The Interconnect Problem: From Emerging Devices to Energy-efficient Networks

Vladimir Stojanović

Integrated Systems Group
Massachusetts Institute of Technology
High-speed links needed everywhere

PC or Console

Backbone Router Rack
What makes it challenging

- Requires sophisticated equalization circuits

High speed link chip

Channel response

<table>
<thead>
<tr>
<th>Frequency [GHz]</th>
<th>Attenuation [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>-10</td>
</tr>
<tr>
<td>4</td>
<td>-20</td>
</tr>
<tr>
<td>6</td>
<td>-30</td>
</tr>
<tr>
<td>8</td>
<td>-40</td>
</tr>
<tr>
<td>10</td>
<td>-50</td>
</tr>
</tbody>
</table>

9” FR4, via stub

26” FR4, via stub

source: Rambus
Chip-to-chip I/O scaling problem

- Bandwidth need grows faster than energy/bit drops
- Creates exponentially increasing I/O power consumption
  - In power constrained systems (like processors and anything inside the box) – this limits the available bandwidth
Parallel off-chip links

- Often share clock generation and synch
- Limited equalization (few taps)
- Most power burned to drive the 50 Ω line
  - Current-mode – 200 mV swing (4 mW off 1 V supply,
    - Data rate independent
    - With receiver and pre-driver, at 10 Gb/s energy budget 500 fJ/bit
  - Voltage-mode – (2 pJ/bit state-of-the-art, dynamic power)
    - Can possibly scale to 500 fJ/bit, but not much further
Convergence of platforms

Only way to meet future system feature set, design cost, power, and performance requirements is by programming a processor array
- Multiple parallel general-purpose processors (GPPs)
- Multiple application-specific processors (ASPs)

1000s of processor cores per die

“The Processor is the new Transistor” [Rowen]
On-chip interconnect: A system perspective

FPGAs, Multi-core/Network Processors, SoCs – using more routing resources

- Existing architectures
  - Designed to relieve the interconnect scaling problem
    - Latency, capacitance, loss
- But, interconnects will consume nearly 50% of chip power
  - Power and Area Efficiency becoming critical

Example MIT RAW numbers

Power Consumption
- Clock 26%
- Network 36%
- Computation 39%
Electronic shared memory switches

- Centralized switch (Bus, Crossbar)
  - Fast, but scales poorly
  - Power hungry
- Distributed switch (e.g. Mesh)
  - Slow, power hungry
  - Other on-chip networks – (fat tree, torus, etc)
    - Compromise between density, power and latency/data rate
- Memory interfaces
  - Power and density limit

- IBM CELL
  - DRAM interface
  - Bus

- Sun Niagara
  - Cross-bar
  - DRAM interfaces
  - Mesh
  - Rings

- Intel Terascale
  - UltraSPARC-Core
  - DRAM interfaces

MIT ISG
CICS
Center for Integrated Circuits and Systems
Massachusetts Institute of Technology
The problem spans many layers

Interconnect Problem

On-chip Network

Off-chip I/O

Network Architecture

Design Optimization

Communication (Eq., Mod, Coding)

Link modeling, Characterization

Circuits

Tx, Rx, Ctrl, Meas

Data Rate Density (Gbps/um)

Energy/Bit (pJ/Bit)

