JOURNAL OF
Optical Communications
and Networking

PINE: Photonic Integrated Networked Energy efficient datacenters (ENLITENED Program) [Invited]

Madeleine Glick,[1,*] Nathan C. Abrams,[1] Qixiang Cheng,[2] Min Yee Teh,[1] Yu-Han Hung,[1] Oscar Jimenez,[1] Songtao Liu,[3] Yoshitomo Okawachi,[1] Xiang Meng,[1] Leif Johansson,[4] Manya Ghobadi,[5] Larry Dennison,[6] George Michelogiannakis,[7] John Shalf,[7] Alan Liu,[8] John Bowers,[3] Alex Gaeta,[1] Michal Lipson,[1] and Keren Bergman[1]

[1]Columbia University, New York, New York 10027, USA
[2]University of Cambridge, Cambridge, UK
[3]Electrical and Computer Engineering Department, University of California, Santa Barbara, California 93106, USA
[4]Freedom Photonics, LLC, Santa Barbara, California 93117, USA
[5]CSAIL, MIT, Cambridge, Massachusetts 02139, USA
[6]NVIDIA Corp., Santa Clara, California 95051, USA
[7]Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA
[8]Quintessent, Santa Barbara, California 93105, USA
*Corresponding author: msg144@columbia.edu

Received 14 July 2020; revised 28 October 2020; accepted 29 October 2020; published 30 November 2020 (Doc. ID 402788)

We review the motivation, goals, and achievements of the Photonic Integrated Networked Energy efficient datacenter (PINE) project, which is part of the Advanced Research Projects Agency–Energy (ARPA-E) ENergy-efficient Light-wave Integrated Technology Enabling Networks that Enhance Dataprocessing (ENLITENED) program. The PINE program leverages the unique features of photonic technologies to enable alternative mega-datacenters and high-performance computing (HPC) system architectures that deliver more substantial energy efficiency improvements than can be achieved through link energy efficiency alone. In phase 1 of the program, the PINE system architecture demonstrated an average factor of 2.2× improvement in transactions/joule across a diverse set of HPC and datacenter applications. In phase 2, PINE will demonstrate an aggressive 1.0 pJ/bit total link budget with high-bandwidth-density dense wavelength-division multiplexing (DWDM) links to enable additional 2.5× or more efficiency gains through deep resource disaggregation.

https://doi.org/10.1364/JOCN.402788

1. INTRODUCTION

The recent explosive growth in data analytics applications that rely on machine learning techniques is leading to a convergence between datacenters and high-performance computing (HPC) systems that are driving an intensely growing need for compute performance. The effective execution performance efficiency of these massive parallel architectures is directly affected by how data moves among the numerous compute, memory, and storage resources, and is dramatically impacted by the growing energy consumption associated with data movement, as well as utilization efficiency of heterogeneous compute and memory resources.

These challenges have motivated the research in the *Photonic Integrated Networked Energy efficient datacenter* (PINE) project, which is part of the Advanced Research Projects Agency–Energy (ARPA-E) *ENergy-efficient Light-wave Integrated Technology Enabling Networks that Enhance Dataprocessing* (ENLITENED) program. The PINE architecture addresses the data movement challenge by leveraging the unique properties of photonics to steer bandwidth to where it is needed rather than over-provisioning network resources, which significantly increases energy consumption. Photonics can also be used to efficiently perform resource disaggregation. PINE is built upon three pillars:

(1) **Disaggregated PINE Architecture**. Our proposed PINE datacenter architecture maximizes the benefits provided by seamlessly integrating low-power silicon photonic (SiP) links and large numbers of embedded low-radix broadband photonic circuit switches to enable inter-multi-chip module (MCM) connectivity and reconfigures them within the same rack for high-performance and low overhead rack-scale resource disaggregation. The sharing of

resources presents an abstract concept of the datacenter as a pool of disaggregated resources that can be reallocated at fine granularity to prevent applications from being bottlenecked on a particular resource type, as well as to prevent underutilization of resources.

(2) **Photonic MCMs with Ultra-Low-Power High "Chip Escape" Bandwidth Density.** System designers find themselves in a narrow box with memory and input/output (I/O) packaging. Running the I/O pins at higher bandwidth incurs a power cost. Many CPU/GPU cores are intrinsically capable of carrying extremely demanding computing tasks, but they do not have the necessary off-chip bandwidth for full and efficient utilization of their resources. The PINE embedded high-bandwidth-density flexible photonic connectivity realized in the MCM active interposer platform enables multi-Tbps chip escape bandwidths.

(3) **Energy-Optimized Dense Wavelength-Division Multiplexing Photonic Connectivity**. The PINE dense wavelength-division multiplexing (DWDM) optical links build on a new generation of components specifically optimized for energy efficiency [1–3]. Our novel light source platform composed of a single, high-power, high-efficiency laser coupled [4] to a multiple wavelength comb generator is used to allocate power for >50 wavelengths.

PINE research is being conducted over two phases, with phase 1 having been completed in 2019 and phase 2 commencing in 2020. Our next-generation PINE phase 2 full system solution builds on phase 1 successes to perform full system-level integration consisting of photonic interconnected MCMs with switching flexibility demonstrating the scaled PINE architecture datacenter prototype under realistic workloads. PINE phase 2 will target further photonic link energy reductions toward 1 pJ/bit, demonstrate MCM chip edge bandwidth densities >5 Tb/s/mm, and reduce the system-wide energy consumption of datacenters by a factor of 2.8×. Phase 2 has an increased focus on deep intra-node disaggregation and in particular artificial intelligence (AI) data analytics applications, which are projected to deliver 5× accelerated execution performance measured by traversed edges per second (TEPS) per watt. Phase 2 PINE will also extend cost-effective supply chain options and commercialization paths for the PINE energy-efficient high-bandwidth-density links and the MCM integration supporting the disaggregated PINE system architecture. The PINE architecture is designed to support diverse emerging data-intensive workloads while optimizing energy efficiency. The overall PINE architecture is summarized in Fig. 1.

Here we provide a review of the PINE program. We start in Section 2 with an overview of the PINE system architecture, bandwidth steering, and our approach to disaggregation. In Section 3, we describe details of our optically interconnected MCMs and SiP switch and high-density couplers. We follow with Section 4 focusing on our DWDM SiP links, simulations, and novel components. Finally, we summarize our conclusions in Section 5.

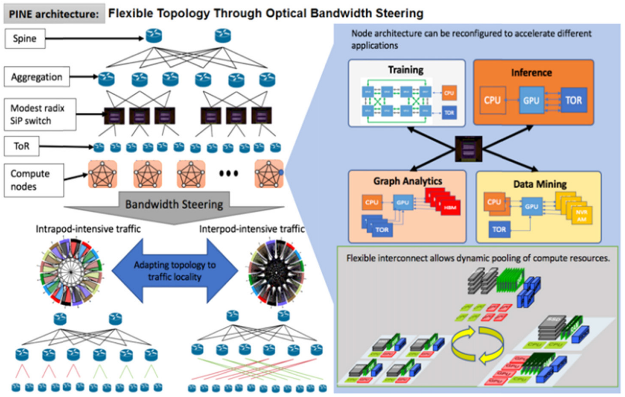
PINE architecture:
Flexible Topology Through Optical Bandwidth Steering
Spine
Aggregation
Modest radix SiP switch
ToR
Compute nodes
Bandwidth Steering
Intrapod-intensive traffic
Interpod-intensive traffic
Adapting topology to traffic locality
Node architecture can be reconfigured to accelerate different applications
Training
CPU
TOR
Inference
CPU
GPU
TOR
Graph Analytics
CPU
GPU
HBM
TOR
Data Mining
CPU
GPU
NVRAM
TOR
Flexible interconnect allows dynamic pooling of compute resources.

Fig. 1. PINE system architecture.

2. PINE SYSTEM ARCHITECTURE

A. PINE Phase 1 Architectural Design and Evaluation

In phase 1, the PINE system architecture focused on system-wide bandwidth steering using a scaled flexible fat tree that demonstrated an average factor of 2.2× improvement in transactions/joule (55% reduction in power per fixed size transaction) and a 20% reduction in average network latency across a diverse set of applications [5]. The fundamental building blocks of PINE's system-wide bandwidth steering are reconfigurable SiP switches. The PINE MCM 2.5D and 3D assembly processes are being developed in the interposer and 3D active interposer platforms in which we have demonstrated high-density electronic/photonic integration [6–8] and the first multi-layer photonic switch platform, which achieved record low cross talk and extinction of >50 dB [9,10].

Our fully integrated comb laser with experimentally demonstrated 41% power pump conversion drives the PINE phase 1 links [11]. The PINE architecture extends scalability with high-efficiency semiconductor optical amplifiers (SOAs) and the first monolithic quantum dot (QD) SOA on silicon demonstrated record wall-plug efficiency (WPE) of 14.2% and on-chip gain of 39 dB [12]. The MCM platform also advanced passive alignment high-density optical fiber chip-IO with demonstrated robustness to temperature and fabrication variations while maintaining a penalty of less than 0.6 dB on the coupling efficiency [13,14].

Phase 1 included a comprehensive cross-layer modeling and energy/performance analysis at the link, node, and full system levels. PINE phase 1 demonstrated an energy-optimized high-bandwidth-density SiP link design with aggregate bandwidth of 640 Gb/s, utilizing a two-stage deinterleaver to divide the wavelength-division multiplexing (WDM) channels into four subgroups, each with 10 cascaded ring resonators modulated at 16 Gb/s (4 × 10 × 16 Gb/s). The link is then operating at 2.2 pJ/bit utilizing the AIM Photonics PDK, custom designed coupler, and the ASIC driver/receiver designed specifically by our own team. The phase 1 receiver is operating at a sensitivity of −17 dBm with 0.75 pJ/bit. PINE phase 2 will build on these successes to further reduce the energy consumption of the receivers by 80%. The phase 2 receiver will be operating at an improved sensitivity of −22.5 dBm, which further reduces the required input laser power. Together with the improved

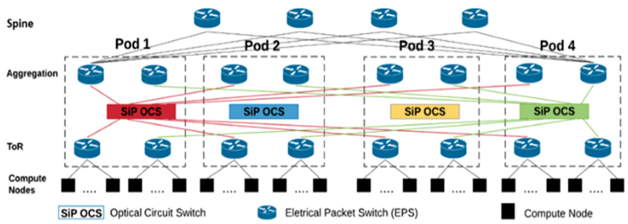

Fig. 2. Bandwidth steering in the flexible fat tree PINE architecture. A reconfigurable fat tree topology enabled by the placement of SiP optical circuit switches (OCSs) between the top-of-rack (ToR) switches at the first level of the topology and aggregation switches (second level) of different pods [5].

custom designed heaters and optical couplers, the phase 2 link will have a reduced energy aiming toward a goal of 1 pJ/bit. In addition, PINE phase 2 will perform system-level integration consisting of photonic interconnected MCMs with switching flexibility demonstrating the PINE architecture under realistic workloads. PINE phase 2 will target photonic links with >1 Tb/s aggregate bandwidth, demonstrate MCM chip edge bandwidth densities >5 Tb/s/mm, and reduce the system-wide energy consumption of datacenters by a factor of 2.8×.

