On-Fiber Photonic Computing

Mingran Yang[*]
MIT

Zhizhen Zhong[*]
MIT

Manya Ghobadi
MIT

ABSTRACT

In the 1800s, Charles Babbage envisioned computers as analog devices. However, it was not until 150 years later that a Mechanical Analog Computer was constructed for the US Navy to solve differential equations. With the end of Moore's Law, photonic computing is revitalizing the promise of analog computing by leveraging photons' speed, bandwidth, and energy efficiency for faster, more efficient, and scalable analog computing systems. This paper argues that the networking community should augment pluggable transponders with photonic computing capabilities to enable a backward-compatible solution for in-network computing. We propose *on-fiber* photonic computing to perform computing operations inside network transponders while the data is in the optical domain. We discuss the components required to enable the seamless integration of computation into the very fabric of optical communication links. We then discuss several use cases of on-fiber photonic computing, including machine learning inference, video encoding, load balancing, and intrusion detection.

CCS CONCEPTS

• **Hardware** → **Networking hardware**; **Emerging optical and photonic technologies**; • **Networks** → **Physical links**; *In-network processing*; Wide area networks;

KEYWORDS

Photonic computing, Network hardware design, Optical Networks, In-network computing

ACM Reference Format:
Mingran Yang, Zhizhen Zhong, and Manya Ghobadi. 2023. On-Fiber Photonic Computing. In *The 22nd ACM Workshop on Hot Topics in Networks (HotNets '23), November 28–29, 2023, Cambridge, MA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3626111.3628177

1 INTRODUCTION

Recent advancements in programmable switches have led to the emergence of in-network computing paradigms [21, 27, 31, 37, 46, 52, 56, 62, 65, 66], wherein computations are executed within the network. In-network computing minimizes latency and optimizes efficiency by implementing applications, such as packet classification, inside network switches/routers. However, today's in-network computing proposals rely on implementing computing operations directly on switch/router ASICs, placing a burden on the available resources of the die, which is already operating at its maximum capacity. As a result, state-of-the-art in-network computing proposals cannot perform the complex operations needed to run various latency-sensitive applications [14, 22, 52].

To address this challenge, prior work has suggested augmenting current router ASICs with hardware accelerators to accommodate more complex operations [1, 21, 22, 52]. These approaches typically involve integrating new compute building blocks into existing router architectures. However, replacing all switches/routers in a network with new ones capable of performing complex in-network operations is challenging, if not impossible. On the one hand, given the chip power and area budget constraints of router ASIC, these new architectures support a limited number of computations apart from packet routing. On the other hand, deploying a new set of ASICs with equivalent reliability and robustness in today's system is economically prohibitive and requires a significant time investment. We argue that a *backward-compatible pluggable* approach is required to address this challenge. To that end, we propose enhancing pluggable optical transponders with photonic computing capabilities.

Photonic computing is an emerging field that utilizes lightwaves to execute a wide range of computation operations, including vector multiplication [19, 50, 71], pattern matching [6, 75], comparison [6], signal processing [34], and logic operations [68]. This paradigm enables fast and energy efficient computing [50, 60, 71]. By encoding and manipulating information in the *analog domain* using light, photonic computing may revolutionize various application domains, such as machine learning acceleration [19, 50, 60, 71], video processing [74], solving complex optimization problems [38], and matching data with pre-determined patterns [6, 75].

This paper argues that pluggable transponders are a prime platform for performing photonic computing inside the network without having to replace networking switches and

[*] Equal contribution

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

HotNets '23, November 28–29, 2023, Cambridge, MA, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0415-4/23/11.
https://doi.org/10.1145/3626111.3628177

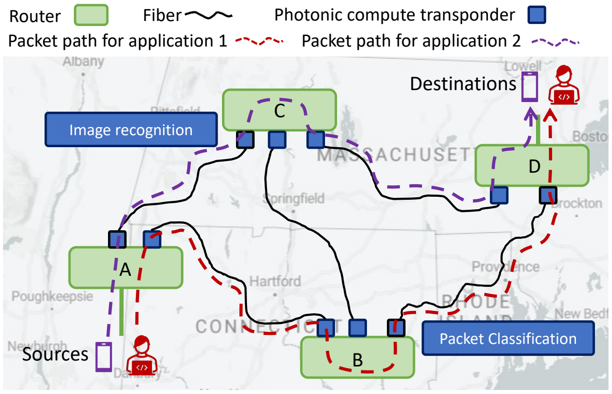

Router
Fiber
Photonic compute transponder
Packet path for application 1
Packet path for application 2
Destinations
Image recognition
C
D
A
B
Sources
Packet Classification

Figure 1: High-level idea of this paper: networks with on-fiber photonic computing.

routers. Optical transponders are ubiquitous in today's wide-area and datacenter networks [43, 44, 48], giving us a unique opportunity to re-purpose them for photonic computing.

Previous state-of-the-art photonic computing proposals ignored backward compatibility with network routers and focused on replacing existing digital computing components with photonic counterparts without considering the network [13, 20, 23, 25, 26, 50, 51, 63, 70, 71]. For instance, Lightning [71] is a photonic-electronic SmartNIC that replaces digital multiplication and accumulation ALUs with photonic cores to serve inference requests while keeping the rest of the digital sub-system intact.

Figure 1 illustrates our proposed vision. In this network, data flows from the source node (site A) to the destination node (site D), traversing intermediate nodes B and C. Unlike conventional networks, we propose performing computation operations, such as packet classification and image recognition, in the optical domain. Our proposed scheme includes enhancing pluggable transponders with photonic computing capabilities to perform computation operations at each node. To support a wide range of applications, our proposal involves configuring transponders with the capability to handle specific computation tasks and route packets to the intermediate nodes based on their computation requirements.

Consider the scenario in Figure 1, in which a user at source site A is sending packets to a user at destination site D. Simultaneously, a cell phone at source site A intends to transmit an image, along with its image recognition result, to another cell phone at destination site D. A photonic computing transponder with packet classification capability is located at site B and another photonic computing transponder with image recognition capability is located at site C. In this case, the packet path for the laptop is $A \longrightarrow B \longrightarrow D$, and the photonic compute transponder at site B performs packet classification computation. Similarly, the packet path for the cell phone is $A \longrightarrow C \longrightarrow D$, and the photonic compute transponder at site C performs image recognition computation.

