Title

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

Journal

Journal of Optical Communications and Networking, 12(12)

ISSN

1943-0620

Authors

Glick, Madeleine
Abrams, Nathan C
Cheng, Qixiang
et al.

Publication Date

2020-12-01

DOI

10.1364/jocn.402788

Peer reviewed

PINE: Photonic Integrated Networked Energy efficient datacenters (ENLITENED)

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]

1Columbia University, New York, New York 10027, USA

2University of Cambridge, Cambridge, UK

3Electrical and Computer Engineering Department, University of California, Santa Barbara, California 93106, USA

4Freedom Photonics, LLC, Santa Barbara, California 93117, USA

5CSAIL, MIT, Cambridge, Massachusetts 02139, USA

6NVIDIA Corp., Santa Clara, California 95051, USA

7Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

8Quintessent, Santa Barbara, California 93105, USA

*Corresponding author: msg144@columbia.edu

We review the motivation, and the achievements of and goals of the PINE ARPA-E ENLITENED program. The PINE program leverages the unique features of photonic technologies to enable alternative mega-datacenters and HPC system architectures that deliver more substantial energy efficiency improvements than could be achieved through link energy efficiency alone. In phase 1, the PINE system architecture demonstrated an average factor of 2.2X improvement in Transactions/Joule across a diverse set of HPC and datacenter applications. In phase 2, PINE will be demonstrating an aggressive 2.2pJ/bit total link budget with high-bandwidth-density DWDM links to enable an additional 2.5x or more efficiency gains through deep resource disaggregation.

1. INTRODUCTION

The recent explosive growth in data analytics applications that rely on machine learning techniques are leading to a convergence between datacenters and high-performance computing (HPC) systems that are driving an intensely growing need for compute performance. The efficiency of these massive parallel architectures is affected by how data moves among

the numerous compute, memory and storage resources, the energy consumption associated with data movement, as well as utilization efficiency of heterogeneous compute and memory resources. The Photonic Integrated Networked Energy efficient datacenter (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 Cluster Architecture**. Our proposed PINE datacenter architecture integrates low-power silicon photonic links and large numbers of embedded low-radix broadband photonic circuit switches to enable inter-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 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 Multi-Chip Modules (MCMs) active interposer platform enables multi-Tbps chip escape bandwidths.
3) **Energy Optimized Dense WDM Photonic Connectivity**. The PINE DWDM optical links build on a new generation of components specifically optimized for energy efficiency [1,2,3]. Our novel light source platform composed of a single, high power, high efficiency laser coupled to a multiple wavelength comb generator is used to allocate power for >50 wavelengths.

Our next-generation PINE phase 2 full system solution builds on phase 1 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 1pJ/bit, demonstrate MCM chip edge bandwidth densities >5Tbps/mm, and reduce the system-wide energy consumption of datacenters by factor of 2.8X. Phase 2 has an increased focus on deep intra-node disaggregation and in particular AI data analytics applications, which projected to deliver 5X 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.

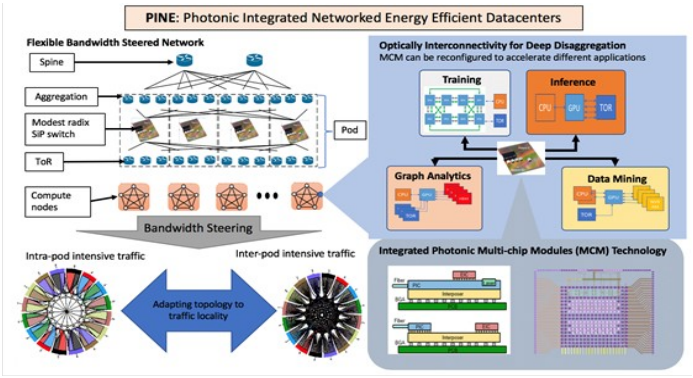
Here we provide a review of the PINE program. We start in Section 2 with an overview of the PINE system architecture, bandwidth steering and our approach to disaggregation. In Section 3, we describe details of our optically interconnected multichip modules and silicon photonic switch and high density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

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 2.2X improvement in transactions/Joule (55% reduction in power per transaction) and a 20% reduction in average network latency across a diverse set of applications. The fundamental building blocks of PINE's system-wide bandwidth steering were reconfigurable optical silicon-in-package switches. The PINE multi-chip module 2.5D and 3D assembly processes were developed in the active interposer platform and demonstrated high density electronic/photonic integration and the first multi-layer photonic switch. PINE phase 1 demonstrated energy optimized high-bandwidth density silicon photonic links with aggregate bandwidth of 640Gb/s and total link energy of 2.2pJ/bit. PINE Phase 2 will build on these successes to 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 >1Tb/s aggregate bandwidth, reduced energy toward 1pJ/bit, demonstrate MCM chip edge bandwidth densities >5Tb/s/mm, and reduce the system-wide energy consumption of datacenters by factor of 2.8X.

In support of the PINE system are multi-chip modules developed in the SUNY active interposer platform that demonstrated high density electronic/photonic integration with edge bandwidth densities of 2.5Tb/s/mm. The first multi-layer micro-ring photonic switch demonstrated in the MCM platform achieved record low crosstalk and extinction of >50dB. Our fully integrated comb laser with experimentally demonstrated 45% power pump conversion drives the PINE phase 1 links with demonstrated aggregate bandwidths of 640Gb/s and total link energy of 2.2pJ/bit. The PINE system extends scalability with high efficiency Semiconductor Optical Amplifiers (SOA)
and the first monolithic quantum dot (QD) SOA on silicon demonstrated record WPE of 14.2% and on-chip gain of 39dB. 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.



