Cost-Effective Optics: Enabling the Exascale Roadmaps

Hot Interconnects 17
August 27, 2009

Alan Benner, bennera@us.ibm.com

IBTA – Chair, ElectroMechanical Working Group
IBM Corp – Sr. Technical Staff Member, Systems & Technology Group
The Question:

- The question motivating this Hot Interconnects Special Session was:

  “Will cost-effective optics fundamentally change the landscape of networking?”
The Answer:

Yes
Yes ..... and No
The Answer:

Yes ..... and No

What exactly did you mean by “..the landscape of networking..” ?
# The Landscape of Networking

<table>
<thead>
<tr>
<th>Physical Link Types</th>
<th>MAN &amp; WAN</th>
<th>Cables – Long</th>
<th>Cables – Short</th>
<th>Backplane / Card-to-Card</th>
<th>Intra-Card</th>
<th>Intra-Module</th>
<th>Intra-chip</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Distinguished by</strong></td>
<td><strong>Length</strong></td>
<td><strong>Typical # lanes per link</strong></td>
<td><strong>Use of optics</strong></td>
<td><strong>Length</strong></td>
<td><strong>Typical # lanes per link</strong></td>
<td><strong>Use of optics</strong></td>
<td><strong>Length</strong></td>
</tr>
<tr>
<td><strong>PHYSICAL</strong></td>
<td><strong>Link Types</strong></td>
<td><strong>Cables – Long</strong></td>
<td><strong>Cables – Short</strong></td>
<td><strong>Backplane / Card-to-Card</strong></td>
<td><strong>Intra-Card</strong></td>
<td><strong>Intra-Module</strong></td>
<td><strong>Intra-chip</strong></td>
</tr>
<tr>
<td>MAN &amp; WAN</td>
<td><strong>Cables – Long</strong></td>
<td>10, - 300 m</td>
<td>1 m - 10 m</td>
<td>0.3 m - 1 m</td>
<td>0.1 m - 0.3 m</td>
<td>5 mm - 100 mm</td>
<td>0 mm - 20 mm</td>
</tr>
<tr>
<td><strong>Typical # lanes per link</strong></td>
<td><strong>Use of optics</strong></td>
<td>Since 80s</td>
<td>Since 90s</td>
<td>2009-2010</td>
<td>2010-2015</td>
<td>Probably after 2015</td>
<td>Later</td>
</tr>
<tr>
<td><strong>Use of optics</strong></td>
<td><strong>Since 80s</strong></td>
<td><strong>Since 90s</strong></td>
<td><strong>Now</strong></td>
<td><strong>2009-2010</strong></td>
<td><strong>2010-2015</strong></td>
<td><strong>Probably after 2015</strong></td>
<td><strong>Later</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Logical Link Types</th>
<th>Internet</th>
<th>Local Area Network</th>
<th>Cluster / Data Center</th>
<th>Storage Area Network</th>
<th>Direct Attach Storage</th>
<th>I/O</th>
<th>Mezzanine Bus</th>
<th>SMP Coherency Bus</th>
<th>Memory Bus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Distinguished by</strong></td>
<td><strong>Function &amp; Link Protocol</strong></td>
<td><strong>Traffic:</strong></td>
<td><strong>Standards:</strong></td>
<td><strong>Traffic:</strong></td>
<td><strong>Standards:</strong></td>
<td><strong>Traffic:</strong></td>
<td><strong>Standards:</strong></td>
<td><strong>Traffic:</strong></td>
<td><strong>Standards:</strong></td>
</tr>
<tr>
<td><strong>Logical</strong></td>
<td><strong>Link Types</strong></td>
<td><strong>Internet</strong></td>
<td><strong>Local Area Network</strong></td>
<td><strong>Cluster / Data Center</strong></td>
<td><strong>Storage Area Network</strong></td>
<td><strong>Direct Attach Storage</strong></td>
<td><strong>I/O</strong></td>
<td><strong>Mezzanine Bus</strong></td>
<td><strong>SMP Coherency Bus</strong></td>
</tr>
<tr>
<td><strong>Interoperability with “Everybody”</strong></td>
<td><strong>Interoperability with “Everybody”</strong></td>
<td><strong>IP</strong></td>
<td><strong>IP</strong></td>
<td><strong>Intra-application, or intra-distributed-application</strong></td>
<td><strong>Shared tech between servers &amp; desktops</strong></td>
<td><strong>Reliability</strong></td>
<td><strong>Reliability &amp; cost vs. DRAM</strong></td>
<td><strong>Comming later</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Standards:</strong></td>
<td><strong>Standards:</strong></td>
<td><strong>Ethernet, ATM, SONET,</strong></td>
<td><strong>Ethernet, WiFi</strong></td>
<td><strong>InfiniBand, 1G Ethernet, 10/40/100Enet</strong></td>
<td><strong>SAS, SATA</strong></td>
<td><strong>Shared tech between servers &amp; desktops</strong></td>
<td><strong>Reliability</strong></td>
<td><strong>Reliability &amp; cost vs. DRAM</strong></td>
<td><strong>Comming later</strong></td>
</tr>
<tr>
<td><strong>100-300m over RJ-45 / CAT5 cabling, or wireless BW &amp; latency to &lt;60 meters</strong></td>
<td><strong>100-300m over RJ-45 / CAT5 cabling, or wireless BW &amp; latency to &lt;60 meters</strong></td>
<td><strong>1G Ethernet, WiFi</strong></td>
<td><strong>InfiniBand, 1G Ethernet, 10/40/100Enet</strong></td>
<td><strong>Fibre Channel</strong></td>
<td><strong>PCI/PCIe</strong></td>
<td><strong>SMP Coherency Bus</strong></td>
<td><strong>SMP Coherency Bus</strong></td>
<td><strong>Memory Bus</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Shared tech between servers &amp; desktops</strong></td>
<td><strong>Shared tech between servers &amp; desktops</strong></td>
<td><strong>Shared tech between servers &amp; desktops</strong></td>
<td><strong>Reliability</strong></td>
<td><strong>Reliability &amp; cost vs. DRAM</strong></td>
<td><strong>Comming later</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Reliability</strong></td>
<td><strong>Reliability</strong></td>
<td><strong>Reliability &amp; cost vs. DRAM</strong></td>
<td><strong>Comming later</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Key Characteristics:**
- Interoperability with “Everybody”
- IP
- Ethernet, ATM, SONET,
The Landscape of Networking – Changed by cost-effective optics?