To realize our vision, there are several research challenges to address. First, we need to carefully design the transponder hardware to accommodate diverse computation tasks within the constraints of a transponder's form factor while facilitating effective coordination between the enhanced transponder and the existing routing hardware. Second, we need a centralized controller to track the information of all the photonic compute transponders and intelligently reconfigure them for various photonic compute tasks based on user demands. Third, we need a compute-communication protocol to enable networking devices and end-hosts to distinguish between compute and non-compute packets while transmitting packets seamlessly through the network.

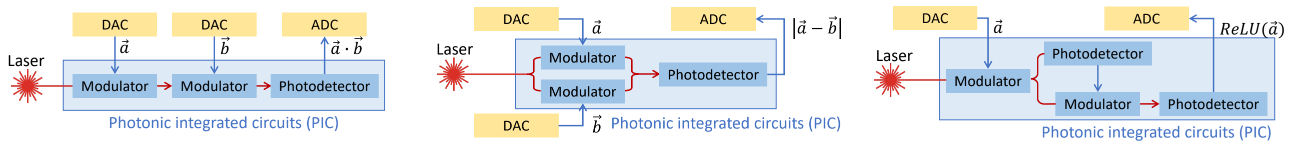
2 BACKGROUND AND MOTIVATION

This section first provides a background on photonic computing (§2.1) and then discusses the benefits of integrating photonic computing into optical transponders to perform complex operations inside the network (§2.2).

2.1 Background on Photonic Computing

Photonic computing provides a wide range of computing primitives, such as vector dot product, pattern matching, and non-linear functions. These computing primitives are achieved with optical devices commonly used in optical communications, like modulators [3] and photodetectors [4]. Modern silicon photonic fabrication enables these devices to be miniaturized and implemented on photonic integrated circuits (PIC) at millimeter scale [9]. In this section, we describe the most relevant photonic computing primitives for in-network computing applications.

P1 **Photonic vector dot product.** With the rapid growth of machine learning, there is a growing interest in achieving fast and energy-efficient vector dot product computations in the photonic domain [17, 50, 71, 72]. Figure 2a illustrates the main technique used to compute the dot product of vectors $\vec{a}.\vec{b}$ in the photonic domain. First, digital-to-analog converters (DACs) transform the digital signals $\vec{a}$ and $\vec{b}$ into analog voltages and apply these voltages to modulators, thereby generating a lightwave with intensity proportional to the analog voltage. Then, the two modulators connected back-to-back produce double-modulated lightwaves, achieving the multiplication of each element $a_i \times b_i$ within the analog domain. The resulting element-wise multiplication is then directed to a photodetector to perform summation $\Sigma_i a_i \times b_i$ [19]. Finally, an analog-to-digital converter (ADC) converts the result of the vector dot product back to digital.



(a) PIC for photonic vector dot product [50]. **(b) PIC for photonic pattern matching [59].** **(c) PIC for photonic nonlinear function [9].**

Figure 2: Photonic computing primitives implemented on a photonic integrated circuit (PIC).

P2 Photonic pattern matching. Another popular primitive in photonic computing is photonic pattern matching [6, 75]. Figure 2b depicts the principle of photonic pattern matching where two phase modulators are encoding the data $\vec{a}$ and target matching pattern $\vec{b}$ in parallel. An optical coupler then combines the phase-modulated light and leverages the wave's interference effect to decide if the data $\vec{a}$ matches the pattern $\vec{b}$. The interfered light intensity is received at the photodetector and converted to digital by an ADC.

P3 Photonic nonlinear function. Photonic nonlinear functions have been experimentally demonstrated using modulators and photodetectors [9, 33]. Figure 2c presents a photonic nonlinear function implementation: by configuring the operating point of the optical modulators in advance, the photodetector's output signal self-modulates its optical copy of the data, essentially creating a ReLU-like function entirely in the optical domain [9]. Although nonlinear functions do not require a huge volume of computations, introducing photonic nonlinear functions together with a photonic vector dot product enables all-optical deep neural network inference. For instance, prior work [9] presents a single-chip photonic deep neural network that integrates photonic vector dot product units and photonic nonlinear units, enabling the complete execution of deep neural network (DNN) inference computations exclusively within the optical domain.

2.2 Benefits of Photonic Computing

Photonic computing operates at significantly higher compute frequency than transistors [23, 26, 63]. Compared to the clock frequency of the state-of-the-art digital accelerators such as TPUs (approximately 1.05 GHz [28]) and GPUs (approximately 1.41 GHz [2]), photonic computing has the potential to improve the speed of computation by orders of magnitude. Moreover, prior work demonstrated the possibility of consuming only 40×10^{-18} Joules for an 8-bit multiply-and-accumulate operation [50]. Compared to the energy consumption of digital accelerators such as TPUs, which consume 7×10^{-14} Joules for an 8-bit multiplication operation, photonic computing can improve the energy efficiency of complex operations.

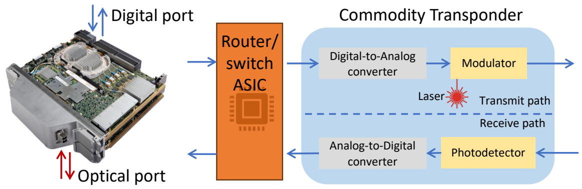


Figure 3: Simplified architecture of a commodity optical transponder in today's WAN [43].

The speed and energy efficiency of photonic computing highlight the advantages of offloading computation-intensive tasks from electronics to photonics. However, computing in the photonics domain relies on optical devices, such as lasers, modulators, photodetectors, DACs, and ADCs. None of these devices is embedded in today's accelerators , switches, or routers, but they are all part of optical transponders in fiber-optic communication wide-area networks (WANs) transmitting data across fiber links. As shown in Figure 3, an optical transponder is a device that converts electrical bits to optical signals to be carried on fiber optics networks and vice versa. The main components of a transponder are laser, optical modulator, photodetector, DAC, and ADC [43, 44].