Furthermore, in Phase 1 we performed a comprehensive cross-layer modeling and energy/performance analysis at the link, node, and full system levels.

Fig. 1. PINE system architecture

PINE System-Wide Bandwidth Steering: During phase 1, our team developed system-wide bandwidth steering to seamlessly reconfigure the interconnect topology to match diverse workload communication patterns [41,4,5,6]. 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. We illustrate the physical topology of our approach in Figure 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 4x4 SiP switches PINE flexible Fat Tree experimental testbed (explained in detail later) running the HPC GTC trace we demonstrated an 62% application execution acceleration using bandwidth steering.

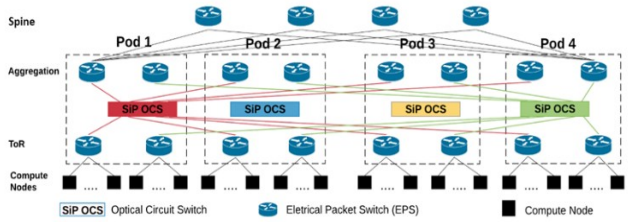


Fig. 2. Bandwidth steering in the flexible Fat Tree PINE architecture. A reconfigurable fat tree topology enabled by the placement of SiP optical circuit switches (OCSs) between the top-of-rack (ToR) switches at the first level of the topology and aggregation switches (second level) of different pods [41].

System level simulations of scaled bandwidth steering in the PINE architecture are performed in our cycle-accurate network simulator Booksim [7], modified to implement bandwidth steering. We use HPC application traces identified by the DOE exascale initiative [8] and publicly-available traces from a Facebook production-level database pod [9]. Placement of tasks on network endpoint is randomized and combined across different applications. This simulates the effects of fragmentation with multiple diverse applications sharing the network. The system-level power and performance models include 16x16 PINE photonic switches and electrical 36-port Mellanox 100Gb/s InfiniBand router with active optical cables [10] and are sized to match the application traces. Figure 3 summarizes the throughput and latency improvements from bandwidth steering. As shown, bandwidth steering improves average network throughput by 1.7× and average network latency by 20%.

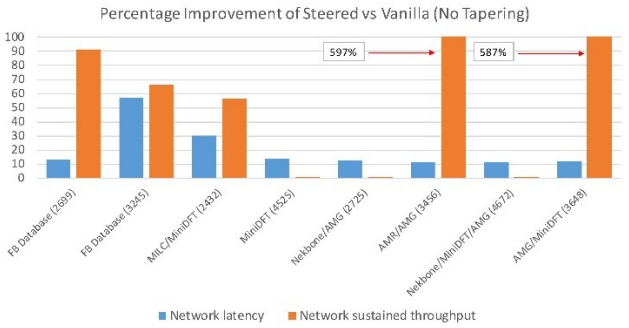


Fig. 3. Average system throughput improvement of the steered fat tree compared to a vanilla fat tree with no tapering (reduction) of top-layer bandwidth [41].

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 [11] and in NERSC's Cori multi-node applications initiate every 17 seconds by average, well above 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 2.2X improvement in Transactions/Joule (55% power consumption per 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 1.7X, and thus reduces compute resource idle power by the same factor. More information can be found in [41].

Building on these successful results, Phase 2 efforts will address the benefits of multi-workload defragmentation, further tapering, 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 servers and reorganizes MCM chips around a reconfigurable network fabric, to 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 the MCM chips . 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 photonic link technologies with efficient comb laser sources reduce the energy consumption and increase bandwidth densities on the system. 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 place stress on compute and interconnect elements and sometimes create 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 over/under utilized 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.

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 either a significant energy efficiency penalty through lower application performance or by bandwidth overprovisioning. 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 more dynamic range for reconfiguring/rewiring nodes, which would in turn deliver a >5X performance-per-Watt advantage for applications with diverse node resource demands. We quantify these gains using the Traversed Edges Per Second (TEPS) per Watt metric.

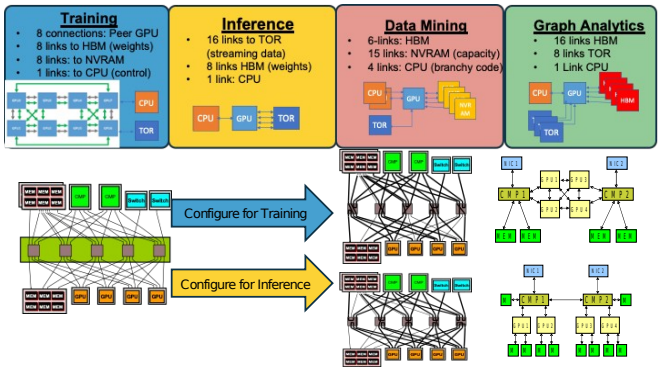


Fig. 4. Schematic of Deep Disaggregation for customized node configurations to support diverse workloads [42].

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 how the 4 canonical workloads have different node organizations to support their requirements. Deep disaggregation enables custom node configurations to be created at job startup 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. For instance, the different phase 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 is 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, GPipe, and DeepSpeed, 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 DNN computation graph separately), demonstrating an increase in training throughput of up to 3.8X. We posit that all of these approaches will benefit from our PINE platform with high network bandwidth.

