8. Input Data path model with optical addition of $z1 = \sum_{m=1}^{M} x[m]$.

Performance of electrical 16-bit adders in 45, 28 and 7nm CMOS process nodes is listed in Table 5. The 7nm values are estimated from 45nm and 28nm values using 10x and 5x energy scaling factors [37], respectively. Energy use of 8-bit adders is ideally half of these values. Energy use is much less than of the TIAs and ADCs listed in Tables 2 and 3, respectively, and is negligible in calculating total energy use of the Input Data path. This is same as in electrical computing where electrical data transfer dominates energy use [8].

$$E_{RX-EtAComp-Add} \approx E_{RX-EtAComp-Add} + E_{Comp-EtAComp-Add} \tag{20}$$

Table 5. CMOS 16-bit adder delay and energy use.

CMOS node nm	Delay (ps) ps	Energy/op (max) fJ	Input bits/op	Rate Gops/s	Energy fJ/bit	Reference
45	1000	496	16	1	31	[38]
45 → 7	40	50	16	25	3	[37,38]
28	448	197	16	2.2	12.5	[41]
28 → 7	30	40	16	33	2.5	[37,41]

4.2. Optical

Figure 8 expands the Fig. 4 Input Data path. It shows M transmitters converting electrical digital data to M optical signal powers by multiplying continuous wave optical power of each of the M wavelengths by electrical signal representations of the data. A WDM MUX passively combines

Optics EXPRESS

the M wavelengths into one fiber. Transmitter a_M, MUX a_W and other optical path losses are not included. For K=1, 0 and 1 binary levels are used, which implements digital optical addition.

$$E_{TX-OtAComp-Add} \approx E_{TX-OtAComp-Add} + E_{Comp-OtAComp-Add} \tag{21}$$

Figure 8 shows a single receiver, comprised of one PD, TIA and ADC. The PD sums the electric field of the M wavelengths and converts the sum to signal current i_{PD}. The TIA and ADC convert the signal current to electrical digital data. TIA output is adjusted to map into the full ADC preamplifier input range. Practical lower limit on the TIA feedback resistor value R/M places a limit on the number of elements M.

Single TIA in Fig. 8 supports the same total photocurrent as M TIAs in Fig. 7. Ideally, TIA noise is proportional to the thermal noise of the feedback resistor, which means it scales with the square root of the resistor value. Therefore, the single TIA output SNR (SNR_{out_TIA}) in Fig. 8 is equal to M times TIA output SNR in Fig. 7.

$$SNR_{out-TIA-Fig.8} = M\ SNR_{out-TIA-Fig.7} \tag{22}$$

The adder output SNR in Fig. 7 is increased by M over the single TIA output SNR (adder input) because the signal sums coherently, while the TIA noise sums in power. Therefore, output SNR due to M TIAs in Fig. 7 is equal to output SNR due to the single TIA in Fig. 8.

Figure 8 shows that the signal current i_{PD} of the single TIA is equal to PD responsivity times the sum of the M wavelength powers. Figure 7 shows that the sum of signal currents of the M TIAs is PD responsivity times the sum of the M wavelength powers, i.e. the same. For canonical configurations, TIA energy use scales with the maximum total input current [40]. Therefore, the M TIAs in Fig. 7 use the same total energy as the single TIA in Fig. 8. TIA implementation constraints, for example by the gain-bandwidth product on the TIA feedback resistor value, place a limit on the number of elements M.

The output SNR of the $(K + \log_2 M)$-bit ADC ($SNR_{ADCout-K+\log_2 M-bit}$) increases 6dB with each bit of resolution increase.

$$SNR_{ADCout-K+\log_2 M-bit} = SNR_{ADCout-K-bit} + 20\log_{10} M\ dB \tag{23}$$

One bit of resolution increase requires 4x lower ADC noise, which with fixed supplies requires 4x increase of the ADC signal capacitor(s) C. This requires g_m (transistor small signal trans-conductance) to increase by 4x to maintain constant g_m/C, which increases ADC energy use by 4x [42]. In general, to increase ADC effective resolution by M bits, or equivalently to increase ADC SNR by $20\log_{10} M$ dB, requires M^2 times the energy.