Optical transponders have separate transmit and receive paths. The transmit path originates from electrical bits, which are then converted to analog voltages by the DAC. Subsequently, these analog voltages are encoded onto light waves using optical modulators. The receive path operates in the opposite direction, whereby incoming optical waves are first decoded into analog voltages through photodetectors and then converted into electrical bits by the ADC.

This paper proposes that network operators enhance optical communication transponders with photonic computing capabilities to perform on-fiber computing. On-fiber computing has several advantages. First, it improves application latency by performing computation inside the network. This property is similar to in-network computing efforts [31, 46, 52, 62] except that the computation is performed on the transponders plugged into routers instead of inside them. Moreover, on-fiber computing does not require replacing router ASICs, thus making it backward compatible for incremental deployment. Compared to conventional in-network computing proposals, on-fiber computing's pluggable transponders offer flexibility and performance improvements.

Second, since the data is already in the optical domain, on-fiber computing does not require constant digital-to-analog conversions, thus saving energy and chip area. In conventional photonic computing proposals, the data conversion between the digital and optical domains requires DACs and ADCs. By directly executing operations on the incoming optical signal, on-fiber photonic computing leverages devices already present in commodity transponders, leading to potentially substantial reductions in both chip area and power consumption.

3 ON-FIBER PHOTONIC COMPUTING

This section describes our proposed on-fiber photonic computing and its networking-related challenges. Computing operations are typically executed above the network stack, while the communication data are carried on fibers beneath the network stack. Connecting these two cross-layer functions is non-trivial, even though they may use the same physical devices. The key question is: "Is it even feasible to perform computation directly on optical data without converting it to the digital domain?" This section discusses three research challenges to enabling on-fiber photonic computing in today's fiber-optics WANs.

- **Data-plane hardware**: How can we augment today's optical transponder hardware to perform photonic computing directly on the incoming optical data?
- **Centralized controller**: What control logic can efficiently allocate computation tasks to photonic compute transponders?
- **Compute-communication protocol:** How can we design a compute-communication protocol to enable devices to distinguish between compute and non-compute packets, while ensuring all packets follow their correct paths in the network?

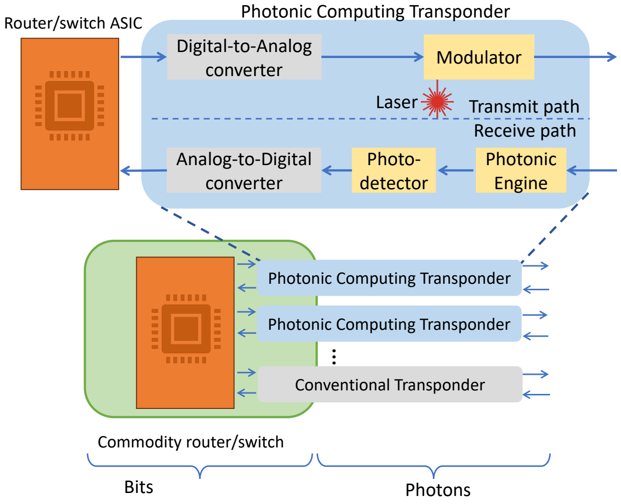

Router/switch ASIC
Photonic Computing Transponder
Digital-to-Analog converter
Modulator
Laser
Transmit path
Receive path
Analog-to-Digital converter
Photo-detector
Photonic Engine
Photonic Computing Transponder
Photonic Computing Transponder
Conventional Transponder
Commodity router/switch
Bits
Photons

Figure 4: Our proposed photonic computing system.

Photonic computing transponder. Figure 4 illustrates our proposed architecture for a photonic compute transponder (top) and showcases the on-fiber photonic computing process (bottom). In our proposed architecture, the transmit path aligns with that of the commodity transponders (as shown in Figure 3), and includes the following stages: ① digital data output from DSP ASIC → ② digital-to-analog converter (DAC) → ③ modulator → ④ optical data output into the fiber. In commodity transponders the receive path involves ① optical data input from the fiber → ② photodetector → ③ analog-to-digital converter (ADC) → ④ digital data input to the DSP ASIC. In contrast, our design augments the receive path with a *photonic engine*. The photonic engine is fabricated on the same chip as the transponder and performs all computations in the analog domain. This chip includes an optical preamble detection module to identify the arrival of a new packet and signal the start of the photonic computing process using techniques described in prior work on all-optical pattern matching [6, 75]. The photonic engine performs the appropriate computation tasks and inserts the results into a predetermined field in the packet header or payload. The output of the photonic engine is a series of lightwaves that we transmit directly to a commodity photodetector, mirroring the process employed by a conventional transponder. We envision that each transponder will contain several photonic computing primitives (§2.1) and perform several computing operations. Service providers will reconfigure each transponder according to the desired operation.

Centralized controller. In a real-world scenario, multiple end-users concurrently utilize on-fiber photonic computing services. As a result, there is a need for a centralized controller to continuously track the status of all photonic compute transponders and dynamically reconfigure them to accommodate a diverse set of photonic computing tasks according to users' demands. Inspired by centralized traffic engineering (TE) algorithms in WANs [10, 67], we propose to use an optimization formulation to allocate photonic computing resources based on demand. The optimization formulation takes user demands in terms of photonic computing task dependency graphs (e.g., a computation DAG) and network topology as input. It then takes the number of transponders at each node as resource constraints. The optimization objective is to satisfy as many compute demands as possible while minimizing the resource utilization of transponders.

Class	Use Case	Current Compute Location	Current Bottleneck(s)	Photonic Computing Primitives	System Design Requirements
C1	Machine learning inference [49, 52, 62, 70, 71]	Servers in the cloud	Poor latency	P1	Packet routing policy, compute graph analyzer, centralized controller
		End hosts / Edge devices	Limited computing resources		
		Routers / Switches / SmartNICs	Limited computing resources		
	Video encoding [45, 53]	End hosts / Edge devices	Limited computing resources	P1	In-network encoding algorithm, packet routing policy, end hosts software stack co-design
C2	IP routing [42]	Routers	Power hungry	P2	Photonic ternary matching hardware, router co-design
	Intrusion detection [69]	Servers in datacenters and SmartNICs	Computing resource hungry	P2	Photonic regular expression matching hardware, Electronic-Photonic co-design
	Data encryption [30, 57]	End hosts / Edge devices	Limited computing resources	P1	Photonic encryption hardware, trust model, in-network encryption algorithm
	Load balancing [58, 73]	Switches	Limited memory for precise load balancing due to replicating entries	P2	Photonic comparator hardware, switch co-design, photonic load balancing algorithm
	Massive MIMO baseband processing [24, 29]	Servers in datacenters	Computing resource hungry	P1 P3	Photonic signal processing hardware and algorithms

Table 1: On-fiber photonic computing use cases. (C1) User-facing applications; (C2) Network functions.