1. *PINE System-Wide Bandwidth Steering*

During phase 1, we developed system-wide bandwidth steering to seamlessly reconfigure the interconnect topology to match diverse workload communication patterns [5–8]. We apply bandwidth steering to keep more traffic at the lower levels of the topology and show that high-level bandwidth can be substantially reduced (taper the connections) with no performance penalty. Effectively, by changing the connectivity of the lower levels, more traffic can directly go to its destination aggregation (middle-level) switch without having to use the top level of the topology. We illustrate the physical topology of our approach in Fig. 2. We only apply bandwidth steering to uplinks as that is sufficient to ensure that traffic using steered connections does not use the top layer of the fat tree. In our 32-node four 4 × 4 SiP switches PINE flexible fat tree experimental testbed (explained in detail below) running the HPC gyrokinetic toroidal code (GTC) trace, we demonstrated a 62% application execution acceleration using bandwidth steering.

System-level simulations of scaled bandwidth steering in the PINE architecture are performed in our cycle-accurate network simulator Booksim [15], modified to implement bandwidth steering. We use HPC application traces identified by the U.S. Department of Energy (DOE) Exascale initiative [16] and publicly available traces from a Facebook production-level database pod [17]. Placement of tasks on network end points is randomized and combined across different applications. This simulates the effects of fragmentation when multiple diverse applications are sharing the network. We pre-configure the network for each simulation and later demonstrate in our experimental testbed the control plane for reconfiguration. The system-level power and performance models include 16 × 16 PINE photonic switches and electrical 36-port Mellanox 100 Gb/s InfiniBand routers with active optical cables [18] and are sized to match the application traces. Figure 3 summarizes the throughput and latency improvements from bandwidth steering. As shown, on average, bandwidth steering improves sustained average network throughput by 1.7× and average network latency by 20%. Throughput improvements demonstrate that bandwidth steering recovers throughput lost in the baseline topology due to non-ideal per-packet load balancing. In addition, bandwidth steering increases throughput in a cost-effective manner without having to over-design for the worst case because any baseline can provide full throughput by increasing link and switch bandwidth. Finally, bandwidth tapering magnifies the benefit of bandwidth steering. More details can be found in [5].

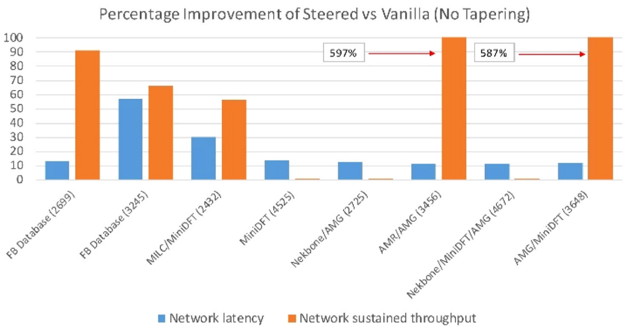

Fig. 3. Average system throughput improvement of the steered fat tree compared to a vanilla fat tree with no tapering (reduction) of top-layer bandwidth [5].

Higher network throughput means that the network can handle proportionally more transactions per second for approximately the same static power (same network components). The SiP optical switch reconfiguration is performed every time an application initiates or terminates in the system. HPC benchmarks typically generate persistent traffic patterns that change slowly [19]; for example, in the National Energy Research Scientific Computing Center's (NERSC) Cori system, multi-node applications are initiated every 17 s on average, much slower than the capabilities of our SiP optical switches, which reconfigure in 20 µs. A key insight is that our efficient SiP optical switches, once configured, impose negligible dynamic power and latency.

The bandwidth tapering enabled by PINE at the top level of the fat tree does not incur a performance penalty and directly increases transactions per second for the same power envelope. Our phase 1 results demonstrate an overall average factor of 2.2× improvement in transactions/joule (55% power consumption per fixed size transaction reduction) across the diverse set of HPC/datacenter applications. Further benefits are gained for communication bandwidth-bound applications, as bandwidth steering directly speeds up application execution time by an average of 1.7×, and thus reduces compute resource idle power by the same factor. More information can be found in [5].

Building on these successful results, phase 2 efforts will address the benefits of advanced reconfigurable network design [20], of multi-workload defragmentation, and a finer grained temporal reconfiguration. Different physical connectivity schemes between electronic and SiP optical switches will be considered and evaluated in the PINE system testbed. Phase 2 will take the bandwidth steering concept deeper into the disaggregated rack.

B. PINE Phase 2: Architecture for Deep Disaggregation

The PINE phase 2 architectural design (Fig. 4) further disaggregates key elements of the traditional datacenter or HPC compute nodes or servers and reorganizes them around a reconfigurable network fabric, conceived to specifically address the stress placed on the system by real-time communication-intensive applications. Embedded photonic switching within the interconnection network steers bandwidth on demand among diverse resources. Deep disaggregation is realized through ultra-high-density assembly of energy-efficient photonic links with GPUs, CPUs, and memory elements in an MCM interposer platform around a unified and reconfigurable photonic network fabric. The MCM interconnects build on PINE ultra-low-power photonic link technologies with efficient comb laser sources to dramatically reduce the energy consumption and increase bandwidth densities system-wide. With flexible interconnectivity, PINE can assign datacenter/HPC resources to workloads with exquisite temporal and size accuracies so that only the required amount of computation power, memory capacity, and interconnectivity bandwidth are made available over the needed time period. This efficient usage of resources reduces the vast amounts of wasted energy consumption of current datacenters, increases return on investment, and simultaneously accelerates time to completion of HPC applications due to more efficient communication between resources allocated to the same application.

Datacenter and HPC workloads show a large diversity in their resource demands: training algorithms for deep learning places stress on compute and interconnect elements and sometimes creates rigid communication patterns, in-memory databases place stress on integrated non-volatile storage bandwidth, and data-intensive analytics place stress on memory capacity and bandwidth. Contemporary architectures strongly compartmentalize storage, interconnect, and memory resources in individual servers that prevent resources from being traded against each other. These inflexible architectures are not able to meet diverse resource requirements for different parts of the workload—forcing datacenter and HPC operators to over-provision system elements that are inefficiently overutilized/underutilized rather than optimized for the task. The PINE architecture builds on numerous, strategically arranged low-to-medium radix optical circuit switches to steer bandwidth on demand. This innovative approach essentially delivers the *optimized connectivity* to the application, thus eliminating over-provisioned energy wasting idle resources. Importantly, the PINE architecture requires only low-to-medium radix switches for low-loss/low-power insertion.

1. Node Level Deep Disaggregation

Existing server chip designs are hard-wired to particular resources. Therefore, they offer virtually no ability to re-provision IO/memory bandwidth to meet application demands. For example, machine learning applications require at least a 3:1 shift in IO/memory bandwidth from inter-GPU connectivity (for training) to off-chip bandwidth (for inference) or else face a significant energy efficiency penalty either through lower application performance or bandwidth over-provisioning. PINE's deep disaggregation approach will deliver this 3:1 shift in the bandwidth/connectivity balance and develop new algorithms to guide reconfiguration of the photonic MCM switch fabric to configure custom nodes from disaggregated resources to meet diverse application demands. Bioinformatics and graph applications require even greater—10:1 or larger—shifts in bandwidth provisioning to meet application requirements. In phase 2, PINE deep disaggregation will deliver a 10:1 or more dynamic range for reconfiguring/rewiring nodes, which would in turn deliver a >5× performance-per-watt advantage for applications with

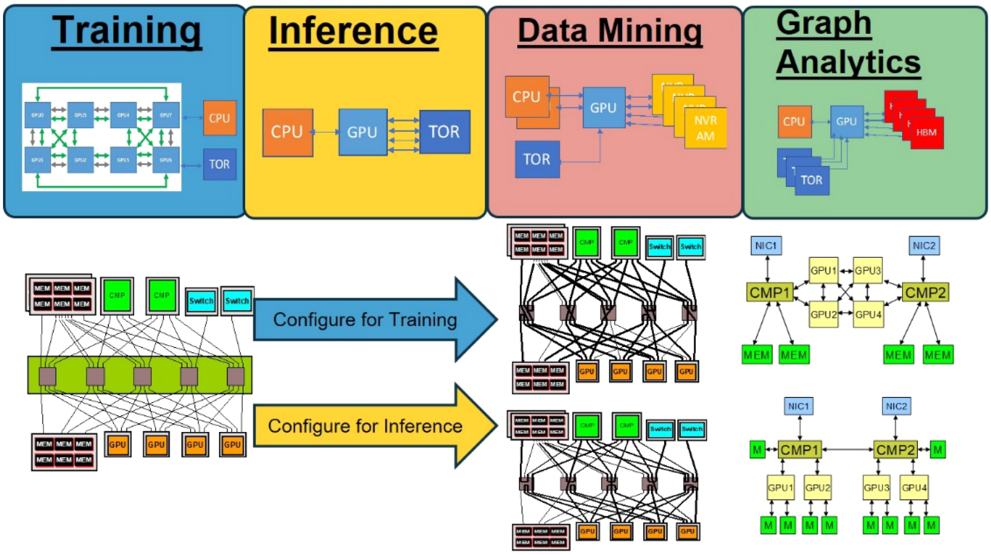

Fig. 4. Schematic of deep disaggregation and customized node configurations to support diverse workloads [49].

diverse node resource demands. We quantify these gains using the TEPS per watt metric.

