Rearchitecting Data Movement in Next-Generation Networked Distributed Systems

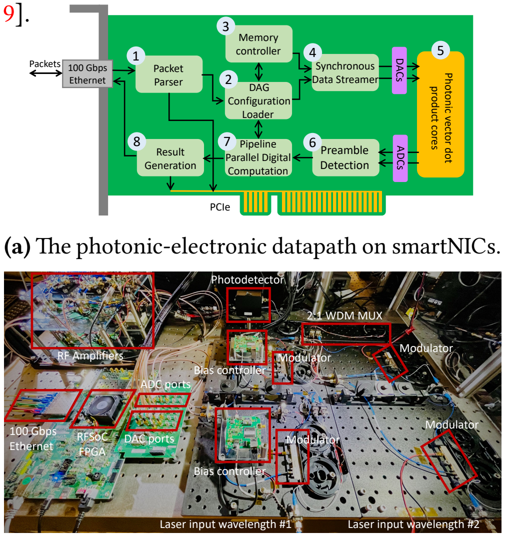
Zhizhen Zhong
zhizhenz@mit.edu

With the explosion of data-intensive applications and the scaling limitations of traditional electronic hardware, computing is at an inflection point. To keep up with the trend, hyperscale networks and data centers have been built to combine a massive number of computing resources in the cloud and connect them to end users, resulting in billions of data packets flooding global network infrastructure. At the same time, the emergence of domain-specific accelerators (e.g., machine learning chips like TPUs, photonic computing) and the trend of resource disaggregation (e.g., far memory) in cloud systems have challenged the traditional norm of CPU-centric system architecture. As a result, **data movement has increasingly become a greater bottleneck in terms of energy and latency than data computation**. This creates a pressing need to rethink and redesign networked systems.

My research aims to transform the design and implementation of **end-to-end data movement pathways** through the co-design between emerging optical technologies and existing computer systems stacks, contributing to a future where data flows across next-generation computer systems with *minimal or zero bottlenecks*. My vision is built on the observation that optics and photonics have the fundamental merits of higher bandwidth and lower energy consumption than their electronic counterparts. Towards this vision, my work takes an *application-centric* view to address several major roadblocks by engineering the unique properties of light and its fundamental particles (photons) in the context of computer networks and systems. First, for the data movement bottleneck at *networking hardware level* among network I/O, memory and emerging domain-specific accelerators like photonic computing, I define **reconfigurable photonic-electronic datapaths** (§1) on network interface cards (smartNICs, DPUs, and IPUs) or optical transceiver modules. Second, for data transmission bottlenecks at *network links level* among distributed machines, I integrate **reconfigurable optical circuit switching (OCS)** (§2) into IP packet switches for traffic engineering (TE) and topology optimization.

My work is uniquely positioned in the systems and networking area because it constantly seeks to extend the community's scope by incorporating emerging technologies from the optics and photonics area. As an old Chinese saying states "the stones of other hills may be used to polish gems here", working across disciplines put me at the vantage point to perform **cross-layer design** to bring the best of both worlds to create new systems with capabilities not possible before. My cross-layer methodology is categorized into: first, *identifying the data movement bottleneck* from the application perspective; then, *going beyond* the conventional view of layering and investigate how different layers of the system interact with each other to propose solutions to improve application performance by touching several layers; finally, *solve the problem* using both theoretical algorithmic design and full system implementation.

I am passionate about building real-world systems from scratch, writing customized software overlays, and producing impactful results for both academia and industry (§3). Making research ideas into practical prototypes with heterogeneous photonic-electronic hardware is challenging. This is because many optical devices are mostly in its "bare-metal" form without reconfigurable software overlays. I view these challenges as **distinctive opportunities to build and evaluate first-of-its-kind systems**. For example, I built the first experimental system capable of serving real-time machine learning inference requests at the record-breaking 4.055 GHz with emerging photonic computing [1, 2]. My research in optical wide-area networks (WANs) was first validated in production networks [3, 4], then globally rolled out at *Meta* to manage traffic spikes during COVID-19 [5]. This line of work later influenced *Tencent* to invest and launch optical fiber cut restoration in prouduction [6]. I also use insights gained from operational experience to ignite new research for distributed machine learning training [7] and WAN TE [8, 9].