Compute-Communication protocol. In today’s WANs, when routers make individual next-hop routing decisions, they typically perform a lookup function on the destination IP address in a routing table to identify the corresponding entry. There is a need for a compute-communication protocol to guarantee that networking devices and end-hosts can distinguish between compute and non-compute packets. We propose enhancing the existing networking protocols and packet header formats to integrate the information on the essential photonic computing primitives a packet requires. Our additional photonic computing packet header is layered on top of the IP header to identify the photonic computing primitive ID. With this additional header information, routers perform next-hop lookup based on two fields: the destination IP address in the IP header and the photonic computing primitive ID specified in the photonic computing header. Since the centralized controller already maintains information on all photonic compute transponders, it serves as the vantage point from which to collect and combine the information from both IP routing and photonic compute routing, subsequently delivering next-hop updates to all routers.

4 USE CASES

This section describes various use cases of on-fiber computing. Table 1 provides a summary of these applications, including user-facing applications, such as machine learning inference [49, 52, 62, 70, 71] and video encoding [45, 53], in addition to a diverse range of network functions, such as IP routing [42], intrusion detection [69], data encryption [30, 57], load balancing [58, 73], and massive MIMO baseband processing [24, 29].

In user-facing applications, end-hosts initiate service requests and send the relevant data to a dedicated processing unit for computation. In contrast, network functions generally involve examining specific fields of the data packets and performing computation operations in the network. Typical computing locations for these applications include servers in the cloud, end-hosts (or edge devices), and in-network locations such as routers, switches, smartNICs, and middleboxes. For the cloud-based scenario, end hosts experience latency bottlenecks due to the packet processing delays at the cloud servers. Meanwhile, edge-based or in-network solutions can only support small models because of their limited computing resources.

On-fiber photonic computing addresses these challenges by performing the required computations while packets traverse the network, thereby achieving low-latency and energy-efficient operations. The fundamental photonic computing primitives include P1 Photonic vector dot product, P2 Photonic pattern matching, and P3 Photonic nonlinear function. There are two primary reasons for the energy savings. First, leveraging energy-efficient photonic devices and avoiding the usage of servers reduce the considerable energy costs associated with CPUs. Second, maintaining data within the photonic domain eliminates the need for costly digital-to-analog conversions.

To support the realization of diverse on-fiber photonic computing systems across a wide range of applications, several system design challenges arise. First, the system design necessitates the development of novel architectures and algorithms to execute computation tasks within the photonic domain. For instance, while prior work introduced the photonic vector dot product primitives, we still need new architecture designs for diverse DNN models and new algorithms to mitigate photonic noise during computation and achieve high accuracy. Second, it is important to formulate a novel packet routing and scheduling policy. In scenarios where multiple end-users demand access to the same photonic compute transponders, this new policy should mitigate congestion and achieve efficient load balancing. Third, these computing tasks require a centralized controller for efficient metadata allocation across the network. For instance, on-fiber machine learning inference requires trained DNN models to be distributed across network devices in advance.

5 DISCUSSION AND LIMITATIONS

Scalability. One advantage of our proposed architecture is that photonic compute transponders can support up to 800 Gbps network bandwidth on one wavelength [12]. This bandwidth can be shared among many users to support their on-fiber computing applications. However, the centralized controller introduces non-trivial scalability challenges. The optimization formulation is fundamentally an integer problem because it needs to decide which photonic computing transponder to use.

Security. Security is progressively gaining significance in modern WANs, with a substantial portion of user communication being end-to-end encrypted. Numerous prior works focus on enabling computation with encrypted data, encompassing a wide range of applications, such as machine learning inference [39], video analytics [41], search systems [18], and deep packet inspection [47]. Our current proposal does not prioritize security, but we believe incorporating the above mechanisms will facilitate performing photonic computing for encrypted optical data over fiber.

Form factor. Our proposed scheme necessitates incorporating supplementary components such as additional photonic components, digital memory, and digital control logic on the transponder, leading to increased chip area and power consumption of transponders. We leave an in-depth analysis of the chip area for future work.

Distributed on-fiber photonic computing. Our scheme suggests performing the required photonic computing task on a single transponder. If the computation task calls for a lot of resources and thus requires the coordination of multiple transponders, we need to deploy and execute the computation task in a distributed manner.

On-fiber photonic computing in datacenters. While this paper focuses on on-fiber photonic computing in WANs, we believe it is extendable to datacenter environments as well. The idea is to deploy the photonic compute transceivers in datacenter switches, similar to the architecture of photonic compute transponders in WANs. When datacenters serve computing request packets from end-users, these photonic compute transceivers undertake a portion of the required computations as packets traverse the datacenter network.

6 RELATED WORK

Analog computing. Charles Babbage laid the foundation of analog computers in the 1800s [8]. Over a century later, the US Navy used the Mechanical Analog Computer (MAC) for flight simulations to solve 4^{th} order differential equations [16]. Recently, there have been several proposals for electrical analog computing chips that use arrays of DACs to encode floating point vectors as electrical voltages and currents [7, 15, 32, 61]. These signals travel through a 2D array of components, such as flash memory, memristors, or transistors, before the result of a matrix-vector product is readout by an ADC. In contrast to these analog computing proposals, we propose distributing photonic computing capabilities across the WAN to execute the computations while data packets are on the fly.

