Can Photonic Interconnects be used for High-Throughput Memory Access in FHE Accelerators?

Dewan Saiham, Mariam Rabadi, Di Wu, and Sazadur Rahman
Department of Electrical and Computer Engineering, University of Central Florida
{dewan.saiham, mariam.rabadi, di.wu, mohammad.rahman}@ucf.edu

***Abstract*—Fully Homomorphic Encryption (FHE) allows computations over encrypted data without sacrificing confidentiality, but its practicality is hindered by high computational demands and memory access constraints. While existing FHE accelerators focus on improving computational efficiency, they are often limited by the insufficient memory bandwidth and inefficient data transfer schemes, leading to significant bottlenecks, especially for processing large amounts of data. In this work, we evaluate whether *OptoLink*, a photonic interconnect architecture, is scalable and capable of providing high bandwidth to overcome these limitations. Leveraging Wavelength Division Multiplexing (WDM) with Space Division Multiplexing (SDM), *OptoLink* achieves an impressive bandwidth of 1.6 TB/s over 128 channels—a 300x improvement over traditional electronic network. Additionally, its ability to efficiently broadcast data and support parallel processing further enhances performance. The broadcasting capability not only enables parallelism but also reduces power consumption in earlier NTT stages, improving overall energy efficiency. With its improved data throughput, scalability, and lower latency, *OptoLink* offers a robust solution capable of satisfying the high data transfer and memory demands of current FHE accelerators.**

***Index Terms*—Fully Homomorphic Encryption, Number Theoretic Transform, Wavelength Division Multiplexing, Memory Acceleration**

I. Introduction

Fully Homomorphic Encryption (FHE) represents a substantial breakthrough in privacy-preserving computing, enabling users to perform calculations on encrypted data without requiring to decrypt it. Secure data processing technology remains critical for applications operating within potentially untrusted environments including cloud computing, financial and healthcare systems, which demand the protection of sensitive data during computations [1–3]. As illustrated in Fig. 1, FHE enables secure computation offloading through data encryption before sending it to a server for processing and receiving the processed result in encrypted form. This process ensures that even if the server is compromised, the sensitive data remains protected throughout the computation because the decryption key is never shared with the server, maintaining the data confidentiality. Large integer and polynomial multiplications, which are fundamental to FHE operations across both integer-based and ring learning with errors (R-LWE) based schemes, are particularly computationally demanding operations that determine how efficient FHE schemes are [4, 5]. Within these schemes, the Number Theoretic Transform (NTT) plays a crucial role in modular polynomial multiplication, accounting for a substantial portion of the computational resources required throughout the FHE process. For example, it represents 51% of the execution time for ciphertext multiplication and 55% of the inference time in homomorphic encryption-based models such as HE ResNet-50 [6, 7]. While NTT reduces the asymptotic complexity of polynomial multiplication from $O(n^2)$ to $O(n \, log \, n)$, where n is the degree of the polynomial, it also introduces challenges in terms of high memory bandwidth and complex access patterns, especially for hardware acceleration [8, 9]. Efforts to accelerate the NTT through various platforms, including FPGA, ASIC, and Compute-in-Memory architectures, have shown promise but remain limited in terms of overall acceleration ratios [10–12]. Resolving these hardware issues is crucial to increasing the efficiency and practicality of FHE in a variety of security-sensitive applications [1–3].

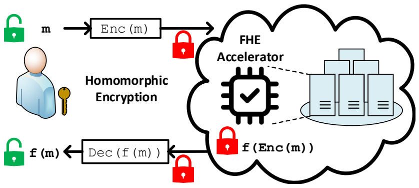

Fig. 1. Computational flow in fully holomorphic encryption (FHE).

Challenge. Parallel and pipelined NTT architectures have been developed to improve the computational efficiency of FHE for specific security parameters [13]. Although these optimizations accelerate computations, they often lack flexibility, which makes them less adaptable across different security levels and hardware configurations. One of the biggest challenges for large-scale NTT accelerator design is to deal with data flow efficiently because of the complicated memory access patterns [14]. NTT computation require continuous movement of polynomial coefficients and twiddle factors from RAM, which induces memory access conflicts, particularly read-after-write conflicts among RAMs and processing elements (PEs) [14, 15]. Pipeline stalls have been used to mitigate these conflicts [16], but they come at the cost of reduced system performance. High-security FHE parameters still impose significant bandwidth demands on memory and interconnects [17]. Current limitations of electronic networks in meeting these demands point to the necessity of novel interconnect solutions that can provide the high throughput requirements for scalable FHE acceleration.

Proposal. Photonic interconnects can be a promising alternative to conventional electronic networks, effectively addressing key issues of FHE acceleration. Unlike electronic networks that rely on resource-intensive multiplexer (MUX) connections in conventional NTT designs, photonic links provide direct, conflict-free data paths, reducing circuit complexity and alleviating memory bandwidth bottlenecks [15]. Furthermore, photonic interconnects exhibit efficient scalability via wavelength-division multiplexing (WDM) and space-division multiplexing (SDM), facilitating one-to-many communication highly suited to high-data-rate transmission [18]. These advantages have been demonstrated in DNN accelerators [19], yet their promise has not been well examined for FHE. In this paper, we seek to determine whether photonic interconnects can indeed alleviate the memory bandwidth limitations in FHE accelerators. That is, we seek to investigate the following research questions.

RQ1: Can photonic interconnects overcome the memory bandwidth limitations of conventional electronic networks in FHE computation?

RQ2: Are photonic interconnects scalable to support complex memory acess patterns due to various FHE security levels and parameters?

RQ3: How do photonic interconnects stand against electronic networks in terms of latency, power consumption and area efficiency?

By evaluating these parameters methodically, this study examines the feasibility of photonic interconnects as a promising option for large-

979-8-3315-2710-5/25/$31.00 © 2025 IEEE

scale FHE computations. To answer the question raised above, this paper introduces *OptoLink*, a photonic interconnect architecture, to overcome memory bandwidth limitations in FHE accelerators. Our key contributions are highlighted below.

