Silicon photonics for next generation computing systems

Tutorial given at the European Conference on Optical Communications, September 22, 2008


Yurii Vlasov
IBM Research Division
Outline

- Hierarchy of interconnects in HPC
  - Active cables for rack-to-rack communications
  - Board level interconnects
  - On-chip optical interconnects
    - CMOS integration challenges
    - Photonic network on a chip
  - Silicon nanophotonics:
    - WDM
    - Light sources
    - Modulators
    - Switches
    - Detectors
- Conclusions

DISCLAIMER. The views expressed in this document are those of the author and do not necessarily represent the views of IBM Corporation.
Hierarchy of interconnects in HPC systems

- Multiple blades on shelf interconnected through an electrical backplane
- Optical interconnects between shelves
  - Interconnects moderated by the switch
- Many Tb/sec off node card and growing (20-50% CGR)
Moore’s law in HPCS

Biggest machine

209 of the top 500 supercomputers have been built by IBM
Sum total performance 14.03596 PF

After Jack Dongarra, top500.org
Roadrunner - #1 in the world

19,872 processors, 1 PFlops

Triblade:
2 Cells
Expansion module
2 Opterons

Area: • 296 racks
3.9 MW Power: 0.35 GF/Watt
Cluster of 18 Connected Units (CU)
104 TB aggregate memory

InfiniBand 4x DDR fat-tree fabric
2-stage fat-tree; all-optical cables
Full bi-section BW within each CU: 384 GB/s
Half bi-section BW among CUs: 3.45 TB/s
Optical interconnects between racks and switches

Infiniband active optical cables

one of 26 switches

From Ken Koch (LANL) presentation at SC2008 conference
Outline

- Hierarchy of interconnects in HPC
- Active cables for rack-to-rack communications
- Board level interconnects
- On-chip optical interconnects
  - CMOS integration challenges
  - Photonic network on a chip
- Silicon nanophotonics:
  - WDM
  - Light sources
  - Modulators
  - Switches
  - Detectors
- Conclusions

DISCLAIMER. The views expressed in this document are those of the author and do not necessarily represent the views of IBM Corporation.
Rack-to-Rack: Parallel Optics in Supercomputers

- Length: ~100m
- # links: ~5-10K
- BW: ~10Gbps/link
- Power: ~50mW/Gb/s/link

Price: few$ per Gbps
Reliability!

MareNostrum (Barcelona) 62TFlops
About 5000 fiber cables

MareNostrum central switch racks:
About 1700 fiber cables/rack today
Examples of silicon photonics integration

4 TRx @ 10Gbps: CMOS integrated: Si modulators+Ge detectors; Mux, VOA, CMOS drivers+TIA/LA; CDR, ADC, monitors, etc. Co-packaged lasers and fiber couplers 2.3W total \(\rightarrow\) \(\sim\) 50pJ/bit

1 TRx @ 10Gbps: CMOS integrated: Si modulators, CMOS drivers+amplifiers; CDR etc. Co-packaged detectors, lasers and fiber couplers 0.4W total \(\rightarrow\) \(\sim\) 40pJ/bit

Can Si Photonics bring $/Gbps down?
Outline

- Hierarchy of interconnects in HPC
- Active cables for rack-to-rack communications
- **Board level interconnects**
  - On-chip optical interconnects
    - CMOS integration challenges
    - Photonic network on a chip
  - Silicon nanophotonics:
    - WDM
    - Light sources
    - Modulators
    - Switches
    - Detectors
- Conclusions

**DISCLAIMER.** The views expressed in this document are those of the author and do not necessarily represent the views of IBM Corporation.
Optics between cards: Optical backplane

Length: ~50cm
# links: ~10K
BW: ~10Gbps/link
Power: ~10mW/Gb/s/link

Price: <1$ per Gbps
Reliability!!

- **Electronics**
  - Fast CMOS designs with low power consumption

- **Opto-electronics**
  - Separate from processor to avoid heat/lifetime issues
    - VCSEL/PD: 2D-arrays (cost)

- **Optical channel**
  - Fibers, fiber flexes
  - PCB embedded waveguides

- **Coupling**
  - Passive alignment and positioning

Courtesy of IBM ZRL – Christoph Berger, Bert Jan Offrein, Martin Schmatz
Multilayer optical waveguides in PCB

4x12 waveguides on 250µm x 250µm grid (proof of feasibility)

Light incoupling

Light outcoupling at end facet

4x12 waveguides

 Courtesy of IBM ZRL - Roger Dangel, Folkert Horst, Bert Offrein
Chip-to-chip link.

