PINE: Photonic Integrated Networked Energy efficient datacenters
Maximizing system-wide energy efficient data movement benefits of photonics: PINE Approach 3 Key Concepts:

<table>
<thead>
<tr>
<th>Energy optimized high bandwidth density links</th>
<th>Optically Connect MCMs – CPU/GPU/Memory</th>
<th>Adaptive connectivity bandwidth steering</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>#2</td>
<td>#3</td>
</tr>
</tbody>
</table>

- **Energy optimized high bandwidth density links**: "Compute" MCM
- **Optically Connect MCMs – CPU/GPU/Memory**: III/V Semiconductor based comb laser
  - Optical fibers
  - Photonic die
  - CPUs, GPUs or router
- **Adaptive connectivity bandwidth steering**: CMP1, CMP2, GPUs, MEMs, NIC1, NIC2

---

[Diagrams and illustrations]
**PINE: Photonic Integrated Networked Energy efficient datacenters**

**Project Management: PI Prof. Bergman, Columbia University**

<table>
<thead>
<tr>
<th>Task 1: PINE System: architectural design, evaluation, and demonstration</th>
<th>Task 2: Optically Interconnected Multi-Chip Modules</th>
<th>Task 3: Dense WDM Silicon Photonic Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>- System wide bandwidth steering</td>
<td>- MCM active interposer integration</td>
<td>- Energy/throughput optimized links</td>
</tr>
<tr>
<td>- Node level deep disaggregation</td>
<td>- Embedded photonic switch fabric</td>
<td>- Integrated comb laser source</td>
</tr>
<tr>
<td>- PINE experimental system testbed</td>
<td>- Low Power EIC</td>
<td>- Quantum dot active devices</td>
</tr>
<tr>
<td></td>
<td>- Robust high density optical IO</td>
<td>- Tech transfer PDK development</td>
</tr>
</tbody>
</table>

**Task 4: PINE Technology to Market - Technologies Supply Chain, Datacenter/HPC Systems**
Key PINE Concept #1: Global Link Optimization
Max Bandwidth Density / Min Energy Design
PhoenixSim: Integrated Link Design and Modeling Environment

Link Optimization

- Sweep FSR of rings
- Channel Spacing
- Modulator characteristics
- Filter array geometry

- Power penalty is sensitive to the alignment of laser and resonance of modulator
- Power penalty exhibits two minima that are related to IL and ER

Ex: Microring Modulator

On resonance

Off resonance

Ex: Design Space of Filter

\[ \kappa = \sin \left( \frac{\pi}{\lambda} \sqrt{\gamma - \gamma_{0}^{2} B(x_{0})} \right) \]

Curvature function (does not depend on the gap)

Other factors: optical modes; coupling gap

Max Data Throughput

Lowest Power/bit
Interleaver

- Ring modulator: 10 Gb/s or 25 Gb/s
- 800G Links: 80 or 32 channels
- More than one de-interleaving stages
Power Consumption Analysis (25Gbs/channel)

-1.5dBm (0.7mW)

Input power per channel

1.7dB
6.5dB

1.7dB
1.7dB 1.9dB

-15.0 dBm

Total: 3.2 pJ/bit
Power Consumption Analysis (10 Gbs/channel)

Input power per channel

-2.5dBm (0.56mW)
WPE: 10%

Sensitivity of Receiver @ 10Gb/s

Total: 2.2 pJ/bit
Energy Optimized Performance – 800G

- Aggregate bandwidth: 800G per link
- 8 groups of 10G x 10 vs 4 groups of 25G x 8
Energy optimized links – Comb laser

- First integrated comb generator (electrical power < 100 mW and AAA battery-power operation)

- Demonstrated long-term (10 days) stability using simple feedback and thermal heaters.

- Designed, simulated, microresonator device for highly efficient WDM source in normal regime.
First Monolithic QD SOA on silicon – record WPE at 14.2%

- Peak gain is occurred at 1315 nm with 39 dB on-chip gain.
- Output saturation power reaches 24 dBm.
- Noise figure as low as 8 dB is demonstrated.