Photonic computing. Photonic computing is an emerging field with the potential to perform fast and energy-efficient computation operations [13, 13, 17, 25, 35, 40, 50, 60, 64, 70]. However, state-of-the-art photonic computing work has focused on developing drop-in computing cores and ignored the impact of optical networks on their systems. Lightning [71] recently proposed a photonic computing datapath to enable machine learning inference on smartNICs. Similarly, IOI [70] proposed processing inference queries using photonic devices inside the network. Yet these approaches still require constant digital-to-analog and analog-to-digital conversions. In this paper, we propose preserving the data in the optical layer during computation and eliminating the extra cost of data conversions.

Active networking. Active network is a network architecture where switches or routers can perform customized computations on packets traverse through them [54]. Early efforts like Capsule [55] and SwitchWare [5] demonstrated the benefits of programmable packet processing within the network datapath. However, there have been security concerns about the radical approach of active networks. Our proposal of on-fiber photonic computing is similar in spirit to active networks, but it allows computing in the physical layer in the optical format without the need to read the packet data. If combined with advanced techniques like homomorphic encryption, our proposal can enable computing on the encrypted optical data, respecting the end-to-end security of the network link.

In-network computing. With the rise of Software-Defined Networking (SDN) [36] and programmable switches [11], many efforts have been made to perform in-network computation on programmable dataplane devices, such as routers and switches [27, 31, 37, 46, 52, 56, 62, 65, 66], smartNICs [49], and FPGAs [21]. However, existing in-network computing paradigms perform digital computation within the network stack using electrical devices. Our vision is to leverage photonic computing and push the computation down to the physical layer beneath the network stack.

7 CONCLUSION

Photonic computing is a powerful technology to perform fast and energy-efficient computation in the analog domain. This paper argues for a paradigm shift wherein the network performs photonic computing while the data is *on fiber*. Our proposal leverages the fact that today's networks already convert digital data to photons using commodity transponders. We discuss the hardware components, centralized controller, and compute-communication protocols required to enable the vision of on-fiber photonic computing.

ACKNOWLEDGMENTS

We thank our shepherd, Paolo Costa, and anonymous HotNets reviewers for their feedback. This work was supported by ARPA-E ENLITENED PINE DE-AR0000843, DARPA FastNICs 4202290027, NSF CNS-2008624, NSF SHF-2107244, NSF CAREER-2144766, NSF PPoSS-2217099, NSF CNS-2211382, NSF FuSe-TG-2235466, Juniper Networks Sponsored University Research Initiative (SURI) award, Sloan fellowship FG-2022-18504, Microsoft PhD fellowship, ACE and CUbiC centers sponsored by SRC and DARPA under the JUMP 2.0 program. This research was also partially sponsored by the United States Air Force Research Laboratory and the Department of the Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of the Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. The nature of this research is not military-related and does not have direct military implications.

REFERENCES

[1] [n. d.]. Tofino Expandable Architecture to Meet 10Tbps-Level Bandwidth Requirements. ([n. d.]). https://www.intel.com/content/www/us/en/products/docs/programmable/baidu-tofino-xa-white-paper.html.

[2] 2021. Nvidia A100 GPU. (2021). https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf.

[3] 2022. 10 GHz Intensity Modulator. (2022). https://www.thorlabs.com/thorproduct.cfm?partnumber=LN81S-FC.

[4] 2022. Thorlabs InGaAs Fixed Gain Amplified Detector, 750 - 1650 nm, DC - 9.5 GHz . (2022). https://www.thorlabs.com/thorproduct.cfm?partnumber=PDA8GS.

[5] D Scott Alexander, William A Arbaugh, Michael W Hicks, Pankaj Kakkar, Angelos D Keromytis, Jonathan T Moore, Carl A Gunter, Scott M Nettles, and Jonathan M Smith. 1998. The SwitchWare active network architecture. *IEEE network* 12, 3 (1998), 29–36.

[6] Fatemeh Alishahi, Kaiheng Zou, Amir Minoofar, Huibin Zhou, Moshe Tur, Jonathan L Habif, and Alan E Willner. 2021. Demonstration of a tunable optical correlation of a 10–15 Gbaud QPSK data signal using nonlinear wave mixing at a remotely controlled node. In *2021 IEEE Photonics Conference (IPC)*. IEEE, 1–2.

[7] Tanner Andrulis, Joel S. Emer, and Vivienne Sze. 2023. RAELLA: Reforming the Arithmetic for Efficient, Low-Resolution, and Low-Loss Analog PIM: No Retraining Required!. In *Proceedings of the 50th Annual International Symposium on Computer Architecture (ISCA '23)*. Association for Computing Machinery, New York, NY, USA, Article 27, 16 pages. https://doi.org/10.1145/3579371.3589062

[8] William Aspray and Michael S. Mahoney. 1991. Computing Before Computers. *Physics Today* 44, 5 (05 1991), 64–65. https://doi.org/10.1063/1.2810115 arXiv:https://pubs.aip.org/physicstoday/article-pdf/44/5/64/8303931/64_2_online.pdf

[9] Saumil Bandyopadhyay, Alexander Sludds, Stefan Krastanov, Ryan Hamerly, Nicholas Harris, Darius Bunandar, Matthew Streshinsky, Michael Hochberg, and Dirk Englund. 2022. Single chip photonic deep neural network with accelerated training. *arXiv preprint arXiv:2208.01623* (2022).

[10] Dhritiman Banerjee and Biswanath Mukherjee. 2000. Wavelength-routed optical networks: Linear formulation, resource budgeting trade-offs, and a reconfiguration study. *IEEE/ACM Transactions on networking* 8, 5 (2000), 598–607.

[11] Pat Bosshart, Glen Gibb, Hun-Seok Kim, George Varghese, Nick McKeown, Martin Izzard, Fernando Mujica, and Mark Horowitz. 2013. Forwarding metamorphosis: Fast programmable match-action processing in hardware for SDN. *ACM SIGCOMM Computer Communication Review* 43, 4 (2013), 99–110.

[12] Di Che. 2022. Ultra-low-complexity map demapper for bandwidth-limited pluggable coherent optics beyond 800g. In *2022 Optical Fiber Communications Conference and Exhibition (OFC)*. IEEE, 01–03.