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.

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. On this testbed we can perform any functionality of an ordinary computing system, including bulk data transfer, 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 4x4 SiP switches and can be configured in both a Dragonfly topology [12] and a Fat-Tree topology [13].

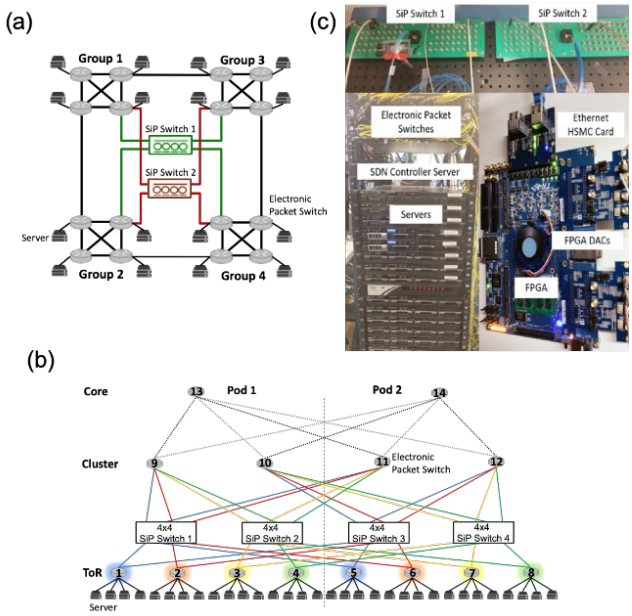


Fig. 5. Fig. 5. Physical testbed networks – a) Dragonfly, b) Fat Tree, and c) implemented hardware [41].

The PINE phase 1 system testbed evaluated bandwidth steering performance of the GTC (Gyrokinetic Toroidal Code) 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 test bed will be substantially scaled to 64 nodes and extended to more accurately reflect real deployed computing systems. Insertion of PINE optically interconnected MCMs with photonic switching capabilities will be performed for 'in-the-path' 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 data center traffic.

3. Optically Interconnected Multi-Chip Modules

A. MCM Interposer

In phase 1 our team developed multi-chip module (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. The active interposer platform developed jointly with SUNY Poly uniquely combines the PIC and interposer into a single integrated substrate. This approach can deliver the best electronic/photonic densities and will become our 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 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 electronic IC CMOS driver chips. The electronic ICs 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 electronic ICs. 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 BGA connections to the PCB.

Fig. 6. 2.5D and 3D MCM integration approaches explored in phase 1.

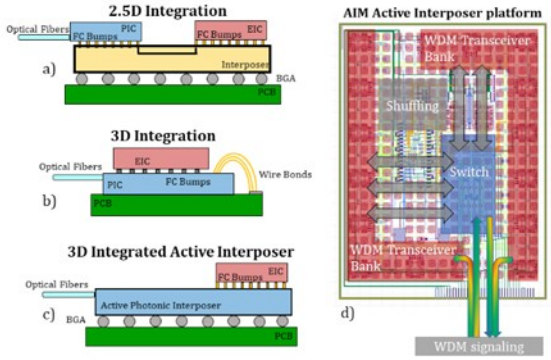


Fig. 6. 2.5D and 3D MCM integration approaches explored in phase 1 (a-c) and the 8 x 8 network-on-chip implemented in the active interposer platform (d).

B. MCM Photonic Switch Fabric (QC input to come)

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 8x8 spatial switch with 4-WDM transmitters, 4-WDM receivers, and EICs flipped on-top, as shown in Figure 7 (top left).

In phase 1 we further designed, fabricated and packaged the first strictly non-blocking optical switch in a switch-and-select (S&S) topology using microring resonators [14, 15], as shown in Figure 7 (top right). The unique Si/SiN multi-layered switch-and-select topology demonstrated breakthrough switch crosstalk suppression and extinction ratio of >50dB and on-chip loss as low <1.8dB. We have also taped out a triple-layered 8x8 microring switch in the same topology with the microscope image shown in Figure 7 (bottom left). The switch chip was packaged using a silicon interposer that can be readily embedded into the MCM platform.

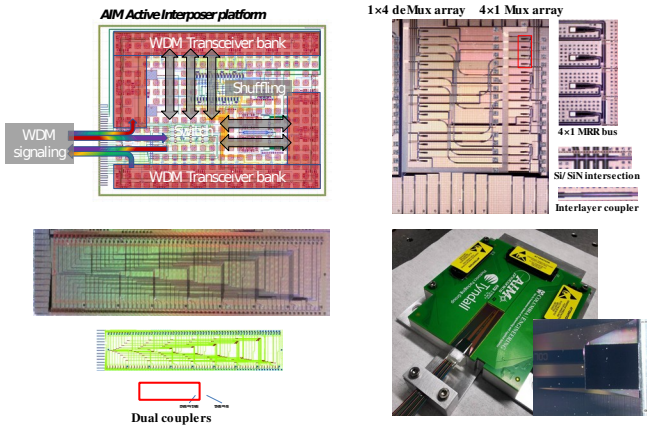


Fig. 7. (top left) GDS layout of a full system-on-chip: 8×8 spatial switch, 4 WDM Tx, 4 WDM Rx, flip-chiped EICs leveraging active interposer platform. (top right) Microscope photo of the fabricated device with insets of the enlarged 4×1 MRR-based spatial multiplexer, the Si/SiN intersections and the interlayer coupler. (bottom left) Microscope photo of the taped out 8×8 triple-layered switch. (bottom right) Packaged switch with a silicon interposer.

