Optical PCB Overview

Jeff Kash, Dan Kuchta, Fuad Doany, Clint Schow, Frank Libsch, Russell Budd, Yoichi Taira, Shigeru Nakagawa, Bert Offrein, Marc Taubenblatt

November, 2009

Electrical BW Bottlenecks → Optics Opportunities

- Electrical Buses become increasingly difficult at high data rates (physics):
  - Increasing losses & cross-talk; Frequency resonant affects
- Optical data transmission:
  - Power Efficiency, much less lossy, not plagued by resonant effects
- Physical size of electrical connections (BGA, connector) limits number of connections
  - Optical density ~10X higher

Rack | Backplane | Module | Card
---|---|---|---
OPTICS | OPTICS | CIRCUIT BOARD | OPTICS

Optical Backplane: 10 Gb/s, 62.5 in pitch
Estimated limit of MCM Electrical Escape
Limit of Electrical Backplane BW
Evolution of Rack-to-Rack Optics in Supercomputers

VCSEL-based Optics has displaced electrical cables today

2002

NEC Earth Simulator
- no optics

IBM Federation Switch for ASCI Purple (LLNL)
- Copper for short-distance links (≤10 m)
- Optical for longer links (20-40m)
- ~3000 parallel links 12+12@2Gb/s/channel

2005

IBM Roadrunner (LLNL) Cray Jaguar(ORNL)

- Introduced in 2008
- Still #1 as of June, 2009
- 4X DDR Infiniband (5Gb/s)
- 55 miles of Active Optical Cables

Exponential Growth in Supercomputing Power

BW requirements must scale with System Performance, ~1B/FLOP (memory & network)

Requires exponential increases in communication bandwidth at all levels of the system ➔ Inter-rack, backplane, card, chip
## The Road to Exascale: Cost and power of a supercomputer

<table>
<thead>
<tr>
<th>Year</th>
<th>Peak Performance</th>
<th>Machine Cost</th>
<th>Total Power Consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>1PF</td>
<td>$150M</td>
<td>2.5MW</td>
</tr>
<tr>
<td>2012</td>
<td>10PF</td>
<td>$225M</td>
<td>5MW</td>
</tr>
<tr>
<td>2016</td>
<td>100PF</td>
<td>$340M</td>
<td>10MW</td>
</tr>
<tr>
<td>2020</td>
<td>1000PF (1EF)</td>
<td>$500M</td>
<td>20MW</td>
</tr>
</tbody>
</table>

**Assumptions:** Based on typical industry trends –
- 10X performance / 4yrs (from top500 chart)
- 10X performance costs 1.5X more
- 10X performance consumes 2X more power

---

## The Road to Exascale: Total bandwidth, cost and power for optics in a machine

<table>
<thead>
<tr>
<th>Year</th>
<th>Peak Performance</th>
<th>(Bidi) Optical Bandwidth</th>
<th>Optics Power Consumption</th>
<th>Optics Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>1PF</td>
<td>0.012PB/s (1.2x10^9Gb/s)</td>
<td>0.012MW</td>
<td>$2.4M</td>
</tr>
<tr>
<td>2012</td>
<td>10PF</td>
<td>1PB/s (10^9Gb/s)</td>
<td>0.5MW</td>
<td>$22M</td>
</tr>
<tr>
<td>2016</td>
<td>100PF</td>
<td>20PB/sec (2x10^9Gb/s)</td>
<td>2MW</td>
<td>$68M</td>
</tr>
<tr>
<td>2020</td>
<td>1000PF (1EF)</td>
<td>400PB/sec (4x10^9Gb/s)</td>
<td>8MW</td>
<td>$200M</td>
</tr>
</tbody>
</table>

**Require >0.2Byte/FLOP I/O bandwidth, >0.2Byte/FLOP memory bandwidth**
- 2008 optics replaces electrical cables (0.012Byte/FLOP, 40mW/Gb/s)
- 2012 optics replaces electrical backplane (0.1Byte/FLOP, 10% of power/cost)
- 2016 optics replaces electrical PCB (0.2Byte/FLOP, 20% of power/cost)
- 2020 optics on-chip (or to memory) (0.4Byte/FLOP, 40% of power/cost)
In HPC space, increased need for and use of optics → cost and power must decrease (per bit unidirectional is shown)

<table>
<thead>
<tr>
<th>Year</th>
<th>Peak Performance</th>
<th>number of optical channels</th>
<th>Optics Power Consumption</th>
<th>Optics Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>1PF</td>
<td>48,000 (@5Gb/s)</td>
<td>50mW/Gb/s (50pJ/bit)</td>
<td>$10/Gb/s</td>
</tr>
<tr>
<td>2012</td>
<td>10PF</td>
<td>2x10^6 (@10Gb/s)</td>
<td>25mW/Gb/s</td>
<td>$1.1/Gb/s</td>
</tr>
<tr>
<td>2016</td>
<td>100PF</td>
<td>4x10^7 (@10Gb/s)</td>
<td>5mW/Gb/s</td>
<td>$0.17/Gb/s</td>
</tr>
<tr>
<td>2020</td>
<td>1000PF (1EF)</td>
<td>8x10^8 (@10Gb/s)</td>
<td>1mW/Gb/s</td>
<td>$0.025/Gb/s</td>
</tr>
</tbody>
</table>

