Floating-point photonic solver using Newton's method

Andrew B. Klein[1], Zheyuan Zhu[1], Guifang Li[1], and Shuo S. Pang[1]
1 CREOL, The College of Optics and Photonics, University of Central Florida, 4304 Scorpius St., Orlando, FL, 32816
Author e-mail address: Andrew.Klein@ucf.edu

Abstract: We present a method for configurable, signed, floating-point encoding and multiplication on limited precision photonic primitives, demonstrating Newton's method with improved accuracy and expanding the dynamic range of the photonic solver by over 200×.

1. Introduction

While CPUs and digital accelerators such as GPUs use floating-point (FLP) format to guarantee high-precision operations, photonic or analog computing uses signal amplitude equivalent to fixed-point (FXP) format, which accommodates limited signal-to-noise ratio (SNR) and reduces data storage requirements. The adoption of FXP encoding guarantees an inherent error for any computing applications, and while in some cases this is not detrimental, it often leads to errors in iterative or high-precision solutions. Our previous work creating a photonic eigensolver showed that using 4-bit FXP precision is not sufficient to achieve FLP-equivalent accuracy [1].

This work demonstrates a method for implementing FLP encoding on a photonic multiplication primitive with 5 signed input levels. The method performs passive significand multiplication and exponent addition operations, promising increased energy efficiency for the same dynamic range. Here, the FLP photonic primitive was used to implement the Newton-Raphson root finding algorithm to find the Golden Ratio. Compared to CPU-based FLP and simulated FXP implementations, the photonic FLP primitive (P-FLP) shows good accuracy and equivalent convergence rate with digital FLP (D-FLP) encoding.

2. Operating Principles

The photonic primitive depicted in Fig. 1(a) consists of a balanced coherent interferometer with independently modulated inputs A and B generating the respective time-division multiplexed (TDM) fields E_A and E_B. The resulting inputs undergo balanced photodetection to produce outputs of signed magnitude $E_A E_B$. The E field amplitude accommodates 5 signed input levels (±2, ±1, and 0). This is the total number of levels for FXP encoding, while FLP encoding increases the effective bitwidth of the system without increasing the signal power or SNR.

In D-FLP encoding, each value A is broken down into a significand s_A and corresponding exponent l_A with a global base β such that $A = s_A \times \beta^{l_A}$. For 32-bit D-FLP encoding, the significand and exponent are recorded within 31 bits using a global base of 2 or 10 [2]. To multiply two numbers, their significands undergo bitwise multiplication, their exponents are added, and an XOR operation is performed on the sign bits [3], such that $A \times B = (s_A \times \beta^{l_A}) \times (s_B \times \beta^{l_B}) = s_A s_B \times \beta^{l_A + l_B}$. Each process requires energy for computation.

In P-FLP encoding, each multiplicand is encoded with a sub-carrier, with amplitude corresponding to the signed significand and modulation frequency corresponding to the exponent: $A = s_A \times \beta^{l_A} \rightarrow s_A \cos(2\pi f_{l_A} t)$, where f_{l_A} is the subcarrier frequency assigned to the exponent l_A and t is the time since the start of the symbol. To multiply two numbers, P-FLP symbols A and B are encoded as time-division multiplexed (TDM) vectors and input to the photonic primitive using directly controlled Mach-Zehnder modulation [1]. Within the primitive, the signed elementwise product is taken at each point along the TDM vectors as shown in Fig. 1(b), resulting in an output $A \times B = s_A \cos(2\pi f_{l_A} t) \times s_B \cos(2\pi f_{l_B} t) = s_A s_B \cos(2\pi f_{l_A} t) \cos(2\pi f_{l_B} t)$ captured via oscilloscope.

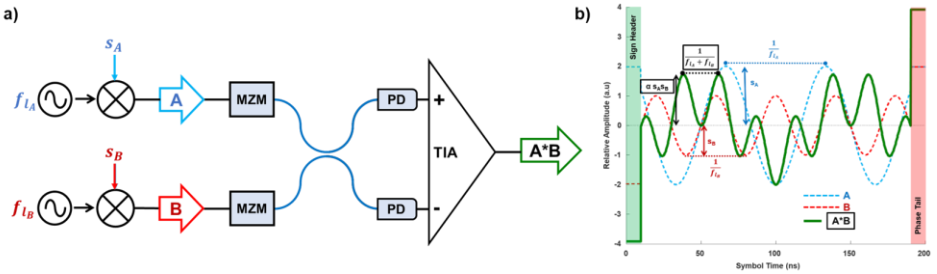

Figure 1. (a) Schematic for P-FLP multiplication. (b) Simulated multiplication $A = 2 \times \beta^{f_{l_A}}$ multiplied by $B = -1 \times \beta^{f_{l_B}}$.

The maximum resulting output frequency will be $f_{l_A} + f_{l_B}$ and can be recovered through a DCT-II transformation, while the sign and significand can be proportionally determined from the amplitude. To make symbol parsing more reliable, a signed header is added to each packet for parity check. Additionally, a tail of constant positive amplitude is added to verify constructive interference. The sub-carrier modulated symbol is synthesized through a field programmable gate array (FPGA, Xilinx XCZU49DR) with two input modulators (JDSU OC-192) can reliably

accommodate carrier frequencies up to 500 MHz, while the output is digitized by a streaming digitizer sampling at 6.25 GSa/sec. Each FLP symbol is repeated for a duration of 200 ns; symbol rate of sub 10 ns can be used in theory. The multiplication, sign operation, and exponent addition are all performed solely through interference; the energy is only expended for signal modulation and detection.