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 superior flexibility [x]. The switch architecture comprises two sections: 1) N switching planes of NxM λ crosspoint matrix for spatial wavelength selection and 2) N comb wavelength aggregators of FSR-matched microring arrays, as illustrated in Figure 8. Each NxM 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.

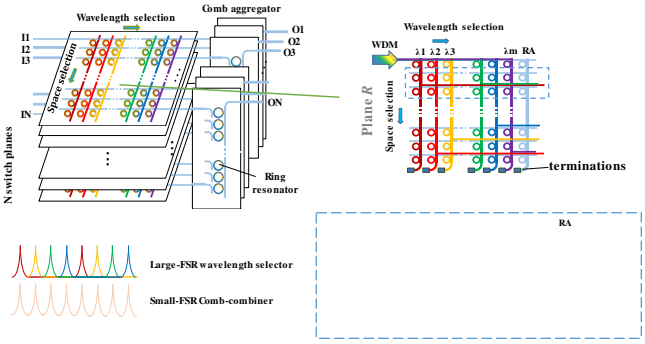


Fig. 8. Proposed space-and-wavelength switch design using arrays of microring-based wavelength selectors and comb aggregators. Insets show the operating principle of the N×Mλ crosspoint matrix and how it fully blocks the first-order in-band crosstalk.

C. Ultra-Low Energy Electronics

In a joint Columbia/NVIDIA effort under phase 2, ultra-low energy electronic ICs will be custom designed and implemented to directly interface with the PICs. These advanced EIC will include drivers for the transmit modulator array, TIAs for the receivers, and heater control circuits and will be fabricated at the TSMC 16nm FinFET process using an MPW run. We will explore circuits for both lower speed optical channels (around 10-16Gb/s) and higher speed optical channels (around 25Gb/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 16nm 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 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 [18-22]. 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 testchip that will be packaged with the PICs. The combined EIC/PIC active interposer will be packaged on an organic substrate and mounted on a printed circuit board.

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 (SMF) and integrated waveguides [16, 17]. The 3D structure, seen in Figure 9, consists of a polymer funnel that routes the incoming fiber (thinned on one end) directly to the facet of a polymer waveguide. At the end of the routing section the fiber and the polymer waveguide are coupled, perfectly aligned, 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.

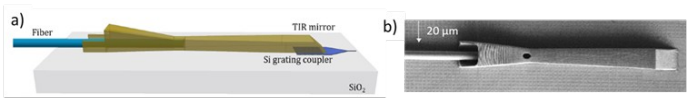


Figure 9: a) Schematic of the robust Plug-and-Play coupler. b) Scanning Electron Microscopy image of our fabricated devices

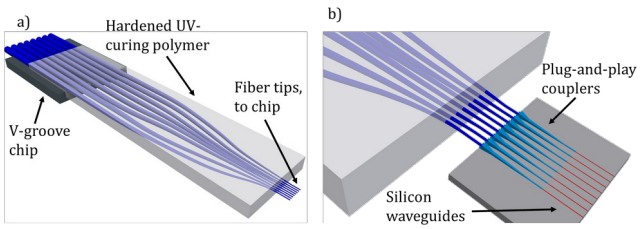


Fig. 10: a) Schematic diagram of the thin fiber array for 8 fibers with a 30 µm pitch. b) Fiber array coupling to an array of 8 plug-and-play couplers written directly on the photonics chip.

Working with these thin fibers allows for high density packaging of couplers with low footprint arrays (down to 30µm pitch). Our results show that our 3D funnel coupler only exerts a 0.05dB 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.

4. Dense WDM Silicon Photonic Links

A. Energy-Throughput Optimized Design

In phase 1, our team developed unique simulation and modeling tools of the PINE links in PhoenixSim [23]. The PhoenixSim environment is built to enable maximization of system performance with optimized energy consumption [24] 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 silicon photonic links using components such as multi-wavelength comb sources, modulators, filters, photodetectors and electrical integrated components (Figure 11a).

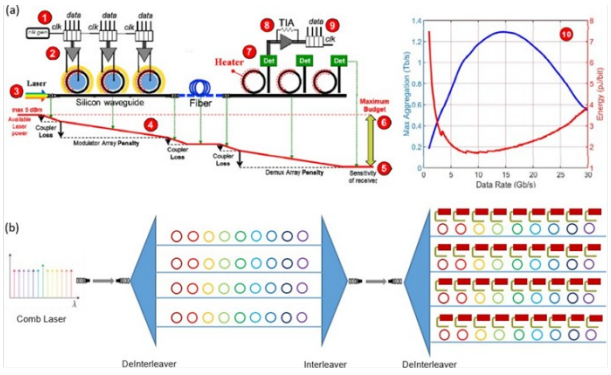


Figure 11: (a) Photonic link models available in Columbia's PhoenixSim software tool, (b) Ultra-low power DWDM link architecture with comb laser source. DWDM comb lines are deinterleaved into four groups, and each group has cascaded microring-based modulators.

A key advantage 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 file for fabrication [24].

The link architecture for phase 2, shown in Figure 11 (b), is sourced by a DWDM comb propagated on-chip where the comb lines are de-interleaved 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 approach enables highly sensitive receivers and for cross-optimization with the optical signal quality, e.g. minimizing the required laser power, in the photonic link [25].