1) Unlike prior works focused only on compute acceleration [14], we identify memory bandwidth as the key bottleneck in FHE and show that compute speedup alone is insufficient.
2) We performed a comprehensive analysis and comparison of photonic and electronic interconnects for FHE use. We are the first, to the best of our knowledge, to consider photonic interconnects for this application and provide an end-to-end analysis of their potential benefits and trade-offs.
3) We propose *OptoLink*, an optical interconnect architecture tailored for FHE, with significantly reduced memory access contention and improved bandwidth for NTT operations. Our architecture is scalable and provides high data rates at reduced power consumption in earlier NTT stages (Sec. III).
4) Using photonics process design kits (PDKs) in combination with electronic-photonic design automation (EPDA) software such as `Synopsys OptSim` and `OptoCompiler`, we develop a scalable $OptoLink$ design for several NTT core designs. As seen from simulations, $OptoLink$ is able to support up to 1.6 TB/s bandwidth using 128 optical channels, with potential to offer even more throughput (Sec.III-E, Sec.IV).

The rest of the paper is structured as follows: Sec.II discusses background, existing limitations, and the motivation behind $OptoLink$. Sec.III details our design and implementation. Sec.IV presents results and analysis, followed by the conclusion in Sec.V.

II. Background and Motivation

In this section we provide a brief description of NTT operation in FHE, existing FHE accelerators, and their limitations.

A. Number Theoretic Transform(NTT)

NTT is a version of the Fast Fourier Transform (FFT) that has been optimized for integer polynomial operations and finite fields, which makes it particularly suitable for cryptographic applications that require exact arithmetic, like lattice-based cryptography. The NTT is taken over a ring, $R_q = \mathbb{Z}_q[x]/(x^n+1)$, with prime modulus q where $q \equiv 1 \mod n$. This guarantees the existence of a primitive n-th root of unity, denoted by ω, such that $\omega^n \equiv 1 \mod q$. The NTT transforms a polynomial $a(x) = \sum_{i=0}^{n-1} a_i x^i$ into a new polynomial representation $\tilde{a}(x)$ using the formula,

$$\tilde{a}_i = \sum_{j=0}^{n-1} a_j \omega^{i \cdot j} \mod q, \quad \text{for } i = 0, 1, \ldots, n-1 \tag{1}$$

where $\omega^{i \cdot j}$ terms are referred to as twiddle factors. These twiddle factors are associated with the powers of the root of unity ω, enabling the NTT to perform convolution over polynomial coefficients directly in the NTT domain. Polynomial multiplication $c(x) = a(x) \cdot b(x)$ can be done in the NTT domain by transforming a and b into their respective NTT representations, multiplying point-wise, and then taking the inverse NTT (INTT) to get back the result in the original domain.

$$c = \text{INTT}(\text{NTT}(a) \circ \text{NTT}(b)), \tag{2}$$

where $\circ$ is the pointwise multiplication of NTT-transformed coefficients. The INTT, which recovers the result from the NTT domain can be represented by,

$$a_j = \frac{1}{n} \sum_{i=0}^{n-1} \tilde{a}_i \cdot \omega^{-i \cdot j} \mod q \tag{3}$$

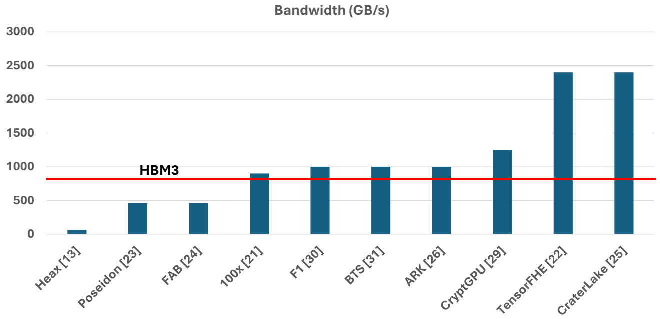

Fig. 2. Bandwidth requirements of state-of-the-art FHE accelerators.

where $\omega^{-i \cdot j}$ is the inverse twiddle factor, scaling by $n^{-1} \mod q$ completes the transformation.

B. Existing FHE Accelerators

Ever since its introduction in 2009 [20], FHE has made significant progress in reducing its initial computational overhead of being $10^9\times$ slower than unencrypted computation. It remains, however, $10,000\times$ to $100,000\times$ slower than conventional computing, which justifies the need for special-purpose hardware accelerators [10]. GPUs, with their parallel processing nature, have enhanced FHE performance by as much as $257\times$ speedups compared to CPUs [21]. Open-source libraries such as cuHE and cuFHE further optimize GPU-based FHE, whereas TensorFHE has shown a $1625.6\times$ speedup compared to CPUs and a $2.9\times$ improvement compared to F1+, comparable to ASIC accelerators [22]. Nevertheless, GPUs are not optimized for FHE and thus consume a lot of power and are inefficient in memory-heavy operations. FPGAs offer greater adaptability for custom FHE implementations like NTT, with solutions such as HEAX and Poseidon achieving over $1000\times$ speedups compared to GPUs [23]. Designs like FAB further enhance FHE acceleration through efficient resource management [24]. ASIC accelerators, tailored to FHE schemes like CKKS and BFV, achieve even superior performance. Bootstrapping hardware and data management optimizations, as in CraterLake [11] and ARK [25], enable deeper computations and reduce bottlenecks, offering orders-of-magnitude improvement over GPUs. ASICs are, however, plagued by large chip area, high power, and enormous memory requirements, rendering them hard to deploy in reality.

C. Limitations of FHE Acceleration Trends

Data Inflation: Memory bandwidth remains a key bottleneck in FHE applications [26], especially in CKKS, as ciphertexts are significantly larger than plaintexts. The inflation causes memory accesses frequently, which computation-oriented optimizations cannot mitigate. **For instance, a chip with** $40,960$ **modular multiplication units at** 2**GHz and** 3**TB/s HBM3 completes computations in** 0.18**ms, but data loading takes** 2.1**ms [25].**

Memory Bandwidth: Irregular memory access patterns also aggravate bandwidth limitations. NTT computations require the storage of massive twiddle factors and intermediate results, which tend to be larger than on-chip caches and cause costly off-chip memory access. Although 4-step FFT/NTT enhance parallelism, they also introduce additional twisting factors, which add memory overhead [27]. Furthermore, the $(n \log n)/2$ butterflies in FFT/NTT pipelines demand substantial hardware resources as n grows, creating dynamic dependencies that static hardware cannot efficiently handle [14].