### PHYSICAL Link Types

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>MAYBE</td>
<td>MAYBE</td>
</tr>
</tbody>
</table>

**Distinguished by Length & Packaging**

<table>
<thead>
<tr>
<th>Length</th>
<th>Multi-km</th>
<th>10, - 300 m</th>
<th>1 m - 10 m</th>
<th>0.3 m - 1 m</th>
<th>0.1 m - 0.3 m</th>
<th>5 mm - 100 mm</th>
<th>0 mm - 20 mm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical # lanes per link</td>
<td>1</td>
<td>1 - 10s</td>
<td>1 - 10s</td>
<td>1 - 100s</td>
<td>1 - 100s</td>
<td>1 - 100s</td>
<td>1 - 100s</td>
</tr>
<tr>
<td>Use of optics</td>
<td>Since 80s</td>
<td>Since 90s</td>
<td>Now</td>
<td>2009-2010</td>
<td>2010-2015</td>
<td>Probably after 2015</td>
<td>Later</td>
</tr>
</tbody>
</table>

### LOGICAL Link Types

<table>
<thead>
<tr>
<th>Internet</th>
<th>Local Area Network</th>
<th>Cluster / Data Center</th>
<th>Storage Area Network</th>
<th>Direct Attach Storage</th>
<th>I/O</th>
<th>Mezzanine Bus</th>
<th>SMP Coherency Bus</th>
<th>Memory Bus</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>MAYBE</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
</tbody>
</table>

**Distinguished by Function & Link Protocol**

<table>
<thead>
<tr>
<th>Key Characteristic</th>
<th>Inter-operability with “Everybody”</th>
<th>100-300m over RJ-45 / CAT5 cabling, or wireless</th>
<th>BW &amp; latency to &lt;60 meters</th>
<th>Domained by FC</th>
<th>Shared tech. between servers &amp; desktops</th>
<th>Shared tech. between servers &amp; desktops</th>
<th>Reliability</th>
<th>Reliability, massive BW, reliability</th>
<th>Reliability &amp; cost vs. DRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use of optics</td>
<td>Since 80s</td>
<td>Maybe Never? (Wireless, Building re-wiring, BW demand)</td>
<td>Since 2000s</td>
<td>Since 90s</td>
<td>Not yet</td>
<td>Scattered</td>
<td>Not yet</td>
<td>Coming</td>
<td>Coming later</td>
</tr>
</tbody>
</table>
Impact of Network performance on system scaling

- Above right: Obligatory reference to Top500.org list, showing that the majority of *processing power* on the list is interconnected with InfiniBand interconnect.
  - Note: Counting by system – rather than by processing power – is unfair, as regards interconnect.
  - A ~2K-CPU “bottom 10” (#490-500) system needs <<1/10th of the interconnect HW vs. a Top 10 system – but they both count as 1 “system”, on the list.