(a) The photonic-electronic datapath on smartNICs.

(b) Our testbed features the first on-NIC photonic computing bypassing CPUs to achieve 4.055 GHz.

Figure 1: Project Lightning: SmartNIC datapath, experimental prototype [1, 2].

1 Accelerated Datapaths at Network Nodes

Disaggregating compute, memory, network I/O and storage resources has emerged as a paradigm shift for data centers to improve efficiency. However, moving data around disaggregated resources to finish a computation task remains a bottleneck to achieve *fungible* and *elastic* task provisioning and allocation. My research takes a systems approach to address this challenge by co-designing the optical layer and electronics parts of the system. In particular, I have worked on the design and implementation of: 1) the first reconfigurable CPU-bypassing *datapath* that handles memory access, analog-digital conversions and conditional logic with directed acyclic graph (DAG) for domain-specific photonic multiply-accumulate (MAC) cores [1, 2], 2) the first smart *transceiver* that performs photonic computing on incoming optical data for machine learning [10, 11, 12, 13] and intrusion detection [14].

Reconfigurable CPU-bypassing datapath for photonic MAC cores. My work, Lightning, is the first system to address the data movement bottleneck between photonic MAC cores and electronic memory and logic components. The root cause of this problem is that photonic MAC cores are inherently passive devices *without any memory or instructions* to control the computation dataflow of complex real-world applications. This problem is worsened when the datapath's control plane is implemented in OS kernel on the CPU. Lightning mitigates these bottlenecks by co-designing the digital and photonic components together for a CPU-bypassing datapath on a smartNIC (Fig. 1a). At the core of Lightning is its *reconfigurable count-action* abstraction. This innovation leverages the unique property that light propagates through the system with a deterministic time of flight. Therefore, a "timer"-like *counter* in the digital domain is capable of triggering the next *action* on behalf of passive photonic components. I built the first experimental prototype for Lightning with an RFSoC FPGA and commodity optical devices (Fig. 1b). To the best of my knowledge, thanks to the fast datapath, our prototype is the **highest-compute-frequency photonic computing system to date** [1]. In testbed experiments serving real-time inference packets, Lightning is 6.6× faster than an Nvidia Triton server with an A100 GPU [2].

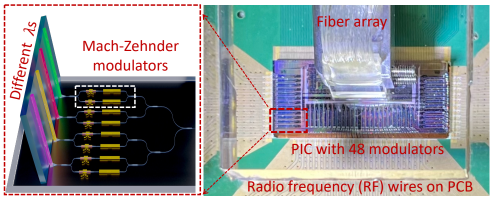


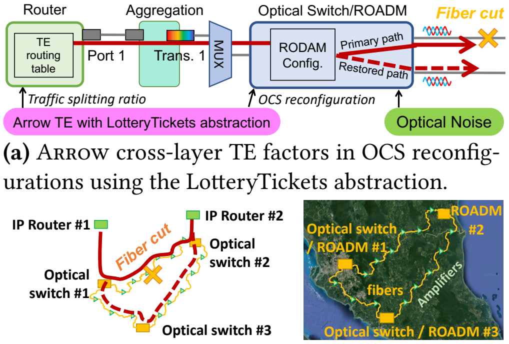
Figure 2: Project OpTCAM: PIC for photonic ternary matching up to 50 GHz [14].

Smart transceiver for in-fiber computing on received optical data. The main devices needed for photonic computing (lasers, modulators, photodetectors, DACs, and ADCs) are already equipped in today's optical transceivers that are widely deployed [12]. Therefore, the networking community is well-positioned to augment optical transceivers with photonic computing capabilities to enable a backward-compatible solution for in-network computing [10, 11, 13]. This line of thought opens up a novel research agenda to enable *in-fiber photonic computing* over fiber links. For example, my collaborators and I made the first step by co-designing the IOI smart transceiver with the Tofino programmable switch to perform in-network optical computing [10]. We further implemented the NetCast edge-computing system on 86 km of deployed fiber in Boston. NetCast enables milliwatt-class edge devices to compute at teraFLOPS rates otherwise reserved for kilowatt-class cloud servers [11, 13]. Our work won the *Best Paper Award* at OECC 2022 [13]. Most recently, I propose to extend in-fiber photonic computing to perform *ternary matching* at several times higher speed and lower power than digital implementations. My preliminary experiments on a photonic integrated circuit (PIC) show that our system, named OpTCAM, is error-free for a million ternary lookups [14]. I am now working on an in-network datapath to support 100 Gbps line-rate intrusion detection on transceivers (Fig. 2).