Dynamic Data Dependency: Key-switching worsens the issues by causing huge memory demands. Decomposition parameter $(dnum)$

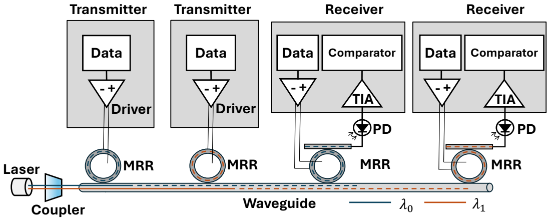

Fig. 3. WDM photonic interconnect linking two transmitters and receivers, working on two different wavelengths of λ_1 and λ_2.

affects computation and memory, where a trade-off must be made between NTT operations and basis conversion. Parallelism approaches such as residue-polynomial-level parallelism is plagued by extra data exchanges and coefficient-level parallelism has latency caused by global NTT communication. Hardware needs to reconfigure dynamically, yet insufficient on-chip memory will result in recurring off-chip access, exacerbating latency and power consumption [17].

With these growing memory requirements, high-bandwidth solutions are required for effective FHE acceleration. Fig. 2 emphasizes the bandwidth requirements of current FHE hardware, and it can be seen that **current electronic interconnects cannot keep up with these data transfer requirements [26].**

III. Methodology

In this section, we consider photonic interconnects as a viable solution to meet the required memory bandwidth of FHE accelerators, design a scalable photonic network architecture, and define an evaluation metric to compare the performance with electronic counterparts.

A. Pathway to TB/s Bandwidth for FHE

To mitigate the bandwidth constraints, emerging chiplet-based FHE accelerators introduced high-bandwidth memory (HBM) technologies such as HBM3 in order to reduce data transfer delay [28]. HBM3, with a 1024-bit datawidth, achieves bandwidths of up to 0.819TB/s per stack [29]. To meet the bandwidth demands of current FHE accelerators, which often require TB/s as seen in Fig. 2, multiple HBM3 stacks with aggregated datawidths in the thousands are typically employed. Photonic interconnects provide an effective way to solve these issues, as explained in the following sections.

Ultra High Bandwidth: With their ability to support ultra-high bandwidths, photonic interconnects represent an attractive alternative for FHE workloads. The proposed *OptoLink* architecture, for example, achieves a bandwidth of 0.8TB/s with only 64 channels which is 16× lower than bitwidth of HBM3 and can scale to meet the demands of all the accelerators listed in Fig. 2. This characteristic reflects *OptoLink*s strength to compete with existing electronic interconnect alternatives while minimizing high datawidth electric interfaces' overhead and complexity.

Flexible Routing: In addition, *OptoLink* overcomes the generality constraints of FHE accelerators through flexible [14] data routing and workload parallelism. Through dynamic multiplexing of data over optical channels, *OptoLink* provides efficient use of resources and minimization of data movement latency between PEs and memory. Leveraging the broadcast property of suggested *OptoLink* network, it further improves flexibility by allowing simultaneous data broadcasting to several PEs. This flexibility allows FHE accelerators to support different computation patterns from high-bandwidth NTT computation to memory-constrained key-switching without heavy architectural changes. The details of the *OptoLink* architecture are elaborated in Sec. III-C, explaining how its architecture achieves ultra-high-speed, low-latency data communication.

Furthermore, experimental results validating the system's performance, scalability, and reliability are presented in Sec. IV. Through *OptoLink*, FHE accelerators can be provided with enough bandwidth to maintain key workloads without compromising adaptability and efficiency, and ultimately, introduce the scalability and viability of FHE to practical, privacy-preserving applications.

B. Photonic Interconnects

Photonic interconnects use silicon photonics to achieve fast and energy-efficient data transmission by replacing traditional electrical signals with light. As shown in Fig. 3, light is produced by a laser that is coupled into an on-chip waveguide, where micro-ring resonators (MRRs) are employed as modulators and filters [30] to modulate electrical data onto selected wavelengths of light. The transmitted signals use the waveguide path to reach the receiver section which contains another MRR array directing the signals to photodetectors (PDs) for electrical signal recovery. In addition to modulation and filtering, receiver MRRs are also utilized as optical tunable splitters for broadcast communication efficiently. These splitters work in a partially resonant state, and they let a portion of the optical power pass through the drop port and the remaining portion through the through port. The carrier concentration within the ring is changed by controlling the bias voltage, thereby modulating the effective refractive index of the waveguide [31]. This dynamic tuning allows for fine-grained control of the way the optical power is split between the through and drop ports, optimizing data distribution among several PEs. An important advantage of photonic interconnects is the application of WDM to allow the simultaneous transmission of multiple data streams via a single waveguide. This technology allows bandwidth capacity to be greatly improved without requiring more interconnects. Existing systems possess the capability to accommodate a maximum of 64 distinct wavelengths, each operating at a rate of 10Gb/s, thereby achieving aggregate throughput levels exceeding 100Gb/s [32, 33]. SDM provides the capability to extend bandwidth potential through implementing multiple parallel waveguides [18].

C. Single OptoLink Channel

Memory controller and NTT modules are connected within a single *OptoLink* channel using photonic interconnect as shown if Fig. 4. The architecture employs WDM to send multiple signals simultaneously through a single waveguide, with each signal assigned a distinct wavelength. Input data such as twiddle factors and coefficients are stored in the memory controller where they are converted to electric signals through digital-to-analog converters (DACs). Transmitter MRRs modulate certain wavelengths onto waveguides according to the converted electrical signals. On the receiving side, tuned MRRs filter out resonant wavelengths, directing each signal to a PD that converts it back into an electrical signal. This signal is then amplified by TIAs and processed by comparators to reconstruct the digital data. Once the data is processed by the NTT module, the results undergo the same demodulation and modulation for return to the memory controller to perform the next stage of the NTT computation. The same wavelengths can be utilized for input and output transmission because distinct waveguides will be utilized for input and output. This strategy of wavelength reuse also lowers the systems power consumption by lowering the number of wavelengths used.