Equalized, 30mV Eye

Equalized, 50mV Eye

Equalized, 90mV Eye

Repead

CNTs

Si-Photonics

MIT ISG

MIT

Cu

[IBM]
Receivers in On-chip Networks

Intel-80 core Terascale

Architectural decision, Jin, HPCA 2007

Application benchmarks, Guo, SLIP 2006

Circuit + wire parameters
Equalized Interconnect in On-chip Networks

Intel-80 core Terascale

Equalized Interconnect

Models and Tools
- No performance and power models
- No modeling/tool framework

Challenges
- Joint Optimization: Circuit + Wire

![Graph showing data rate density vs. energy/bit](image)

```plaintext
\[ w = \left[ w_1, w_2, \ldots, w_{n_{VDD}} \right] \]

\[ h_1 \]

width, spacing

\[ D \]

transmit bits

receive bits

\[ \theta \]

No performance and power models
No modeling/tool framework
Joint Optimization: Circuit + Wire

Models and Tools

Challenges
Equalized Interconnect

No Equalization

Feed Forward Equalization (FFE)

FFE + Decision Feed Back Equalization (DFE)
Joint Optimization Problem: Communication + Circuits + Wires

- Large design space
  - Circuit parameters ($W_{LCM}$, $V_p$, $V_s$), wire dimensions ($W$, $S$), sample delay ($T_d$), equalization coefficients ($W_1$, $W_2$, $W_3$, $y_1$) (>400k points)

Connecting Performance and Power Models

- Need accurate channel model
  - Channel = Wire + Devices & Circuit

- Eye estimation
  - Sample time
  - Equalization coefficients
Result: M8, 10mm

Equalized Interconnect:
- \(x10\) more energy efficient for the same data rate density and latency
- Crosstalk doubles energy cost

Repeater and equalized interconnect

Worst case eye and LMSE method
- Auto generated spice netlist based on the parameters.
- Spice : 190 sec
- Formula-based : 6.5m sec
On-chip Link Test-chip

- 90nm test chip
  - IBM process
  - DDR muxing
- 8 Gb/s max
  - 4 Gb/s/um
- 0.5 pJ/bit @ 8 Gb/s
  - Eye > 100mV
  - 400 fJ/bit Tx
  - 100 fJ/bit TIA+DFE

- Trades-off data-rate density and energy-efficiency
What else can we do?

- Try new device technology
  - Carbon Nanotubes
  - Silicon Photonics
Estimating the impact of CNTs

Create Models for Circuits

CNT Process Characterization

Extract Tradeoff Curves

Make Informed Architecture Choices
Effective Resistivity of CNTs

- CNT bundles that are densely packed, fully & ideally contacted have a resistivity \( \sim 7x \) lower than Cu at 22nm.
- Non-ideal contact resistance amortized over length of nanotube:
  - Insignificant for lengths > \( \sim 1000 \) gate pitches.
Effective Resistivity: Non-idealities

\[ \rho_{EFF} = \frac{d^2}{k} \left( \frac{h}{4e^2 L_0} + \frac{R_{cont}}{L} \right), \]

where \( L > L_0 \) and \( L_0 = C_\lambda d \).

- Current growth limitations result in only ~2x lower resistivity than copper.
Scaling Interconnect: Design Space

- Consider rescaling the interlayer dielectric (ILD) stackup
  - Scale width: maintain minimum wire pitch
  - Increase ILD height (H), decreasing wire thickness (T): maintain constant wire bandwidth

- Integration of vertically aligned CNTs closer to realization

Copper OR CNT Interconnects

Copper OR CNT Vias
CNTs Require Rescaling

- Range of sub-optimally sized wires is greater if CNTs are used with the same cross-section as copper.
Scaling both width and height result in almost 33% energy savings for the same delay.

Characterization Limitations

**Measurement Limitations:**
- **Large impedance mismatch**
  - Measurement variance > than measured value
- **Large test parasitics**
  - Limit bandwidth of input signals
  - Need to de-embed, much larger than CNT

**Physical Limitations:**
- Cumbersome, difficult, unrepeatable test setups
- Limited CMOS integration
- Limited number of measurements