Length: ~1cm
# links: ~100K
BW: ~1Tbps/link
Power: <10mW/Gb/s/link

Price: <0.1$ per Gbps
Reliability !!!

IBM Terabus project

16 Channels TRX1 ⇔ TRX2

TRX1

Waveguides in PCB

TRX2

240 Gb/s bi-directional
9 mW/Gb/s per unidirectional link

Courtesy of IBM Terabus team – Jeff Kash, Clint Schow, Fuad Doany et al
Penetration of optics into HPC

Single HPC machine will contain a similar number of parallel optical channels as currently exists today in all telecommunications links worldwide.

All future dates and specifications are estimations only. Subject to change without notice.

Courtesy of M. Taubenblatt (IBM)
Outline

- Hierarchy of interconnects in HPC
- Active cables for rack-to-rack communications
- Board level interconnects
- On-chip optical interconnects
  - CMOS integration challenges
  - Photonic network on a chip
- Silicon nanophotonics:
  - WDM
  - Light sources
  - Modulators
  - Switches
  - Detectors
- Conclusions

DISCLAIMER. The views expressed in this document are those of the author and do not necessarily represent the views of IBM Corporation.
On-chip optical interconnects

Length: ~0.1-0.3cm
# links: ~100K
BW: ~1Tbps/link
Power: <1mW/Gb/s/link

Price: <<0.01$ per Gbps
Reliability !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

Footprint
Power
Bandwidth
Loss
Latency

Cell processor
(IBM/Toshiba/Sony)
9 cores, 256 GFlops,
Si photonics vs Si nanophotonics

Rack-to-rack; board-to-board

Replace Cu with fibers
Main driver: cost/bit
~10 Tx on a single chip
>40Gbps aggregate

1. Leverage CMOS fab as much as possible to reduce the cost/bit
2. Since high-performance CMOS is not required, previous gen CMOS fab is preferred (e.g. 130nm)
3. No change in server architecture
4. Footprint, power, loss is not as important – relevant comparison with Cu cables
5. Direct competition with other technologies (e.g. InP)

On-chip

Goal: Replace Cu with Si
Main driver: power/bit
~1000 Tx on a single chip
>1Tbs aggregate is a must

1. Integration within high-performance CMOS stack requires aggressive scaling of footprints
2. Advanced CMOS is a must (e.g. 45nm)
3. In 3D CMOS stack can provide significant architectural advantages.
4. Cost is not as important; relevant comparison with CMOS Cu BEOL wiring
5. Direct competition with low power BEOL Cu
On-chip global interconnects

For 130nm, 1.2V

Logic: ~0.1-1pJ/bit

Interconnects:
  on-chip: 3pJ/bit (3mW/Gbps)
  off-chip: 40-60pJ/bit (40mw/Gbps)

- M1 pitch scaling → RC delay is increasing exponentially
- Severe bandwidth limitations
- In a link with repeaters: power/Gbps/length
- Energy/bit scaling slows
- Bandwidth needs growth faster for multicore processors
- Exponentially growing power consumption
- BGA pitch scaling is limited – off-chip bandwidth bottleneck

After ITRS 2007
20Tflops on a chip

Biggest machine

Processor

0.5B/FLOP Interconnect bandwidth (TB/s)
Outline

- Hierarchy of interconnects in HPC
- Active cables for rack-to-rack communications
- Board level interconnects
- On-chip optical interconnects
  - CMOS integration challenges
  - Photonic network on a chip
- Silicon nanophotonics:
  - WDM
  - Light sources
  - Modulators
  - Switches
  - Detectors
- Conclusions

DISCLAIMER. The views expressed in this document are those of the author and do not necessarily represent the views of IBM Corporation.
Silicon nanophotonic waveguides

- Ultra-high optical confinement mode x-section 0.1\(\mu\)m\(^2\)

- With 2\(\mu\)m BOX losses ~1-2dB/cm

- Sharp bend (R~3\(\mu\)m) loss <0.001dB/turn

- Chromatic dispersion (10\(^3\) of D in fibers)

- Nonlinear optical effects (SPM, XPM, FWM, solitons, etc.)