[13] Qixiang Cheng, Jihye Kwon, Madeleine Glick, Meisam Bahadori, Luca P Carloni, and Keren Bergman. 2020. Silicon photonics codesign for deep learning. *Proc. IEEE* 108, 8 (2020), 1261–1282.

[14] Sharad Chole, Andy Fingerhut, Sha Ma, Anirudh Sivaraman, Shay Vargaftik, Alon Berger, Gal Mendelson, Mohammad Alizadeh, Shang-Tse Chuang, Isaac Keslassy, Ariel Orda, and Tom Edsall. 2017. dRMT: Disaggregated Programmable Switching. In *ACM SIGCOMM*.

[15] Chaoqun Chu, Yanzhi Wang, Yilong Zhao, Xiaolong Ma, Shaokai Ye, Yunyan Hong, Xiaoyao Liang, Yinhe Han, and Li Jiang. 2020. PIM-prune: Fine-grain DCNN pruning for crossbar-based process-in-memory architecture. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 1–6.

[16] A.B. Clymer. 1993. The mechanical analog computers of Hannibal Ford and William Newell. *IEEE Annals of the History of Computing* 15, 2 (1993), 19–34. https://doi.org/10.1109/85.207741

[17] Devin Coldewey. 2023. Lightmatter's photonic AI hardware is ready to shine with $154M in new funding. (May 2023). https://techcrunch.com/2023/05/31/lightmatters-photonic-ai-hardware-is-ready-to-shine-with-154m-in-new-funding/.

[18] Emma Dauterman, Eric Feng, Ellen Luo, Raluca Ada Popa, and Ion Stoica. 2020. DORY: An encrypted search system with distributed trust. In *Proceedings of the 14th USENIX Conference on Operating Systems Design and Implementation*. 1101–1119.

[19] Johannes Feldmann, Nathan Youngblood, Maxim Karpov, Helge Gehring, Xuan Li, Maik Stappers, Manuel Le Gallo, Xin Fu, Anton Lukashchuk, Arslan Sajid Raja, et al. 2021. Parallel convolutional processing using an integrated photonic tensor core. *Nature* 589, 7840 (2021), 52–58.

[20] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja, J. Liu, C. D. Wright, A. Sebastian, T. J. Kippenberg, W. H. P. Pernice, and H. Bhaskaran. 2021. Parallel convolutional processing using an integrated photonic tensor core. *Nature* 589, 7840 (2021), 52–58. https://doi.org/10.1038/s41586-020-03070-1

[21] N. Gebara, P. Costa, and M. Ghobadi. 2021. PANAMA: In-network Aggregation for Shared Machine Learning Clusters. In *Proc. Conference on Machine Learning and Systems (MLSys)*. 1–16.

[22] Nadeen Gebara, Alberto Lerner, Mingran Yang, Minlan Yu, Paolo Costa, and Manya Ghobadi. 2020. Challenging the stateless quo of programmable switches. In *Proceedings of the 19th ACM Workshop on Hot Topics in Networks*. 153–159.

[23] Heedong Goh and Andrea Alù. 2022. Nonlocal Scatterer for Compact Wave-Based Analog Computing. *Phys. Rev. Lett.* 128 (Feb 2022), 073201. Issue 7. https://doi.org/10.1103/PhysRevLett.128.073201

[24] Junzhi Gong, Anuj Kalia, and Minlan Yu. 2023. Scalable Distributed Massive {MIMO} Baseband Processing. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. 405–417.

[25] Ryan Hamerly, Liane Bernstein, Alexander Sludds, Marin Soljačić, and Dirk Englund. 2019. Large-scale optical neural networks based on photoelectric multiplication. *Physical Review X* 9, 2 (2019), 021032.

[26] Philip Jacobson, Mizuki Shirao, Kerry Yu, Guan-Lin Su, and Ming C. Wu. 2022. Hybrid Convolutional Optoelectronic Reservoir Computing for Image Recognition. *Journal of Lightwave Technology* 40, 3 (2022), 692–699. https://doi.org/10.1109/JLT.2021.3124520

[27] Xin Jin, Xiaozhou Li, Haoyu Zhang, Robert Soulé, Jeongkeun Lee, Nate Foster, Changhoon Kim, and Ion Stoica. 2017. NetCache: Balancing Key-Value Stores with Fast In-Network Caching. In *Proceedings of the 26th Symposium on Operating Systems Principles (SOSP '17)*.

[28] Norman P. Jouppi, Doe Hyun Yoon, Matthew Ashcraft, Mark Gottscho, Thomas B. Jablin, George Kurian, James Laudon, Sheng Li, Peter Ma, Xiaoyu Ma, Thomas Norrie, Nishant Patil, Sushma Prasad, Cliff Young, Zongwei Zhou, and David Patterson. 2021. Ten Lessons From Three Generations Shaped Google's TPUv4i : Industrial Product. In *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. 1–14. https://doi.org/10.1109/ISCA52012.2021.00010

[29] Minsung Kim, Davide Venturelli, and Kyle Jamieson. 2019. Leveraging quantum annealing for large MIMO processing in centralized radio access networks. In *Proceedings of the ACM special interest group on data communication*. 241–255.

[30] Sam Kumar, Yuncong Hu, Michael P Andersen, Raluca Ada Popa, and David E Culler. 2019. JEDI: many-to-many end-to-end encryption and key delegation for IoT. In *Proceedings of the 28th USENIX Conference on Security Symposium*. 1519–1536.

[31] ChonLam Lao, Yanfang Le, Kshiteej Mahajan, Yixi Chen, Wenfei Wu, Aditya Akella, and Michael Swift. 2021. ATP: In-network Aggregation for Multi-tenant Learning. In *18th USENIX NSDI 21*. USENIX Association, 741–761.

[32] Can Li, Zhongrui Wang, Mingyi Rao, Daniel Belkin, Wenhao Song, Hao Jiang, Peng Yan, Yunning Li, Peng Lin, Miao Hu, et al. 2019. Long short-term memory networks in memristor crossbar arrays. *Nature Machine Intelligence* 1, 1 (2019), 49–57.

[33] Gordon HY Li, Ryoto Sekine, Rajveer Nehra, Robert M Gray, Luis Ledezma, Qiushi Guo, and Alireza Marandi. 2022. All-optical ultrafast ReLU function for energy-efficient nanophotonic deep learning. *Nanophotonics* (2022).

