Emerging Silicon Nanophotonic Networks: Time to Bridge the Gap with System Designers

Davide Bertozzi
University of Ferrara (Italy) - Temporary Guest Scientist at IHP Microelectronics (Germany)
Trends in Extreme HPC

• Evolution of the top10 in the last six years:
  • Average total compute power:
    • 0.86 PFlops → 21 PFlops
    • ~24x increase
  • Average node compute power:
    • 31GFlops → 600GFlops
    • ~19x increase
  • Average number of nodes
    • 28k → 35k
    • ~1.3x increase

  Node compute power main contributor to performance growth

  Node compute power may keep scaling thanks to customization

<<Like 1980s, great time for architects!>>
(John L. Hennessy & David A. Patterson, Turing Lecture, ISCA 2018)
Trends in Extreme HPC

What about Connectivity?

• Top 10 average node level evolutions:
  • Average node compute power:
    • 31GFlops $\rightarrow$ 600GFlops
    • $\sim$19x increase
  • Number of nodes: $\sim$1.3x
  • Total Compute power: $\sim$24x
  • Average bandwidth available per node
    • 2.7GB/s $\rightarrow$ 7.8GB/s
    • $\sim$3.2x increase
  • Average byte-per-flop ratio
    • 0.06 B/Flop $\rightarrow$ 0.01 B/Flop
    • $\sim$6x decrease
  • Sunway TaihuLight (#1) shows 0.004 B/Flop !!

Growing gap in interconnect bandwidth might cause aggregate execution performance not to keep up with available compute power!
Interconnect Power Concern

Data from 28nm NVIDIA chips

- Computation will be relatively inexpensive in terms of energy over communication
- Bandwidth should be increased within tighter and tighter power budgets

Source: W.Dally

Source: S.Borkar
The Communication Hierarchy

A lot of work is going on at the upper layers of the interconnection hierarchy:

- A lot of activity: PCIe, GEN-Z, OpenCAPI, CCIX, Ethernet, InfiniBand, ...

But surprisingly criticalities are showing up even in the lowest layer (chip-scale communications)

**EMERGING Network-on-Chip CRITICALITIES:**

- Latency sensitivity of the multi-hop fabric
- Bandwidth criticalities for future kilo-core chips
- The power overhead for moving bits around
- Non-seamless scaling to off-chip comm.

**WE NEED A GAME CHANGER!**

A lot of work is going on at the upper layers of the interconnection hierarchy:

- A lot of activity: PCIe, GEN-Z, OpenCAPI, CCIX, Ethernet, InfiniBand, ...

But surprisingly criticalities are showing up even in the lowest layer (chip-scale communications)
Silicon Photonics: Game Changer?

Silicon photonics uses co-integration techniques of optical components and/or transceivers with standard CMOS manufacturing process.
Silicon Photonics: Game Changer?

Silicon photonics is delivering integrated optical transceivers and holds promise of bringing optical communications closer to and deeper into the processing node.

Integrated optical transceivers

Electrical transceivers

Conventional hop-by-hop data movement

Flattened end-to-end data movement

Courtesy of K. Bergman
Silicon Photonics: Game Changer?

Silicon photonics is delivering integrated optical transceivers and holds promise of bringing optical communications closer to and deeper into the processing node.
Silicon Photonics: Game Changer?

Silicon photonics holds promise of integrated optical transceivers and of bringing optical communications closer to and deeper into the processing node.

**Requirements for that to happen**

- Divide cost by 1.5 orders of magnitude at least
- Improve energy efficiency by one order of magnitude at least
- Efficient integration solutions with electronics
- Improve system-ability of the technology

**Improving Technology Maturity**

&

**Architecture and system-level design**
Mind the Gap

The gap between system-level designers and technology developers is huge!

- Architecture design points stem directly from designers’ intuition
- Descriptive information at different abstraction layers are mixed
- Designs are difficult to compare with one another
- The application of well-known optimization techniques is difficult
- No consistent methodologies to explore the design space
- Most of the design space still largely unknown
Mind the Gap

The gap between system-level designers and technology developers is huge!

- Architecture design points stem directly from designers’ intuition
- Descriptive information at different abstraction layers are mixed
- Designs are difficult to compare with one another

The application of well-known optimization techniques is difficult

- No consistent methodologies to explore the design space
- Most of the design space still largely unknown

Golden age of ONOC assessment (~2008-2012)

Example of optical parameters used in early-stage analyses:

- Optical fiber loss: 0.5e-5 dB/cm
- Coupler loss: 0.5 - 1 dB
- Splitter loss: 0.2 dB
- Non-linearity loss at 30 mW: 1 dB
- Modulator insertion loss: 1 dB
- Waveguide loss: 2.5 dB/cm
- Waveguide crossing loss: 0.05 dB
- Filter through loss: 1e-4 - 1e-3 dB
- Filter drop loss: 1 dB
- Photodetector loss: 1 dB
- Laser efficiency: 30-50%
- Receiver sensitivity: -20 dBm

Estimated power savings with nanophotonic networks

Early-stage ONOC analysis: inflated expectations
The gap between system-level designers and technology developers is huge!