D. Scalable OptoLink Network Architecture

The *OptoLink* architecture in Fig. 5 uses five distinct waveguides to connect to four NTT modules in order to carry twiddle factors and data. *Waveguides 1* and *2* transport coefficients to the NTT modules, while *Waveguides 3* and *4* transfer the required twiddle factors. After

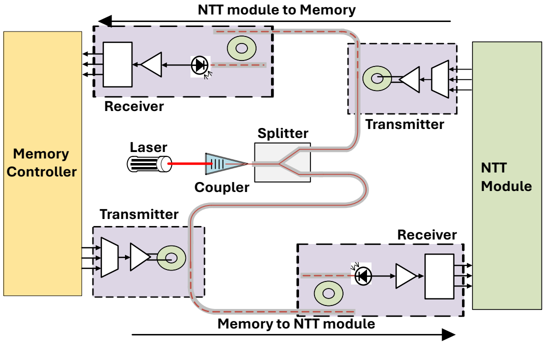

Fig. 4. WDM-based single channel *OptoLink* network interconnecting Memory and NTT module. The network transfers data and twiddle factors to the NTT Module and sends computed outputs back to the memory.

processing, the outputs are sent back to the memory controller via *Waveguide 5*. The system utilizes two groups of wavelengths to control data flow: $\lambda_1 - \lambda_{16}$ for input coefficients and twiddle factors, and $\lambda_{17} - \lambda_{24}$ for output transmission. Each waveguide carries a single bit per channel, while multiple optical channels can function in parallel via SDM for rapid data transfer. For example, for an 128-channel *OptoLink* system, there will be 128 waveguides in parallel.

Three distinct ways of communication between memory and NTT modules in *OptoLink*—unicast, broadcast, and multicast—offer flexible data transfer appropriate for various computation stages. For an $n = 8$ polynomial, as illustrated in Fig. 5, each NTT module needs distinct input coefficients, which requires unicast transmission. In this mode, separate data streams are concurrently sent to different NTT modules through *Waveguide 1*, using wavelengths $\lambda_1 - \lambda_4$. In this scenario, all receiver MRRs are in an on-resonance state, ensuring that no signal passes through the through port, thus directing data solely to the intended modules. The first NTT stage employs common twiddle factors which allows broadcast-mode data distribution for all NTT modules. In this mode, receiver MRRs are tuned to a partially resonant state, allowing optical signals to be evenly split between the through and drop ports. For example, wavelength λ_9 is designated for broadcast transmission from memory to all NTT modules via *Waveguide 3*. As the computation progresses to the second stage, NTT modules exhibit distinct data dependencies—NTT 1 and 2 share one set of twiddle factors, while NTT 3 and 4 share another. This requires multicast transmission, which combines unicast and broadcast techniques. Wavelength λ_9 is used for broadcast communication to NTT modules 1 and 2, whereas wavelength λ_{10} is assigned to NTT modules 3 and 4. *OptoLink*'s broadcasting feature also allows for flexible routing by dynamically adjusting communication modes through receiver MRR tuning, enabling data to be directed to any NTT module as required. This capability overcomes the limitations of traditional static hardware, which lacks adaptability [15]. As the polynomial size n increases, the demand for hardware resources grows, and the data dependencies in NTT computations can vary greatly [14]. The data path adjustment capability of *OptoLink* can be operated in real time for effective computational requirements adaptation and smooth scalability improvement while conventional architectures demand costly modifications for handling changes.

The design of *OptoLink* incorporates scalable features which fulfill rising computational demands for FHE accelerators. The architecture reaches high throughput levels through increased optical channels and SDM and WDM implementation. Research indicates that one waveguide can process 64 wavelength multiplexing points which produces valuable bandwidth improvements [33]. The tunability of MRR does allow every transmitter to connect to more than one receiver, resulting in fewer modulators, but overall power goes up with the complete setup of off-chip lasers, MRRs, and photodetectors added. A trade-off between power efficiency and scalability and supporting high-throughput FHE workload tolerance is therefore necessary. The *OptoLink* architecture is able to meet these challenges successfully, meeting bandwidth and scalibitlity demands while being versatile across applications.

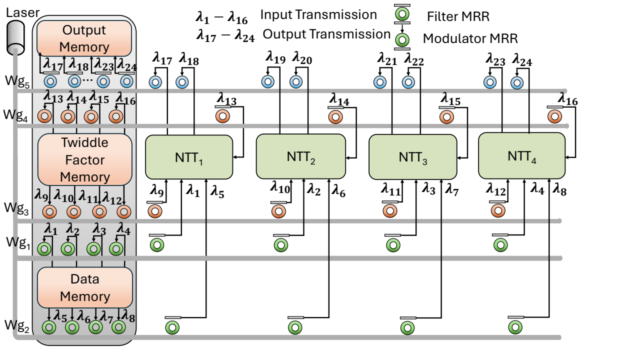
Laser
$\lambda_1 - \lambda_{16}$ Input Transmission
$\lambda_{17} - \lambda_{24}$ Output Transmission
Filter MRR
Modulator MRR
Output Memory
Twiddle Factor Memory
Data Memory
Wg_5
Wg_4
Wg_3
Wg_1
Wg_2
λ_{17} λ_{18} ... λ_{23} λ_{24}
λ_{13} λ_{14} λ_{15} λ_{16}
λ_9 λ_{10} λ_{11} λ_{12}
λ_1 λ_2 λ_3 λ_4
λ_5 λ_6 λ_7 λ_8
NTT_1
NTT_2
NTT_3
NTT_4
λ_{17} λ_{18}
λ_{19} λ_{20}
λ_{21} λ_{22}
λ_{23} λ_{24}
λ_{13}
λ_{14}
λ_{15}
λ_{16}
λ_9 λ_1 λ_5
λ_{10} λ_2 λ_6
λ_{11} λ_3 λ_7
λ_{12} λ_4 λ_8

Fig. 5. Schematic of the *OptoLink* architecture interfacing with four NTT modules where $n = 8$. Wavelengths $\lambda_1 - \lambda_{16}$ are used for input transmission and $\lambda_{17} - \lambda_{24}$ are used for output transmission for broadcast communication.