The optical link budget includes anticipated coupling losses, component insertion losses, WDM channel crosstalk at the given receiver sensitivity, and the required margin and BER for the NRZ modulation format [26,27]. The comb laser is set to have an overall 8% wall plug efficiency (WPE), with 20% pump laser efficiency and 40% comb conversion efficiency. The PINE phase 2 link energy budget detailed in Table 1, exceeds the program metrics to deliver 1Tb/s of aggregate bandwidth with less than 1.0pJ/bit.

Table 1: PINE Link Energy Budget

PINE Link	Phase 1	Phase 2
Aggregate Bandwidth	640 Gb/s	1Tb/s
Laser power	0.62 pJ/bit	0.26 pJ/bit
Modulator driver	0.3 pJ/bit	0.12 pJ/bit
Receiver amplifier	0.75 pJ/bit	0.21 pJ/bit
Heaters	0.23 pJ/bit	0.23 pJ/bit
Interposer host interface	0.25 pJ/bit	0.18 pJ/bit
Total link budget	**2.2 pJ/bit**	**1.0 pJ/bi**

B. Dense WDM 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 ultrahigh precision. Such a source is ideal for wavelength-division multiplexing (WDM) applications since the spacing of all 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 a microheater [28]. 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 the successful 1) development of an integrated comb source, 2) the 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 Figure 12.

To increase the comb source power-per-line we plan to build on our recent advances with comb formation in the normal GVD regime [29-31]. For precise control over the strength and spectral position of the mode crossings, we utilize a coupled-ring geometry based on the Vernier effect [32]. We used integrated platinum heaters on each of the rings to tune the spectral position of the mode interaction to our pump wavelength of 1559 nm [33], and mode interaction results in splitting of the resonance. We automate the comb generation process by controlling the integrated heaters and are able to deterministically generate the normal GVD comb. Figure 13 shows the generated comb spectrum that has a high pump-to-comb conversion efficiency of 41% to the 38 generated comb lines, each with >100 μW of power (Fig. 13). The FSR of the generated comb is 201.6 GHz.

The proposed link architecture for Phase 2 is based on a comb source with 100- GHz comb spacing. For a pump wavelength of 1300nm and 200mW of pump power, we achieve a pump-to-comb conversion efficiency of 42% to 36 comb lines separated by 100GHz, each with power above 0dBm [34]. We will further optimize the power uniformity and conversion efficiency of the generated comb lines by tailoring the GVD and mode-crossing strength. In addition, we will develop the integration of the

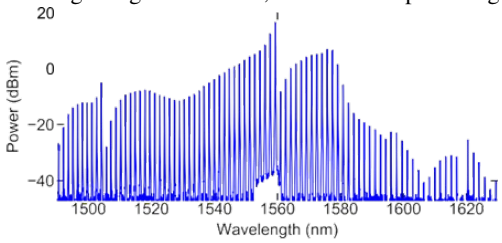


Figure 13: Measured normal GVD comb spectrum with a comb line spacing of 201.6 GHz and a 41% pump to comb conversion efficiency. 18 lines with powers >1mW, 38 lines >100 μW, and 51 lines >50 μW

comb source with the SiN hybrid laser. We will optimize the overall GVD of the system 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.

Foundry Compatible Comb Resonators: Si3N4 ring resonators which 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 Si3N4, however efforts to generate high Q frequency combs have been challenging [36-38]. In phase 2 we will develop PECVD Si3N4 films by addressing scattering and absorption losses, providing a platform for achieving low loss, crack-free Si3N4 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 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 cascaded MZI (Figure 14) will employ light splitters based on multimode mode 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 [35], ensures that no higher order modes are excited in these wider waveguides.

In Figure 15 we show (a) the transmission of a traditionally designed MZI based on 500nm wide single mode waveguide and (b) the newly

designed based on 1.2µm 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 5nm in the waveguide width than when using standard waveguides.

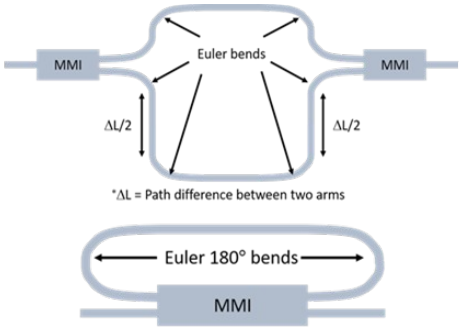


Figure 14: Designs of MZI and Racetrack WDM structures with reduced sensitivity to fabrication variations.

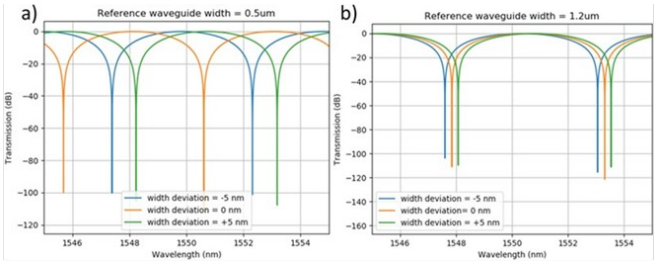


Figure 15: (a) Transmission of a traditionally designed MZI based on 500 nm wide single mode waveguide and (b) MZI based on preliminary designs composed of the newly designed 1.2 µm wider waveguide.