Architecture design points stem directly from designers’ intuition

Descriptive information at different abstraction layers are mixed

Designs are difficult to compare with one another

The application of well-known optimization techniques is difficult

No consistent methodologies to explore the design space

Most of the design space still largely unknown

How to change the through of disillusionment into a slope to enlightenment?
A Framework to Bridge the Gap

**Goal:**
Bridge the gap between developers of emerging devices and circuit & system designers, thus coupling emerging interconnect technologies and architectures with digital systems and working out novel system-level design concepts.

**Focus:**
- Photonically-integrated chip-scale parallel computing
- Their coupling with off-chip memory sub-systems

**Methodology:**
- Addressing the horizontal integration gap
- Addressing the vertical integration gap
OPTICAL NETWORKS-ON-CHIP

BACKGROUND
Optical NoC Initiator

4-stage modulator

Wavelength-division multiplexed input signal

Electrical Signal
Optical NoC Target
Wavelength-Routed Optical NoCs

Main feature: static allocation of channels to source-destination pairs

- No Time spent in routing and arbitration
- All-optical interconnect solution
- Performance predictability
- All-to-all communications can take place concurrently
- Hard to scale to a large number of cores

The topology needs to avoid interference of same-wavelength carriers
Wavelength-Routed Optical NoCs
Main feature: static allocation of channels to source-destination pairs

Better topologies exist, that reuse the same set of 4 wavelengths across all initiators
Better topologies exist, that reuse the same set of 4 wavelengths across all initiators.

State-of-the-art «Snake» topology.
A major source of overhead of optical NoCs comes from static power. Insertion loss (and laser power requirements) depends on the connectivity pattern.
Horizontal Integration Challenge

Processor(s)  Li  NoC  Optical Network  DRAM  GPU
Target Architecture

Solutions such as 3D or 2.5D integration allow for the separation of both electronic and photonic processes and open the door to a fully dedicated process optimization for the photonic die.
System View

Top level

Source: IBM
System View

Not Just E/O and O/E Converters, but an Architecture Integration Challenge

Data rate adaptation
(De-)Serialization
Flow control
Clock Resynchronization
Message-dependent deadlock avoidance
1) Data rate adaptation

Clock speed

\[ [0.5 \div 3] \text{ GHz} \]

Modulation Rate

\[ \geq 10\text{GHz} \]
2) (De-)Serialization

Architecture Integration Challenge

32/64/
128/256
bits

Optical bitstream
3) Flow control

Buffer size is a function of the round trip time for full-throughput operation
4) Clock Resynchronization

Architecture Integration Challenge

4) Clock Resynchronization

Data $\geq 10$GHz
5) Message-dependent deadlock avoidance
The bridge is a complex block taking care of key functional tasks for architecture correct operation, built on top of a multi-technology platform and supporting GHz-range signaling rates.
Bridge Configuration

One of the key challenges consists of overcoming the inherent serial nature of optical communications.

- Increasing the signalling rate of optical channels
- Increasing the bit-level parallelism (WDM)

A combination thereof

Implications over the SerDes Architecture, hence over the performance-power trade-off of the bridge

Research Goal:
Explore and Characterize the Configuration Space of the Bridge

Pay Attention: CMOS cannot achieve arbitrary speeds!
Technology Partitioning

16 3D-stacked computation clusters
16x16 optical NoC (ONoC)

Array of off-chip laser sources

In static-power dominated technologies like silicon photonics, operation at high transmission rates may become a priority to cut down on pJ/bit

Better performing technologies than CMOS may be required in the back-end of the bridge

We select IHP 130nm SiGe BiCMOS (SG13S)
- fT/Fmax=250 GHz / 340 GHz
- 3.3V I/O CMOS, 1.2V logic CMOS
- 5 thin metal layers, 2 thick ones

Target logic family:
- 2.5V compatible ECL
- A Cell library provides std cell gates
- Logic synthesis from HDL enabled (Synopsys DC)

40/28nm ultra-low power industrial technology library

Similar technologies provide monolithic integration of optical components with the BiCMOS process
Bridge Architecture