1 meter waveguide with 0.3dB/cm


1.3Tbps through a single waveguide

B. Lee et al, IEEE PTL, 20, 398 (2008)
CMOS front end (FEOL) photonic integration

Advantages:
- Most dense integration
- Very high-performance (lowest power) photonic devices
- Same mask set, standard processing
- Same design environment (e.g. Cadence)
- Seamless use of W plugs and M1 Cu metal

Disadvantages:
- Needs modifications to CMOS
  - Ge, Ge doping, T constraints
- Scalability with CMOS is not guaranteed
  - Specialty SOI wafer
  - e.g. 45nm HiKMG, no poly-Si
- Temperature stability
- Xtalk (electrical and optical)
Waveguides in bonded SOI, Detection and modulation in III-V (bonded)
SiON, poly-Ge deposited on of BEOL
Poly-Si, poly-Ge deposited on of BEOL

Advantages:
- No CMOS FEOL modifications
- No CMOS contamination
- Potentially scalable with technology
- Integration with III-V high performance photonics

Disadvantages:
- Needs additional masks/processing
- Severe T constraints for processing
- Low performance Si/Ge devices, high losses, dark currents etc.
- III-V die to wafer bonding
- Difficult to utilize Cu interconnects
- Additional high AR vias to CMOS required
- Difficult to integrate with BGA
- Temperature stability
- Xtalk (electrical and optical)
3D integration

After J.Knickerbocker et al, ECTC 2006

After A.Topol et al, IBM JRD 2006

ITRS 2007

TSV

10^8 IO/cm^2

Wafer 1

Wafer 2

IBM water jet


3um pitch
Outline

- Hierarchy of interconnects in HPC
- Active cables for rack-to-rack communications
- Board level interconnects
- On-chip optical interconnects
  - CMOS integration challenges
  - **Photonic network on a chip**
- Silicon nanophotonics:
  - WDM
  - Light sources
  - Modulators
  - Switches
  - Detectors
- Conclusions

**DISCLAIMER.** The *views expressed* in this document are those of the author and *do not necessarily represent.* the *views of IBM Corporation.*
Vision for 22nm CMOS (circa 2018) - 10 TFLOPs on a 3D chip

- 36 “Cell” chip (~300 cores)

System level study: IBM, Columbia, Cornell, UCSB

Co-PIs:
Jeff Kash (IBM)
Keren Bergman (Columbia)
Yurii Vlasov (IBM)

- Logic plane: ~300 cores
- Memory plane: ~30GB eDRAM
- Photonic plane: On-Chip Optical Network
  - >70Tbps optical on-chip
  - >70Tbps optical off-chip

Photonic layer is not only connecting various cores, but also routes the traffic

All future dates and specifications are estimations only. Subject to change without notice.
On-chip optical network

Message is encoded in N WDM channels in a bit-parallel manner

Requires switching all the WDM traffic at once

No optical buffering. Switching of the whole message once per communication. Resembles circuit switch network.
Network on a chip: electronics vs photonics

75Tb/s off-chip & on chip switching

- **Electronics**
  - Power is bandwidth x length dependent
  - On chip: buffer, receive and re-transmit every single bit at every switch
  - Off-chip: even more power hungry (50Ω) and bandwidth limited by BGA count

  Electronic network ~500W

- **Optics**
  - Power independent of bitrate and length
  - Modulate/receive ultra-high bandwidth data stream once — no re-transmit (~15x communications power savings)
  - Off-chip and on-chip power and bandwidth are equivalent (~40x chip interconnect power savings)
  - Broadband switch fabric is nearly free in power dissipation → highly scalable

  Photonic network <80W

Power efficient computing. More FLOPs per Watt

All specifications are estimations only. Subject to change without notice.
Outline

- Hierarchy of interconnects in HPC
- Active cables for rack-to-rack communications
- Board level interconnects
- On-chip optical interconnects
  - CMOS integration challenges
  - Photonic network on a chip
- Silicon nanophotonics:
  - WDM
  - Light sources
  - Modulators
  - Switches
  - Detectors
- Conclusions

DISCLAIMER. The views expressed in this document are those of the author and do not necessarily represent the views of IBM Corporation.
Wavelength multiplexers

- Number of channels
- Xtalk
- Footprint
- Insertion loss
- Temperature tolerances
- Fabrication tolerances

- AWG
- MMI
- Lattice filters
- Ring resonator filters
- Echelle grating

Si thermooptic coefficient
\[ \Delta n/dT = 2 \times 10^{-4} \text{ K}^{-1} \]


non-IBM μ-processor @3.3 GHz, 1.4 Vdd

±15°C
Large GVD, large DGD
Relatively small variations of the wg width
Large phase errors
Large crosstalk

- 16-channel AWG, 200GHz
- 200µm x 500µm area
- 3dB insertion loss
- 15dB Xtalk
Echelle grating

- 650x170um
- Cross-talk 19db
- Insertion loss 3dB

“safe” design

- 250x200um
- Cross-talk 17db
- Insertion loss 4dB

“extreme” design

No active tuning. As-fabricated!