2 Reconfigurable Topology over Network Links

For decades, "high-volume optical transmission with fine-grained electrical packet processing" has been the dominant model for computer networks. Operators used to take the optical layer as a static resource when provisioning IP topology, then route traffic demands flexibly at the packet level. My work challenges this scheme and proposes to *dynamically adjust IP topology and capacity by reconfiguring optical wavelengths* through optical circuit switching (OCS). My results showed that this paradigm shift allows operators to improve affected data movement path from failures [3, 4], demand fluctuations [15, 16], and electrical buffer bottlenecks [17, 18].

Reconfigurable topology to restore capacity from fiber cuts. Fiber cuts are undesirable events because (*i*) each fiber carries multiple IP links with several Tbps of capacity, and (*ii*) fiber cuts take tens of hours to repair. My work, Arrow, restores failed bandwidth resources by reconfiguring the light from the cut fibers to healthy fibers without over-provisioning. However, there are many *possible OCS reconfiguration options* for the broken fiber. Taking all of them into a cross-layer optical-IP optimization is computationally intractable while taking only one of them is sub-optimal. We propose a novel abstraction, called *LotteryTickets*, which represents *offline*-calculated OCS reconfiguration options, to participate in TE's *online* computation (Fig. 3a). Our pre-computed *LotteryTickets* significantly reduces the problem search space and makes cross-layer TE feasible. Our results show Arrow supports 2.0×–2.4× more demand without compromising 99.99% availability [3]. We conducted a field trial at Meta's production optical WAN (Fig. 3b), and our experiment paper was presented as a *Postdeadline Paper* at OFC 2021 [4].



(a) Arrow cross-layer TE factors in OCS reconfigurations using the LotteryTickets abstraction.

(b) Field trial of fiber cut restoration in Meta.

Figure 3: Project Arrow & Bow: system design and production trial at Meta [3, 4].

Reconfigurable bandwidth during traffic fluctuations. During my Ph.D., I was among the early researchers who worked on reconfigurable optical networks [15, 16, 19]. I designed and implemented *LPSplitting*, the first system to reconfigure multi-hop optical wavelengths into multiple shorter wavelengths while adjusting the modulation format to augment IP-layer topology following the Shannon capacity law [15]. We present the key insight using a *Pareto-front* analysis that reconfigurable IP-over-optical network topology has a fundamental tradeoff between potential traffic interruption and IP-layer capacity augmentation [16].

End-to-end optical datapath without buffer. Another pillar of my Ph.D. work is on designing large-scale *bufferless* networks by replacing electrical packet switches with time-synchronized OCS switches for latency-sensitive applications in data centers [17, 20] and power-grid communication networks [18]. The insight behind this line of work is to *simplify* the network fabric to only perform *forwarding without buffering* with circuit switching on the optical layer, and offload fine-grained packet *queueing* and *scheduling* to end hosts. To address the problem of limited port count in OCS and build an all-optical switching fabric at scale, I proposed a novel periodical optical switching scheme with time synchronization [18]. This work won the *1st Place Best Poster Award* at OECC 2017 [18].

3 Research Impact

The type of researchers that I regard as my role model are those "innovation bilinguals", who have a deep understanding in both academia and industry to generate impactful research. Therefore, I spent a gap year at *Meta* as a research consultant between my Ph.D. and postdoc, and continue to collaborate with *Tencent* and *Thorlabs* to foster partnerships between academia and industry.

Academia-industry partnership. My research distinguishes itself from others by carefully aligning *assumptions* with practical situations such that the algorithms and systems are ready to be deployed for real-world impacts. For example, while working at Meta, I collected traffic traces from hundreds of real-world training jobs running on production clusters. My measurement insights led to the conceptualization of TopoOpt, a system that co-optimizes network topology and parallelization strategy for distributed training jobs [7]. My long-term collaboration with *Tencent* in optical WAN first pushed the deployment of Arrow's optical restoration [3] in production [6], then proved a long-time hypothesis that it is feasible to predict fiber cuts if optical-layer telemetry data is collected every few seconds. This insight led to a new system, PreTE, that leverages statistical prediction and hypothesis testing for WAN TE [8]. Another most recent work, MegaTE, extends TE systems to involve millions of virtual instance endpoints by flipping the TE control loop from the conventional top-down approach to the bottom-up asynchronous query [9].