**CNT Characterization Test-chip**

- **Horizontally grown SWCNTs**
  - Grow CNTs on separate substrate
  - Transfer to testchip using resist
  - Use ball bonder, or metal deposition to make final pad contact

- **Vertically grown CNTs**
  - Create compatible CNT masks that align with test interface
  - Create contacts at the end of CNT chips
  - Align & mate using optical IR aligner
Extracting CNT Data

- Any pair of transceivers acts like a bi-directional link/equivalent time scope

- Sweep time & voltage to measure step response waveforms

- Extract $S$-parameters, delay, edge rates of both CMOS & CML signals
What else can we do?

- Try new device technology
  - Carbon Nanotubes
  - Silicon Photonics
Manycore interconnect bottlenecks

- Slow on-chip mesh
- On-chip memory controllers
  - Power and density limited memory BW
  - At most 3-6 Tb/s in next few years
  - Need to move them off-chip
    - Use Si-Photonics
Optical system architecture

- **Message path**
  - Electrical mesh to reach the appropriate access point
  - Core waveguide to the switch matrix
  - Statically routed through the switch matrix
  - DIMM waveguide and optic fiber to reach hub chip
  - Routed through hub chip to correct DRAM chip
  - Returns through the separate response networks

- **Si-photonic network**
  - Faster, more energy-efficient communication across the chip
  - Exports memory controllers and switching to DRAM hub chips
  - Overcomes power and density limitations on memory bandwidth
Optical multi-group system architecture

- Multi-group architecture
  - Can help alleviate on-chip interconnect bottleneck
  - Breaks the single on-chip electrical mesh into several groups
    - Each with its own smaller mesh
    - Each group still has one AP for each DIMM and thus can access all of memory
    - Since there are more APs each AP is narrower (uses less λs)
  - Uses optical network as a very efficient global crossbar
  - Hub networks now include a crossbar and arbitration

Potentially 40-80 Tb/s
Traffic generator modeling GUPS access pattern

For rough comparison the Cray X1E with 248 cores (1.13GHz) sustains 1.8 GUPS, while the Cray Red Storm/XT3 cluster with 12,960 cores (2.4GHz) sustains 30 GUPS.

http://icl.cs.utk.edu/hpcc
Link components

- 64 Tx (0.01 dB/ring)
- 64 Rx (0.01 dB/ring)
- PD (0.1 dB)
- SM fiber 10 cm
- Memory – buffer chip
- Multi-core chip

Components:
- 64 x λ laser source
- Coupler (1 dB)
- Insertion (1 dB)
- Core
- 1:32 splitter (0.2 dB/split)
- 2 cm (1 dB/cm)
- 0.9 cm (1 dB/cm)
- Drop (2.5 dB/drop)
- 0.05 dB/cross.
- 64 Crossings
- 64 filter pairs (through loss 0.01 dB/ring)
- 64 filter pairs (through loss 0.01 dB/ring)
- Drop (2.5 dB/drop)
- 2 cm
## Optical power budget

<table>
<thead>
<tr>
<th>Component</th>
<th>Preliminary Design</th>
<th>Power loss</th>
<th>Optimized Design</th>
<th>Power loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coupler loss</td>
<td>1 dB/coupler</td>
<td>3 dB</td>
<td>1 dB/coupler</td>
<td>3 dB</td>
</tr>
<tr>
<td>Splitter loss</td>
<td>0.2 dB/split</td>
<td>1 dB</td>
<td>0.2 dB/split</td>
<td>1 dB</td>
</tr>
<tr>
<td>Non-linearity</td>
<td>1 dB</td>
<td>1 dB</td>
<td>1 dB</td>
<td>1 dB</td>
</tr>
<tr>
<td>Through loss</td>
<td>0.01 dB/ring</td>
<td>3.17 dB</td>
<td>0.01 dB/ring</td>
<td>3.17 dB</td>
</tr>
<tr>
<td>Modulator Insertion loss</td>
<td>1 dB</td>
<td>1 dB</td>
<td>0.5 dB</td>
<td>0.5 dB</td>
</tr>
<tr>
<td>Crossing loss</td>
<td>0.2 dB/crossing</td>
<td>12.8 dB</td>
<td>0.05 dB/crossing</td>
<td>3.2 dB</td>
</tr>
<tr>
<td>On-chip waveguide loss</td>
<td>5 dB/cm</td>
<td>20 dB</td>
<td>1 dB/cm</td>
<td>4 dB</td>
</tr>
<tr>
<td>Off-chip waveguide loss</td>
<td>0.5e-5 dB/cm</td>