With modifications to accommodate deployment on the photonic primitive, the Newton-Raphson [4] method takes the form of Algorithm 1 for a precalculated function $f(x)$ with $D(x) = 1/f'(x)$, and initial guess x_0.

Algorithm 1. Modified Newton-Rhapson Method

$f_q(x) \leftarrow quantize(f(x))$; $D_q(x) \leftarrow quantize(D(x))$;
for k in *range(N):*
$$x_k \leftarrow x_{k-1} - (f_q(x_{k-1}) \times D_q(x_{k-1}))$$

The modified Newton-Raphson Method is well suited to find roots of quadratic equations so long as the root is not also the vertex. The application demonstrated herein is to find the Golden Ratio φ, defined as the value $\varphi = a/b$ such that $a/b = (a+b)/a$, which can be changed to the quadratic equation $\varphi^2 - \varphi - 1 = 0$.

3. Results and Discussion

Algorithm 1 was used to solve the Golden Ratio quadratic equation $\varphi^2 - \varphi - 1 = 0$ using P-FLP multiplication over 15 iterations with an initial guess of $x_0 = 1.2$ (Fig. 2(a)). The set of exponents corresponding to the distinct f_l were chosen as {-4, 1}, with global base $\beta_{FLP} = 3$ to match the signed 5-level inputs of the photonic primitive. This result was compared to the single-precision D-FLP multiplication as well as base-2 FXP calculation simulated with 129 signed input levels. For P-FLP quantization, each value in the precalculated f(x) and D(x) was rounded to its nearest value in the P-FLP encoding schema. For FXP quantization, each value in f(x) and D(x) was encoded as in Equation 1, corresponding to a global base $\beta_{FXP} = 2$, and where N is the number of positive FXP levels.

$$f_q(x_i) = round\left(\frac{N \times f(x_i)}{2^{floor(\log_2(\max(|f(x)|)))}}\right) \times 2^{\text{floor}(\log_2(\max(|f(x)|))) - \text{round}(\log_2(N))} \quad (1)$$

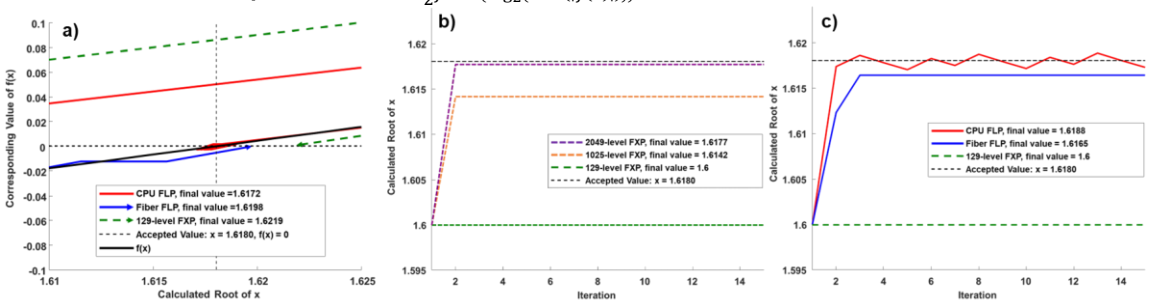

Figure 2. Golden Ratio Calculation. (a) Convergence comparison, initial guess x_0=1.6. (b) Iteration progression, initial guess x_0=1.2. (c) Simulated FXP iteration, initial guess x_0=1.6.

Fig. 2(a) shows that the P-FLP implementation converges to $\varphi = 1.6198$, a result within 0.11% of the accepted value. Compared to the 0.05% error of the double-precision D-FLP on CPU and the 0.24% error of the 129-level FXP implementation, P-FLP offers a significant advantage for limited S/N encoding and approaches the accuracy required for scientific computing. Additionally, all methods show the same rate of convergence.

However, as the model converges closely to the expected value, the number of available FXP levels must be dramatically increased to match the accuracy of the P-FLP implementation. As shown in Fig. 2(b), the D-FLP implementation cannot iterate with $x_0 = 1.6$ if 129-level FXP encoding is used, showing that P-FLP encoding offers immediate advantages for photonic computing in expanding the accuracy and effective bitwidth. Fig. 2(c) further shows that 2049 signed FXP levels must be used to exceed the accuracy of the P-FLP implementation. It is reasonable to anticipate that with a better S/N ratio and enough bands to accommodate more than 6 exponents, the P-FLP implementation could soon match or exceed the performance of the D-FLP implementation.

4. Summary

In summary, we have demonstrated a method for implementing FLP multiplications in photonic computing, using the properties of the analog signal to perform sign operations, significand multiplication, and exponent addition without expending energy. Using Newton's method to find the Golden Ratio, P-FLP multiplications produced an accuracy of ±0.11%, which exceeded the accuracy of 1025-level FXP multiplications. This effectively increased the bitwidth of the system by over 200x without increased requirements for SNR.

[1] Zhu et al., "Sparse coherent photonic processor for solving eigenmode problems", in IPC 2023 Orlando Conference Proceedings
[2] "IEEE Standard for Floating-Point Arithmetic," in IEEE Std 754-2019 (Revision of IEEE 754-2008), vol., no., pp.1-84, 22 July 2019
[3] Rafiquzzaman, M. Fundamentals of Digital Logic and Microcomputer Design. Germany: Wiley, 2005.
[4] Kendall E. Atkinson, An Introduction to Numerical Analysis, (1989) John Wiley & Sons, Inc.