- Below right: Top500 data, filtered to show the impact of interconnect on system performance.
  - Net: Replacing 1Gigabit Ethernet with InfiniBand improves overall performance by 45% - 75% on the same nodes.
  - Larger systems see more impact – no surprise
  - Linpack depends relatively *little* on interconnect performance – other apps see more or less impact

- For 10Gb Ethernet, I have been unable to find any data, except for <<1K cores – doesn’t even make the Top500 list.
  - I’ve been waiting since 2003 for a Top500 10GbE system – 10GbE has been too expensive versus IB & 1GbE – performance not good, either
  - ?Maybe *this* year?
Roadmaps to Exascale have been well explored in the DARPA/IPTO study “ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems”, by Peter Kogge (former IBM Fellow, now at Notre Dame), et. al. Available at http://www.nd.edu/~kogge/reports.html

A couple key points:
- “The single most difficult and pervasive challenge perceived by the study group dealt with energy, namely,...energy per operation”
- “[The] energy in data transport will dwarf the traditional computational component in future Exascale systems....particularly so for the largest data center class.”

Data transport is *the* energy problem.
- Perhaps 200x more energy needed to transport a bit from a nearest-neighbor chip than to operate on it.
  - Energy needed for a floating-point operation (~’13-’16): 0.1-0.05 pJ/bit
  - Energy needed for data transport (on-card, ~6”,~’13-’16): 2-10 pJ/bit
- Across a big system (50meter diameter, multi-stage network, with routers and multiple transceiver hops), the factor may be 1000x.

Summarizing:
- Other than some software and reliability issues associated with 100-Million to 1Billion-way CPU parallelism, scaling to Exascale is hard but straightforward -- *unless* data needs to move between chips.
  - Yes, 100 Million to Billion-way

However, data does need to move between chips.
Next Steps, 2010-2013: Practical Petascale Blue Waters System

- **Target: #1 productivity supercomputer in 2011:**
  - 1-2 PetaFLOP/s Sustained (~10 PF Peak)

- **Selected Statistics:**
  - More than 200,000 cores.
  - More than 1 PetaByte of memory.
  - More than 10 PetaBytes of user disk storage.
  - More than 0.5 Exabyte of archival storage.
  - Up to 400 Gbps external connectivity.

- **Uses: Modeling Very Complex Systems**
  - Cells, Organs, and Organisms
  - Hurricanes, (incl. storm surge, ..)
  - Galaxy formation in early universe
  - Effect of Sun’s corona on Earth’s ionosphere
  - Design: Aircraft, Jet engines, motors, fusion,
  - Atom-level New materials design
  - …

- **Reference:** [www.ncsa.uiuc.edu/BlueWaters/](http://www.ncsa.uiuc.edu/BlueWaters/)
### Assumptions: Based on typical historical trends (see, e.g., top500.org and green500.org):

- 10X performance, 4 years later, costs 1.5X more dollars
- 10X performance, 4 years later, consumes 2X more power

<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>

Acknowledgment: J. Kash
How much optics, and at what cost?

<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 (1.2×10^5Gb/s)</td>
<td>0.012MW</td>
<td>$2.4M</td>
</tr>
<tr>
<td>2012</td>
<td>10PF</td>
<td>1PB/s (10^7Gb/s)</td>
<td>0.5MW</td>
<td>$22M</td>
</tr>
<tr>
<td>2016</td>
<td>100PF</td>
<td>20PB/sec (2×10^8Gb/s)</td>
<td>2MW</td>
<td>$68M</td>
</tr>
<tr>
<td>2020</td>
<td>1000PF (1EF)</td>
<td>400PB/sec (4×10^9Gb/s)</td>
<td>8MW</td>
<td>$200M</td>
</tr>
</tbody>
</table>