<td>~ 0 dB</td>
<td>0.5e-5 dB/cm</td>
<td>~ 0 dB</td>
</tr>
<tr>
<td>Drop loss</td>
<td>2.5 dB/drop</td>
<td>5 dB</td>
<td>1.5 dB/drop</td>
<td>3 dB</td>
</tr>
<tr>
<td>Photodetector loss</td>
<td>0.1 dB</td>
<td>0.1 dB</td>
<td>0.1 dB</td>
<td>0.1 dB</td>
</tr>
<tr>
<td>Receiver sensitivity</td>
<td>-20 dBm</td>
<td>-20 dBm</td>
<td>-20 dBm</td>
<td>-20 dBm</td>
</tr>
<tr>
<td>Power per wavelength</td>
<td>26.07 dBm</td>
<td>26.07 dBm</td>
<td>-1.03 dBm</td>
<td>-1.03 dBm</td>
</tr>
<tr>
<td>(0.40 W)</td>
<td>(0.40 W)</td>
<td>(0.78 mW)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Power required at source</td>
<td>3.3 kW</td>
<td>3.3 kW</td>
<td>6.38 W</td>
<td>6.38 W</td>
</tr>
</tbody>
</table>
## Data transmission latency

<table>
<thead>
<tr>
<th>Component</th>
<th>Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serializer/Deserializer (50ps each)</td>
<td>50ps</td>
</tr>
<tr>
<td>Modulator driver latency</td>
<td>108ps</td>
</tr>
<tr>
<td>Through latency (2.5ps/adjacent channel)</td>
<td>7.5ps</td>
</tr>
<tr>
<td>Drop latency (20ps/drop)</td>
<td>60ps</td>
</tr>
<tr>
<td>Waveguide latency (106.7ps/cm)</td>
<td>427ps</td>
</tr>
<tr>
<td>SM fiber latency (48.3ps/cm)</td>
<td>483ps</td>
</tr>
<tr>
<td>Photodetector+TIA latency</td>
<td>200ps</td>
</tr>
<tr>
<td>Total latency</td>
<td>1.385ns</td>
</tr>
</tbody>
</table>

- Total latency 14 bit times
  - Less than 4 clock cycles (with 2.5 GHz core clock)
  - Almost equally split among
    - Fiber, waveguide, and Tx/Rx circuits
Electrical back-end

- **Power** < 250 fJ/bit
- **Area** 0.02% for 4000 links
  - ~200 cells (0.5um x 0.2 um)
  - 20 (um)^2 per link
MIT Eos1 65 nm test chip

Texas Instruments standard 65 nm bulk CMOS process

First ever photonic chip in sub-100nm CMOS

Automated photonic device layout

Monolithic integration with electrical modulator drivers

2 mm x 2 mm die
16 ring modulators
8 modulator drivers
84 ring filters
~10 cm of waveguides
waveguide crossings
8 MZ modulators
12 photo detectors
Ring modulator
Digital driver
Paperclips
Waveguide crossings
M-Z test structures

Two-ring filter
Vertical coupler grating
One-ring filter
Photo detector
4 ring filter banks
65 nm lithography great for rings!

Ring filter layout in Cadence

SEM photos after dielectric etch

Poly-Si ring

Poly-Si heater

MIT ISG

Massachusetts Institute of Technology
Frequency response of 4-channel filter

FSR is ~16nm (2.7THz)
The scans are
1268nm - 1276nm
at 0.05nm intervals

120 GHz spacing
5nm ring radius step

Sensitivity
1THz/nm poly H
30GHz/nm poly W

First 100 GHz spaced bank in sub-100nm, first in bulk CMOS, first in poly Si
- Enables 60-120 wavelengths/waveguide ( >200 Gb/s/um data rate density)
Conclusion

- Interconnects are very complex micro-communication systems
- Cross-layer design approach is needed to solve the on-chip and off-chip interconnect problem
  - Equalized interconnects extend copper
    - 10x in energy-efficiency with proper circuit-wire co-design
  - CNTs further improve the performance by 2x
    - With proper bundle resizing and ILD scaling
  - Si Photonics improves the throughput by 10-20x
    - Unifies on-chip and off-chip interconnects
    - Requires a different network architecture (electrical for switching, optical for energy-efficient transport)
Acknowledgments

- Franz Kaertner, Rajeev Ram, Judy Hoyt, Krste Asanovic, Henry Smith, Erich Ippen, Martin Schmidt
- Jason Orcutt, Anatoly Khilo, Ben Moss, Charles Holzwarth, Hanqing Lee, Christopher Batten, Ajay Joshi, Fred Chen, Byungsub Kim
- DARPA UNIC program – PM Jag Shah
- Texas Instruments – Dennis Buss and Tom Bonifield
- FCRP Interconnect Focus Center