Transmitter side

Receiver side

Gateway
Bridge Architecture – Transmitter Side

Optimization: only one set of buffers for all destinations

1 transmission module for each target
(15 of them in a 16x16 ONoC)

One virtual channel
for each message class
to avoid (message-dependent) deadlock

Bit-level Parallelism

Source-synchronous communication

Network Interface Frequency
=f(Modulation rate)

Modulation Rate (e.g., 10 Gbps)
Bridge Architecture: Receiver Side

Source-synchronous communication

Receiver module 1
1 receiver module for each transmitter

Network Interface Frequency = f(transmission frequency)
Modulation Rate (e.g., 10 Gbps)
Flow Control

Credit-based flow control:
- Reuses the datapath
- Exploits low dynamic power of ONoCs
- No round-trip timing assumptions

Can fire only if credits available
2:1 Mux Cell is the main building block

Master D-Latch

Slave D-Latch

Input Data

Output Data

Transmission frequency = twice the input clock

Lower PLL frequency

PERFECT BINARY TREE STRUCTURE

- \( M = \log_2(N) \) Stages working at halved speed with respect to one another
- The number of building blocks per stage is inversely proportional to the operating frequency ➔ Energy savings
Flexibility

More parallelism: remove stages from the right

Scale up: add more stages to the left

This architecture is very flexible, it can easily span a wide bridge configuration space.
Experimental Results

Bridge Front-End Architecture

CMOS
Two process nodes
(bulk 40nm or 28 nm FD-SOI)

(De)-Serialization + Transceivers
CMOS

Opto-electronics + ONOC

Clock domains

Experimental Results

(De)-Serialization + Transceivers
CMOS

ECL 130nm

Partitioning options due to multi-stage nature of the serializer

CMOS

Experimental Results

Two process nodes
(bulk 40nm or 28 nm FD-SOI)
Experimental Results

Target Data Rate for source-destination connection: @25 Gbit/s

Target Data Rate for source-destination connection: @40 Gbit/s

Bridge Front-End Architecture

Fully CMOS

DP1
DP2
DP3
DP4
DP5

Bridge Front-End Architecture

Fully CMOS

DP: Design point

FD-SOI 28nm

CMOS 40nm

ECL 130nm
Experimental Results

Energy-per-bit (pJ/bit)

- 130nm ECL-CMOS 40nm
- 130nm ECL-28nm FDSOI CMOS
- X : Not feasible

1 channel x 25 Gbit/s
2 channels x 12.5 Gbit/s
4 channels x 6.25 Gbit/s
1 channel x 40 Gbit/s
2 channels x 20 Gbit/s
4 channels x 10 Gbit/s

Fully-CMOS
Hybrid CMOS-ECL

25 Gbit/s
40 Gbit/s

-31%
-84%
Experimental Results

16x16 \( \lambda \)-Router Topology
16mm x 16mm optical layer
Experimental Results

Energy efficiencies in the ballpark of 1 to 2 pJ/bit are possible with more WDM channels, a trend that higher signaling speeds exacerbate.

100% bandwidth utilization.
**Experimental Results**

**Network-Level Trade-Offs**

<table>
<thead>
<tr>
<th>TSV</th>
<th>Laser Sources</th>
<th>Total Power [W]</th>
<th>SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25 Gbit/s 40 Gbit/s</td>
<td>25 Gbit/s 40 Gbit/s</td>
<td></td>
</tr>
<tr>
<td>1 channel</td>
<td>1216 2176</td>
<td>32 32</td>
<td>65.6 83.4</td>
</tr>
<tr>
<td>2 channels</td>
<td>1216 2176</td>
<td>48 48</td>
<td>7.4 79.7</td>
</tr>
<tr>
<td>4 channels</td>
<td>2176 2176</td>
<td>80 80</td>
<td>9.6 11.01</td>
</tr>
</tbody>
</table>

- **Optical parallelism comes with cost and signal integrity concerns!**
Vertical Integration Challenge
The Design Space

ONoC topology design points stem directly from designers’ intuition

The design space is currently largely unknown

HOW TO «SYNTHESIZE» THE MOST EFFICIENT ONoC SOLUTION FOR THE REQUIREMENTS OF THE CONNECTIVITY PROBLEM AT HAND?

Major Requirements:
- start from a high-level description, operate on abstractions and refine them into an actual implementation with components from a technology library.
Can we extend the paradigms and methodologies of EDA to the context of emerging silicon nanophotonic interconnection networks?
Can we extend the paradigms and methodologies of EDA to the context of emerging silicon nanophotonic interconnection networks?

Design automation should not determine which technology to pursue.

