PINE: Photonic Integrated Networked Energy efficient datacenters

Keren Bergman
Columbia University

October 24, 2018
PINE Team
PINE: Key Drivers

Key Concept 1: Energy Optimized WDM Links
Key Concept 2: Multi-Chip Modules with Photonic Connectivity
Key Concept 3: Bandwidth Steering (Reprogrammable Network)
Maximizing the data movement benefits of photonics:

<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>

| ![Image 1] | ![Image 2] | ![Image 3] |

- III/V Semiconductor based comb laser
- CPUs, GPUs or router
- Optical fibers
- Photonic die
- "Compute" MCM
- "Storage" MCM
Microsoft – Dataset

- Microsoft Azure data centers:
  - Compute, Storage, HPC, GPU clusters
- Total number of data centers: **134**
- Total number of switches: **85,000**
- Total number of links: **10,000,000**
- Packet loss, link utilization, buffer sizes, and traffic patterns

- **Keep the utilization low**
- **Hope for no major packet losses**

- **Packet loss is bad**
- **Paradox**
- **Billions of packet discards every day**

**Key questions**

**1. Architecture:** Which layer has high packet drops/link utilization?

**2. Application:** Do certain types of traffic exchange/drop more?

**3. Temporal:** Can we predict periods of high loss/utilization?

**Manya Ghobadi**
Programmable physical layer photonic connectivity

- Efficient networks ($/Gbps)
- Network size
- PINE programmable interconnect

- Treat the physical layer as a black box

- Cost & Complexity
- Simple solutions
- Intelligent solutions
Key PINE Concept #1: Energy Optimized WDM Links

**PhoenixSim**: a multi-layer simulation and modeling software solution that enables design automation and interactive design space exploration over the physical, networking and application layers.

<table>
<thead>
<tr>
<th>PINE Power Analysis</th>
<th>PINE Improvements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Link Consumption</td>
<td>2.2 pJ/bit</td>
</tr>
<tr>
<td></td>
<td>1.3 pJ/bit</td>
</tr>
<tr>
<td>Link Power Margin</td>
<td>3.0 dB</td>
</tr>
<tr>
<td></td>
<td>3.0 dB</td>
</tr>
<tr>
<td>Link Consumption (no margin)</td>
<td>1.9 pJ/bit</td>
</tr>
<tr>
<td></td>
<td>1.2 pJ/bit</td>
</tr>
</tbody>
</table>
Optimizing power consumption

Energy Optimized Driver

Thermal Control ‘tapless’

Energy Optimized Receiver

Optical Power (µW)

ΔI (µA)
Minimizing link losses and penalties

Optical Coupling:
Robust, passive, highly misalignment tolerant

Parameter optimization

[Image: Diagram of optical coupling from electronics to silicon waveguide to electronics]

Laser
Silicon waveguide

From Electronics
Photonic Elements
To Electronics

Det

Optimum Curve

Total Filter Penalty (dB)

10 Gb/s
50 Channels
10 Channels

Quality Factor (x 1000)

[Bahadori et al., Optical Interconnects 2015]
Energy/Throughput co-Optimization

- 5.1 dBm (0.31 mW) Laser Power Required

WPE: 30%

Power Analysis

<table>
<thead>
<tr>
<th>Power Analysis</th>
<th>PINE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Link Consumption</td>
<td>2.24 pJ/bit</td>
</tr>
<tr>
<td>Link Power Margin</td>
<td>3.0 dB</td>
</tr>
<tr>
<td>Link Consumption (no margin)</td>
<td>1.93 pJ/bit</td>
</tr>
</tbody>
</table>

Sensitivity of Receiver @ 10Gb/s

Laser Power Required

Energy/bit:
- 59%
- 23%
- 8%
- 10%
Laser: Fully Integrated Dense WDM Comb Source

- SiN micro-resonators generate combs → multi-wavelength source
- Compact, low-power comb sources
- Power consumption < 100 mW
- Operated with AAA battery.

Gaeta, Lipson
Energy Efficient Sources

Quantum dot comb laser - UCSB

Record high performance Fabry-Perot QD lasers with lowest lasing threshold and longest lifetime of more than a million hours at 35°C.
First QD mode locked lasers directly grown on Si

- 20 GHz fundamental mode locked laser with a wide spectrum and low jitter
  - Threshold: 42 mA to 58 mA as the SA reverse bias voltage increases from 0 V to 2 V
  - Wide mode locking area: $I_{\text{gain}}$ from 75 mA to 200 mA and $V_{SA}$ from 1 V to 5 V.
  - Lowest timing jitter reported for semiconductor passively mode locked lasers (82.7 fs integrated from 4 – 80 MHz)
1280nm DFB WDM laser array performance

- High optical power
- High efficiency singlemode laser
- Developed for WDM Array
  - >200mW
- HR/AR design:
- ~36% Wall-plug efficiency @ 20C

SBIR Collaborative Project (R. Carlson): Photonic Memory Controller Module (P-MCM)
Passive – High Density Optical Fiber IO