To increase ADC effective resolution by $\sqrt{M}$ bits, or equivalently to increase the ADC SNR by $10\log_{10} M$ dB, requires M times the energy. This is the same increase in SNR as from summing the output of M ADCs in Fig. 7, whose total energy use is M times that of a single K-bit resolution ADC. Therefore, when operating at the same output SNR, the M ADCs in Fig. 7 use the same total energy as the single $(K + \log_2 M)$-bit ADC in Fig. 8.

For non-return to zero (NRZ) modulation, each of the receivers in Fig. 7 does not require a full ADC, only a limiting amplifier (LA). The receiver in Fig. 8 still needs a linear TIA and $(1 + \log_2 M)$-bit ADC. For simplicity, energy use of a LA is assumed approximately equal to that of a full 1-bit ADC. Therefore, M NRZ receivers in Fig. 7 use the same energy as a single NRZ receiver in Fig. 8.

4.3. Summary

Transmitters in Figs. 7 and 8 have the same input SNR, operate at the same modulation rate and optical power, and their modulators use the same energy, resulting in same energy use.

$$E_{TX-EtAComp-Add} = E_{TX-OtAComp-Add} \tag{24}$$

Optics EXPRESS

Receivers in Figs. 7 and 8 have the same output SNR, operate at the same data sampling rate, and their ADCs and TIAs use the same energy, resulting in same energy use.

$$E_{RX-EtAComp-Add} = E_{RX-OtAComp-Add} \tag{25}$$

Therefore, optical and electrical Input Data paths computing vector element sum employing addition operation use the same total energy.

$$E_{Total-OtAComp-Add} = E_{TX-OtAComp-Add} + E_{RX-OtAComp-Add} \tag{26}$$

$$\begin{aligned} E_{Total-EtAComp-Add} &= E_{TX-EtAComp-Add} + E_{RX-EtAComp-Add} \\ &= E_{Total-OtAComp-Add} \end{aligned} \tag{27}$$

In other words, a single M-input electrical adder negligibly contributes to its associated optical data transfer energy use, just as a single M-input optical adder negligibly contributes to its associated optical data transfer energy use.

4.4. Optical implementation considerations

A significant design challenge of the optical path in Fig. 8 is matching the optical power of all the wavelengths to be within the precision of the data. The power values must be within one least significant bit (LSB) of the desired K-bit precision. This is very difficult and requires complex calibration. For example, in volume optics manufacturing, to achieve reasonable yield of datacom WDM transmitters, all the wavelength optical powers are within 3dB of each other; i.e. the ratio of the highest to lowest optical power is 2 or less, which translates to best case 1-bit of uncalibrated precision. In very high-volume manufacturing matching is 4.5dB, which is less than 1-bit of uncalibrated precision.

A significant design challenge of the receiver in Fig. 8 is that the resolution of the ADC is $\log_2$M bits greater than of the ADC in Fig. 7. The difficulty of implementing increased ADC precision, places a significant limit on the overall optical computing precision.

There are other approaches to implementing optical addition. If signals are coherent and have the same polarization state, their electric fields will sum or subtract when combined.

Another implementation of binary addition is used in a 4-bit optical arithmetic logic unit (ALU) [21]. The ALU core is ~2mm by 2mm; about half the reported chip area. A 7nm CMOS 8-bit electrical ALU is under 2um by 2um. That's an area ratio of over 1 million, which shows the challenge of competing with fine geometry CMOS. The power consumption of the 4-bit optical ALU scaled to 7nm and operating at 4GHz is reported as ~0.5mW. This is an optimistic result because functions like clock distribution, which is typically half the energy use of a computing system, and laser source are not included. The complete power consumption of a 90nm CMOS 64-bit ALU electrical computing core is reported as 300mW when operating at 4GHz [43]. Scaling 90nm to 7nm reduces power by 30x [37], which results in 10mW. A 4-bit ALU is ~16x lower than a 64-bit ALU, which results in ~0.6mW 4-bit electrical ALU power. This is close to the ~0.5mW 4-bit optical ALU power. Both results do not include memory access, which dominates energy use in complete computing systems [8]. The 4-bit optical ALU reference shows another problem with many optical computing proposals which is focus on the non-dominant energy use elements.