The PINE photonic interconnected MCM node uniquely enables a flexible "photonic fabric" that can strategically rewire disaggregated components to form "nodes" on-the-fly to meet diverse workload requirements. Indeed, it blurs the boundary between the node and the rack. In phase 2, we will focus on increasing bandwidth flexibility. Figure 4 shows schematically how the four canonical workloads have different node organizations to support their requirements. Deep disaggregation enables custom node configurations to be created at job start-up to support those diverse workload requirements with energy-optimized connectivity.

Phase 2 will focus on the algorithmic approach to tune the energy-performance optimization of the workloads. We will develop the intra-node photonic switched flexible connectivity architecture as well as outline the control plane necessary to make runtime configuration possible in a scalable manner inside a rack. For instance, the different phases of machine learning applications can be dynamically assembled to deliver needed GPU/memory connectivity bandwidths. A common approach to distributed training is data parallelism where the training data are distributed across multiple workers (e.g., GPU, TPU, CPU). In data parallel training, workers need to communicate their model parameters after each training iteration. This can be done in a variety of ways, including parameter servers, ring-allreduce, tree-reduce, and hierarchical all-reduce. However, there is a rapid increase in model and pipeline parallelism training, motivated by the rapid increase in the computation and memory requirements of neural network training. The size of deep learning models has been doubling about every 3.5 months. Many models, such as Google's Neural Machine Translation and NVIDIA's Megatron, no longer fit on a single device and need to be distributed across multiple GPUs. To train such models, model parallelism (and hybrid data-model parallelism) approaches partition the model (and data) across different workers. Model parallelism is an active area of research, with various model partitioning techniques. For example, pipeline parallel approaches, such as PipeDream [21], GPipe [22], and DeepSpeed [23], have emerged as a sub-area of model parallelism. Recent work explores optimizing over a large space of fine-grained parallelization strategies [e.g., parallelizing each operator in a deep neural network (DNN) computation graph separately], demonstrating an increase in training throughput of up to 3.8×. We posit that all of these approaches will benefit from our PINE platform with high network bandwidth.

Another part of our study is the reconfigurability possible as a function of cost and complexity of the enabling optical components and the control plane. Since we are focusing on rack-level disaggregation, we expect that we can demonstrate significant benefits with low to modest complexity using low-radix optical switches and simple control planes.

With our industry partners, we will be able to evaluate potential supply chain options for practical deployments to deliver high-value benefits of bandwidth steering and deep disaggregation in datacenter and HPC applications.

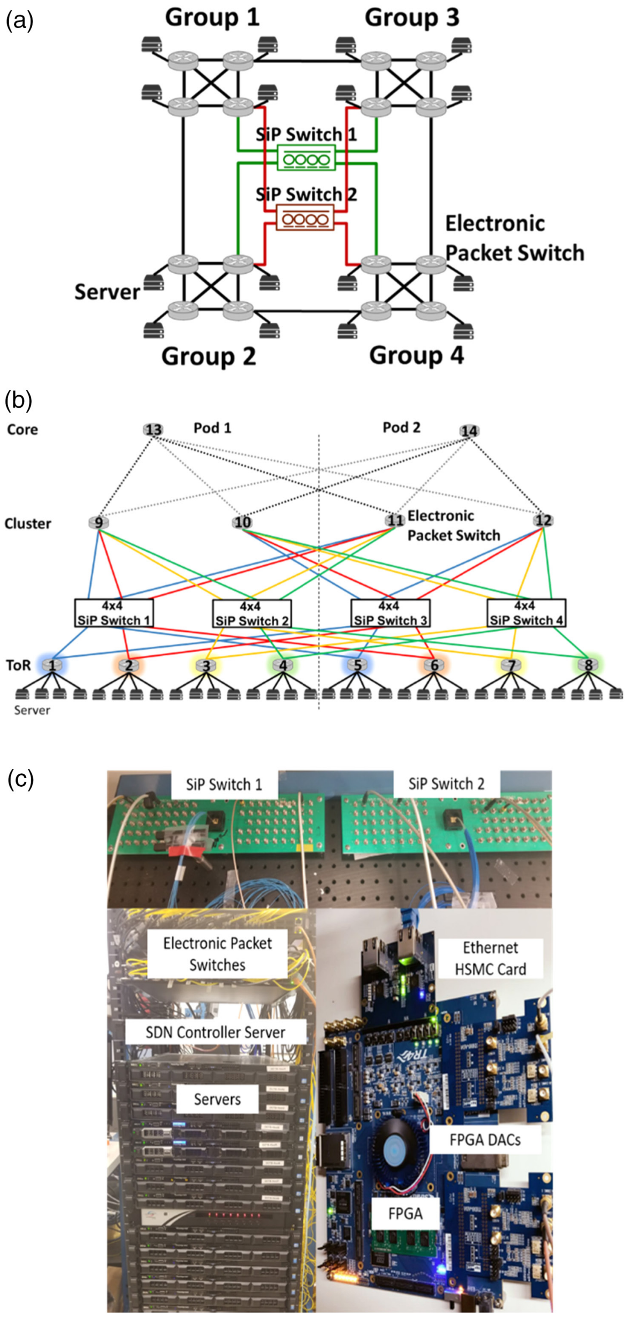

Fig. 5. Physical testbed networks: (a) dragonfly, (b) fat tree, and (c) implemented hardware [5].

C. PINE Experimental System Implementation

To explore the feasibility of the PINE bandwidth steering concept using SiP switches, we built an HPC/datacenter testbed that integrates our SiP switches with a traditional electronically packet switched environment composed of commercial servers and packet switches, as well as field-programmable gate arrays (FPGAs) and software-defined networking (SDN) controller servers to implement reconfiguration [5]. On this testbed, we can perform any functionality of an ordinary computing system, including bulk data transfer, virtual machine (VM) migrations, and HPC benchmark applications. By integrating the SiP switches into the network and performing bandwidth steering to optimize the network topology, phase 1 demonstrated significant performance improvements through reduction in the total execution time of the application compared to running on traditional network topologies, which translates to a proportional improvement in energy consumption. The 32-node PINE HPC/datacenter testbed (Fig. 5) includes four 4 × 4 SiP switches and can be configured in both a dragonfly topology [24] and a fat tree topology [25].

The PINE phase 1 system testbed evaluated the bandwidth steering performance of the GTC HPC benchmark application and showed significant performance improvements in execution time acceleration by over 62% with tapering compared to the standard fat tree. In phase 2, the testbed will be substantially scaled to 64 nodes and extended to more accurately reflect real deployed computing systems. Once an application is placed on the physical nodes, we expect the optimal connectivity based on the expected traffic pattern to change. Therefore, in phase 2, we propose reconfiguring topology in the order of tens of seconds, which is the average arrival time for a new application [5]. During each reconfiguration epoch, the network controller will derive an expected traffic pattern based on the current job placements in the network, and realize the logical connectivity such that the overall topology is optimized for the expected traffic pattern. Insertion of PINE optically interconnected MCMs with photonic switching capabilities will be performed for validation. The scaled system will allow more realistic traffic experiments on the impact of link congestion, latency, and performance across a broad range of HPC, machine learning applications using built-in GPUs within the servers, and datacenter traffic.

3. OPTICALLY INTERCONNECTED MULTI-CHIP MODULES

A. MCM Interposer

In phase 1 we developed MCM transceivers to provide tight integration of the photonics with the driving electronics. Several different MCM prototypes were developed in 2.5D and 3D interposer platforms [Figs. 6(a) and 6(b)]. The active interposer platform developed jointly with SUNY Poly uniquely combines the photonic integrated circuit (PIC) and interposer into a single integrated substrate [Fig. 6(c)]. This approach can deliver the best electronic/photonic densities and will become a main path for phase 2 integration of the link and switch photonics with the custom electronic integrated circuits (EICs). The transition from 2.5D to the 3D active interposer will be a joint effort between Columbia/SUNY Poly and NVIDIA.

The active interposer combines the functionalities of the PIC and interposer into a single die, allowing photonic components to be fabricated and directly integrated with through silicon vias (TSVs) and additional metal redistribution layers to allow connectivity on both the front and back side of the active interposer. In addition to the interposer layers, the active interposer will contain all of the features found in the PIC process, allowing fabrication and routing of active and passive photonic devices. The top side of the active interposer will be used to provide connectivity to the EIC complementary metal–oxide–semiconductor (CMOS) driver chips. The EICs will be flip-chipped on top of the active interposer using copper pillars. Copper pillars will be grown at wafer scale on both the active interposer and EICs. The top side of the active interposer will also provide a platform for integration into a compute node, such as a CPU, GPU, memory, or VLSI chip. The backside of the active interposer will be reserved for ball grid array (BGA) connections to the printed circuit board (PCB). Figure 6(d) shows an example of an 8 × 8 network-on-chip implemented in the active interposer platform. The fully packaged MCM transceiver prototype, mounted on the fanout PCB, is shown in Fig. 7(a), the MCM transceiver utilizing a passive silicon interposer for 2.5D integration is shown in Fig. 7(b).

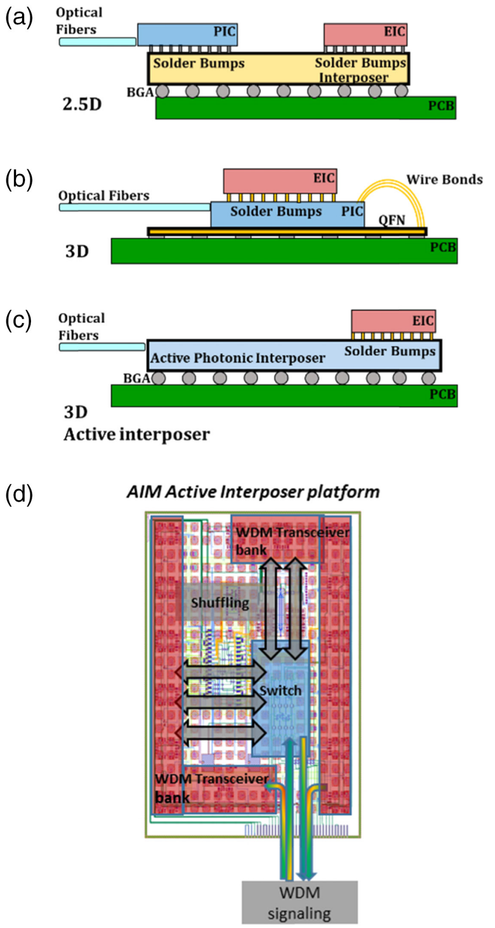

Fig. 6. (a)–(c) 2.5D and 3D MCM integration approaches explored in phase 1; (d) 8 × 8 network-on-chip implemented in the active interposer platform.