Open-source optical hardware. For a long time, photonic technologies had a high barrier of entry, making this emerging technology less accessible to general researchers without optics and photonics backgrounds. Meanwhile, the promises of photonic computing and networking have attracted a lot of attentions. To bridge this gap, I made my Lightning prototype's software code and hardware design open-source on Github [1]. I hope to build a community of developers and practitioners to work together to democratize photonic technologies. Now, more than 50 researchers from worldwide institutions (e.g., *Purdue*, *ETH Zurich*, *U. Cambridge*, *UNSW*, etc.) have signed up for it (Fig. 4).

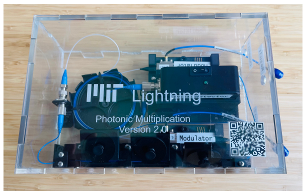


Figure 4: Lightning's open-source dev kit at https://lightning.mit.edu.

4 Future Work

Moving forward, I plan to continue my research to realize my overarching goal of transforming data movement pathways in large-scale computer systems. In particular, I aim to develop "a full technology stack" from physical-layer innovations to abstractions and automated compilers, all the way to applications and software systems for high-performance photonic-electronic systems (Fig. 5).

The need for a reconfigurable datapath controller. My work, Lightning, first scratched the surface of this topic by designing a fast datapath among NIC, photonic MAC cores, and memory [1, 2]. I plan to expand my research agenda and collaborate with colleagues in other domains (e.g., GPUs, accelerators, novel memory and storage devices) to investigate the datapath design among different units (Fig. 5). For example, a direct datapath between photonic MAC cores and GPUs will allow them to exchange data with minimal overhead to work collaboratively on the same task (e.g., a new type of parallel DNN training with hybrid hardware). With the existence of NIC-memory datapath (e.g., RDMA) and NIC-GPU datapath (e.g., GPU-Direct), I think the time has come to unify the efforts in the community and design a reconfigurable datapath controller inside the end hosts. This new datapath controller will be a foundational building block for future disaggregated systems.

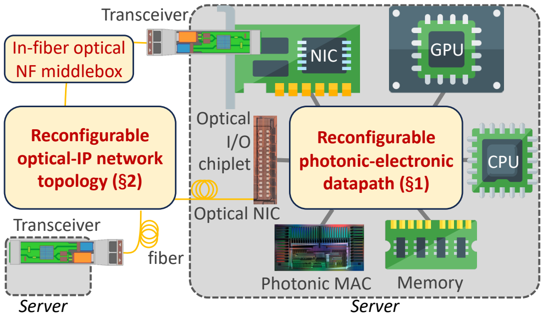

Transceiver
In-fiber optical NF middlebox
NIC
GPU
Reconfigurable optical-IP network topology (§2)
Optical I/O chiplet
Reconfigurable photonic-electronic datapath (§1)
CPU
Optical NIC
Transceiver
fiber
Photonic MAC
Memory
Server
Server

Figure 5: Rearchitecting data movement for photonic-electronic computer systems.

In-fiber optical network middleboxes. As discussed in §1 and my recent work [10, 11, 12], transceivers are the prime platform for performing photonic computing inside the network without having to replace or modify Ethernet switches. This paradigm shift promises the benefits of reduced analog-digital and optical-electrical conversions because of the capability to perform photonic computing directly on incoming optical data. I plan to build such in-fiber middleboxes that accelerates resource-demanding network functions (NFs) like intrusion detection and load balancing in the optical layer as data moves through the system.

A new optical NIC with a photonic network stack. Emerging data-intensive applications require *dynamic* and *fungible* provisioning of disaggregated compute and memory resources beyond the boundary of commodity servers. Recently, the emergence of optical I/O chiplets has enable chips to communicate directly using light. I plan to leverage this hardware technology leap and design a new network interface that allows optical signals from the OCS to directly enter the aforementioned datapath controller to access remote compute and memory resources bypassing all electronic network stacks on the NIC (Fig. 5). Therefore, a new network stack for the optical layer is needed. I will work on designing a novel *photonic network stack* (e.g., photonic header matching [14], error correction, etc.) that enables "RDMA over optical I/O" or "optically-connected far memory" without commodity NICs.