Design automation can lead to concrete evaluation of a new technology.
Front-End Synthesis Methodology

I. Switching Primitives Representation
II. Technology Mapping
III. Selection of modulation carriers
IV. Netlist connectivity

Can we understand all topology design points in the context of a unified design framework?

Can we populate the design space of wavelength-routed optical NoC topologies?
The filter is used to implement both the drop function at initiator side and the add one at target side.
SYNTHESIS METHODOLOGY

1. **Wavelength Resolution**
   
   Each channel of the WDM input signal should be resolved so to be routed to a different output.

   ![Wavelength Resolution Graph (WRG) for a generic 4x4 WRONoC.]

   \[ \lambda_i = \lambda_{i} \]

2. **Technology Mapping**
   
   E.g., Grouping the 1x2 DFs into compact 2x2 photonic switching elements (PSEs), from a technology library!

   ![Technology Mapping Diagram]

3. **Symbolic Wavelength Assignment**
   
   Assign a resonant wavelength to the MRRs.

   ![Symbolic Wavelength Assignment Diagram]

4. **Topology Connection**
   
   Draw the topology logic scheme. It’s a \( \lambda \)-router! However, it is optimized wrt baseline: only 3 resonator types!

   ![Topology Connection Diagram]

   Constraint: avoid conflicts!
Our synthesis methodology can potentially populate the complete design space of WRONoC topologies by spanning all possible technology mappings, subject to the constraints of each stage for legal solutions.

Only with 2x2 PSEs, the number of WRONoC topologies in the design space amounts to $[(n-1)!]^n$.

A 4x4 WRONoC topology can be implemented in 1296 different ways.

These crossings should be considered as apparent, since at this stage we are drawing the logic topology, not the physical one.
Back-End Synthesis Methodology

I. Front-End Methodology

II. Device Parameter Selection

III. Placement and routing

IV. LOGIC TOPOLOGY

V. Physical Design

VI. PHYSICAL TOPOLOGY
What is the exact radius length of the MRR, typically in the range 5-20\(\mu\)m?

What is the exact value of the \(n\) wavelengths used by each initiator in an \(n \times n\) wavelength-routed optical NoC?

What is the maximum bit-level communication parallelism on the I/O optical channels? Not just cost and reliability, but also feasibility!

This is not just a refinement step, due to the ROUTING FAULT concern: It has implications on network-level throughput and scalability.

\[ \Rightarrow \text{Parallelism is 6 in } PSE_x \]
What is the exact radius length of the MRR, typically in the range 5-20um?

What is the exact value of the n wavelengths used by each initiator in an n x n wavelength-routed optical NoC?

What is the maximum bit-level communication parallelism on the I/O optical channels? Not just cost and reliability, but also feasibility!

This is not just a refinement step, due to the ROUTING FAULT concern: It has implications on network-level throughput and scalability.
What is the exact radius length of the MRR, typically in the range 5-20um?

What is the exact value of the n wavelengths used by each initiator in an n x n wavelength-routed optical NoC?

What is the maximum bit-level communication parallelism on the I/O optical channels? Not just cost and reliability, but also feasibility!

This is not just a refinement step, due to the ROUTING FAULT concern: It has implications on network-level throughput and scalability.

Available parallelism is 64 in PSE_x
Available parallelism is 75 in PSE_y
Available parallelism is 9 in PSE_z
What is the exact radius length of the MRR, typically in the range 5-20um?

What is the exact value of the n wavelengths used by each initiator in an n x n wavelength-routed optical NoC?

What is the maximum bit-level communication parallelism on the I/O optical channels? Not just cost and reliability, but also feasibility!

This is not just a refinement step, due to the ROUTING FAULT concern: It has implications on network-level throughput and scalability

As topology size increases, the proliferation of filter types and wavelength channels may limit the availability of non-overlapped transmission peaks, which may cause the topology to be practically infeasible

Parallelism and Scalability Limitations
PARAMETER UNCERTAINTY

There exists a post-fabrication variation scenario that ends up in a routing fault

Even without overlapping, proximity raises optical crosstalk concerns

Conservative design-for-reliability constraint:
Let us assign device parameters and state an achievable bit-level parallelism such that routing faults will not take place under any variability scenario

We modelled the Ring radius/wavelength channel selection problem subject to routing fault avoidance as a Constrained Optimization Problem, and used ASP as declarative technology.
Back-End Synthesis Methodology

I. Front-End Methodology
II. Device Parameter Selection
III. Placement and routing
IV. Logic Topology
V. Physical Design
VI. Physical Topology
Placement and Routing

Electronic P&R tools cannot be reused here

We propose PROTON+, a tool for automatic placement and routing of ONoC topologies (Collaboration with prof. Schlichtmann at TU Munich)

The tool tries to strike a good balance between **crossing losses** and **propagation losses**, which might be conflicting objectives

Minimize waveguide length  Minimize no. of crossings

Lots of unexpected waveguide crossings (which burden on the static power budget)
Physical Design Space Exploration

\[
\min_{x, r} \max_{p \in \mathcal{P}} \alpha \cdot L_p(x, r) + \beta \cdot C_p(x, r)
\]

Our objective functions minimize the insertion loss across the lossiest path. This indirectly limits total laser power.