B. MCM Photonic Switch Fabric

Embedding the photonic switch fabric within the MCM platform provides the optical domain reconfigurability for the PINE deep disaggregation. In phase 1, we leveraged the AIM active interposer platform and developed the first prototype of an optical network-on-chip, which monolithically integrates an 8 × 8 spatial switch with four WDM transmitters, four WDM receivers, and EICs flipped on top [Fig. 6(d)]. In phase 1, we further designed, fabricated, and packaged

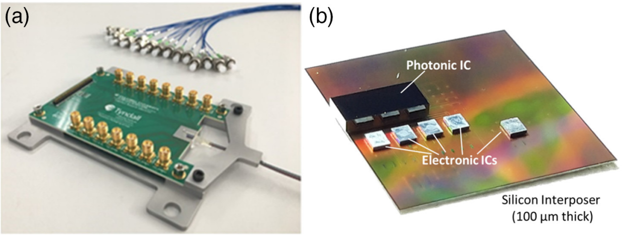

Fig. 7. (a) Fully packaged MCM transceiver prototype, mounted on the fan-out PCB; (b) MCM transceiver utilizing a passive silicon interposer for 2.5D integration.

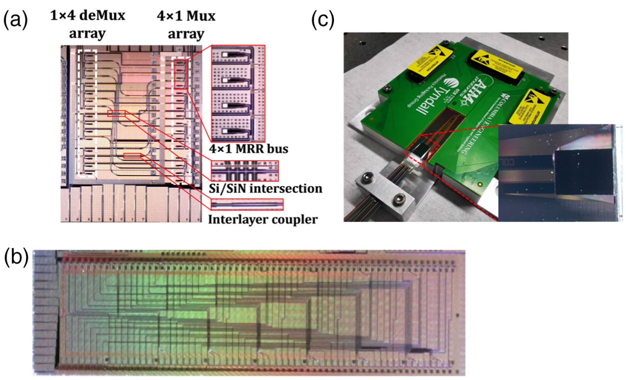

Fig. 8. (a) Microscope photo of the fabricated device with insets of the enlarged 4 × 1 MRR-based spatial multiplexer, the Si/SiN intersections, and the interlayer coupler. (b) Microscope photo of the taped out 8 × 8 triple-layered switch. (c) Packaged switch with a silicon interposer.

the first strictly non-blocking optical switch in a switch-and-select (S&S) topology using microring resonators [9,10], as shown in Fig. 8(a). The unique Si/SiN multi-layered S&S topology demonstrated breakthrough switch cross talk suppression and extinction ratio of >50 dB and on-chip loss as low <1.8 dB. We have also taped out a triple-layered 8 × 8 microring switch in the same topology with the microscope image shown in Fig. 8(b). The switch chip was packaged using a silicon interposer that can be readily embedded into the MCM platform.

In phase 2, we will scale the topology to support >100 × 100 connectivity with microring resonators and develop bandwidth steering in both the spatial and spectral domains, for increased flexibility [26]. The switch architecture comprises two sections: (1) N switching planes of an N × M λ crosspoint matrix for spatial wavelength selection and (2) N comb wavelength aggregators of free-spectral range (FSR)-matched microring arrays, as illustrated in Fig. 9. Each N × M λ switching plane consists of colored arrays of microrings in N rows and (M + 1) columns, in which the wavelength selection is handled by the first row while the space selection is actuated by the rings in each column. The comb aggregators apply FSR-matched large ring elements. This will enable a scalable switch fabric that combines switching in the space domain with wavelength-selectivity to define fine-grained connectivity for node disaggregation in both the physical port and wavelength channel.

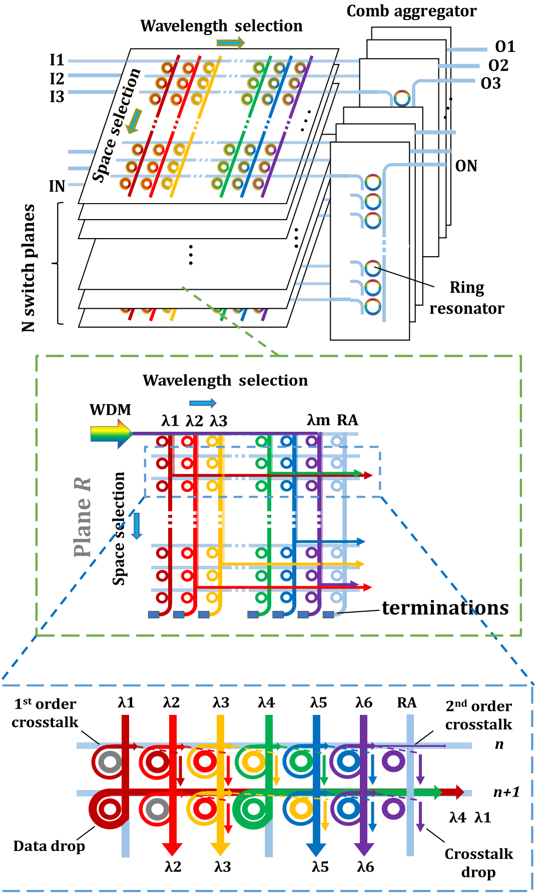

Fig. 9. Proposed space-and-wavelength switch design using arrays of microring-based wavelength selectors and comb aggregators. Insets show the operating principle of the N × Mλ crosspoint matrix and how it fully blocks the first-order in-band cross talk.

C. Ultra-Low Energy Electronics

In a joint Columbia/NVIDIA effort under phase 2, ultra-low energy EICs will be custom designed and implemented to directly interface with the PICs. These advanced EICs will include drivers for the transmit modulator array, transimpedance amplifiers (TIAs) for the receivers, and heater control circuits and will be fabricated at the TSMC in the 16 nm FinFET process using a multi-project wafer (MPW) run. We will explore circuits for both lower speed optical channels (around 10–16 Gb/s) and higher speed optical channels (around 25 Gb/s) with the goal of optimizing across the entire system architecture. Based on phase 1 exploration, lower speed channels give better energy/bit although higher speeds result in higher bandwidth density. The 16 nm FinFET process is ideally suited for these speeds and will provide a straightforward path for future technology transfer. We plan two tape-outs of the test chip—one in year 1, and one in year 2. Refinements will be made based on the measurement results of the first tape-out.

The most energy-efficient electrical communication links are those that can be implemented with the simplest possible architectures and circuitry. Bundled-data, clock forwarding

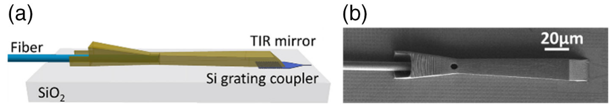

Fig. 10. (a) Schematic of the robust plug-and-play coupler; (b) scanning electron microscopy image of our fabricated devices.

architectures are thus almost universally employed in energy-efficient short-reach electrical interconnect. In our overall system architecture, we will forward a shared clock for each group of eight optical data channels and use this clock to sample data at the receiver avoiding expensive clock distribution and clock recovery circuits. We have extensively researched similar clock forwarding techniques for short-reach electrical links [27–30]. This clock forwarding technique allows for very simple transmit circuitry in which the data is serialized, clocked on a common clock, and sent to the modulator driver. On the receive side, the data from the TIA receiver is sampled using the forwarded clock and deserialized. We will implement these circuits along with data generator and checker circuits on an EIC test chip that will be packaged with the PICs. The combined EIC/PIC active interposer will be packaged on an organic substrate and mounted on a PBC.

D. Robust High-Density Optical IO

In phase 1, we developed a 3D photonic structure for a robust, passive, and simultaneous mechanical and optical coupling between single-mode fibers (SMFs) and integrated waveguides [13,14]. The 3D structure, seen in Fig. 10, consists of a polymer funnel that routes the incoming fiber (thinned on one end) directly to the facet of a polymer waveguide. The fiber is routed into the funnel independently of its exact position relative to the center of the funnel (20 μm diameter), enabling extremely wide alignment tolerance. When fully inserted, the fiber is tightly held at the smaller end of the funnel routing section. At this point, the fiber and the polymer waveguide are perfectly aligned, optically coupled, and mechanically held in place. Their optical modes are designed to spatially match in order to obtain a high coupling efficiency. The coupler fabrication is done using a 3D two-photon nanoprinting tool to polymerize an epoxy-based photoresist on top of fabricated waveguide chips. Standard optical fibers are thinned down on one end to 10 μm (a process that is also available commercially) and then directly inserted into the funnel.

Working with these thin fibers allows for high-density packaging of couplers with low footprint arrays (down to 30 μm pitch) as in Fig. 11. Our results show that our 3D funnel coupler only exerts a 0.05 dB penalty, which means our plug-and-play coupler has a minimal effect on the device coupling efficiency while allowing compatibility with standard automated alignment tools.

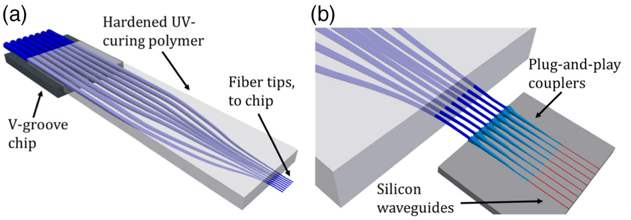

Fig. 11. (a) Schematic diagram of the thin fiber array for eight fibers with a 30 μm pitch. (b) Fiber array coupling to an array of eight plug-and-play couplers written directly on the photonics chip.

4. DWDM SILICON PHOTONIC LINKS

A. Energy-Throughput Optimized Design

In phase 1, we developed unique simulation and modeling tools of the PINE links in PhoenixSim [31]. The PhoenixSim environment is built to enable maximization of system performance with optimized energy consumption [32] and can be used for cross-layer physical-parameter photonics design from ring radii, to channel bit rates and modulation formats. We have performed comprehensive cross-layer modeling and energy/performance analysis for energy-efficient DWDM SiP links using components such as multi-wavelength comb sources, modulators, filters, photodetectors, and electrical integrated components [Fig. 12(a)].