As discussed in Section 2.3, low precision optical computing partial implementations are often compared to high precision electrical computing complete implementations like ones using GPUs and TPUs. A nanophotonic accelerator using micro-disk-based adders and shifters is reported as achieving 10x to 100x improvement in energy use over conventional GPUs and TPUs [22]. Also reported is 42.4mW power consumption for the central computing-block made up of sixteen 16-bit optical adders operating at 12.8GHz, which is 13pJ/bit. Yet this is ~5x higher energy use than the 7nm CMOS 16-bit electrical adders listed in Table 5.

Optics EXPRESS

The above nanophotonic accelerator enhanced with photonic local storage registers reports 20x to 600x improvement in energy use over GPUs and TPUs [23]. Also reported is 1060mW power consumption for the central computing-block made up of twenty-five 16-bit optical adders operating at 50GHz, which is 53pJ/bit. Yet this is ~20x higher energy use than the 7nm CMOS 16-bit electrical adders listed in Table 5. Both nanophotonic accelerator implementations show that computing-block energy use is minor.

5. Inner product

NxM-element matrix $Y[n,\ m]$ is shown electrically and optically multiplied by M-element vector $x[m]$ in Figs. 9 and 10, respectively, resulting in N-element vector $z[n]$.

$$z[n] = Y[n,\ m]\ x[m] \tag{28}$$

Figure 9 is a combination of Figs. 5 and 7. Figure 10 is a combination of Figs. 6 and 8. Input K-bit wide data is read in-parallel/sequentially from Data Storage, matrix vector multiplied to compute output data, and written sequentially to Data Storage. WDM implementation constraints place a limit on the number of elements M.

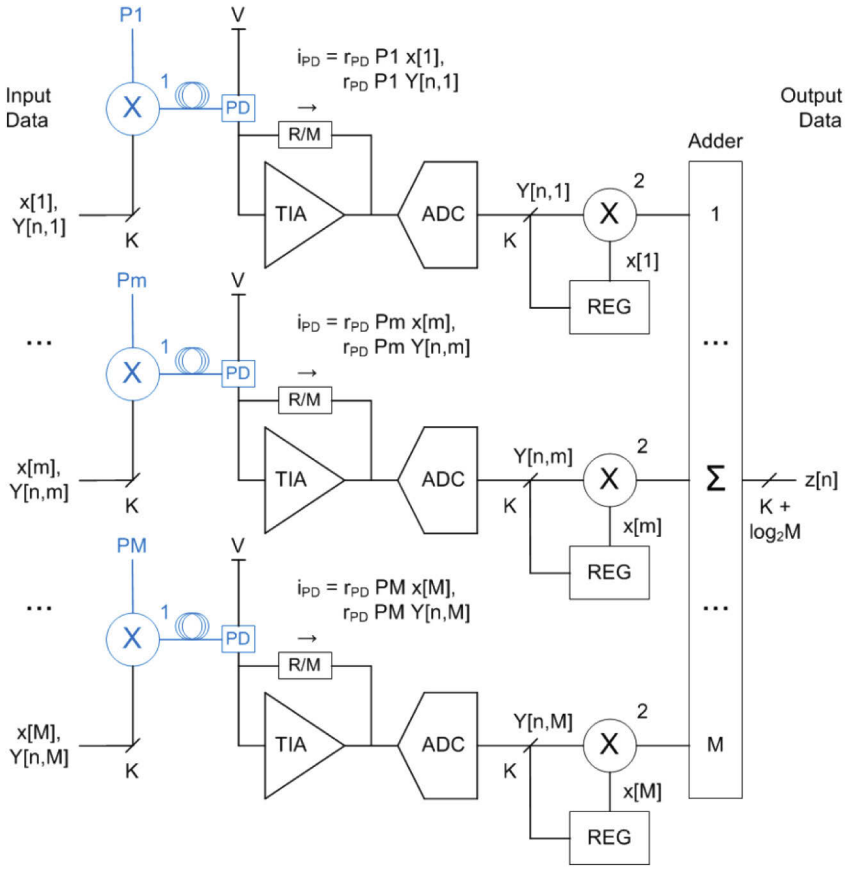

Fig. 9. Input Data path model with electrical inner product of $z[n] = Y[n,\ m]\ x[m]$.