- **Target >0.2Byte/FLOP I/O bandwidth plus >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 system power/cost)
  - 2016 optics replaces electrical PCB (0.2Byte/FLOP, 20% of system power/cost)
  - 2020 optics on-chip (or to memory) (0.4Byte/FLOP, 40% of system power/cost)

Acknowledgment: J. Kash
## Cost and Power per bit (unidirectional)

<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>$2 \times 10^7$ (@10Gb/s)</td>
<td>25mW/Gb/s</td>
<td>$1.1/Gb/s</td>
</tr>
<tr>
<td>2016</td>
<td>100PF</td>
<td>$4 \times 10^7$ (@10Gb/s)</td>
<td>5mW/Gb/s</td>
<td>$0.17/Gb/s</td>
</tr>
<tr>
<td>2020</td>
<td>1000PF (1EF)</td>
<td>$8 \times 10^8$ (@10Gb/s)</td>
<td>1mW/Gb/s</td>
<td>$0.025/Gb/s</td>
</tr>
</tbody>
</table>

### Future directions for optical cables:
- Lower cost (reducing >60%/year)
- Much more BW (increasing >210%/year)
- Much lower power (improving >45%/year)

### Variety of methods for reaching these targets
- Higher bitrates: 10-20-20 Gb/s per channel
- Smaller footprint for O/E modules
- Move optics closer to logic
- New technologies

Acknowledgment: J. Kash
Can electrical transmission solve the data transport problem?
Can electrical transmission solve the data transport problem?

- No
Can electrical transmission solve the data transport problem?

- **No**

**Hard limits due to**
- Attenuation at high frequencies
- Connectors: physical bandwidth density relative to silicon BW density at acceptable reliability
- Size and weight of high-bandwidth copper cables (even with active copper)
- Power efficiency:
  - Typical standard (IB, 10GbE, FC) link budget is for ~20-25 dB of loss between Tx and Rx – i.e., >99% of transmitted power is lost before it reaches the receiver -- *Terribly* energy inefficient.

Can optical transmission solve the data transport problem?

- Yes
Can optical transmission solve the data transport problem?

- Yes – with a *lot* of work.
Can optical transmission solve the data transport problem?

- Yes – with a *lot* of work.

- There are a lot of questions yet to be answered:  (i.e., ➔R&D Opportunities Galore⇐)
  - How are the optical components packaged?
  - What media?
    - Multi-mode glass fiber is completely dominant today for <80m optical data center & cluster links
    - Single-mode would be great, other than XCVR and connector costs
    - Polymer optical fibers? – Maybe for high channel count?
    - Polymer Waveguides very attractive for on-card & Mid⇒Edge links – can loss be reduced more?
  - What wavelength / wavelengths?
    - Shortwave 850 nm (VCSEL) is dominant today, in US – may have more troubles at >>10 Gb/s
    - Interesting work happening at 980 nm and 1100 nm (mostly Japan)
    - 1310 nm is common for short single-mode
    - 1500-1600 nm range most flexible for “fancier” optics (WDM, silicon photonics, single-mode,...)
  - What Transmitter & Receiver Architectures?
    - Direct-modulated? (VCSELs or Horizontal-cavity?)
    - Modulators? (Silicon modulators? (MZ or Rings?) SOAs? How are supply lasers attached? )
    - Receivers – III-V PINs? APDs? Ge on silicon?
  - What’s the best bit-rate per channel
  - Where does optics make sense within the systems?
Vision for 2012: Packaging of optical interconnects is critical

| 2012 | 25mW/Gb/s | $1.1/Gb/s |

- Reach cost target with simplified packaging and higher volumes?
- Better to put optics close to logic rather than at the card edge
  - Avoids power, distortion & cost of electrical link on each end of optical link
  - Breaks through pin-count limitation of multi-chip modules (MCMs)

Optics on-card:

- Operation at 10 Gb/s: equalization required
  - ~2cm 50Ω traces
- Operation to >15 Gb/s: no equalization required
  - ~2cm 50Ω traces

Bandwidth limited by # of pins

- NIC
- Ceramic
- Organic card
- 1cm Flex
- Laser+driver IC
- Fiber

Optics on-MCM

- Opto module
- Laser+driver IC
- Optical bulkhead connector


- But this will increase cabling costs…. @$1/fiber end, optical connector cost is $0.60/Gb/s
- Will that make CWDM economically attractive (III-V or Si photonics)?
Vision for 2016: Optical PCB’s: The Terabus Project

<table>
<thead>
<tr>
<th>Year</th>
<th>Power</th>
<th>Cost/GB/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>5mW/Gb/s</td>
<td>$0.17/Gb/s</td>
</tr>
</tbody>
</table>

Dense Hybrid Integration: demonstrate a low-cost packaging approach compatible with conventional PCB manufacturing and surface-mount board assembly

- 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
  - We are working on MM VCSEL interconnects, can Si photonics compete?