A key advance of the PhoenixSim design platform under phase 1 is the full integration with the Synopsys Photonic Solutions, a commercial simulation software for photonic design and fabrication. Such integration allows for flexible design including multi-physics effects from the photonic components including detailed structural geometry, doping levels, and the frequency response of the modulator. The integration with Synopsys tools provides a direct path from design to fabrication and commercialization. The link model designed in PhoenixSim, based on foundry PDK and custom designed components, can directly generate layout GDS files for fabrication [32].

The link architecture for phase 2, shown in Fig. 12(b), is sourced by a DWDM comb propagated on-chip where the comb lines are deinterleaved into four groups passed to the cascaded microring resonator modulator array. Our design for ultra-low-power circuitry leverages the strong dependence of energy consumption on the ratio between data rate and transistor transit frequency, with lower per-channel data rates yielding more than proportional power savings since the scaling of power dissipation is strongly supra-linear. This

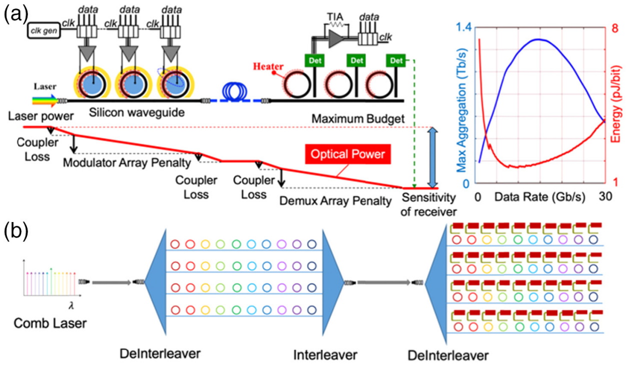

Fig. 12. (a) Photonic link models available in the PhoenixSim software tool. (b) Ultra-low-power DWDM link architecture with a comb laser source. DWDM comb lines are deinterleaved into four groups, and each group has cascaded microring-based modulators.

approach enables highly sensitive receivers and allows for cross-optimization with the optical signal quality, e.g., minimizing the required laser power, in the photonic link [33].

The optical link budget includes anticipated coupling losses, component insertion losses, WDM channel cross talk at the given receiver sensitivity, and the required margin and BER for the non-return-to-zero (NRZ) modulation format [34,35]. The comb laser is set to have an overall 8% WPE, with 20% pump laser efficiency and 40% comb conversion efficiency.

The energy consumption of the photonic link depends on the laser source, modulators, receivers, and their associated electronics. The link budget is architecture-based, which is shown in Fig. 12(a). There is a trade-off between the data rate of individual modulators and the photonic link architecture to achieve the optimal power efficiency. For example, to maintain the target aggregate bandwidth of 640 Gb/s, the photonic link with modulators at a data rate that is lower than 10 Gb/s has a much more complex architecture and thus higher propagation loss. On the other hand, a link with modulators at a data rate above 20 Gb/s leads to simpler architecture, but the energy required for modulation and detection becomes dominant. Based on our PhoenixSim simulation, the data rate at 16 Gb/s leads to the optimal power efficiency for the given link architecture shown in Fig. 12(b).

B. DWDM Integrated Comb Laser

Recent developments of microresonator-chip-based frequency combs offer the prospect of controllably generating many single-frequency components that are evenly spaced to ultra-high precision [36]. Such a source is ideal for WDM applications because the spacing of all of the frequency components can readily be fixed to a specified frequency grid by stabilizing the microresonator, which the Columbia group has demonstrated at low heater powers with an integrated microheater [37]. Such an approach is in contrast to using an equal number of single-frequency laser sources in which each laser must be stabilized to maintain its frequency on the grid, which adds substantial complexity and required power. Our efforts during phase 1 included (1) successful development of an integrated comb source, (2) realization of high pump-to-comb conversion efficiency using the silicon nitride (SiN) platform, and (3) full integration into a transceiver module as shown in Fig. 13.

Previously, the Columbia team has demonstrated the integration of a III–V RSOA with a high Q SiN microresonator for a fully integrated comb source based on dissipative Kerr solitons [38]. However, soliton Kerr combs suffer from low pump-to-comb conversion efficiencies [39,40] and the sech^2 spectral profile resulting in an exponential falloff of the power in the comb lines away from the pump. To increase the comb source power-per-line, we plan to build on our recent advances with comb formation in the normal group-velocity dispersion (GVD) regime [11,41]. Such combs take advantage of a localized change in the microresonator cavity dispersion, which can be achieved through mechanisms such as pump modulation and mode interactions [42]. Mode interactions between different modes result in mode splitting at the frequency degeneracy point resulting in the creation of a localized anomalous GVD region, which is critical for the generation of modulation instability sidebands critical for initial comb formation. In our scheme, we use a coupled-ring geometry based on the Vernier effect [43] to induce the mode interaction, which enables precise control over the strength and spectral position of the mode crossings. The spatial separation between the two microresonators controls the coupling strength, and the ratio between the microresonator radii controls the periodicity of the mode interaction. We used integrated platinum heaters on each of the rings to tune the spectral position of the mode interaction to our pump wavelength [44]. We automate the comb generation process by controlling the integrated heaters and are able to deterministically generate a phase-locked normal GVD comb. Figure 14 shows the generated comb spectrum, which has a high pump-to-comb conversion efficiency of 41% to the 38 generated comb lines, each with >100 μW of power [11]. The FSR of the generated comb is 201.6 GHz.

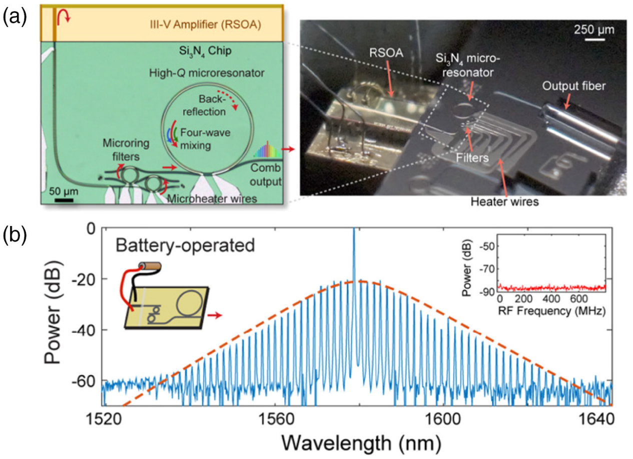

Fig. 13. (a) Microscope image of an integrated comb source (left) and photograph of a fully integrated comb source (right). A III–V reflective semiconductor optical amplifier (RSOA) is coupled to the SiN chip, which consists of two Vernier microring filters, which enables wavelength tuning, and a high Q microresonator, which acts as the narrowband backreflector and the nonlinear resonator for comb generation. The SiN microresonator is tuned using integrated platinum heaters, and the output is collected using an optical fiber [38]. (b) Soliton comb spectrum generated using a AAA battery supplying 98 mW of electrical power.

The proposed link architecture for phase 2 is based on a comb source with 100 GHz comb spacing. Our numerical modeling predicts, for a pump wavelength of 1300 nm and 200 mW of pump power, a pump-to-comb conversion efficiency of 42% to 36 comb lines separated by 100 GHz, each with power above 0 dBm. In addition, we will develop the integration of the comb source with the SiN hybrid laser similar to the previous demonstration with the soliton Kerr comb described in Fig. 13 [38]. We will further optimize the power uniformity and conversion efficiency of the generated comb lines by tailoring the GVD and mode-crossing strength to deliver a high-efficiency comb source generating 64 or 40 channels for 16 Gb/s and 25 Gb/s per channel modulation rates, enabling a 1 Tb/s link. The overall bandwidth can be achieved by increasing the density of the comb lines or by increasing the overall number of comb lines that are generated.

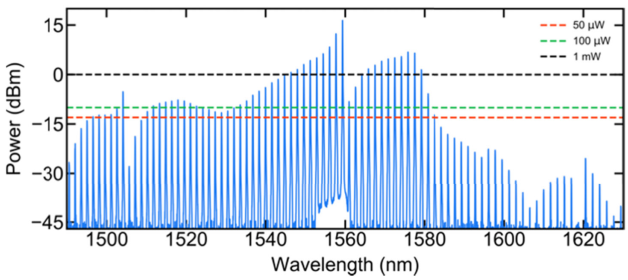

Fig. 14. Measured normal GVD comb spectrum with a comb spacing of 201.6 GHz [42]. The pump-to-comb conversion efficiency is 41%. The black dashed line corresponds to comb line powers >1 mW (18 lines), the green dashed line corresponds to powers >100 μW (38 lines), and the red dashed line corresponds to powers >50 μW (51 lines).

The major advance of a normal GVD comb is the high pump-to-conversion efficiency. We are currently investigating the bandwidth over which this can be generated, and preliminary modeling has shown it is possible to generate over 100 channels with sufficient comb power. In addition, we are investigating designs to implement microresonators with smaller FSRs down to 50 GHz to realize denser comb spacings. For both of these approaches, the generation bandwidth along with the pump-to-comb efficiency is being investigated.

1. Foundry Compatible Comb Resonators

Si_3N_4 ring resonators that can generate frequency combs (high Q) have been almost solely made using low-pressure chemical vapor deposition (LPCVD), a high temperature process not supported by foundries. Plasma-enhanced chemical vapor deposition (PECVD) is a standard, low temperature, commercial process for depositing Si_3N_4; however, efforts to generate high Q frequency combs have been challenging [45–47]. In phase 2, we will develop PECVD Si_3N_4 films by addressing scattering and absorption losses, providing a platform for achieving low loss, crack-free Si_3N_4 films suitable for frequency comb generation with foundry compatible process.