E. Implementation Platform and Parameter Selection

To evaluate the performance of the proposed *OptoLink* architecture, we established photonic interconnects between NTT modules and off-chip memory, to tackle the communication bottlenecks identified in an FPGA-based FHE accelerator, named HEAX [15]. HEAX's memory-to-NTT complexities are used here to highlight the limitations of traditional electronic networks and the necessity of photonic solutions for data transfer rate optimization and latency minimization. In order to obtain a proper estimate of *OptoLink*, we used `Synopsys OptoCompiler` to incorporate critical photonic parameters such as detector responsivity, modulator insertion loss, and coupling efficiency (see Table I). Using these parameters we calculated the needed laser power, to facilitate proper signal transmission despite optical defects. We employed `Synopsys Design Compiler` to examine the timing behavior, power consumption, and area usage of the electrical components as well. To explore scalability across various computational workloads, we conducted evaluations with diverse NTT module configurations and bitwidth, namely configurations with 4, 8, and 16 modules. This detailed study provided us with keen insight into *OptoLink*'s ability to meet the increasing needs of FHE acceleration without compromising efficiency.

F. Evaluation Metric (R)

To evaluate the performance of the electronic network and *OptoLink* in a fair way, we present an evaluation metric R. This metric reflects trade-offs among bitrate, latency, and power consumption. Because there are huge differences in bitrate and power consumption

TABLE I
PHOTONIC PARAMETERS CONSIDERED IN *OptoLink*.

Component	***Value***
Laser Source	5 dB
Coupler	1 dB
Splitter	0.2 dB
Waveguide	1 dB/cm
Ring Drop	0.7 dB
Ring Through	0.01 dB
Photodetector	0.5 dB

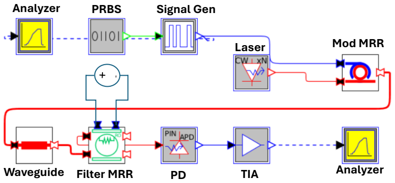

Fig. 6. Simulation setup for a single channel of the *OptoLink* system. A PRBS generates random signals transmitted over a $1000\mu m$ channel, operating at a wavelength of $1550nm$.

between photonic and electronic interconnects, this function gives a normalized metric to compare efficiency of the system

$$R = \frac{\text{Bitrate}}{\text{Latency} \times \text{Power}} \tag{4}$$

where Bitrate is the data transmission rate, Latency is the time taken for transferring data, and Power is the total power consumption of the system. A higher R value indicates a more efficient interconnect system, which balances high throughput with power overhead. Our target is to maximize the value of R for *OptoLink*, to make sure that it outperforms electronic networks.

IV. Results and Analysis

A. Timing Analysis

To examine the timing performance of the *OptoLink* network, we conducted data transmission experiments using two optical channels, simulated with `Synopsys OptoCompiler`. A pseudo-random bit sequence (PRBS) generator transmitted the data at a rate of 10Gb/s, with 6.4ns taken to transmit one sequence. The modulator MRRs encoded the data onto specific wavelengths—1550nm for channel 1 and 1551nm for channel 2—before transmission. Fig. 7(c), Fig. 7(d) are the data received on channel 1 and channel 2, respectively, confirming secure data transfer and integrity in the *OptoLink* system. We operate within the $1500 - 1600$ nm wavelength region and to make sure there is no crosstalk and interference we have a channel spacing of 0.5 nm. The gain of the TIA is set to $5k\Omega$.

One of the key benefits of *OptoLink* is its very low latency. Transmission of data through a $1000\mu m$ waveguide incurs only a 10ps delay, which is much lower than the 3.04ns required by traditional electronic networks. This decrease in propagation time is indicative of *OptoLink*'s better efficiency at high-speed applications. With this $10ps$ data transfer time, the data rate for a single *OptoLink* channel is calculated at 100Gb/s or 12.5GB/s. Extending this configuration to 128 channels yields an aggregate bandwidth of 1.6TB/s, thereby achieving the TB/s bandwidth necessary for FHE operations as seen on Fig 2. In an 192-channel implementation, *OptoLink* offers 2.4TB/s, on par with the NVIDIA A100 [34]. With even further scaling, it is as much as 12.8TB/s with 1024 channels, several orders of magnitude ahead of electrical networks, which can barely achieve 42.1GB/s at the same bitwidth (Table II). Notably, to achieve *OptoLink*'s 1.6TB/s bandwidth using electronic connections would require a highly unrealistic 4864-bit data sequence. This rapid data movement is critical to FHE workloads, lowering memory-to-compute latency and eliminating bottlenecks.

B. Power Analysis

The power consumption of the *OptoLink* system is largely determined by laser source and MRRs. To estimate the total power consumption of the system, we used the following equation:

$$P_{\text{total}} = P_{\text{laser}} + P_{\text{TX}} + P_{\text{RX}} \tag{5}$$

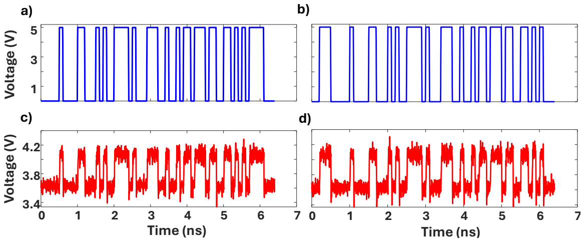

Fig. 7. (a-b) Input electrical signals applied to modulator MRRs in two distinct *OptoLink* channels for specified wavelengths. (c-d) Corresponding signals after conversion by the PD and amplification by the TIA in each channel, read for the same wavelengths.