OS and compiler support for photonic-electronic systems Cross-layer design on heterogeneous systems with photonic and electronic hardware requires interoperability and reconfigurability of both parts. However, it usually requires expert knowledge in both fields to design effective abstractions and write domain-specific language for cross-layer reconfigurability, upon which developers build applications. I believe this process should and must be automated with advanced compiler technologies and novel data plane operating system (OS) support. I will work on building compilers that convert applications in high-level frameworks (e.g., PyTorch) into machine code with customized OS support that runs on emerging photonic-electronic hardware platforms.

Data movement as an explicit application-defined primitive. Most of the software frameworks and collective communication libraries are structured with compute or memory usage as its core primitive, without flexible support for data movement to be dynamically reconfigured by user applications. For a long time, data movement has been treated as underlying auxiliary infrastructure without application-level programmability. With the all the aforementioned novel hardware and systems, I think it is time to add data movement as *an explicit application-defined primitive* such that programmers can customize data movement just as GPU programming. I plan to work on developing this software library on top of hardware and make it open-source to the community.

Data-driven design automation and runtime decision-making. To match and eventually outperform state-of-the-art digital electronic platforms (e.g., GPUs), photonic-electronic computer systems need to scale up to integrate hundreds of thousands of photonic components with electronic circuits. The complexity of designing such a system is beyond human capability and requires design automation tools to facilitate smart decisions on critical design parameters (e.g., number of wavelengths, placement and routing of optical devices, etc.). I plan to address this problem by using a data-driven approach to navigate the huge design space of both electronics and photonics. at the design phase and fast reconfiguration decision making at the operation stage to build practical large-scale photonic-electronic computer systems.

Reference (* indicates equal contribution)

[1] **Zhizhen Zhong**, Mingran Yang, Jay Lang, Christian Williams, Liam Kronman, Alexander Sludds, Homa Esfahanizadeh, Dirk Englund, and Manya Ghobadi. Lightning: A Reconfigurable Photonic-Electronic SmartNIC for Fast and Energy-Efficient Inference. *ACM SIGCOMM 2023*, pages 452–472, 2023.

[2] **Zhizhen Zhong**, Mingran Yang, Jay Lang, Dirk Englund, and Manya Ghobadi. Demo: First Demonstration of Real-Time Photonic-Electronic DNN Acceleration on SmartNICs. *ACM SIGCOMM (Demos) 2023*, pages 1173–1175, 2023.

[3] **Zhizhen Zhong**, Manya Ghobadi, Alaa Khaddaj, Jonathan Leach, Yiting Xia, and Ying Zhang. ARROW: restoration-aware traffic engineering. *ACM SIGCOMM 2021*, pages 560–579, 2021.

[4] **Zhizhen Zhong**, Manya Ghobadi, Maximilian Balandat, Sanjeevkumar Katti, Abbas Kazerouni, Jonathan Leach, Mark McKillop, and Ying Zhang. BOW: First Real-World Demonstration of a Bayesian Optimization System for Wavelength Reconfiguration. *Optical Fiber Communication (OFC) Conference*, 2021. (**Postdeadline Paper**).

[5] Yiting Xia, Ying Zhang, **Zhizhen Zhong**, Guanqing Yan, Chiun Lin Lim, Satyajeet Singh Ahuja, Soshant Bali, Alexander Nikolaidis, Kimia Ghobadi, and Manya Ghobadi. A Social Network Under Social Distancing:{Risk-Driven} Backbone Management During COVID-19 and Beyond. *USENIX NSDI 2021*, pages 217–231, 2021.

[6] Congcong Miao, **Zhizhen Zhong**, Ying Zhang, Kunling He, Fangchao Li, Minggang Chen, Yiren Zhao, Xiang Li, Zekun He, Xianneng Zou, and Jilong Wang. FlexWAN: Software Hardware Co-design for Cost-Effective and Resilient Optical Backbones. *ACM SIGCOMM 2023*, pages 319–332, 2023.