C. Robust DWDM Filters

In our architecture, Mach–Zehnder interferometer (MZI) filter trees will feed narrow FSR ring resonator cascades, which further filter for the desired wavelength and bandwidth. Today, due to their small dimensions and tight bends, high-confinement single-mode silicon waveguides are highly sensitive to fabrication errors and, in particular, to variations in waveguide width. For example, a width variation as small as 5 nm (within the specs of many foundries today) in the arms of an unbalanced MZI will induce a 250 GHz frequency shift of the whole transmission. High-power consumption thermal tuners are then required to compensate for this undesired fabrication-induced result. Phase 2 will include WDM structures that are tolerant to fabrication variations of 2 GHz for 5 nm waveguide width variation, *eliminating the need for high-power thermal heaters*. Both ring resonators and a cascaded MZI (Fig. 15) will employ light splitters based on multimode interference devices (MMI), shown to be robust to fabrication variations and bends based on wide waveguides where the mode interacts minimally with the sidewalls. The Euler bends where the radius of curvature is adiabatically increased along the length of the bend [34] ensure that no higher order modes are excited in these wider waveguides.

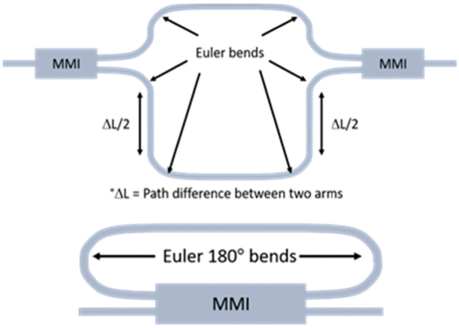

Fig. 15. Designs of MZI and racetrack WDM structures with reduced sensitivity to fabrication variations.

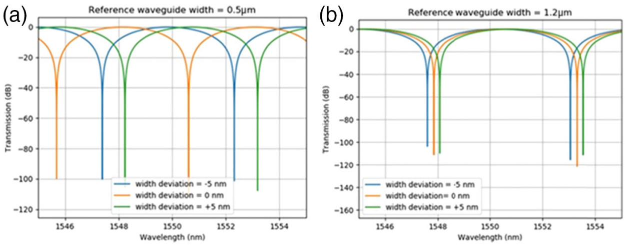

Fig. 16. Transmission of (a) a traditionally designed MZI based on 500 nm with a single mode waveguide and (b) an MZI based on preliminary designs composed of the newly designed 1.2 μm wider waveguide.

In Fig. 16(a), we show the transmission of a traditionally designed MZI based on a 500 nm wide single mode waveguide, and in Fig. 16(b), we show the newly designed MZI based on a 1.2 μm wider waveguide. For both cases, we change the width by ±5 nm and observe the transmission shifts in wavelength. The newly designed MZI with wider waveguide width has lower sensitivity due to changes of 5 nm in the waveguide width as compared to standard waveguides.

D. Quantum Dot Active Devices

During phase 1 of ENLITENED, UCSB has investigated SOAs and mode-locked comb lasers using QD gain material directly grown on silicon as part of the PINE energy-bandwidth optimized optical link thrust. The effort has resulted in several novel demonstrations and performance records.

1. World's First SOA Directly Grown on Silicon

SOAs made with QD gain material have several advantages compared to bulk or quantum well (QW) counterparts in terms of effective gain bandwidth, saturated output power (SOP), and lower noise figure. Phase 1 SOAs demonstrated

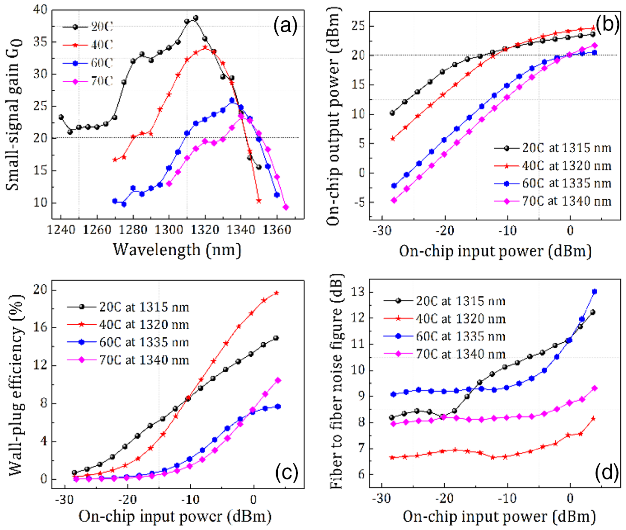

Fig. 17. Si-based QD-SOA performance comparison under different stage temperatures: (a) on-chip small signal gain as a function of wavelength, (b) on-chip output power as a function of on-chip input power, (c) wall-plug efficiency as a function of on-chip input power, (d) fiber-to-fiber noise figure as a function of on-chip input power ($I_{gain} = 750$ mA).

record performance with on-chip small signal gain as high as 39 dB, noise figure 6.6 dB, and saturation output power of 24 dBm (Fig. 17) [12]. Such high-performance SOAs are an enabler of optically switched networks by compensating for insertion loss in optical switches and increasing the operating link margin. The performance demonstrated here compares favorably in all figures of merit against standalone native substrate SOAs and heterogeneously integrated devices on a SiPs platform (detailed comparison in [12]). This suggests direct growth on Si could provide a highly scalable manufacturing platform for not only as-grown SOAs but also for heterogeneous integration via wafer bonding and hybrid integration via edge coupling.

2. *World's First Mode-Locked Comb Lasers Directly Grown on Silicon*

Phase 1 results include a 20 GHz mode-locked comb laser using a chirped QD active region design to increase gain bandwidth, producing >58 wavelengths of usable power within 3 dB of uniformity (and 80 lines within 10 dB) [48]. Comb sources are critical for highly parallel DWDM link architectures, which is the most promising route to energy-efficient, bandwidth dense links.

Building off of phase 1, the relevant QD technology and associated design/processes/learnings will be transferred from UCSB to Quintessent to enable a direct commercialization path for datacenter and HPC customers at the conclusion of phase 2. Quintessent's primary focus within PINE phase 2 will be on QD SOA development and maturation, with the QD comb source being incubated for commercialization and available as needed for risk mitigation to the project's primary comb approach. Quintessent will help standardize high wavelength count sources for future energy-efficient, bandwidth dense datacom optics such as that being developed under PINE.

5. CONCLUSION

Energy-optimized optical link technology is essential for improving the energy efficiency of interconnects for datacenters and HPC. However, given that the interconnect accounts for 15%–25% of the total system power, the opportunity for system efficiency improvements is limited if photonics is treated strictly as a more efficient "wire replacement" technology. In order to achieve significant energy reduction of the interconnection network as well as more efficient compute resource utilization in datacenters and HPC, the system must be dealt with as a whole, building on and taking advantage of unique features of novel components and energy-optimized links. We have summarized the PINE approach that can reduce energy consumption by more than 2× based on SiPs and bandwidth steering at the architecture level and opportunities for even further improvements through effective resource disaggregation that is enabled by the high bandwidth density and distance independence provided by photonic link technologies.

Funding. Advanced Research Projects Agency-Energy (DE-AR0000843).

REFERENCES

1. Q. Cheng, M. Bahadori, M. Glick, S. Rumley, and K. Bergman, "Recent advances in optical technologies for data centers: a review," Optica **5**, 1354–1370 (2018).
2. Y. Shen, X. Meng, Q. Cheng, S. Rumley, N. Abrams, A. Gazman, E. Manzhosov, M. Glick, and K. Bergman, "Silicon photonics for extreme scale systems," J. Lightwave Technol. **37**, 245–259 (2019).
3. Q. Cheng, M. Glick, and K. Bergman, "Optical interconnection networks for high performance systems," in *Optical Fiber Telecommunications VII*, A. Willner, ed. (Academic, 2019) pp. 785–825.
4. M. Mashanovitch, S. Fryslie, B. Buckley, K. Guinn, G. Morrison, and L. A. Johansson, "High-power, efficient DFB laser technology for RF photonics links," in *IEEE Avionics and Vehicle Fiber-Optics and Photonics Conference (AVFOP)*, Portland, Oregon (2018).
5. G. Michelogiannakis, Y. Shen, M. Y. Teh, X. Meng, B. Aivazi, T. Groves, J. Shalf, M. Glick, M. Ghobadi, L. Dennison, and K. Bergman, "Bandwidth steering in HPC using silicon nanophotonics," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '19)*, Denver, Colorado (Association for Computing Machinery, 2019), paper 41.
6. N. C. Abrams, Q. Cheng, M. Glick, M. Jezzini, P. Morrissey, P. O'Brien, and K. Bergman, "Silicon photonic 2.5D integrated multichip module receiver," in *Conference on Lasers and Electro-Optics* (2020).
7. N. C. Abrams, Q. Cheng, M. Glick, M. Jezzini, P. O'Brien, and K. Bergman, "Silicon photonic 2.5D multi-chip module transceiver for high-performance data centers," J. Lightwave Technol. **38**, 3346–3357 (2020).
8. N. C. Abrams, Q. Cheng, M. Glick, E. Manzhosov, M. Jezzini, P. Morrissey, P. O'Brien, and K. Bergman, "Design considerations for multi-chip module silicon photonic transceivers," Proc. SPIE **11308**, 113080I (2020).
9. Q. Cheng, L. Y. Dai, N. C. Abrams, Y.-H. Hung, P. E. Morrissey, M. Glick, P. O'Brien, and K. Bergman, "Ultralow-crosstalk, strictly non-blocking microring-based optical switch," Photon. Res. **7**, 155–161 (2019).