Future Vision: optically-enabled MCMs (Multi-Chip modules)
Vision for 2020: Optically connected 3-D Supercomputer Chip

<table>
<thead>
<tr>
<th>Year</th>
<th>1mW/Gb/s</th>
<th>$0.025/Gb/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>2020</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **36 “Cell” 3-D chip**
- Silicon photonics layer integrated with high performance logic and memory layers
- Layers separately optimized for performance and yield

**Logic plane** ~300 cores, ~5TF (36 “supercores”)

**Memory plane** ~30GB eDRAM

**Photonic plane** On-Chip Optical Network
- >20 Tbps (bidirectional) optical on-chip (between supercores)
- >20 Tbps optical off-chip

Photonic layer not only connects the multiple cores, but also routes the traffic

System level study:
IBM, Columbia, Cornell, UCSB
Si Photonics for \textit{all} of off-chip comm links
- Electrical link to optical transceiver cannot be a transmission line, or power budget is broken – must be integrated on or next to processing or routing chip
- On-chip network links stay electrical

Strawman Target: \((20+20)\) Tb/s optical off-chip bandwidth
- \(\approx 40\) W per chip for optics at 1 mW/Gb/s
- \(\approx 1000\) per chip for optics at \$0.025/Gb/s
- \(4000\) 10 Gb/s channels or \(2000\) 20 Gb/s channels (~200 fibers or waveguides with DWDM)
- Other factors TBD

(Please see the previous “Can optical transmission solve…?” slide…)
Acknowledgements


- IBM’s Terabus Polymer Waveguide Interconnects & Silicon Photonics programs, and PERCS/HPCS program are partially supported by DARPA

Summary

▪ The Question:

“Will cost-effective optics fundamentally change the landscape of networking?”

▪ The Discussion:

ƒRegardless of potential slow-downs in Moore’s law (as stated – transistors/chip) or “Moore’s law” (price/performance), the even-faster System-level improvement rate will continue.

nBetter packaging, better cooling, 3-D, better interconnect, better system design, etc, etc., …

ƒThe only exception is electrical chip-chip I/O & interconnect – which can’t continue at this rate, at viable energy-efficiency (pJ/bit at required distances).

nHard physical & signal integrity barriers in attenuation & physical size vs. silicon chips

▪ The Short Answer:

ƒNet: Yes

Cost effective optics will fundamentally change the landscape of networking ---- by making it *possible* for the rest of the IT landscape to keep on improving as it has in the past ---- with a *lot* of work.
Backups
Exponential growth in system-level performance: almost 2x/year

- Systems-level improvements have been much faster than chip-level Moore’s-law
  - Parallel system performance increasingly comes from high-level interconnection of increasingly-parallel chips & boxes
- BW requirements must scale with system performance, ~1B/FLOP (memory + networks)
  - Requires exponential increases in communication bandwidth at all levels of the system

Note: Top500’s artificial benchmark (Linpack) has interconnect requirements midway between many “real” supercomputing apps & data center apps ➔ similar trends & CAGRs apply to data centers.

- CPU Trend: ~50-60% (2x/18 mo.)
- Parallel System Trend: (~90%) = CPU trend + increased parallelism
- System Level Performance
  - 10X performance every 3.5-4 years
  - 1PFlop=10^{15} Flop
  - Roadrunner 1PFlop
  - 06/13/2008 http://www.top500.org
  - 1PFlop=10^{15} Flop
  - 06/13/2008 http://www.top500.org

Performance Development

System Level Performance

- Cluster/Parallel: ~90% CAGR, continuing
- Box: 70-80% CAGR, continuing
- Uniprocessor: 50% CAGR, slowing
- Transistors & Pkg: 15%-20% CAGR, slowing