[34] Weilin Liu, Ming Li, Robert S Guzzon, Erik J Norberg, John S Parker, Mingzhi Lu, Larry A Coldren, and Jianping Yao. 2016. A fully reconfigurable photonic integrated signal processor. *Nature Photonics* 10, 3 (2016), 190–195.

[35] Weichen Liu, Wenyang Liu, Yichen Ye, Qian Lou, Yiyuan Xie, and Lei Jiang. 2019. Holylight: A nanophotonic accelerator for deep learning in data centers. In *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 1483–1488.

[36] Nick McKeown, Tom Anderson, Hari Balakrishnan, Guru Parulkar, Larry Peterson, Jennifer Rexford, Scott Shenker, and Jonathan Turner. 2008. OpenFlow: enabling innovation in campus networks. *ACM SIGCOMM computer communication review* 38, 2 (2008), 69–74.

[37] Rui Miao, Hongyi Zeng, Changhoon Kim, Jeongkeun Lee, and Minlan Yu. 2017. SilkRoad: Making Stateful Layer-4 Load Balancing Fast and Cheap Using Switching ASICs. In *Proceedings of the 2017 ACM SIGCOMM Conference (SIGCOMM '17)*.

[38] Microsoft. 2023. Project AIM (Analog Iterative Machine). (2023). https://www.microsoft.com/en-us/research/project/aim/.

[39] Pratyush Mishra, Ryan Lehmkuhl, Akshayaram Srinivasan, Wenting Zheng, and Raluca Ada Popa. 2020. DELPHI: a cryptographic inference service for neural networks. In *Proceedings of the 29th USENIX Conference on Security Symposium*. 2505–2522.

[40] Jiaxin Peng, Yousra Alkabani, Shuai Sun, Volker J Sorger, and Tarek El-Ghazawi. 2020. Dnnara: A deep neural network accelerator using residue arithmetic and integrated photonics. In *Proceedings of the 49th International Conference on Parallel Processing*. 1–11.

[41] Rishabh Poddar, Ganesh Ananthanarayanan, Srinath Setty, Stavros Volos, and Raluca Ada Popa. 2020. Visor: Privacy-preserving video analytics as a cloud service. In *Proceedings of the 29th USENIX Conference on Security Symposium*. 1039–1056.

[42] Yakov Rekhter, Jon Crowcroft, and Brian E. Carpenter. 1997. IPv4 Address Behaviour Today. RFC 2101. (Feb. 1997). https://doi.org/10.17487/RFC2101

[43] Kim Roberts, Douglas Beckett, David Boertjes, Joseph Berthold, and Charles Laperle. 2010. 100G and beyond with digital coherent signal processing. *IEEE Communications Magazine* 48, 7 (2010), 62–69.

[44] Kim Roberts, Qunbi Zhuge, Inder Monga, Sebastien Gareau, and Charles Laperle. 2017. Beyond 100 Gb/s: capacity, flexibility, and network optimization. *Journal of Optical Communications and Networking* 9, 4 (2017), C12–C24.

[45] Michael Rudow, Francis Y Yan, Abhishek Kumar, Ganesh Ananthanarayanan, Martin Ellis, and KV Rashmi. 2023. Tambur: Efficient loss recovery for videoconferencing via streaming codes. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. 953–971.

[46] Amedeo Sapio, Marco Canini, Chen-Yu Ho, Jacob Nelson, Panos Kalnis, Changhoon Kim, Arvind Krishnamurthy, Masoud Moshref, Dan Ports, and Peter Richtarik. 2021. Scaling Distributed Machine Learning with In-Network Aggregation. In *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*. USENIX Association,

785–808. https://www.usenix.org/conference/nsdi21/presentation/sapio

[47] Justine Sherry, Chang Lan, Raluca Ada Popa, and Sylvia Ratnasamy. 2015. Blindbox: Deep packet inspection over encrypted traffic. In *Proceedings of the 2015 ACM conference on special interest group on data communication*. 213–226.

[48] Rachee Singh, Manya Ghobadi, Klaus-Tycho Foerster, Mark Filer, and Phillipa Gill. 2018. RADWAN: rate adaptive wide area network. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*. 547–560.

[49] Giuseppe Siracusano, Salvator Galea, Davide Sanvito, Mohammad Malekzadeh, Gianni Antichi, Paolo Costa, Hamed Haddadi, and Roberto Bifulco. 2022. Re-architecting traffic analysis with neural network interface cards. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*. 513–533.

[50] Alexander Sludds, Saumil Bandyopadhyay, Zaijun Chen, Zhizhen Zhong, Jared Cochrane, Liane Bernstein, Darius Bunandar, P Ben Dixon, Scott Hamilton, Matthew Streshinsky, Ari Novack, Tom Baehr-Jones, Michael Hochberg, Manya Ghobadi, Ryan Hamerly, and Dirk Englund. 2022. Delocalized Photonic Deep Learning on the Internet's Edge. *Science* 378, 6617 (2022), 270–276. https://doi.org/10.1126/science.abq8271

[51] Alexander Sludds, Ryan Hamerly, Saumil Bandyopadhyay, Zhizhen Zhong, Zaijun Chen, Liane Bernstein, Manya Ghobadi, and Dirk Englund. 2022. Demonstration of WDM-Enabled Ultralow-Energy Photonic Edge Computing, In Optical Fiber Communication Conference (OFC) 2022. *Optical Fiber Communication Conference (OFC) 2022*, Th3A.3. https://doi.org/10.1364/OFC.2022.Th3A.3

[52] Tushar Swamy, Alexander Rucker, Muhammad Shahbaz, Ishan Gaur, and Kunle Olukotun. 2022. Taurus: a data plane architecture for per-packet ML. In *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*. 1099–1114.

[53] Vivienne Sze, Madhukar Budagavi, and Gary J Sullivan. [n. d.]. High efficiency video coding (HEVC). Springer.

[54] David L Tennenhouse, Jonathan M Smith, W David Sincoskie, David J Wetherall, and Gary J Minden. 1997. A survey of active network research. *IEEE communications Magazine* 35, 1 (1997), 80–86.