10. Q. Cheng, S. Rumley, M. Bahadori, and K. Bergman, "Photonic switching in high performance datacenters [Invited]," Opt. Express **26**, 16022–16043 (2018).
11. B. Y. Kim, Y. Okawachi, J. K. Jang, M. Yu, X. Ji, Y. Zhao, C. Joshi, M. Lipson, and A. L. Gaeta, "Turn-key, high-efficiency Kerr comb source," Opt. Lett. **44**, 4475–4478 (2019).
12. S. Liu, J. Norman, M. Dumont, D. Jung, A. Torres, A. C. Gossard, and J. E. Bowers, "High-performance O-band quantum-dot semiconductor optical amplifiers directly grown on a CMOS compatible silicon substrate," ACS Photon. **6**, 2523–2529 (2019).
13. O. A. Jimenez Gordillo, M. A. Tadayon, Y.-C. Chang, and M. Lipson, "3D photonic structure for plug-and-play fiber to waveguide coupling," in *Conference on Lasers and Electro-Optics* (Optical Society of America, 2018), paper STh4B.7.
14. O. A. Jimenez Gordillo, S. Chaitanya, Y. Chang, U. Dave, A. Mohanty, and M. Lipson, "Plug-and-play fiber to waveguide connector," Opt. Express, **27**, 20305–20310 (2019).
15. N. Jiang, D. U. Becker, G. Michelogiannakis, J. Balfour, B. Towles, D. E. Shaw, J. Kim, and W. J. Dally, "A detailed and flexible cycle-accurate network-on-chip simulator," in *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)* (2013), pp. 86–96.
16. "Characterization of the DOE Mini-apps," https://portal.nersc.gov/project/CAL/overview.htm.
17. A. Roy, H. Zeng, J. Bagga, G. Porter, and A. C. Snoeren, "Inside the social network's (datacenter) network," in *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication (SIGCOMM '15)* (ACM, 2015), pp. 123–137.
18. Mellanox, "1U EDR 100 Gb/s InfiniBand switch systems hardware user manual models," Technical report SB7700/SB7790 (2015).
19. K. J. Barker, A. Benner, R. Hoare, A. Hoisie, A. K. Jones, D. K. Kerbyson, D. Li, R. Melhem, R. Rajamony, E. Schenfeld, S. Shao, C. Stunkel, and P. Walker, "On the feasibility of optical circuit switching for high performance computing systems," in *Proceedings of the 2005 ACM/IEEE Conference on Supercomputing (SC '05)*, 2005, p. 16.
20. M. Y. Teh, Z. Wu, and K. Bergman, "Flexspander: augmenting expander networks in high-performance systems with optical bandwidth steering," J. Opt. Commun. Netw. **12**, B44–B54 (2020).
21. D. Narayanan, A. Harlap, A. Phanishayee, V. Seshadri, N. R. Devanur, G. R. Ganger, P. B. Gibbons, and M. Zaharia, "PipeDream: generalized pipeline parallelism for DNN training," in *Proceedings of the 27th ACM Symposium on Operating Systems Principles, (SOSP '19)*, Hunstville, Ontario, Canada (Association for Computing Machinery, 2019), pp. 1–15.
22. Y. Huang, Y. Cheng, A. Bapna, O. Firat, D. Chen, M. Chen, H. Lee, J. Ngiam, Q. V. Le, and Y. Wu, "GPipe: efficient training of giant neural networks using pipeline parallelism," in *Advances in Neural Information Processing Systems* (2019), pp. 103–112.
23. S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He, "ZeRO: memory optimizations toward training trillion parameter models," arXiv:1910.02054 (2019).
24. Y. Shen, S. Rumley, K. Wen, Z. Zhu, A. Gazman, and K. Bergman, "Accelerating of high performance data centers using silicon photonic switch-enabled bandwidth steering," in *European Conference on Optical Communication (ECOC)* (2018).
25. Y. Shen, M. H. N. Hattink, P. Samadi, Q. Cheng, Z. Hu, A. Gazman, and K. Bergman, "Software-defined networking control plane for seamless integration of multiple silicon photonic switches in Datacom networks," Opt. Express **26**, 10914–10929 (2018).
26. Q. Cheng, M. Bahadori, M. Glick, and K. Bergman, "Scalable space-and-wavelength selective switch architecture using microring resonators," in *Conference on Lasers and Electro-Optics*, OSA Technical Digest (Optical Society of America, 2019), paper STh1N.4.
27. J. W. Poulton, J. M. Wilson, W. J. Turner, B. Zimmer, X. Chen, S. S. Kudva, S. Song, S. G. Tell, N. Nedovic, W. Zhao, and S. R. Sudhakaran, "A 1.17-pJ/b, 25-Gb/s/pin ground-referenced single-ended serial link for off- and on-package communication using a process- and temperature-adaptive voltage regulator," IEEE J. Solid-State Circuits **54**, 43–54 (2018).
28. J. W. Poulton, W. J. Dally, X. Chen, J. G. Eyles, T. H. Greer, S. G. Tell, J. M. Wilson, and C. T. Gray, "A 0.54 pJ/b 20 Gb/s ground-referenced single-ended short-reach serial link in 28 nm CMOS for advanced packaging applications," IEEE J. Solid-State Circuits **48**, 3206–3218 (2013).
29. J. M. Wilson, W. J. Turner, J. W. Poulton, B. Zimmer, X. Chen, S. S. Kudva, S. Song, S. G. Tell, N. Nedovic, W. Zhao, and S. R. Sudhakaran, "A 1.17 pJ/b 25 Gb/s/pin ground-referenced single-ended serial link for off- and on-package communication in 16 nm CMOS using a process- and temperature-adaptive voltage regulator," in *IEEE International Solid-State Circuits Conference-(ISSCC)* (IEEE, 2018), pp. 276–278.
30. W. J. Turner, J. W. Poulton, J. M. Wilson, X. Chen, S. G. Tell, M. Fojtik, T. H. Greer, B. Zimmer, S. Song, N. Nedovic, and S. S. Kudva, "Ground-referenced signaling for intra-chip and short-reach chip-to-chip interconnects," in *IEEE Custom Integrated Circuits Conference (CICC)* (IEEE, 2018).
31. S. Rumley, M. Bahadori, K. Wen, D. Nikolova, and K. Bergman, "PhoenixSim: crosslayer design and modeling of silicon photonic interconnects," in *Proceedings of the 1st International Workshop on Advanced Inter-connect Solutions and Technologies for Emerging Computing Systems* (ACM, 2016), article 7.
32. M. Bahadori, S. Rumley, D. Nikolova, and K. Bergman, "Comprehensive design space exploration of silicon photonic interconnects," J. Lightwave Technol. **34**, 2975–2987 (2016).
33. K. Bergman, S. Rumley, N. Ophir, D. Nikolova, R. Hendry, Q. Li, K. Padmara, K. Wen, and L. Zhu, "Silicon photonics for Exascale systems," in *Optical Fiber Communication Conference* (Optical Society of America, 2014), paper M3E–1.
34. N. Ophir, C. Mineo, D. Mountain, and K. Bergman, "Silicon photonic microring links for high-bandwidth-density, low-power chip I/O," IEEE Micro **33**, 54–67 (2013).
35. K. Padmaraju and K. Bergman, "Resolving the thermal challenges for silicon microring resonator devices," Nanophotonics **3**, 269–281 (2014).
36. A. L. Gaeta, M. Lipson, and T. J. Kippenberg, "Photonic-chip-based frequency combs," Nat. Photonics **13**, 158–169 (2019).
37. C. Joshi, J. K. Jang, K. Luke, X. Ji, S. A. Miller, A. Klenner, Y. Okawachi, M. Lipson, and A. L. Gaeta, "Thermally controlled comb generation and soliton mode locking in microresonators," Opt. Lett. **41**, 2565–2568 (2016).
38. B. Stern, X. Ji, Y. Okawachi, A. L. Gaeta, and M. Lipson, "Battery-operated integrated frequency comb generator," Nature **562**, 401–405 (2018).
39. C. Bao, L. Zhang, A. Matsko, Y. Yan, Z. Zhao, G. Xie, A. M. Agarwal, L. C. Kimerling, J. Michel, L. Maleki, and A. L. Willner, "Nonlinear conversion efficiency in Kerr frequency comb generation," Opt. Lett. **39**, 6126–6129 (2014).
40. J. Jang, Y. Okawachi, X. Ji, C. Joshi, M. Lipson, and A. L. Gaeta, "Universal conversion efficiency scaling with free-spectral-range for soliton Kerr combs," in *Conference on Lasers and Electro-Optics* (2020), paper JTu2F.32.
41. A. Fülöp, M. Mazur, A. Lorences-Riesgo, Ó. B. Helgason, P.-H. Wang, Y. Xuan, D. E. Leaird, M. Qi, P. A. Andrekson, A. M. Weiner, and V. Torres-Company, "High-order coherent communications using mode-locked dark pulse Kerr combs from microresonators," Nat. Commun. **9**, 1598 (2018).
42. V. E. Lobanov, G. Lihachev, T. J. Kippenberg, and M. L. Gorodetsky, "Frequency combs and platicons in optical microresonators with normal GVD," Opt. Express **23**, 7713–7721 (2015).
43. N. Kobayashi, K. Sato, M. Namiwaka, K. Yamamoto, S. Watanabe, T. Kita, H. Yamada, and H. Yamazaki, "Silicon photonic hybrid ring-filter external cavity wavelength tunable lasers," J. Lightwave Technol. **33**, 1241–1246 (2015).
44. S. A. Miller, Y. Okawachi, S. Ramelow, K. Luke, A. Dutt, A. Farsi, A. L. Gaeta, and M. Lipson, "Tunable frequency combs based on dual microring resonators," Opt. Express **23**, 21527–21540 (2015).
45. E. A. Douglas, P. Mahony, A. Starbuck, A. Pomerene, D. C. Trotter, and C. T. DeRose, "Effect of precursors on propagation loss for plasma-enhanced chemical vapor deposition of SiN_x:H waveguides," Opt. Mater. Express **6**, 2892–2903 (2016).

46. L. Wang, W. Xie, D. V. Thourhout, H. Yu, and S. Wang, "Nonlinear silicon nitride waveguides based on a PECVD deposition platform," Opt. Express **26**, 9645–9654 (2018).
47. X. Ji, S. P. Roberts, and M. Lipson, "High quality factor PECVD Si_3N_4 ring resonators compatible with CMOS process," in *Conference on Lasers and Electro-Optics* (Optical Society of America, 2019), paper SM2O.6.
48. S. Liu, X. Wu, D. Jung, J. C. Norman, M. J. Kennedy, H. K. Tsang, A. C. Gossard, and J. E. Bowers, "High-channel-count 20 GHz passively mode-locked quantum dot laser directly grown on Si with 4.1 Tbit/s transmission capacity," Optica **6**, 128–134 (2019).
49. K. Bergman, J. Shalf, G. Michelogiannakis, S. Rumley, L. Dennison, and M. Ghobadi, "PINE: an energy efficient flexibly interconnected photonic data center architecture for extreme scalability," in *IEEE Optical Interconnects Conference (OI)*, Santa Fe, New Mexico (2018), pp. 25–26.

Madeleine Glick received her Ph.D. in physics at Columbia University for research on electro-optic effects of GaAs/AlGaAs quantum wells. She is currently Senior Research Scientist at Columbia University. From 1992 to 1996, she was a Research Associate with CERN, Geneva, Switzerland, as part of the Lightwave Links for Analogue Signal Transfer Project for the Large Hadron Collider. From 2002–2011, she was Principal Engineer at Intel (Intel Research Cambridge UK, Intel Research Pittsburgh) leading research on optical interconnects for computer systems. Her research interests are in applying photonic devices and interconnects to computing systems. Dr. Glick is a Senior Member of IEEE and OSA.