Optics EXPRESS

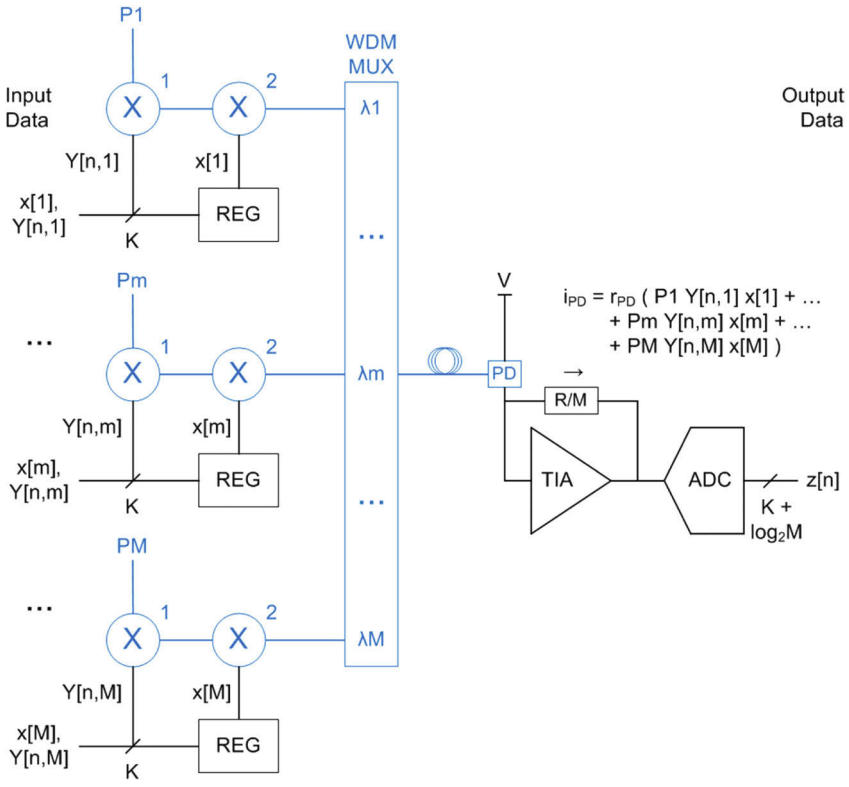

Fig. 10. Input Data path model with optical inner product of $z[n] = Y[n,\ m]\ x[m]$.

5.1. Electrical

Figure 9 shows M one modulator transmitters converting electrical digital data to M optical signal powers by multiplying continuous wave optical power of each of the M wavelengths by electrical signal representations of the data. Transmitter a_M and other optical path losses are not included. Also not included is optional WDM MUX and DEMUX, each with a_W path loss, combining and separating the M wavelengths into and from one fiber, if a single fiber link is desired.

Figure 9 shows M receivers, each comprised of one PD, TIA and ADC. Each PD converts optical signal power to signal current i_{PD}. Each TIA and ADC convert the signal current to electrical data. TIA output is adjusted to fully map into the ADC preamplifier input range. Performance of high-speed CMOS TIAs and ADCs is listed in Tables 2 and 3, respectively.

Performance of electrical multiplier-adders in 14 and 7nm CMOS process nodes is listed in Table 6. The first set of 7nm values are derived from the power consumption sum of the 16-bit multiplier and adder [37,38] listed in Tables 4 and 5, respectively. The second set of 7nm values are estimated from the 8-bit multiplier-adder 14nm values using 1.4x energy scaling factor [37]. Energy use is consistent with that of many commercial high-speed digital equalizers. It is much less than of the TIAs and ADCs listed in Tables 2 and 3, respectively, and is negligible in calculating total energy use of the Input Data path.

$$E_{RX-EtAComp-InnerP} \approx E_{RX-EtAComp-InnerP} + E_{Comp-EtAComp-InnerP} \tag{29}$$

An implementation of the M-input adder in Fig. 9 is a summing tree using M-1 two-input adders, or approximately one two-input adder per multiplier. The simplest implementation reuses latches of each two-input adder resulting in $\log_2 M$ pipeline stages.