[55] David L Tennenhouse and David J Wetherall. 1996. Towards an active network architecture. *ACM SIGCOMM Computer Communication Review* 26, 2 (1996), 5–17.

[56] Muhammad Tirmazi, Ran Ben Basat, Jiaqi Gao, and Minlan Yu. 2020. Cheetah: Accelerating Database Queries with Switch Pruning. In *Proceedings of the 2020 ACM SIGMOD Conference (SIGMOD '20)*.

[57] Marten Van Dijk, Craig Gentry, Shai Halevi, and Vinod Vaikuntanathan. 2010. Fully homomorphic encryption over the integers. In *Advances in Cryptology–EUROCRYPT 2010: 29th Annual International Conference on the Theory and Applications of Cryptographic Techniques, French Riviera, May 30–June 3, 2010. Proceedings 29*. Springer, 24–43.

[58] Erico Vanini, Rong Pan, Mohammad Alizadeh, Parvin Taheri, and Tom Edsall. 2017. Let it flow: Resilient asymmetric load balancing with flowlet switching.. In *NSDI*, Vol. 17. 407–420.

[59] LM Walpita, SC Wang, and WSC Chang. 1988. Integrated-optic Mach-Zehnder rf phase comparator. *Applied optics* 27, 18 (1988), 3772–3773.

[60] Tianyu Wang, Shi-Yuan Ma, Logan G Wright, Tatsuhiro Onodera, Brian C Richard, and Peter L McMahon. 2022. An optical neural network using less than 1 photon per multiplication. *Nature Communications* 13, 1 (2022), 123.

[61] Qiangfei Xia and J Joshua Yang. 2019. Memristive crossbar arrays for brain-inspired computing. *Nature materials* 18, 4 (2019), 309–323.

[62] Zhaoqi Xiong and Noa Zilberman. 2019. Do switches dream of machine learning? toward in-network classification. In *Proceedings of the 18th ACM workshop on hot topics in networks*. 25–33.

[63] Xingyuan Xu, Mengxi Tan, Bill Corcoran, Jiayang Wu, Andreas Boes, Thach G Nguyen, Sai T Chu, Brent E Little, Damien G Hicks, Roberto Morandotti, et al. 2021. 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature* 589, 7840 (2021), 44–51.

[64] Javier Yanes. 2020. Optical Computing: Solving Problems at the Speed of Light. (Feb. 2020). https://www.bbvaopenmind.com/en/technology/future/optical-computing-solving-problems-at-the-speed-of-light/.

[65] Mingran Yang, Alex Baban, Valery Kugel, Jeff Libby, Scott Mackie, Swamy Sadashivaiah Renu Kananda, Chang-Hong Wu, and Manya Ghobadi. 2022. Using trio: juniper networks' programmable chipset-for emerging in-network applications. In *Proceedings of the ACM SIGCOMM 2022 Conference*. 633–648.

[66] Yifan Yuan, Omar Alama, Jiawei Fei, Jacob Nelson, Dan RK Ports, Amedeo Sapio, Marco Canini, and Nam Sung Kim. 2022. Unlocking the Power of Inline {Floating-Point} Operations on Programmable Switches. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*. 683–700.

[67] Hui Zang, Jason P Jue, Biswanath Mukherjee, et al. 2000. A review of routing and wavelength assignment approaches for wavelength-routed optical WDM networks. *Optical networks magazine* 1, 1 (2000), 47–60.

[68] Yi Zhang, Yadong Wang, Yunyun Dai, Xueyin Bai, Xuerong Hu, Luojun Du, Hai Hu, Xiaoxia Yang, Diao Li, Qing Dai, et al. 2022. Chirality logic gates. *Science Advances* 8, 49 (2022), eabq8246.

[69] Zhipeng Zhao, Hugo Sadok, Nirav Atre, James C Hoe, Vyas Sekar, and Justine Sherry. 2020. Achieving 100gbps intrusion prevention on a single server. In *Proceedings of the 14th USENIX Conference on Operating Systems Design and Implementation*. 1083–1100.

[70] Zhizhen Zhong, Weiyang Wang, Manya Ghobadi, Alexander Sludds, Ryan Hamerly, Liane Bernstein, and Dirk Englund. 2021. IOI: In-network Optical Inference. In *Proceedings of the ACM SIGCOMM 2021 Workshop on Optical Systems*. 18–22.

[71] Zhizhen Zhong, Mingran Yang, Jay Lang, Christian Williams, Liam Kronman, Alexander Sludds, Homa Esfahanizadeh, Dirk Englund, and Manya Ghobadi. 2023. Lightning: A Reconfigurable Photonic-Electronic SmartNIC for Fast and Energy-Efficient Inference. In *ACM SIGCOMM 2023 Conference*.

[72] Hailong Zhou, Jianji Dong, Junwei Cheng, Wenchan Dong, Chaoran Huang, Yichen Shen, Qiming Zhang, Min Gu, Chao Qian, Hongsheng Chen, et al. 2022. Photonic matrix multiplication lights up photonic accelerator and beyond. *Light: Science & Applications* 11, 1 (2022), 30.

[73] Junlan Zhou, Malveeka Tewari, Min Zhu, Abdul Kabbani, Leon Poutievski, Arjun Singh, and Amin Vahdat. 2014. WCMP: Weighted Cost Multipathing for Improved Fairness in Data Centers. Article No. 5. https://dl.acm.org/doi/10.1145/2592798.2592803

[74] Tiankuang Zhou, Wei Wu, Jinzhi Zhang, Shaoliang Yu, and Lu Fang. 2023. Ultrafast dynamic machine vision with spatiotemporal photonic computing. *Science Advances* 9, 23 (2023), eadg4391.

[75] Morteza Ziyadi, Mohammad Reza Chitgarha, Salman Khaleghi, Amirhossein Mohajerin-Ariaei, Ahmed Almaiman, Joe Touch, Moshe Tur, Carsten Langrock, Martin M Fejer, and Alan E Willner. 2014. Tunable optical correlator using an optical frequency comb and a nonlinear multiplexer. *Optics Express* 22, 1 (2014), 84–89.