D. Quantum Dot Active Devices

During phase 1 of ENLITENED, UCSB has been investigating semiconductor optical amplifiers and mode-locked comb lasers using quantum dot (QD) gain material directly grown on silicon as part of the PINE energy-bandwidth optimized optical link thrust. The effort has been very successful, resulting in several novel demonstrations and performance records:

•World's first semiconductor optical amplifier (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 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. 16) [39]. 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.

•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) [40]. Comb sources are critical for highly parallel DWDM link architectures which is the most promising route to energy efficient, bandwidth dense links.

In addition, as part of Tech-to-Market efforts in Phase 1 including continued engagement with industry partners, we gathered that there is significant commercial interest in the quantum dot technology developed by UCSB. To support ARPA-E's Tech-to-Market mission of bringing technology to the field to create tangible impact, and to fulfill ENLITENED Phase 2's vision of creating clear technology transition paths, in Phase 2 the development of this technology will be carried out by Quintessent Inc.: a start-up specifically spun out to serve as a commercialization vehicle for the quantum dot technology at UCSB.

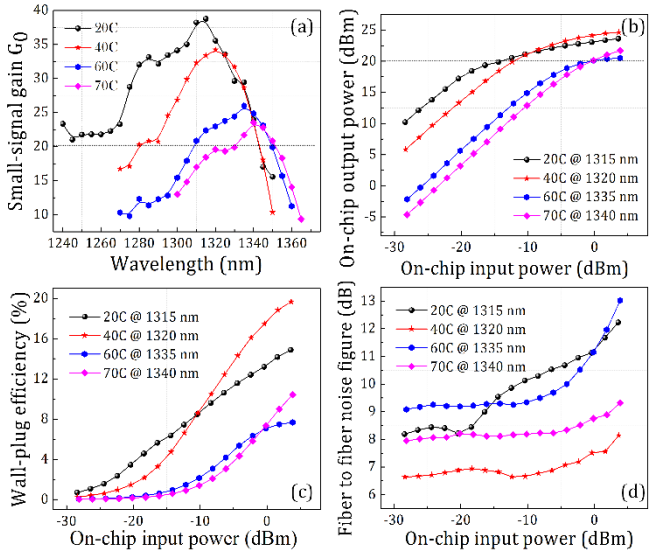


Figure 16: Si-based QD-SOA performance comparison under different stage temperatures (a) on-chip small signal gain as a function of wavelength (b) on-chip output power as a function of on-chip input power (c) wall-plug efficiency as a function of on-chip input power (d) fiber-to-fiber noise figure as a function of on-chip input power (I_{gain} = 750 mA).

Building off of UCSB's success in 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 has joined the newly formed CW-WDM MSA to help standardize high wavelength count sources for future energy efficient, bandwidth dense datacom optics such as that being developed under PINE.

4. Conclusions

Energy optimized optical link technology is essential for improving the energy efficiency of interconnects for datacenters and HPC. However, given 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 data centers 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 greater than 2x based on silicon photonics and bandwidth steering at the architecture and opportunities for even further improvements through effective resource disaggregation that is enabled by the high bandwidth density and distance-independence provided by of photonic link technologies.

Funding Information. Advanced Research Projects Agency-Energy (ARPA-E) (ENLITENED); U.S. Department of Energy (DE-AR0000843)