Nathan C. Abrams received his B.S., M.S., and Ph.D. degrees in electrical engineering from Columbia University in the City of New York in 2014, 2016, and 2020, respectively. His research interests relate to silicon photonic integration for optical interconnects.

Qixiang Cheng (Member, IEEE) received a B.S. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2010, and a Ph.D. degree from the University of Cambridge, Cambridge, U.K., in 2014. He then joined the Shannon Laboratory, Huawei, China, where he researched future optical computing systems. From September 2016, he was a Research Scientist with the Lightwave Research Laboratory, Columbia University, New York, NY, USA. He has been appointed as University Lecturer in photonic devices and systems at the University of Cambridge, since January 2020. His current research interests focus on system-wide photonic integrated circuits for optical communication and optical computing applications, including a range of optical functional circuits, such as packet-, circuit-, and wavelength-level optical switch fabrics, massively parallel transceivers, optical neural networks, and optical network-on-chip.

Min Yee Teh is currently a Ph.D. student in electrical engineering at Columbia University. Since 2018, he has been a researcher with Google's Networks Infrastructure team. His research interests include reconfigurable optical networks, robust traffic, and topology optimization methods.

Yu-Han Hung is currently a postdoctoral research scientist with Lightwave Research Laboratory, Columbia University, USA. Dr. Hung received M.S. and Ph.D. degrees from National Cheng Kung University studying nonlinear dynamics of semiconductor lasers and their applications in 2012 and 2016, respectively. He then stayed as a postdoctoral research fellow for two years before he moved to Columbia University as a postdoctoral research scientist. Dr. Hung's research interests lie in academic investigation and industrial applications of silicon photonic (SiP)-integrated high-performance computing (HPC) systems design and SiP-based devices/interconnects for datacenters and HPC systems.

Oscar Jimenez received a M.Sc. degree in optics from INAOE, Mexico. He joined Lipson Nanophotonics Group at Columbia University as a Ph.D. student in 2015. His research is focused on developing robust interfaces between integrated photonic chips and optical fibers.

Songtao Liu received a B.E. degree (Hons.) in electronic information science and technology from Henan University, Kaifeng, China, in 2012, and a Ph.D. degree in microelectronics and solid state electronics from the University of Chinese Academy of Sciences, Beijing, China, 2017. His Ph.D. dissertation was on the monolithically integrated InP-based mode-locked lasers. He is currently a Post-Doctoral Researcher with the University of California, Santa Barbara, CA, USA. His research interests are in the field of photonic integrated circuits, with an emphasis on monolithically integrated mode-locked lasers, semiconductor optical amplifiers, and narrow linewidth tunable lasers both on III–V and silicon platforms.

Yoshitomo Okawachi received B.S. and M.Eng. degrees in engineering physics in 2002 and 2003, and the M.S. and Ph.D. degrees in applied physics in 2006 and 2008 from Cornell University, Ithaca, NY, USA. He is currently a Research Scientist in the Department of Applied Physics and Applied Mathematics at Columbia University. His research areas include optical frequency comb generation in silicon-based waveguides and microresonators, coherent computing based on degenerate optical parametric oscillation in microresonators, parametric nonlinear interactions in photonic devices, slow light, and all-optical signal processing using space–time duality techniques. Dr. Okawachi is a member of The Optical Society. He is the recipient of the 2017 Tingye Li Innovation Prize. He has served on the CLEO, LAOP, FiO, and ACPC subcommittees and is a referee for 26 peer-reviewed journals. He is currently an associate editor for *Optics Letters* and Vice Chair of the OSA Integrated Photonics and Ultrafast Optical Phenomena Technical Groups. He was the 2017 Ambassador for The Optical Society.

Xiang Meng received his Ph.D. degree in electrical engineering from Columbia University, New York, NY, USA in 2017. His research interests include scientific parallel computing and numerical analysis on emerging photonic devices, ranging from nanolasers and nano-sensors to high-speed optical transceivers and energy-efficient photonic interconnects, mainly for applications in high-performance communication, computing, and datacenter platforms.

Leif Johansson received his Ph.D. degree in 2002 from University of London and is a co-founder and the CTO of Freedom Photonics. Dr. Johansson has 20 years of experience related to photonic integrated circuits, widely tunable lasers, radio-frequency photonics, analog optical communications, radio over fiber systems, and coherent optical communications. Before founding Freedom Photonics, Dr. Johansson has been involved in the development of RF photonic devices and subsystems at University of California, Santa Barbara, and at Agility Communications (now Lumentum). Dr. Johansson has authored or co-authored over 260 publications in the areas of RF photonics and photonic integrated circuits and has given numerous invited and contributed talks at various international conferences. He has served on technical committees for the IEEE Microwave Photonics Conference, Optical Fiber Communications Conference, and IEEE Photonics Conference.

Manya Ghobadi received her Ph.D. in computer science from the University of Toronto in 2013. She is an assistant professor at the EECS department at MIT. Before MIT, she was a researcher at Microsoft Research and a software engineer at Google Platforms. Dr. Ghobadi is a computer systems researcher with a networking focus and has worked on a broad set of topics, including datacenter networking, optical networks, transport protocols, and network measurement. Her work has won the best dataset award and best paper award at the ACM Internet Measurement Conference (IMC) as well as the Google research excellent paper award.

Larry Dennison holds Ph.D., M.S., and B.S. degrees from the Massachusetts Institute of Technology. Dr. Dennison joined NVIDIA in September of 2013 and leads the Network Research Group. His current research interests include large networks of GPUs, switch micro-architectures, network-on-chip, and photonic interconnects. At, NVIDIA, he was the principal investigator for the DesignForward project, which was responsible for several GPU shared-memory concepts such as NVSHMEM and NCCL. His team proposed development of a GPU shared memory fabric and developed the first NVSwitch architecture. Prior to NVIDIA, he worked on software systems such as high-performance distributed applications, database scaling for the cloud, and software-defined networking. He also architected and led the development of the ASIC chipset for the Avici Terabit Router, which utilized a 3D toroidal network. At BBN, Dr. Dennison was the principal investigator for MicroPathfinder, a wearable computer that connected to other wearables over a very low-power RF network.

George Michelogiannakis received B.Sc. and M.Sc. degrees with honors from the University of Crete, Greece, and a Ph.D. from Stanford University in 2012 where he was selected for the Stanford Graduate Fellowship. He is now a research scientist in the computer architecture group at Berkeley Laboratory. His latest work focuses on the post Moore's law era looking into specialization, emerging devices (transistors), memories, photonics, and 3D integration. He is also currently working on optics and architecture for HPC and datacenter networks.

John Shalf received his B.S. and M.S. in electrical and computer engineering from Virginia Tech in 1992. He is department head for computer science at Lawrence Berkeley National Laboratory and leads the computer architecture group there. Prior to that, he was deputy director of hardware technology for the DOE Exascale Computing Project (ECP).

Alan Y. Liu (S'13–M'16) received a Ph.D. degree in electronic and photonic materials from the University of California, Santa Barbara. He is currently the CEO of Quintessent, which he cofounded to commercialize quantum dot based lasers and photonic integrated circuits. He was previously a consultant at Booz Allen Hamilton and advised clients on various photonics R&D programs.

John E. Bowers (F'94) received M.S. and Ph.D. degrees from Stanford University. He was with AT&T Bell Laboratories. He is currently the Director of the Institute for Energy Efficiency, University of California, Santa Barbara. He is also a Professor with the Department of Electrical and Computer Engineering, University of California, and the Department of Materials, University of California. His research interests are primarily concerned with silicon photonics, optoelectronic devices, optical switching and transparent optical networks, and quantum dot lasers. He is a member of the National Academy of Engineering and the National Academy of Inventors. He is a fellow of OSA and the American Physical Society. He was a recipient of the IEEE Photonics Award, the OSA/IEEE Tyndall Award, the IEEE LEOS William Streifer Award, and the South Coast Business and Technology Entrepreneur of the Year Award.

Alex Gaeta received his Ph.D. in 1991 in optics from the University of Rochester. He joined the faculty in the Department of Applied Physics and Applied Mathematics at Columbia University in 2015, where he is the David M. Rickey Professor. Prior to this, he was a professor in the School of Applied and Engineering Physics at Cornell University for 23 years. He has published more than 250 papers in quantum and nonlinear photonics. He cofounded PicoLuz, Inc. and was the founding Editor-in-Chief of *Optica*. He is a Fellow of the OSA, APS, and IEEE, is a Thomson Reuters Highly Cited Researcher, and was awarded the 2019 Charles H. Townes Medal from OSA.

Michal Lipson (SM'07–F'13) is a Eugene Higgins Professor of Electrical Engineering and Professor of Applied Physics at Columbia University. She received her Ph.D. in Physics in the Technion in 1998. Following a postdoctoral position in MIT in the Material Science department from 1998 to 2001, she joined the School of Electrical and Computer Engineering at Cornell University and was named the Given Foundation Professor of Engineering at the School of Electrical and Computer Engineering in 2012. In 2015, she joined Columbia University. She is a member of the American Academy of Arts and Science and a member of the National Academy of Science. She has been awarded the NAS Comstock Prize in Physics, the MacArthur Fellowship, the Blavatnik Award, The Optical Society's R. W. Wood Prize, the IEEE Photonics Award, and the Erna Hamburger Award.

Keren Bergman is the Charles Batchelor Professor of Electrical Engineering at Columbia University where she also serves as the Faculty Director of the Columbia Nano Initiative. Prof. Bergman received a B.S. from Bucknell University in 1988, and a M.S. in 1991 and Ph.D. in 1994 from MIT, all in electrical engineering. At Columbia, Prof. Bergman leads the Lightwave Research Laboratory encompassing multiple cross-disciplinary programs at the intersection of computing and photonics. Prof. Bergman serves on the Leadership Council of the American Institute of Manufacturing (AIM) Photonics, leading projects that support the institute's silicon photonics manufacturing capabilities and Datacom applications. She is the recipient of the 2016 IEEE Photonics Engineering Award and is a Fellow.