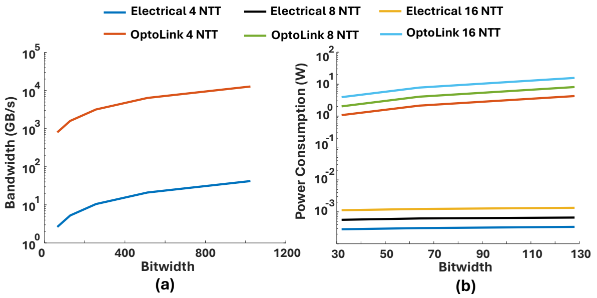

Fig. 8. Comparison between electronic network and *Optolink* for different number of NTT cores and bitwidth. (a) Bitwidth vs Bandwidth and (b) Bitwidth vs Power Consumption.

where P_{laser} is the power dissipation of the laser source, and P_{TX} and P_{RX} is the power dissipation of the transmitter and receiver sections, respectively. According to the specified values, P_{TX} and P_{RX} are estimated to be $0.9mW$ and $0.6mW$ per channel, respectively [35]. The power consumption associated with the *OptoLink* system, along with that of an electronic network system in various configurations, is presented in Table III. For an *OptoLink* system consisting of 128 channels connected in parallel to 4 NTT modules, the estimated power consumption is around $3.91W$. Each channel in this configuration consists of 24 transmitters and receivers each, which contributes significantly to the power consumption. The power usage of the *OptoLink* system grows linearly with the number of NTT cores, at $7.82W$ for 8 NTT cores and $15.63W$ for 16 NTT cores. This is due to the higher number of transmitters and receivers per channel required for larger number of NTT cores. Also for different bitwidths, the number of *OptoLink* channels that operate in parallel increases to accommodate data transmission through SDM, leading to higher overall power consumption at larger bitwidths. In comparison, the power consumed by the electronic network in facilitating data transfer via 128-bit configurations utilizing 4, 8, and 16 NTT modules is considerably low, at $336.99\mu W$, $661.74\mu W$, and $1332.31\mu W$, respectively. This reduced power consumption is due to the absence of devices utilized in the generation and processing of optical signals, which are essential components of the *OptoLink* system. Table III also provides further insights into systems' power consumption configured with 32-bit and 64-bit architectures, where the same trend can be seen.

TABLE II
Bitrate comparison of Electronic Network and *OptoLink*

	Electronic Network		OptoLink	
Bitwidth	***Latency***	***Bitrate***	***Latency***	***Bitrate***
32	3.04ns	1.32GB/s	10ps	0.4TB/s
64	3.04ns	2.63GB/s	10ps	0.8TB/s
128	3.04ns	5.26GB/s	10ps	1.6TB/s

TABLE III
POWER CONSUMPTION FOR ELECTRONIC NETWORK AND OPTOLINK

Bitwidth	NTT Cores	Power Consumption	
		Electronic Network (μW)	*OptoLink* (W)
32	4	283.89	1.07
	8	562.44	2.12
	16	1121.9	4.23
64	4	308.18	2.02
	8	619.29	4.04
	16	1232.19	8.09
128	4	336.99	3.91
	8	661.74	7.82
	16	1332.31	15.63

TABLE IV
POWER COMPARISON OF *OptoLink* IN DIFFERENT NTT STAGES FOR 16 NTT CORES. POWER SAVING IS MORE PROMINENT IN EARLIER STAGES.

Bitwidth	*NTT*				
	Stage 1	*Stage 2*	*Stage 3*	*Stage 4*	*Stage 5*
32	3.38 W	3.48 W	3.68 W	4.09 W	4.91 W
64	6.75 W	6.96 W	7.37 W	8.19 W	9.82 W
128	13.5 W	13.92 W	14.7 W	16.37 W	19.65 W

Table IV presents the power consumption of a 16-NTT core architecture, calculated for different bitwidths and NTT stages. The broadcasting operation of *OptoLink* achieves a noticeable reduction in power consumption in Stage 1 as opposed to Stage 5. In particular, Stage 1 has an average power consumption 31.2% lower than that of Stage 5 demonstrating the power efficiency of optical data broadcasting. The use of a shared data stream that is split among an array of NTT modules results in fewer wavelengths being required, which also lessens the number of lasers, transmitters, and receivers, resulting to reduced total power consumption. Compared to larger polynomial NTT operations, the energy efficiency is greatly improved as more computations can be carried out with one broadcast, making *OptoLink* a highly effective tool for large-scale FHE applications.

C. Area Analysis

The area requirements for the proposed $OptoLink$ architecture were evaluated by comparing the space occupied by conventional electronic networks with that of the photonic components integral to $OptoLink$. Using a $32nm$ technology library, we estimated the area for the electronic networks, enabling precise measurement based on realistic process design parameters. For a 128-bit NTT configuration, the area requirements for electronic networks scale nearly linearly with the number of NTT units. Specifically, configurations with 4, 8, and 16 NTT units occupied areas of $3097.3\mu m^2$, $5741.2\mu m^2$, and $11861.9\mu m^2$, respectively. In contrast, $OptoLink$'s photonic data transmission components necessitate a larger area due to the photonic elements involved. Prior research suggests that each photonic transmitter or receiver occupies approximately $0.0096mm^2$ per wavelength [36]. Critical to wavelength-selective modulation in our architecture, MRRs add to the area requirements; with a typical MRR radius of $5\mu m$, the total area contribution from MRRs is estimated to be around $0.01mm^2$ [37].

D. Evaluation metric (R) Analysis

A new evaluation metric was defined in Eq. 4 was created to guarantee fair assessment between *OptoLink* and the electronic network. With a 32-bit channel using 4 NTT cores *OptoLink* produces an R value reaching up to 3.79×10^{22} but the electronic network stops at 1.53×10^{21}. As shown in Fig. 9 *OptoLink* displays a superior efficiency in high-bandwidth workload handling when compared to other alternatives. The data shows *OptoLink* delivers better performance than the electronic network across every configuration because it generates superior R values throughout all tested NTT core amounts and bitwidth values. *OptoLink* maintains superior performance because its high bandwidth and minimal delay establishes an effective answer for extensive FHE acceleration operations. Fig. 9 shows the results which demonstrate that OptoLink outperforms traditional electronic networks by providing superior scalability.