Optics EXPRESS

Table 6. CMOS 8-bit and 16-bit multiplier-adder delay and energy use.

CMOS node nm	Delay (ps) ps	Energy/op (max) fJ	Input bits/op	Rate Gops/s	Energy fJ/bit	Reference
45 → 7	58	367	16	17.5	23	[37,38]
14	18	222	8	56	28	[44]
14 → 7	11	159	8	90	20	[37,44]

5.2. Optical

Figure 10 shows M two modulator transmitters converting electrical digital data to M optical signals by multiplying continuous wave optical power of each of the M wavelengths by electrical signal representations of the data. A WDM MUX passively combines the M wavelengths into one fiber. Transmitter $a_M{}^2$, MUX a_W and other optical path losses are not included. For K=1, 0 and 1 binary levels are used, which implements digital optical inner product.

Optical path losses require compensation with optical or electrical amplification, which may significantly increase energy use. If a single fiber link is used for the Fig. 9 data path, the total loss through its optical components is $a_M\ a_W{}^2$. This is approximately equal to $a_M{}^2\ a_W$, the total loss through the optical components in Fig. 10. Therefore, not including energy use of amplification required to compensate for optical path losses, does not affect relative energy use comparison of Figs. 9 and 10 Input Data paths.

Figure 2 shows the data sequencing. Modulator 2 operates at a much lower rate than modulator 1, and its energy use is negligible in comparison.

$$E_{TX-OtAComp-InnerP} \approx E_{TX-OtAComp-InnerP} + E_{Comp-OtAComp-InnerP} \tag{30}$$

Figure 10 shows a single receiver, comprised of one PD, TIA and ADC. The PD sums the electric field of the M wavelengths and converts the sum to signal current i_{PD}. The TIA and ADC convert the signal current to electrical digital data. TIA output is adjusted to fully map into the ADC preamplifier input range. Practical lower limit on the TIA feedback resistor value R/M places a limit on the number of elements M.

5.3. Summary

Transmitters in Figs. 9 and 10 have the same input SNR, operate at the same modulation rate and optical power, and their modulators use the same energy, resulting in same energy use.

$$E_{TX-EtAComp-InnerP} = E_{TX-OtAComp-InnerP} \tag{31}$$

Receivers in Figs. 9 and 10 have the same output SNR, operate at the same data sampling rate, and their ADCs and TIAs use the same energy, resulting in same energy use.

$$E_{RX-EtAComp-InnerP} = E_{RX-OtAComp-InnerP} \tag{32}$$

Therefore, optical and electrical Input Data paths computing matrix vector product employing inner product operation use the same total energy.

$$E_{Total-OtAComp-InnerP} = E_{TX-OtAComp-InnerP} + E_{RX-OtAComp-InnerP} \tag{33}$$

$$\begin{aligned} E_{Total-EtAComp-InnerP} &= E_{TX-EtAComp-InnerP} + E_{RX-EtAComp-InnerP} \\ &= E_{Total-OtAComp-InnerP} \end{aligned} \tag{34}$$

Fully parallel computing can be implemented by replicating N times all the blocks in Figs. 9 and 10. Since the same total operations are performed by the same blocks, serial and parallel computing energy use is the same.

Optics EXPRESS

6. Conclusion

This paper shows that optical Transmitters in the Input Data paths with electrical and optical type A Computing, like in Figs. 3 and 4, respectively, use the same energy. Further, the energy use by optical type A Computing, like in Fig. 4, is shown to be negligible compared to energy use by the optical Transmitter preceding it. Similar results are used to support many optical computing proposals.

This paper also shows that optical Receivers in the Input Data paths with electrical and optical type A Computing, like in Figs. 3 and 4, respectively, use the same energy. Further, energy use by electrical type A Computing, like in Fig. 3, is shown to be negligible compared to energy use by the optical Receiver preceding it. Such results are missing in many optical computing proposals, and lead to incorrect energy use conclusions.