CAGR = Compound Annual Growth Rate

Roadrunner

1PFlop=10^{15} Flop
Evolution of Rack-to-Rack Optics in Supercomputers

**2002**

- NEC Earth Simulator
  - no optics

**2005**

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

**2008: 1PF/s**

- IBM Roadrunner (LLNL)
  - Cray Jaguar (ORNL)

- InfiniBand
- 3 miles of optical cables, longest = 60m

*http://www.nccs.gov/jaguar/
*http://www.lanl.gov/roadrunner/
Device Challenges for On-Chip Si Photonics

**Silicon Photonics Today**

- **Ring modulator**
  - Radius 30µm

- **WDM filters**
  - 1.4mm x 0.6mm = 0.85mm²
  - Add. power for tuning

- **Deflection Switch**
  - Macroscopic

- **On-chip optical amplifier**
  - Macroscopic
  - Not integrated

- **Optical Coupling**
  - Hand-aligned
  - Low numbers

**Future: Integrated 3-D chip**

- **Radius 6µm**
- **50x100µm = 0.005mm²**
  - No active feedback, accuracy is defined by fabrication
- **30° temp range**

- **Area 25X**
  - Power 10X

- **Area 150X**
  - No active tuning
  - Thermally stable

- **Nanoscale**
  - No active tuning
  - Thermally stable

- **Mass manufacturable III-V on Si**
  - (~1000/chip)
  - Wafer scale

- **Mass manufacturable alignment (~200 fibers)**
  - Package in presence of C4s and heat sink
Device Challenges for On-Chip Si Photonics

Silicon Photonics Today

- **Ring modulator**
  - Radius 30 μm
  - Area 25X
  - Power 10X

Goal: Replace Cu cables with fibers
- **Main driver:** cost/bit
- **~100 Gb/s → ~10 links**
- Direct competition with other technologies traditionally perceived as expensive (e.g. InP)

On-chip optical amplifier
- **Macroscopic**
- **Not integrated**

Optical Coupling
- **Hand-aligned**
- **Low numbers**

Future: Integrated 3-D chip

- **Goal:** Replace Cu wiring with Si photonics
- **Main drivers:** power/bit, cost/bit
- **>>10 Tbps → 1000’s of links/chip**
- Direct competition with “free” BEOL Cu

- **Radius 6 μm**
- **Area 150X**
- **No active tuning**
- **Thermally stable**

Macroscopic
**30° temp range**

- **Mass manufacturable III-V on Si (~1000/chip)**
- **Wafer scale**

- **Optical Gain Block**
  - Mass manufacturable alignment (~200 fibers)
  - Package in presence of C4s and heat sink
Supercomputers are used to gain more insight into complex systems

- Improve understanding – **significantly larger scale, more complex and higher resolution models; new science applications**
- Multiscale and multiphysics – **From atoms to mega-structures; coupled applications**
- Shorter time to solution – **Answers from months to minutes**
Supercomputer performance is improving steadily

An example:

Weather Simulation

Supercomputers = CPUs + DIMMs + Power + Cooling + **Interconnect**

Supercomputers & High-End Servers have many many cables. A few real-world scenarios:
Optics will play an increasing role in supercomputers as they approach the Exascale.

Parallel optical interconnects are fast replacing copper cables today:
- Steady cost reduction (~60%/yr) is critical to wider adoption, including optical backplane circa 2012.
- Single wavelength multimode VCSEL-based links appear to be lowest cost and lowest power.
  - Readily extensible to 10-20 Gb/s, perhaps 5mW/Gb/s.
  - Position optics near logic for largest benefits.

If cost can be further lowered, optically-enabled circuit cards based on polymer waveguides will be deployed, circa 2012-2016:
- Optical board manufacturing “ecosystem” needs to evolve.
- Work today is single wavelength multimode VCSEL-based, could migrate to CWDM.
- Or even singlemode DWDM with cheap Si photonics.

Optics directly on the chip for on- and off-chip global interconnects is a future possibility:
- Drive is power savings for communications.
- Still at an early stage, basic building blocks being developed.
- Needs to be approached from a systems level, not individual devices.

Challenge/Question: How/when/where can/should “computer-com” industry evolve from multimode VCSEL/MMF links to singlemode WDM?

Take-home:
Optical interconnect for supercomputers and other high-end systems could be growing at >200% CAGR (deployed Gb/s), if cost can be improved at 60% CAGR ($/Gb/s) and power can be improved at 45% CAGR (mw/Gb/s) at the same time.

We’re banking on this happening.