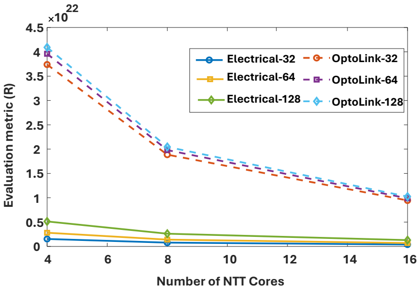

Fig. 9. Evaluation metric (R) comparison between electronic network and $Optolink$ for different number of NTT cores.

E. Discussion

The results refer to the high timing performance of *OptoLink*, where 128 parallel photonic channels achieve 1.6TB/s, while electronic networks achieve just 5.26GB/s for the same bitwidth. This ultra-low latency is highly desirable for FHE, reducing bottlenecks and enabling high-speed data communication. The advantages come, however, at the expense of higher power consumption. *OptoLink*'s broadband capability saves power by minimizing redundant transmission, which lowers power consumption by 31.2% at initial NTT stages. This is especially useful with large polynomial operations as data streams shared between NTT cores cut down on active optical components. Despite the increased power requirements, *OptoLink* obtains a better R value than an electronic network, demonstrating its superior data transmission efficiency. The R value analysis shows that *OptoLink* is consistently superior to electronic networks for leveraging high throughput and low latency for large-scale high-performance FHE acceleration. With these results, it is evident that the questions posed in Sec. I regarding scalability, bandwidth, and efficiency have been addressed, showing that OptoLink can be a viable solution. For large-scale FHE workloads, these results make *OptoLink* a viable alternative for existing electronic networks.

V. CONCLUSION

OptoLink resolves major issues in current FHE accelerators by employing photonic interconnects. It can achieve picosecond latency, which is much lower than that of electronic networks. Additionally, due to the broadcast capability of OptoLink, the energy consumption during initial NTT stages is lowered due to fewer numbers of wavelengths being used. *OptoLink*'s total bandwidth of 1.6 TB/s across 128 channels at 100 Gb/s per channel allows it to handle large ciphertexts. Although photonic components introduce power and area overhead, the R value of OptoLink is higher than of electronic networks. This higher R value, showcases better performance. There is a lot of work going on in recent times to develop power-efficient MRRs, which will reduce the power consumption in future implementations. In short, *OptoLink* is a scalable, high-speed and energy-efficient interconnect solution for FHE accelerators. Future research will focus on further optimizing power and area to enable it for practical next-generation privacy-preserving computing.

VI. ACKNOWLDGEMENTS

ChatGPT has been used for grammar checking and correction.

References

[1] J. H. Cheon, M. Kim, and M. Kim, "Optimized search-and-compute circuits and their application to query evaluation on encrypted data," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 1, pp. 188–199, 2015.

[2] A. Chatterjee and I. Sengupta, "Sorting of fully homomorphic encrypted cloud data: Can partitioning be effective?" *IEEE Transactions on Services Computing*, vol. 13, no. 3, pp. 545–558, 2017.

[3] A. Abdallah and X. S. Shen, "A lightweight lattice-based homomorphic privacy-preserving data aggregation scheme for smart grid," *IEEE Transactions on Smart Grid*, vol. 9, no. 1, pp. 396–405, 2016.

[4] C. Gentry and S. Halevi, "Implementing gentry's fully-homomorphic encryption scheme," in *Annual international conference on the theory and applications of cryptographic techniques*. Springer, 2011, pp. 129–148.

[5] J. W. Bos, K. Lauter, J. Loftus, and M. Naehrig, "Improved security for a ring-based fully homomorphic encryption scheme," in *Cryptography and Coding: 14th IMA International Conference, IMACC 2013, Oxford, UK, December 17-19, 2013. Proceedings 14*. Springer, 2013, pp. 45–64.

[6] S. S. Roy, F. Turan, K. Jarvinen, F. Vercauteren, and I. Verbauwhede, "Fpga-based high-performance parallel architecture for homomorphic computing on encrypted data," in *2019 IEEE International symposium on high performance computer architecture (HPCA)*. IEEE, 2019, pp. 387–398.

[7] B. Reagen, W.-S. Choi, Y. Ko, V. T. Lee, H.-H. S. Lee, G.-Y. Wei, and D. Brooks, "Cheetah: Optimizing and accelerating homomorphic encryption for private inference," in *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2021, pp. 26–39.

[8] J. Fan and F. Vercauteren, "Somewhat practical fully homomorphic encryption," *Cryptology ePrint Archive*, 2012.

[9] A. López-Alt, E. Tromer, and V. Vaikuntanathan, "On-the-fly multiparty computation on the cloud via multikey fully homomorphic encryption," in *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, 2012, pp. 1219–1234.

[10] N. Samardzic, A. Feldmann, A. Krastev, S. Devadas, R. Dreslinski, C. Peikert, and D. Sanchez, "F1: A fast and programmable accelerator for fully homomorphic encryption," in *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture*, 2021, pp. 238–252.

[11] N. Samardzic, A. Feldmann, A. Krastev, N. Manohar, N. Genise, S. Devadas, K. Eldefrawy, C. Peikert, and D. Sanchez, "Craterlake: a hardware accelerator for efficient unbounded computation on encrypted data," in *Proceedings of the 49th Annual International Symposium on Computer Architecture*, 2022, pp. 173–187.

[12] D. Li, A. Pakala, and K. Yang, "Mentt: A compact and efficient processing-in-memory number theoretic transform (ntt) accelerator," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 30, no. 5, pp. 579–588, 2022.

[13] P. Duong-Ngoc *et al.*, "Efficient k-parallel pipelined ntt architecture for post quantum cryptography," in *2020 International SoC Design Conference (ISOCC)*, 2020, pp. 212–213.