F. Horst et al., SPIE Europe, Feb. 2008
Prospects for ultimate integrated CMOS photonics

Die-to-die variation of a single ring

<table>
<thead>
<tr>
<th></th>
<th>WIDTH</th>
<th>GAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVERAGE</td>
<td>0.4802</td>
<td>0.1986</td>
</tr>
<tr>
<td>3SIGMA</td>
<td>0.0053</td>
<td>0.0087</td>
</tr>
<tr>
<td>RANGE</td>
<td>0.0065</td>
<td>0.011</td>
</tr>
<tr>
<td>MEASURED</td>
<td>22</td>
<td>24</td>
</tr>
</tbody>
</table>

Wafer-to-wafer variation of all rings

<table>
<thead>
<tr>
<th></th>
<th>WIDTH</th>
<th>GAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVERAGE</td>
<td>0.4843</td>
<td>0.1860</td>
</tr>
<tr>
<td>3SIGMA</td>
<td>0.0077</td>
<td>0.0090</td>
</tr>
<tr>
<td>RANGE</td>
<td>0.0168</td>
<td>0.0122</td>
</tr>
<tr>
<td>MEASURED</td>
<td>346</td>
<td>140</td>
</tr>
</tbody>
</table>

ITRS CMOS tolerances for different technologies

<table>
<thead>
<tr>
<th>CMOS generation</th>
<th>Litho</th>
<th>Total tolerance 3σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>130nm</td>
<td>248nm, OPC</td>
<td>40nm</td>
</tr>
<tr>
<td>65nm</td>
<td>193nm, OPC</td>
<td>35nm</td>
</tr>
<tr>
<td>45nm</td>
<td>193nm, OPC</td>
<td>16nm</td>
</tr>
<tr>
<td>22nm</td>
<td>EUV</td>
<td>1.2nm</td>
</tr>
</tbody>
</table>

Scaling of Si CMOS is done – performance will only increase with new generations of CMOS
Ring resonator WDM filters

40x70um (6 channels, 400GHz)
Cross-talk 30db
Insertion loss 1.3dB

F.Xia et al Optics Express (2007).

✓ 2 nm wide error-free operating window @ 10 Gbps
✓ Temperature independent operation ±15°C

Y. Vlasov, W. M. J. Green, and F. Xia, Nature Photonics 2008
Outline

- Hierarchy of interconnects in HPC
- Active cables for rack-to-rack communications
- Board level interconnects
- On-chip optical interconnects
  - CMOS integration challenges
  - Photonic network on a chip
- Silicon nanophotonics:
  - WDM
  - Light sources
  - Modulators
  - Switches
  - Detectors
- Conclusions

DISCLAIMER. The views expressed in this document are those of the author and do not necessarily represent the views of IBM Corporation.
Light generation

On-chip light source

- **Pro:**
  - No off-chip coupling necessary
  - Compact package
  - Efficient use of optical power

- **Con:**
  - III-V on Si?
  - Added on-chip power
  - Fast, compact directly modulated laser have to be developed
  - T control
  - FIT, Lifetime, Serviceability

Off-chip light source

- **Pro:**
  - Serviceable, replaceable
  - Power is off the on-chip equation
  - Any cheap CW (eg DFB)
  - T stability,

- **Con:**
  - Multiple on-chip fiber couplers
  - Packaging, alignment
  - Massive on-chip power distribution grid
  - Additional source of Xtalk, loss, etc.
On-chip lasers

- AlGaInAs membrane bonded on SOI wafer
  - Cavity defined by silicon waveguide (no critical alignment)
  - Length ~1mm

- 7.5-µm device coupled to Si wire
- CW lasing with $I_{th} = 0.6$mA, voltage $V_{th} = 1.5-1.7$V,


J. Van Campenhout et al. OFC 2007, PDP
Fiber couplers for off-chip lasers

D. Taillaert et al., IEEE JQE, 38, 949 (2002)


Wafer-level testing
Surfaces are occupied (heat sink, BGA)?