Gain fitting

\[ G = G_0 \frac{1 + P_{in}/P_s}{1 + G_0 P_{in}/P_s}, \]

1315 nm measured
1315 nm fitted

On-chip amplifier gain (dB)

Saturation power: 24 dBm

On-chip input power (dBm)

Gain mapping

On-chip gain mapping

On-chip input power (dBm)

Fiber to fiber noise figure (dB)

Noise figure

\[ N = \frac{2 \cdot P_{ASE}}{G h v B_0} + \frac{1}{G} \]
Passive Alignment – High Density Optical Fiber IO

• Plug-and-play fiber to waveguide coupling

• Demonstrated robustness to temperature and fabrication variations.
• Temperature fluctuations of >80 degrees have penalty of less than 0.6 dB coupling efficiency.
• Designed and fabricated the funnel for inverse taper coupling.
Plug-and-Play high-density chip IO coupling

- Edge coupling between patterned polymer waveguide matched to Silicon mode delocalized in the polymer cladding.
- Compatible with plug-and-play coupler arrays approach.
PINE Scaling chip ‘escape’ bandwidth density

NVIDIA NVSwitch

- 18 NVLink 2.0 ports → 9 per long edge top/bottom
- 50GB/s per port (25GB/s each Tx/Rx)
- 1 NVLink ~ 2mm of linear edge
- 50GB/s per 2mm → 200Gb/s/mm

PINE Dense WDM Silicon Photonic:
- 250um fiber pitch
- 8 fiber links ~ over 2mm linear edge
- 80 λs per fiber link; each λ at 10Gb/s = 800 Gb/s per link
- 6.4 Tb/s per 2mm → 3.2 Tb/s/mm

Each fiber ‘pin’ carries 80 λs (scalable)

Fiber pitch 250um
Key PINE Concept #2: Ubiquitous Optically Connected-MCM

- OC-MCM: Optically Connected Multi-chip Module
  - Optical communication among interposers/MCM
  - Enables fully flexible and scalable architectures

Approach: Deep Disaggregation
...network of Resources
Interconnected 2.5D/3D node (MCM)
PINE v1 TI Transceiver Prototype

- 2.5D on interposer packaging process
- PICs and EICs assembled at Tyndall
PINE Version 1 2.5D Packaging

- Packaging completed both version 1 prototypes (TI and Cisco TIAs)
- TI (left) and Cisco with DC breakout board (right)
- Packaging iterations improved coupling robustness
PINE v2 2.5D Transceiver Trench Prototype

- Requires larger distance between EIC and PIC (trench keep out area and 300 um separation suggested by Tyndall)
- Optical fanout from PIC trench to dicing channel

Columbia/Cisco co-design
Custom designed EICs (28nm CMOS)
Scalable MRR space-and-wavelength switch fabric architecture

- FSR dominant by the ring radius
  - tolerant to fabrication variations

- \( N \times M \lambda \) in Crossbar layout for each plane
  - Spatial selection in the column
  - Wavelength selection in the row
  - Wavelength aggregated using a large comb ring
Scalable 16x16 Multilayer Photonic Switching Fabric

- A 16×16 switch using AIM MPW
- Modular 12 units of 4x4 switch blocks
- Total 384 microring cells and 240 on-chip Ge PDs
PINE Key Concept #3: Adaptive, Flexible Connectivity with Bandwidth Steering

Applications Traffic → Un-steered connectivity → Bandwidth Steered Connectivity
In a standard Fat Tree, ToR EPSs are only connected to their parent aggregation EPSs.
Inter-pod traffic (defragmentation) needs to traverse (+ hops) through spine to connect between pods.
Spine often oversubscribed, energy consuming large routers.
Flexible Fat Tree: Insertion of SiP Switches – bandwidth steering