[14] J. Zhang *et al.*, "Sok: Fully homomorphic encryption accelerators," *ACM Computing Surveys*, 2022.

[15] M. S. Riazi *et al.*, "Heax: An architecture for computing on encrypted data," in *Proceedings of the twenty-fifth international conference on architectural support for programming languages and operating systems*, 2020, pp. 1295–1309.

[16] A. C. Mert *et al.*, "A Flexible and Scalable NTT Hardware : Applications from Homomorphically Encrypted Deep Learning to Post-Quantum Cryptography," in *Design, Automation & Test in Europe (DATE)*, 2020, pp. 346–351.

[17] M. Zhou, Y. Nam, P. Gangwar, W. Xu, A. Dutta, K. Subramanyam, C. Wilkerson, R. Cammarota, S. Gupta, and T. Rosing, "Fhemem: A processing in-memory accelerator for fully homomorphic encryption," *arXiv preprint arXiv:2311.16293*, 2023.

[18] K. Bergman *et al.*, "Photonic Network-on-Chip Design," *Springer*, 2014. [Online]. Available: http://www.springer.com/series/7236

[19] Y. Li *et al.*, "Spacx: Silicon photonics-based scalable chiplet accelerator for dnn inference," in *IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2022, pp. 831–845.

[20] C. Gentry, *A fully homomorphic encryption scheme*. Stanford university, 2009.

[21] W. Jung, S. Kim, J. H. Ahn, J. H. Cheon, and Y. Lee, "Over 100x faster bootstrapping in fully homomorphic encryption through memory-centric optimization with gpus," *IACR Transactions on Cryptographic Hardware and Embedded Systems*, pp. 114–148, 2021.

[22] S. Fan, Z. Wang, W. Xu, R. Hou, D. Meng, and M. Zhang, "Tensorfhe: Achieving practical computation on encrypted data using gpgpu," in *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2023, pp. 922–934.

[23] Y. Yang, H. Zhang, S. Fan, H. Lu, M. Zhang, and X. Li, "Poseidon: Practical homomorphic encryption accelerator," in *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2023, pp. 870–881.

[24] R. Agrawal, L. de Castro, G. Yang, C. Juvekar, R. Yazicigil, A. Chandrakasan, V. Vaikuntanathan, and A. Joshi, "Fab: An fpga-based accelerator for bootstrappable fully homomorphic encryption," in *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2023, pp. 882–895.

[25] J. Kim, G. Lee, S. Kim, G. Sohn, M. Rhu, J. Kim, and J. H. Ahn, "Ark: Fully homomorphic encryption accelerator with runtime data generation and inter-operation key reuse," in *2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2022, pp. 1237–1254.

[26] L. de Castro, R. Agrawal, R. Yazicigil, A. Chandrakasan, V. Vaikuntanathan, C. Juvekar, and A. Joshi, "Does fully homomorphic encryption need compute acceleration?" *arXiv preprint arXiv:2112.06396*, 2021.

[27] D. H. Bailey, "Ffts in external or hierarchical memory," *The journal of Supercomputing*, vol. 4, pp. 23–35, 1990.

[28] A. Aikata, A. C. Mert, S. Kwon, M. Deryabin, and S. S. Roy, "Reed: Chiplet-based accelerator for fully homomorphic encryption," *Cryptology ePrint Archive*, 2023.

[29] JEDEC. (2024, November) Jedec publishes hbm3 update to the high bandwidth memory (hbm) standard. Accessed: 2024-11-17. [Online]. Available: https://www.jedec.org/news/pressreleases/jedec-publishes-hbm3-update-high-bandwidth-memory-hbm-standard

[30] W. Bogaerts, P. De Heyn, T. Van Vaerenbergh, K. De Vos, S. Kumar Selvaraja, T. Claes, P. Dumon, P. Bienstman, D. Van Thourhout, and R. Baets, "Silicon microring resonators," *Laser & Photonics Reviews*, vol. 6, no. 1, pp. 47–73, 2012.

[31] E. Peter, A. Thomas, A. Dhawan, and S. R. Sarangi, "Active microring based tunable optical power splitters," *Optics Communications*, vol. 359, pp. 311–315, 2016.

[32] S. Werner, J. Navaridas, and M. Luján, "Designing low-power, low-latency networks-on-chip by optimally combining electrical and optical links," in *2017 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2017, pp. 265–276.

[33] R. Morris, A. K. Kodi, and A. Louri, "Dynamic reconfiguration of 3d photonic networks-on-chip for maximizing performance and improving fault tolerance," in *2012 45th Annual IEEE/ACM International Symposium on Microarchitecture*. IEEE, 2012, pp. 282–293.

[34] J. Choquette, W. Gandhi, O. Giroux, N. Stam, and R. Krashinsky, "Nvidia a100 tensor core gpu: Performance and innovation," *IEEE Micro*, vol. 41, no. 2, pp. 29–35, 2021.

[35] Y. Li, K. Wang, H. Zheng, A. Louri, and A. Karanth, "Ascend: A scalable and energy-efficient deep neural network accelerator with photonic interconnects," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 7, pp. 2730–2741, 2022.

[36] Y. Thonnart, M. Zid, J. L. Gonzalez-Jimenez, G. Waltener, R. Polster, O. Dubray, F. Lepin, S. Bernabé, S. Menezo, G. Parès *et al.*, "A 10gb/s si-photonic transceiver with 150μw 120μs-lock-time digitally supervised analog microring wavelength stabilization for 1tb/s/mm 2 die-to-die optical networks," in *2018 IEEE International Solid-State Circuits Conference-(ISSCC)*. IEEE, 2018, pp. 350–352.

[37] G. Li, X. Zheng, H. Thacker, J. Yao, Y. Luo, I. Shubin, K. Raj, J. E. Cunningham, and A. V. Krishnamoorthy, "40 gb/s thermally tunable cmos ring modulator," in *The 9th International Conference on Group IV Photonics (GFP)*. IEEE, 2012, pp. 1–3.