Where \( L_p \) and \( C_p \) are approximate functions of path lengths and no. of crossings

**Placement**: Non-linear optimization problem solved with an IPM

**Routing**: adaptation of the Lee’s algorithm «Maze Router»

By setting the weights of the objective function, the best physical mapping for the technology/topology at hand can be achieved.
Layout of 16x16 λ-Router with PROTON+

- Ins. Loss max = 44 dB
- 255 crossings on the critical path
- 28636 μm waveguide length on the critical path
- 24425 sec of CPU time (Intel Core 2 Quad CPU with 8GB RAM running at 2.33GHz)
Layout of 16x16 λ-Router with PROTON+

- Ins. Loss max = 44dB
- 255 crossings on the critical path
- 28636um waveguide length on the critical path
- 24425 sec of CPU time
  (Intel Core 2 Quad CPU with 8GB RAM running at 2.33GHz)

- Hubs
- Memory controller
TU Munich’s Placement and Routing Tools

Prof. Ulf Schlichtmann

PROTONv2.0 (PLATON)
- Proton v2.0 (PLATON) implements a **force-directed placement algorithm**
- Better computation times, better insertion losses
- PLATON is well-suited for large-scale topologies, rather than for small-scale ones

Maximum insertion loss (dB)

- Manual
- PROTON [Boos ICCAD’13]
- PLATON

Computation time

- PROTON
- PLATON
Exploring the Design Space

We exhaustively generated all 4x4 WRONoC topologies and mapped them with Proton+

There is a large variability in the design space: from 18 to 39 crossings!

- This raises the issue of placement-aware logic topology synthesis, completely new discipline for optical NoCs.
- \(\lambda\)-Router and snake proposed in literature are not the best topologies from the critical path length viewpoint!
- Design automation helps to get the most out of a technology
Scalability

We performed device parameter selection to assess scalability of generic topologies

<table>
<thead>
<tr>
<th>Radius selection range and Incremental step</th>
<th>Fabrication options</th>
<th>( R_{\min} )</th>
<th>( R_{\text{step}} )</th>
<th>( R_{\max} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>R_{opt}</td>
<td>5( \mu )m</td>
<td>1( \mu )m</td>
<td>25( \mu )m</td>
<td></td>
</tr>
<tr>
<td>R'_{opt}</td>
<td>5( \mu )m</td>
<td>1( \mu )m</td>
<td>30( \mu )m</td>
<td></td>
</tr>
<tr>
<td>R''_{opt}</td>
<td>5( \mu )m</td>
<td>0.25( \mu )m</td>
<td>30( \mu )m</td>
<td></td>
</tr>
</tbody>
</table>

Radius Tolerance: 10nm

Laser uncertainty: 0.5nm

Ideal Fabrication
- With overly fine step and large rings, the upper bound is roughly a 60x60 topology, with limited parallelism though!
- Achievable parallelism most sensitive to the incremental step of MRR radii

Conservative PROCESS VARIATIONS
- Only 4x4 ONoCs are certainly feasible
- Multiple wavelength selection options useless if uncertainty ranges are not reduced accordingly
Conclusions

• High-performance computing systems will be soon again interconnect-limited. Emerging technologies can be game changers.

• Time for a concrete evaluation of emerging silicon nanophotonic networks in small-scale systems. How? By bridging the gap with system designers.

• Horizontal integration gap:
  • ENoC-ONoC bridge key to determining configuration of optical connections (data rate, parallelism), and its energy efficiency.
  • 1-2 pJ/bit communication can be realistically targeted at 40 Gbps connection rate, with 4 WDM channels@10Gbps in parallel (bridge in 28 nm CMOS). Signal integrity is an issue.

• Vertical integration gap:
  • Design methods have been developed to populate the largely unknown design space of wavelength routed topologies.
  • Early-stage complete cross-layer synthesis methodology defined.
  • More energy-efficient topologies than existing ones in literature have been «synthesized».
  • Design automation: an enabler for emerging technologies.
Acknowledgement