Application 1
Application 2
Application 3

Pre-Publication Accepted to SC19: G. Michelogiannakis et al. “Bandwidth Steering for HPC using Silicon Nanophotonics”
Flexible Fat Tree: Direct connectivity with bandwidth steering
Bandwidth Steered Flex Fat Tree: Throughput Improvement

Transactions per second

- SiP Switches 16x16

- Communication-intense traces higher benefits:
  - Neckbone/AMG: 13x
  - AMG/MiniDFT: 10x
  - AMR/AMG: 7.8x
  - Facebook (2699): 2.6x
  - Facebook (3245): 66%
  - MILC/MiniDFT: 56%
Bandwidth Steered Flex Fat Tree: Latency Improvement

- All benchmarks: 5% to 20% average hop reduction

- Communication-intense benchmarks: 28% to 87% average latency reduction
  – Due to less congestion
Bandwidth Steering Architecture – Fat Tree Topology: PINE Prototype System Testbed

- 32 compute nodes composed of VMs on 16 servers, with 10G NICs
- EPSs virtually partitioned from two OpenFlow PICA8 Ethernet packet switches (48 10G SFP+ ports)
- Bandwidth Steering – Demonstrated with four 4x4 SiP switches

Pre-Publication Accepted to SC19: G. Michelogiannakis et al. “Bandwidth Steering for HPC using Silicon Nanophotonics”
System Experimental Results – Standard Fat Tree vs Flexible Fat Tree

- Operating skeletonized GTC application with MPI
- Standard Fat Tree:
  - all upper layer links used
  - runtime = 72 sec
- Flexible PINE Fat Tree:
  - direct lower layer connectivity
  - only 4 upper layer links utilized
  - runtime = 56 sec
Flexible Fat Tree: Removal of Upper Layer Links

- PINE Flexible Fat Tree – provides direct connectivity for traffic at the lower layer
- Remove spine layer links to reduce energy consumption
System Results: Standard Fat Tree vs Flexible Fat Tree with Links Removed

- **Flexible Fat Tree:**
  - direct lower layer connectivity
  - Can remove upper layer EPSs = reduced energy consumption

- **Standard Fat Tree:** all remaining upper layer links congested
- runtime = 115 sec (vs 72 sec)

- **Flexible Fat Tree:**
  - unaffected by removal of links, remains at 56 sec runtime

69% execution time difference
ENLITENED – PINE T2M Highlights

• Our **industry partners** have shown their commitment to the PINE technology and vision by continuing and increasing their involvement in the PINE project and proposal for Phase 2.
  – NVIDIA is increasing participation in collaboration in PINE Phase 2.
  – Cisco is continuing collaboration with Columbia device design and know-how towards Cisco products. Developing energy efficient metal-oxide-semiconductor (MOS)-based ring and racetrack modulators targeting 10G and 25G NRZ.

• **Patents and IP:**
  • Patents have been filed for the Columbia comb and couplers
  • Patent pending UCSB
    – Dislocation Glide Suppression for Misfit Dislocation Free Heteroepitaxy
  • A disclosure has been submitted through LBNL
    – Bandwidth Steering in HPC Using Silicon Nanophotonics

• **Start-ups**
  • Three **start-ups**, based on PINE technology in different phases of formation;
    – Columbia comb laser
    – Columbia couplers
    – UCSB quantum dot based lasers and SOAs (now Quintessent)
ENLITENED – PINE T2M Highlights

PUBLICATIONS PRODUCED UNDER PROJECT

Journals

ENLITENED – PINE T2M Highlights

Conference proceedings

Conference proceedings cont.


- Qixiang Cheng, Liang Yuan Dai, Meisam Bahadori, Nathan C. Abrams, Padraic Morrissey, Madeleine Glick, Peter O'Brien, Keren Bergman, “Si/SiN Microring-Based Optical Router in Switch-and-Select Topology”, accepted to ECOC 2018, paper We1C.3