T Shoji et al, EL Dec. 2002
V.Almeida et al, OL Aug. 2002
Y.Vlasov et al, OE Nov. 2003

Chip edge

Fiber coupler array

100μm

Courtesy of IBM
Edge-coupling, no surfaces
Packaging?
Outline

- Hierarchy of interconnects in HPC
- Active cables for rack-to-rack communications
- Board level interconnects
- On-chip optical interconnects
  - CMOS integration challenges
  - Photonic network on a chip
- Silicon nanophotonics:
  - WDM
  - Light sources
  - Modulators
  - Switches
  - Detectors
- Conclusions

DISCLAIMER. The views expressed in this document are those of the author and do not necessarily represent the views of IBM Corporation.
Modulator with external CW laser
PIN modulator

Forward bias

High efficiency
$V\pi \cdot L = 0.036 \text{V}\cdot\text{cm}$

Low RF cutoff
$\sim 0.5\text{GHz}$

Diffusion and recombination in forward bias


Fast RF

- To work in reverse bias
- To increase the length
- Use traveling wave amplifier

40 Gbps open eye
~6 Vpp is combined with 3VDC
few mm long

A. Liu et al, IPNRA 2007, paper IMD3

- To work in forward bias
- Pre-emphasize signal

12.5 Gbps open eye
~16 Vpp pre-emphasized

Low power

✓ Resonant enhancement
✓ Slower than the cavity lifetime

1.8V @ 10Gbps; 85fJ/bit

M. Watts et al OFC 2008 PDP paper


Power scaling

✓ Start from non-resonant MZI
✓ IBM measured 5pJ/bit
✓ Linear (almost) scaling with F
On-chip interconnects fundamental tradeoffs: Footprint vs Power vs Bandwidth

On-chip design phase-space:
- Low Power ↓ 30X
- Small Footprint ↓ 30X
- Large Bandwidth ↑ 20X
- Small Loss ↓ 10X
- Fast response ↑ 4X

Example:
Modulator design strategies – stretch to achieve Low Power

1. Low power → resonant enhancement
   However high-Q resonance → limited BW

2. Low power → decrease serial resistance
   However higher doping → increased losses

3. Low power → increase interaction length
   However footprint is large

All these requirements contradict to each other – innovative engineering is necessary to achieve a tradeoff
Outline

- Hierarchy of interconnects in HPC
- Active cables for rack-to-rack communications
- Board level interconnects
- On-chip optical interconnects
  - CMOS integration challenges
  - Photonic network on a chip
- Silicon nanophotonics:
  - WDM
  - Light sources
  - Modulators
  - **Switches**
  - Detectors
- Conclusions

**DISCLAIMER.** The *views expressed* in this document are those of the author and *do not* necessarily *represent* the *views of IBM Corporation.*
Broadband Deflection Switch

Design Metrics:

- High throughput – 1Tbps
- Low latency – few ns
- Ultra-compact footprint < 0.001 mm²
- Low crosstalk < -25 dB
- High extinction ratio > 10 dB
- Temperature independent operation ±15°C
Microring Comb Switch

- Ultra-compact footprint
- Multi-channel operation
- Low-power, high-speed switching

**Drawback:**
- Not broadband: Sensitive to temperature fluctuations, fabrication tolerances

Switching Characteristics

Free carrier injection into central ring:
- Drop port transmission suppressed by >15dB
- Through port transmission increases to zero loss

Outline

- Hierarchy of interconnects in HPC
- Active cables for rack-to-rack communications
- Board level interconnects
- On-chip optical interconnects
  - CMOS integration challenges
  - Photonic network on a chip
- Silicon nanophotonics:
  - WDM
  - Light sources
  - Modulators
  - Switches
  - Detectors
- Conclusions

DISCLAIMER. The views expressed in this document are those of the author and do not necessarily represent the views of IBM Corporation.
Challenges for on-chip nanophotonics

Current Gen. CMOS

CMOS scaling

Area 2X smaller
Power per circuit 2.8X lower

+1 Gen. CMOS

Photonics scaling

Tele- and Data-com

Best available integrated photonics
(either InP or CMOS Si/Ge)

Area 10-20X
Power 30-100X

BW ~40-1000Gbps
Power ~ 50pJ/bit
Area ~0.5cm²

Equivalent to 5-10 generations of CMOS
Summary

- Requirements from system level should define the requirements and performance metrics for individual devices
- These requirements are very tough and almost impossible to meet
- Nevertheless at the moment it looks that there is a room for innovation