- Plug-and-play fiber to waveguide coupling

Mode matching between fiber and polymer waveguide. 
Ultra-Low Loss: Average coupling efficiency for 18 devices: -0.05 dB

Variations of ± 4µm on fiber mode diameter have a penalty of less than 1dB on the coupling efficiency.
PINE Scaling chip ‘escape’ bandwidth density

- 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
- 64 λ.s per fiber link; each λ at 10Gb/s = 640 Gb/s per link
- 5.12 Tb/s per 2mm → 2.56 Tb/s/mm
Key PINE Concept #2: Ubiquitous Optically Connected-MCM

- Optical communication among interposers/MCM
- Universal interface - Builds on recent industry efforts (Gen-Z, OpenCAPI, CCIX)
- Enables fully flexible and scalable architectures

Approach: network of resources ...rather than a network of servers

Fiber carrying 0.5 - 1 Tb/s

High-Density fiber coupling array (~ 1-10 TB/s)

Example: memory module:
6 packages, 48GB, 1.5 TB/s
One package: 8GB, 0.25 TB/s
Multi-Chip Modules with Photonic Connectivity

- SUNY-Poly CNSE capabilities for fabrication of 2.5D and 3D interposers.
- The interposer is leveraged to develop electronic / photonic multi-chip modules (MCM).
Active interposer technology combines PIC and interposer into a single platform.

- Shortens RF paths
- **Switch**: 8x8 MZI
- **RX EIC**: TI TIAAs
- Switch, modulators, laser integration - full Network-on-Chip
PINE: 3D MCM Integration

- 3D integration provides a platform for lower parasitics along the critical path
- Single channel RX and clock (minimize RX parasitics between EIC and PIC)
- Active interposer combines benefits of both approaches
High-Performance Photonic Switch Fabrics

- Highly Scalable, Small Footprint, Switch-and-Select MRR Topology
- Reduced Cross-Talk and Loss Overhead

Leverage 2-layer Si/SiN integration

Switch-and-select topology fully blocks the first-order crosstalk
Triple-layer 8x8 MRR Switch

- Thermo-optic and electro-optic MRR-based 8x8 switch-and-select switches
- Additional SiN transfer layer inserted
- Further reduce insertion loss, reduced power consumption
Disaggregated Data Center Architecture

Current server

Current rack

Disaggregated rack

Pool and compose

However... Inter Node Bandwidth (10 GB/s) << RAM Bandwidth (100 GB/s – 1 TB/s)
PINE Key Concept #3: Adaptive, Flexible Connectivity with Bandwidth Steering → Deep Disaggregation

• Introduce optical switches in the OC-MCM topology
  – 8x8 realizable with today’s technology
  – Tens of switches can be collocated on a single chip

• Flexibly assembled nodes

• Transparent for packets
  – Low-Latency direct connectivity
  – Energy efficient
Structure “Assembled” with Flexible Interconnect
Bandwidth Steering – System Modeling

Baseline Architecture

Network Parameters

Traffic Pattern

BookSim

PINE Architecture Generator

PINE Architecture Configurator

Flexible interconnect layer

Saturation load

Baseline

Configuration 1-1

Conf. 1-N

Configuration k-1

Conf. k-N

PINE arch #1

PINE arch #k
Bandwidth Steering – System Modeling

Map to Physical Hardware

• Optical switches steer channel connections
Bandwidth Steering Results

- **Bandwidth Increase compared to Vanilla (%):**
  - Facebook 2699: 46.3%
  - Facebook 3245: 6.0%

- **Average packet latency decrease compared to Vanilla (%):**
  - Facebook 2699: 35.2%
  - Facebook 3245: 12.7%

- **Maximum packet latency decrease compared to Vanilla (%):**
  - Facebook 2699: 31.7%
  - Facebook 3245: 5.6%

- **Achieved packet injection rate (%):**
  - Ethernet Infiniband: 0.2%
  - NVLINK PINE 7x7: 6.2%
  - PINE 14x14: 8.8%
PINE Bandwidth Steering Architecture: Prototype System Testbed

- 64- node system arranged in Dragonfly topology.
- Bandwidth steering performed through multiple embedded 4x4 silicon photonic switches
- Operates various HPC benchmarks such as GTC, HPCG, MiniFE
Bandwidth Steering: accelerated ‘time to solution’

Prototype Topology

GTC Traffic Matrix (from simulations)

Strong neighbor to neighbor group traffic

Experimental Results – Comparison of Performance
Summary:

- Data Movement is Critical to any Future Performance Scaling
  - Power Consumption
  - Bandwidth Density (and Cost)

- Photonics: System-Wide Tb/s per ‘wire’ and 1 pJ/bit
  - Ultra bandwidth dense WDM photonic links
  - Energy efficiency and cost via co-integration 2.5D and 3D platforms
  - High bandwidth Optically Connected Memory/GPUs/CPUs

- Deeply disaggregated Architectures
  - Optical connectivity for flexibly assembled interconnectivity topologies

- Computer architecture landscape is changing rapidly - Data Analytics, AI
  - Optical bandwidth steering, adaptable architectures for scalability
  - Ultimate energy efficiency – use only required resources for needed time period