- Table is based on historical trends for HPCs
- To get optics to millions of units in HPC, need ~$1/Gb/s unidirectional
  - Cost targets continue to decrease with time below that
- Power is OK for 2012, then sharp reductions will be needed

HPC driving volume optics → Higher volumes → lower Cost

A Single machine in the next few years could be similar to today’s WW parallel optics

- WW volume in 2008
- HPC single machine
- Commercial use
- e.g. 100GE
- MareNostrum
- ASCI Purple
- Roadrunner

Industry trend derived roadmap, not IBM product plans
Optical Printed Circuit Boards and Components: Enabling mass manufacturing

**Electronics:** Wires and discretes …

**Optics:** Fibers and modules…

… to Printed Circuit Boards with electrical components … to integrated waveguides on PCBs with optical components

---

Optical waveguide interconnects:

The Terabus project and related work

Future Vision: optically-enabled MCM’s

- Low-density, conventional electrical interface for power & control
- High-density, wide and fast optical interfaces for data I/O
- Much higher off-module bandwidth at low cost in $$ and power

<table>
<thead>
<tr>
<th>Circa 2014-2016</th>
<th>Power</th>
<th>Cost</th>
<th>Datarate</th>
<th>Density</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;10mw/Gbps (EOE)</td>
<td>&lt;$0.25/Gbps (TRx + on-card optics)</td>
<td>20Gbps/channel</td>
<td>2 Tbps/cm² (on module)</td>
<td>&lt; 10 FIT per channel</td>
</tr>
</tbody>
</table>
Full Terabus Link (985-nm): 2 Transceiver Optomodules on Optocard

**TRX1:**
16TX + 16RX

16 Channels TRX1 → TRX2 at 10Gb/s + 16 Channels TRX1 ← TRX2 at 10Gb/s

Optical interconnect density exceeds electrical:

Another reason for optics – Electrical Packaging is Running out of Pins – Optics Can Help

For each 10Gb/s signal you need:
- 3 or 4 electrical pins (differential pair) on 800µm pitch
- 1 OE array element (250µm pitch)

Optics density exceeds electrical by more than 10X
Comparison of Electrical and Optical Interconnect Density

Electrical Packaging is mature
- # Signal Pins & BW/pin not increasing at rate of silicon
- All-electrical packages will not have enough pins or per-pin BW

Chokepoint is at the module-to-circuit board
- Optics density is much greater than electrical there

Advantages of Polymer Waveguide Technology compared to Parallel Fiber Optics

- Integrated mass manufacturing
  - Board, sheet, film level processing of optical interconnects
  - Lower assembly, waveguide jumper costs
  - Costs for wide busses should scale better

- Simple assembly
  - Avoid fiber handling
  - Similar assembly procedures as for electrical components and boards
  - Establish electrical and optical connections simultaneously
  - Avoid separate optical layer (if integrated with board)

- New or compact functionality supporting new architectures
  - Shuffles, Crossings, splitters, …
  - Enabler for multi-drop splitting & complex re-routing that is expensive in fiber

- Higher density, waveguide pitch < 125 um (best future fiber pitch)
  - Higher bandwidth density, less signal layers

- Cost
  - Reduced Optics module cost AND jumper/connector costs both important
  - Possible Lower Maintenance costs
Optical PCB Roadmap

Initial FOCUS

2012 - 2014
Discrete Active Optical Flex assemblies are easily replaced, less disruptive

Optical waveguides/ modules on PCB (could be flex laminate) once technology matures to sufficient reliability

Optical Waveguides embedded in card w/optical vias are a complete replacement for all high speed copper lines

2014 - 2016

Use of active waveguide flex and fiber cable in systems

TODAY

Drawer

MCM

Electrical from MCM to back of drawer, electrical to backplane and to electrical connector for active optical cable

Backplane

MCM

Drawers attach to electrical backplane (PCB)

Rack to Rack

Active Optical fiber based cable: Rugged (may be routed under floor), 3-100m lengths

FUTURE with OPCB

Drawer

Active waveguide flex from MCM to back of drawer, may include channel shuffles

Backplane

Optical cable based optical backplane, may be passive waveguide flex cables or fiber cables (initially more likely). May include channel shuffles, typically < 1m-2m

Rack to Rack

Passive Optical fiber based cable: Rugged (may be routed under floor), 3-100m lengths
Optical Printed Circuit Boards

- IBM Research has invested heavily in the past 5 years in Optical printed circuit board technology based on multi-mode polymer waveguides
  - Partially funded by the US Government (Terabus program)
- We believe this technology will be needed to provide the needed BW for future server generations, allow highly integrated electrical-optical links and provide a path to much lower cost optical links.
- We are interested in establishing a market eco-system that will provide components, standards and specifications for this technology.