References

1. [1] Qixiang Cheng, Meisam Bahadori, Madeleine Glick, Sébastien Rumley, and Keren Bergman, "Recent advances in optical technologies for data centers: a review," Optica, 5, 1354-1370 (November 2018).
2. [2] Yiwen Shen, Xiang Meng, Qixiang Cheng, Sebastien Rumley, Nathan Abrams, Alexander Gazman, Evgeny Manzhosov, Madeleine Glick, Keren Bergman, [Invited] "Silicon Photonics for Extreme Scale Systems," IEEE/OSA Journal of Lightwave Technology, 37, 245-259 (January 2019).
3. Optical Fiber Telecommunications VII 2019: (Book Chapter) Optical Interconnection Networks for High Performance Systems, Qixiang Cheng, Madeleine Glick, Keren Bergman.
4. M. Adda and A. Peratikou. Routing and fault tolerance in z-fat tree. IEEE Transactions on Parallel and Distributed Systems, 28(8):2373–2386, Aug 2017.
5. [5] Jung Ho Ahn, Nathan Binkert, Al Davis, Moray McLaren, and Robert S. Schreiber. Hyperx: Topology, routing, and packaging of efficient large-scale networks. In Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, SC '09, pages 41:1–41:11, 2009.
6. [6] Mohammad Al-Fares, Alexander Loukissas, and Amin Vahdat. A scalable, commodity data center network architecture. In Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication, SIGCOMM '08, pages 63–74. ACM, 2008.
7. [7] Nan 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 2013 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), pages 86–96, April 2013.
8. [8] Characterization of the DOE mini-apps. https://portal.nersc.gov/project/CAL/doeminiapps. htm. Accessed: 2019-02-16.
9. [9] Arjun Roy, Hongyi Zeng, Jasmeet Bagga, George Porter, and Alex 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, pages 123–137. ACM, 2015.
10. [10] Mellanox 1U EDR 100Gb/s InfiniBand switch systems hardware user manual models: SB7700/SB7790. Technical report, 2015.
11. [11] 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 SC '05: Proceedings of the 2005 ACM/IEEE Conference on Supercomputing, pages 16–16, Nov 2005.
12. [12] 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 2018 European Conference on Optical Communication (ECOC), pages 1–3, Sep. 2018.
13. [13] Yiwen Shen, Maarten H. N. Hattink, Payman Samadi, Qixiang Cheng, Ziyiz Hu, Alexander Gazman, and Keren Bergman. Software-defined networking control plane for seamless integration of multiple silicon photonic switches in datacom networks. Opt. Express, 26(8):10914–10929, Apr 2018.
14. [14] Qixiang Cheng, Liang Yuan Dai, Nathan C. Abrams, Yu-Han Hung, Padraic E. Morrissey, Madeleine Glick, Peter O'Brien, and Keren Bergman. Ultralow-crosstalk, strictly non-blocking microring-based optical switch. Photon. Res., 7(2):155–161, Feb 2019.
15. [15] Qixiang Cheng, Sébastien Rumley, Meisam Bahadori, and Keren Bergman, "Photonic switching in high performance datacenters [Invited]," Opt. Express 26, 16022-16043(2018).
16. [16] Oscar A. Jimenez Gordillo, Mohammad Amin Tadayon, You-Chia Chang, and Michal Lipson. 3d photonic structure for plug-and-play fiber to waveguide coupling. In Conference on Lasers and Electro-Optics, page STh4B.7. Optical Society of America, 2018.
17. [17] O. A. Jimenez Gordillo, S. Chaitanya, Y. Chang, U. Dave, A. Mohanty, and M. Lipson. Plug-and-play fiber to waveguide connector. 2018.
18. [18]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, S. R. and Sudhakaran, S.R., 2018. 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 Journal of Solid-State Circuits, 54(1), pp.43-54.
19. [19] John W. Poulton, William J. Dally, Xi Chen, John G. Eyles, Thomas H. Greer, Stephen G. Tell, John M. Wilson, and C. Thomas 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 Journal of Solid-State Circuits 48, no. 12 (2013): 3206-3218.
20. [20] 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, 2018, February. A 1.17 pJ/b 25Gb/s/pin ground-referenced single-ended serial link for off-and on-package communication in 16nm CMOS using a process-and temperature-adaptive voltage regulator. In 2018 IEEE International Solid-State Circuits Conference-(ISSCC) (pp. 276-278). IEEE.
21. [21] 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, 2018, April. Ground-referenced signaling for intra-chip and short-reach chip-to-chip interconnects. In 2018 IEEE Custom Integrated Circuits Conference (CICC) (pp. 1-8). IEEE.
22. [22] T. Gray et al. DOE FastForward 2 Memory Quarter 5 Report. Technical report, NVIDIA, 2015.
23. [23] Sébastien Rumley, Meisam Bahadori, Ke Wen, Dessislava Nikolova, and Keren 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, page 7. ACM, 2016.
24. [24] Meisam Bahadori, Sébastien Rumley, Dessislava Nikolova, and Keren Bergman. Comprehensive design space exploration of silicon photonic interconnects. Journal of Lightwave Technology, 34(12):2975–2987, 2016.
25. [25] Keren Bergman, Sébastien Rumley, Noam Ophir, Dessislava Nikolova, Robert Hendry, Qi Li, Kishore Padmara, Ke Wen, and Lee Zhu. Silicon photonics for exascale systems. In Optical Fiber Communication Conference, pages M3E–1. Optical Society of America, 2014.