Energy use by math intensive programmable computing tasks, like matrix vector product, is dominated by data transfer, for either optical or electrical computing. By itself, the optical and electrical computing energy use is negligible compared to optical transmitter and receiver energy use, respectively, and does not affect energy use totals. For math intensive programmable operations like in machine learning applications, switching to optical from electrical computing does not lower energy use. This is in stark contrast to significant energy use reduction by switching to optical from electrical data transfer.

In this paper, the comparison of optical and electrical computing energy use is apples-to-apples, which constrains the optical data transfer to be the same. However, optical computing restricts modulation and coding choices, which prevents minimizing energy use through link parameter optimization. Electrical computing does not impose such restrictions and can have fully optimized optical link parameters, which results in significantly lower data transfer energy use than for optical computing. This is basic communication theory, which is not rederived in this paper.

Acknowledgments. The author would like to thank Dr. Po Dong, Arash Farhoodfar, Prof. Boris Murmann, Dr. Roberto Rodes, Prof. Clint Schow, Dr. Peter Winzer, Nelson Wright, Prof. S.J. Ben Yoo, and Optics Express Reviewers for their careful review and detailed comments.

Disclosures. The author declares no conflicts of interest.

Data availability. Data underlying the results presented in this paper are available in Refs. [27–44].

References

1. D. Hartree, "The ENIAC: An Electronic Computing Machine," Nature **158**(4015), 500–506 (1946).
2. M. Horowitz, "Computing's energy problem (and what we can do about it)," in *IEEE International Solid-State Circuits Conference Digest of Technical Papers*, pp. 10–14, February 2014.
3. N. Verma, H. Jia, H. Valavi, M. Ozalay, L. Chen, B. Zhang, and P. Deaville, "In-memory computing: Advance and prospects," IEEE Solid-State Circuits Mag. **11**(3), 43–55 (2019).
4. B. Murmann, "Mixed-Signal Computing for Deep Neural Network Inference," IEEE Trans. VLSI Syst. **29**(1), 3–13 (2021).
5. D. Schneider, "Deep Learning at the Speed of Light," IEEE Spectrum **58**(1), 28–29 (2021).
6. J. Koetsier, "Photonic Supercomputer For AI: 10X Faster, 90% Less Energy, Plus Runway For 100X Speed Boost," *Forbes*, April 2021.
7. S. Magar, E. Caudel, and A. Leigh, "A Microcomputer with Digital Signal Processing Capability," in *International Solid-State Circuits Conference Digest of Technical Papers*, pp. 32–33, (1982).
8. V. Sze, Y. Chen, T. Yang, and J. Emer, "How to Evaluate Deep Neural Network Processors, TOPS/W (alone) Considered Harmful, ISSCC 2020 Tutorial," *IEEE Solid-State Circuits Magazine*, pp. 28–41, (2020).
9. C. Sun, M. Wade, Y. Lee, J. Orcutt, L. Alloatti, M. Georgas, A. Waterman, J. Shainline, R. Avizienis, S. Lin, B. Moss, R. Kumar, F. Pavanello, A. Atabaki, H. Cook, A. Ou, J. Leu, Y.-H. Chen, K. Asanovic, R. Ram, M. Popovic, and V. Stojanovic, "Single-chip Microprocessor That Communicates Directly Using Light," Nature **528**(7583), 534–538 (2015).
10. J. Enoch and V. Lakshminarayanan, "Duplication of unique optical effects of ancient Egyptian lenses from the IV/V Dynasties; lenses fabricated ca 2620-2400 BC or roughly 4600 years ago," Opthalmic and Physiology Optics **20**(2), 126–130 (2000).
11. B. Yoo and D. Miller, "Nanophotonic computing: scalable and energy-efficient computing with attojoule nanophotonics," in *IEEE Photonics Society Summer Topical Meeting Series*, pp. 1–2, July 2017.