[7] Weiyang Wang, Moein Khazraee, **Zhizhen Zhong**, Manya Ghobadi, Zhihao Jia, Dheevatsa Mudigere, Ying Zhang, and Anthony Kewitsch. TopoOpt: Co-optimizing Network Topology and Parallelization Strategy for Distributed Training Jobs. *USENIX NSDI 2023*, pages 739–767, 2023.

[8] Congcong Miao and **Zhizhen Zhong**. PreTE: Traffic Engineering with Predictive Failures. *MIT Technical Report (under submission)*, pages 1–17, 2023.

[9] Congcong Miao*, **Zhizhen Zhong***, Yunming Xiao*, Feng Yang*, Senkuo Zhang*, Yinan Jiang, Zizhuo Bai, Chaodong Lu, Jingyi Geng, Zekun He, Yachen Wang, Xianneng Zou, and Chuanchuan Yang. MegaTE: Extending WAN Traffic Engineering to Millions of Endpoints in Virtualized Cloud. *ACM SIGCOMM 2024*, pages 1–14, 2024.

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

[11] 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. Delocalized Photonic Deep Learning on the Internet's Edge. *Science*, 378(6617):270–276, 2022.

[12] Mingran Yang*, **Zhizhen Zhong***, and Manya Ghobadi. On-Fiber Photonic Computing. *ACM HotNets*, 2023.

[13] Alexander Sludds, Ryan Hamerly, Saumil Bandyopadhyay, Zaijun Chen, **Zhizhen Zhong**, Liane Bernstein, Manya Ghobadi, and Dirk Englund. WDM-Enabled Photonic Edge Computing. *OptoElectronics and Communications Conference (OECC)*, pages 1–3, 2022. (**Best Paper Award**).

[14] Saumil Bandyopadhyay*, **Zhizhen Zhong***, Yuqin Duan, Alexander Sludds, Ryan Hamerly, and Dirk Englund. OpTCAM: An Optical Ternary Matching System. *MIT Technical Report (in preparation)*, pages 1–8, 2023.

[15] **Zhizhen Zhong**, Jipu Li, Nan Hua, Gustavo B Figueiredo, Yanhe Li, Xiaoping Zheng, and Biswanath Mukherjee. On QoS-assured degraded provisioning in service-differentiated multi-layer elastic optical networks. In *IEEE Global Communications Conference (GLOBECOM)*, 2016.

[16] **Zhizhen Zhong**, Nan Hua, Massimo Tornatore, Jialong Li, Yanhe Li, Xiaoping Zheng, and Biswanath Mukherjee. Provisioning short-term traffic fluctuations in elastic optical networks. *IEEE/ACM Transactions on Networking*, 27(4):1460–1473, 2019.

[17] Ruijie Luo, Yufang Yu, Nan Hua, **Zhizhen Zhong**, Jialong Li, Xiaoping Zheng, and Bingkun Zhou. Achieving ultralow-latency optical interconnection for high performance computing (HPC) systems by joint allocation of computation and communication resources. *Optical Fiber Communications (OFC) Conference*, pages 1–3, 2019. (**Top-Scored Paper**).

[18] **Zhizhen Zhong**, Nan Hua, Zhu Liu, Wenjing Li, Yanhe Li, and Xiaoping Zheng. Evolving optical networks for latency-sensitive smart-grid communications via optical time slice switching (OTSS) technologies. *Opto-Electronics and Communications Conference (OECC)*, pages 1–3, 2017. (**1st Place Best Poster Award**).

[19] **Zhizhen Zhong**, Nan Hua, Massimo Tornatore, Yao Li, Haijiao Liu, Chen Ma, Yanhe Li, Xiaoping Zheng, and Biswanath Mukherjee. Energy efficiency and blocking reduction for tidal traffic via stateful grooming in IP-over-optical networks. *Journal of Optical Communications and Networking*, 8(3):175–189, 2016.

[20] **Zhizhen Zhong**, Nan Hua, Yufang Yu, Zhongying Wu, Juhao Li, Haozhe Yan, Shangyuan Li, Ruijie Luo, Jialong Li, Yanhe Li, and Xiaoping Zheng. Throughput scaling for MMF-enabled optical datacenter networks by time-slicing-based crosstalk mitigation. *Optical Fiber Communication (OFC) Conference*, pages M2E–5, 2018.