26. [26] Noam Ophir, Christopher Mineo, David Mountain, and Keren Bergman. Silicon photonic microring links for high-bandwidth-density, low-power chip I/O. IEEE Micro, 33(1):54–67, 2013.
27. [27] Kishore Padmaraju and Keren Bergman. Resolving the thermal challenges for silicon microring resonator devices. Nanophotonics, 3(4-5):269–281, 2014.
28. [28] Chaitanya Joshi, Jae K. Jang, Kevin Luke, Xingchen Ji, Steven A. Miller, Alexander Klenner, Yoshitomo Okawachi, Michal Lipson, and Alexander L. Gaeta. Thermally controlled comb generation and soliton modelocking in microresonators. Opt. Lett., 41(11):2565–2568, Jun 2016.
29. [29] Xiaoxiao Xue, Pei-Hsun Wang, Yi Xuan, Minghao Qi, and Andrew M. Weiner. Microresonator Kerr frequency combs with high conversion efficiency. Laser & Photonics Reviews, 11(1):1600276, 2017.
30. [30] Jae K. Jang, Yoshitomo Okawachi, Mengjie Yu, Kevin Luke, Xingchen Ji, Michal Lipson, and Alexander L. Gaeta. Dynamics of mode-coupling-induced microresonator frequency combs in normal dispersion. Opt. Express, 24(25):28794–28803, Dec 2016.
31. [31] Sven Ramelow, Alessandro Farsi, Stéphane Clemmen, Jacob S. Levy, Adrea R. Johnson, Yoshitomo Okawachi, Michael. R. E. Lamont, Michal Lipson, and Alexander L. Gaeta. Strong polarization mode coupling in microresonators. Opt. Lett., 39(17):5134–5137, Sep 2014.
32. [32] 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. Journal of Lightwave Technology, 33(6):1241–1246, Mar 2015.
33. [33] Steven A. Miller, Yoshitomo Okawachi, Sven Ramelow, Kevin Luke, Avik Dutt, Alessandro Farsi, Alexander L. Gaeta, and Michal Lipson. Tunable frequency combs based on dual microring resonators. Opt. Express, 23(16):21527–21540, Aug 2015.
34. [34] T. Herr, V. Brasch, J. D. Jost, I. Mirgorodskiy, G. Lihachev, M. L. Gorodetsky, and T. J. Kippenberg. Mode spectrum and temporal soliton formation in optical microresonators. Phys. Rev. Lett., 113:123901, Sep 2014.
35. [35] Matteo Cherchi, Sami Ylinen, Mikko Harjanne, Markku Kapulainen, and Timo Aalto. Dramatic size reduction of waveguide bends on a micron-scale silicon photonic platform. Opt. Express, 21(15):17814–17823, Jul 2013.
36. [36] E. A. Douglas, Patrick Mahony, Andrew Starbuck, Andy Pomerene, Douglas C. Trotter, and Christopher T. DeRose. "Effect of precursors on propagation loss for plasma-enhanced chemical vapor deposition of SiN x: H waveguides." Optical Materials Express 6, no. 9 (2016): 2892-2903.
37. [37] L. Wang, W. Xie, D. V. Thourhout, H. Yu, and S. Wang, Optics Express 26, 9645-9654.
38. [38] Xingchen Ji, Samantha P. Roberts, and Michal Lipson. High quality factor PECVD Si3N4 ring resonators compatible with CMOS process. In Conference on Lasers and Electro- Optics, page SM2O.6. Optical Society of America, 2019.
39. [39] Songtao Liu, Justin Norman, Mario Dumont, Daehwan Jung, Alfredo Torres, Arthur C. Gossard, John E. Bowers, 'High-Performance O-Band Quantum-Dot Semiconductor Optical Amplifiers Directly Grown on a CMOS Compatible Silicon Substrate', ACS Photonics 2019 6 (10), 2523-2529.
40. [40] Songtao Liu, Xinru Wu, Daehwan Jung, Justin C. Norman, M. J. Kennedy, Hon K. Tsang, Arthur C. Gossard, and John E. Bowers. High-channel-count 20 GHz passively mode-locked quantum dot laser directly grown on Si with 4.1Tbit/s transmission capacity. Optica, 6(2):128–134, Feb 2019.
41. George Michelogiannakis, Yiwen Shen, Min Yee Teh, Xiang Meng, Benjamin Aivazi, Taylor Groves, John Shalf, Madeleine Glick, Manya Ghobadi, Larry Dennison, and Keren Bergman. 2019. Bandwidth steering in HPC using silicon nanophotonics. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '19). Association for Computing Machinery, New York, NY, USA, Article 41, 1–25. DOI:https://doi.org/10.1145/3295500.3356145
42. 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," 2018 IEEE Optical Interconnects Conference (OI), Santa Fe, NM, 2018, pp. 25-26, doi: 10.1109/OIC.2018.8422036.
Nathan
43. 41 N. C. Abrams, Q. Cheng, M. Glick, M. Jezzini, P. Morrissey, P. O'Brien, and K. Bergman, Silicon Photonic 2.5D Integrated Multi-Chip Module Receiver*, Conference on Lasers and Electro-Optics*, May 2020.
44. 42 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, Journal of Lightwave Technology, March 2020.
45. 43 N. C. Abrams, Q. Cheng, M. Glick, E. Manzhosov, M. Jezzini, P. Morrissey, P. O'Brien, K. Bergman, Design Considerations for Multi-Chip Module Silicon Photonic Transceivers, Photonics West, February 2020.

First A. Author (M'76–SM'81–F'87) and the other authors may include biographies at the end of regular papers. This author became a Member (M) of IEEE in 1976, a Senior Member (SM) in 1981, and a Fellow (F) in 1987. The first paragraph may contain a place and/or date of birth (list place, then date). Next, the author's educational background is listed. The degrees should be listed with type of degree in what field, which institution, city, state, and country, and year degree was earned. The author's major field of study should be lower-cased.

The second paragraph uses the pronoun of the person (he or she) and not the author's last name. It lists military and work experience,

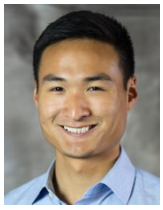
including summer and fellowship jobs. Job titles are capitalized. The current job must have a location; previous positions may be listed without one. Information concerning previous publications may be included. Try not to list more than three books or published articles. The format for listing publishers of a book within the biography is: title of book (city, state: publisher name, year) similar to a reference. Current and previous research interests end the paragraph.

The third paragraph begins with the author's title and last name (e.g., Dr. Smith, Prof. Jones, Mr. Kajor, Ms. Hunter). List any memberships in professional societies. Finally, list any awards and work for committees and publications. If a photograph is provided, the biography will be indented around it. The photograph is placed at the top left of the biography. Personal hobbies will be deleted from the biography.

Songtao Liu received the B.E. degree (Hons.) in electronic information science and technology from Henan University, Kaifeng, China, in 2012,

and the 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.

John E. Bowers (F') received the 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.



Alan Y. Liu (S'13 – M'16) received the Ph.D. degree in electronic and photonic materials from the University of California, Santa Barbara. He is currently the CEO of Quintessent, which he co-founded 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 Shalf received his BS and MS 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, John was deputy director of hardware technology for the DOE Exascale Computing Project (ECP).

George Michelogiannakis received the BSc and MSc degrees with honors from the University of Crete, Greece, and PhD from Stanford University in 2012 where he was selected for the Stanford Graduate Fellowship. George 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.**

Manya Ghobadi received her PhD 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. Manya is a computer systems researcher with a networking focus and has worked on a broad set of topics, including data center 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 Google research excellent paper award.

Dr. 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 3-D 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.
.