Optics EXPRESS

12. A. Tait, T. de Lima, E. Zhou, A. Wu, M. Nahmias, B. Shastri, and P. Prucnal, "Neuromorphic Photonic Networks Using Silicon Photonic Weight Banks," Sci. Rep. **7**, 7430 (2017).
13. M. Nazirzadeh, M. Shamsabardeh, and B. Yoo, "Energy-Efficient and High-Throughput Nanophotonic Neuromorphic Computing," in *Conference on Lasers and Electro-Optics (CLEO) OSA Technical Digest*, no. ATh3Q.2, May 2018.
14. J. Feldmann, N. Youngblood, C. Wright, H. Bhaskaran, and W. Pernice, "All-optical Spiking Neurosynaptic Networks with Self-learning Capabilities," Nature **569**(7755), 208–214 (2019).
15. D. Sarrazin, H. Jordan, and V. Heuring, "Digital fiber optic delay line memory," Appl. Opt. **29**(5), 627–637 (1990).
16. D. Hillerkuss, M. Winter, M. Teschke, A. Marculescu, J. Li, G. Sigurdsson, K. Worms, S. Ezra, N. Narkiss, W. Freude, and J. Leuthold, "Simple all-optical FFT scheme enabling Tbit/s real-time signal processing," Opt. Express **18**(9), 9324–9340 (2010).
17. Y. Shen, N. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljacic, "Deep Learning with Coherent Nanophotonic Circuits," Nat. Photonics **11**(7), 441–446 (2017).
18. T. Ishihara, J. Shiomi, N. Hattori, Y. Masuda, A. Shinya, and M. Notomi, "An Optical Neural Network Architecture based on Highly Parallelized WDM-Multiplier-Accumulator," in *IEEE/ACM Workshop on Photonics-Optics Technology Oriented Networking, Information and Computing Systems (PHOTONICS)*, pp. 15–21, November 2019.
19. A. Mehrabian, M. Miscuglio, Y. Alkabani, V. Sorger, and T. El-Ghazawi, "A Winograd-based Integrated Photonics Accelerator for Convolutional Neural Networks," IEEE J. Sel. Top. Quantum Electron. **26**(1), 1–12 (2020).
20. V. Bangari, B. Marquez, H. Miller, A. Tait, M. Nahmias, T. Ferreira de Lima, H. Peng, P. Prucnal, and B. Shastri, "Digital Electronics and Analog Photonics for Convolutional Neural Networks (DEAP-CNNs)," IEEE J. Sel. Top. Quantum Electron. **26**(1), 1–13 (2020).
21. Z. Ying, C. Feng, Z. Zhao, S. Dhar, H. Dalir, J. Gu, Y. Cheng, R. Soref, D. Pan, and R. Chen, "Electronic-photonic arithmetic logic unit for high-speed computing," Nat. Commun. **11**(1), 2154 (2020).
22. 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.
23. 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.
24. F. Sunny, A. Mirza, M. Nikdast, and S. Pasricha, "CrossLight: A Cross-Layer Optimized Silicon Photonic Neural Network Accelerator," Early-stage research not yet peer reviewed, pp. 1–6, February 2021.
25. F. Yu, K. Sun, Q. Yu, and A. Beling, "High-Speed Evanescently-Coupled Waveguide Type-II MUTC Photodiodes for Zero-Bias Operation," J. Lightwave Technol. **38**(24), 6827–6832 (2020).
26. K. Nozaki, S. Matsuo, T. Fujii, K. Takeda, A. Shinya, E. Kuramochi, and M. Notomi, "Femtofarad optoelectronic integration demonstrating energy-saving signal conversion and nonlinear functions," Nat. Photonics, vol. **13**, no. 7, pp. 454–459, (2019).
27. C. Schow, A. Rylyakov, C. Baks, F. Doany, and J. Kash, "25-Gb/s 6.5-pJ/bit 90-nm CMOS-Driven Multimode Optical Link," IEEE Photonics Technol. Lett. **24**(10), 824–826 (2012).
28. L. Szilagyi, J. Pliva, R. Henker, D. Schoeniger, J. Turkiewicz, and F. Ellinger, "A 53-Gbit/s Optical Receiver Frontend With 0.65 pJ/bit in 28-nm Bulk-CMOS," IEEE J. Solid-State Circuits **54**(3), 845–855 (2019).
29. K. Li, S. Liu, X. Ruan, D. Thomson, Y. Hong, F. Yang, L. Zhang, C. Lacava, F. Meng, W. Zhang, P. Petropoulos, F. Zhang, and G. Reed, "Co-design of a differential transimpedance amplifier and balanced photodetector for a sub-pJ/bit silicon photonics receiver," Opt. Express **28**(9), 14038–14054 (2020).
30. K. Lakshmikumar, A. Kurylak, M. Nagaraju, R. Booth, R. Nandwana, J. Pampanin, and V. Boccuzzi, "A Process and Temperature Insensitive CMOS Linear TIA for 100 Gb/s/ λ PAM-4 Optical Links," IEEE J. Solid-State Circuits **54**(11), 3180–3190 (2019).
31. H. Li, G. Balamurugan, J. Jaussi, and B. Casper, "A 112 Gb/s PAM4 Linear TIA with 0.96 pJ/bit Energy Efficiency in 28 nm CMOS," in *IEEE 44th European Solid State Circuits Conference (ESSCIRC)*, pp. 238–241, September 2018.
32. B. Xu, Y. Zhou, and Y. Chiu, "A 23 mW 24GS/s 6b Time-Interleaved Hybrid Two-Step ADC in 28 nm CMOS," IEEE J. Solid-State Circuits **52**(4), 1091–1100 (2017).
33. Z. Zhang, L. Wei, J. Lagos, E. Martens, Y. Zhu, C.-H. Chan, J. Craninckx, and R. Martins, "A Single-Channel 5.5 mW 3.3GS/s 6b Fully Dynamic Pipelined ADC with Post-Amplification Residue Generation," in *International Solid-State Circuits Conference Digest of Technical Papers*, pp. 254–256, February 2020.
34. M. Zhang, Y. Zhu, C. Chan, and R. Martins, "A 4× Interleaved 10GS/s 8b Time-Domain ADC with 16× Interpolation-Based Inter-Stage Gain Achieving >37.5 dB SNDR at 18 GHz Input," in *International Solid-State Circuits Conference Digest of Technical Papers*, pp. 252–254, February 2020.
35. Y. Lyu and F. Tavernier, "A 1GS/s Reconfigurable BW 2nd-Order Noise-Shaping Hybrid Voltage-Time Two-Step ADC Achieving 170.9 dB FoM," in *Symposia on VLSI Technology & Circuits, Virtual Conference*, no. CD2.4, June 2020.
36. M. Pisati, F. De Bernardinis, P. Pascale, and C. Nani, "A Sub-250 mW 1-to-56Gb/s Continuous-Range PAM-4 42.5 dB IL ADC/DAC-Based Transceiver in 7 nm FinFET," in *IEEE International Solid- State Circuits Conference Digest of Technical Papers*, pp. 116–118, February 2019.
37. A. Stillmaker and B. Baas, "Scaling equations for the accurate prediction of CMOS device performance," Integration **58**, 74–81 (2017).

Optics EXPRESS

38. Q. Xie, X. Lin, S. Chen, M. Dousti, and M. Pedram, "Performance Comparisons between 7 nm FinFET and Conventional Bulk CMOS Standard Cell Libraries," IEEE Trans. Circuits Syst. II **62**(8), 761–765 (2015).
39. 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.
40. B. Razavi, "The Transimpedance Amplifier (A Circuit for All Seasons)," IEEE Solid-State Circuits Mag. **11**(1), 10–97 (2019).
41. A. Vatanjou, E. Lte, T. Ytterdal, and S. Aunet, "Ultra-low Voltage and Energy Efficient Adders in 28 nm FDSOI Exploring Poly-biasing for Device Sizing," Microprocessors and Microsystems **56**, 92–100 (2018).
42. B. Murmann, "A/D Converter Trends: Power Dissipation, Scaling and Digitally Assisted Architectures," in *IEEE Custom Integrated Circuits Conference*, pp. 105–112, September 2008.
43. S. Mathew, M. Anders, B. Bloechel, and T. Nguyen, "A 4-GHz 300-mW 64-bit integer execution ALU with dual supply voltages in 90-nm CMOS," IEEE J. Solid-State Circuits **40**(1), 44–51 (2005).
44. 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 14 nm CMOS," in *IEEE International Solid-State Circuits Conference Digest of Technical Papers*, pp. 104–105, February 2018.