Optical Interconnects
Outline

Interconnection Networks
  ◦ Terminology
  ◦ Topology basics
  ◦ Examples of interconnects for
    • Real HPC systems (Cray Jaguar, IBM’s Blue Gene/Q)
    • Data Centers (DC)
  ◦ Traffic profiles of HPC and DC

Optical Interconnects
  ◦ Motivation
  ◦ Building blocks
  ◦ Architecture examples for all packaging hierarchy levels:
    • Rack-to-rack
    • On-board and board-to-board
    • On-Chip
  ◦ Sum-up – issues
Interconnection Networks: What is an interconnection network?

- Parallel systems need the processors, memory, and switches to be able to communicate with each other
  - The connections between these elements define the interconnection network
Interconnection Networks: Terminology

- **Node**
  - Can be either processor, memory, or switch

- **Link**
  - The data path between two nodes (Bundle of wires that carries a signal)

- **Neighbor node**
  - Two nodes are neighbors if there is a link between them

- **Degree**
  - The degree of a node is the number of its neighbors

- **Message**
  - Unit of transfer for network clients (e.g. cores, memory)

- **Packet**
  - Unit of transfer for network
Interconnection Networks: Basics

- **Topology**
  - Specifies way switches are wired
  - Affects routing, reliability, throughput, latency, building ease

- **Layout and Packaging Hierarchy**
  - The nodes of a topology are mapped to packaging modules, chips, boards, and chassis, in a physical system

- **Routing**
  - How does a message get from source to destination
  - Static or adaptive

- **Flow control and Switching paradigms**
  - What do we store within the network?
  - Entire packets, parts of packets, etc?
  - Circuit switching vs packet switching

- **Performance**
  - Throughput, latency. Theoretically and via simulations
Interconnection Networks: Topology

- Direct topology and indirect topology
  - In direct topology: every network client has a switch (or router) attached
  - In indirect topology: some switches do not have processor chips connected to them, they only route

- Static topology and dynamic topology
Interconnection Networks: Topology

Examples (Direct topologies):

- **6-linear array**
- **6-ring**
- **6-ring arranged to use short wires**

**Diagrams:**
- 2D 16-Mesh
- 2D 16-Torus
- 3D 8-Cube
Interconnection Networks: Topology

- Examples (Indirect topologies):

  - Fat Trees: Fatter links (really more of them) as you go up
  - Trees
  - 8-node butterfly
  - clos
Interconnection Networks: Topology

- Theoretical topology evaluation metrics:
  - **Bisection width**: the minimum number of wires that must be cut when the network is divided into two equal sets of nodes.
  
  ![Bisection Diagram]

  - **Bisection Bandwidth**: The collective bandwidth over bisection width
  - **Ideal Throughput**: throughput that a topology can carry with perfect flow control (no idle cycles left on the bottleneck channels) and routing (perfect load balancing). Equals the input bandwidth that saturates the bottleneck channel(s) for given traffic pattern. For uniform traffic (bottleneck channels = bisection channels):
    - **Network Diameter**
    - **Average Distance** (for given traffic pattern). For uniform traffic: $D_{avg} = \frac{1}{N^2} \sum_{x,y} distance(x,y)$
    - **Average zero Load Latency** (related to average distance)

- Simulations
  - Throughput, average latency vs offered traffic (fraction of capacity) for different traffic patterns
Interconnection Networks: topology design trade-offs

- Topologies with small diameter and large bisection bandwidth: greater path diversity, allow more traffic to be exchanged among nodes/routers (=better throughput)
- But, topologies with large node degree: fixed number of pins partitioned across a higher number of adjacent nodes. Thinner channels: greater serialization latency.
Interconnection Networks: Topology selection

- The quality of an interconnection network should be measured by how well it satisfies the communication requirements of different target applications.
- On the other hand, problem-specific networks are inflexible and good “general purpose” networks should be opted for.
## Topologies in Real (old) HPC Machines

<table>
<thead>
<tr>
<th>System</th>
<th>Topology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red Storm (Opteron + Cray network, future)</td>
<td>3D Mesh</td>
</tr>
<tr>
<td>Blue Gene/L</td>
<td>3D Torus</td>
</tr>
<tr>
<td>SGI Altix</td>
<td>Fat tree</td>
</tr>
<tr>
<td>Cray X1</td>
<td>4D Hypercube*</td>
</tr>
<tr>
<td>Myricom (Millennium)</td>
<td>Arbitrary</td>
</tr>
<tr>
<td>Quadrics (in HP Alpha server clusters)</td>
<td>Fat tree</td>
</tr>
<tr>
<td>IBM SP</td>
<td>Fat tree (approx)</td>
</tr>
<tr>
<td>SGI Origin</td>
<td>Hypercube</td>
</tr>
<tr>
<td>Intel Paragon (old)</td>
<td>2D Mesh</td>
</tr>
<tr>
<td>BBN Butterfly (really old)</td>
<td>Butterfly</td>
</tr>
</tbody>
</table>

Many of these are approximations: E.g., the X1 is really a “quad bristled hypercube” and some of the fat trees are not as fat as they should be at the top.
**HPC systems**

- **HPC** (High Performance Computing) system or supercomputer: a computer with a high-level computational capacity compared to a general-purpose computer.

- The speed of supercomputers is measured and benchmarked in "FLOPS" (FLoating point Operations Per Second), and not in terms of "MIPS" (Million Instructions Per Second), as is the case with general-purpose computers.

- The **TOP500** project (www.top500.org) ranks and details the 500 most powerful (non-distributed) computer systems in the world.
HPC systems

Some HPC systems that made it to the top of the TOP500 lists:
Interconnection Networks: Real HPC (Cray Jaguar)

- Cray –“Jaguar”:
  - 3D torus network
  - **Blade**: 4 network connections
  - **Cabinet(Rack)**: 192 Opteron Processors – 776 Opteron Cores, 96 nodes
  - **System**: 200 cabinets
  - Linpack performance of 1.759 petaflop/s

A single Node

- Node: 73.6 GF, 16 GB
- Blade: 294 GF, 64 GB
- Rack: 7.06 TF, 1.54 TB
- System: 1382 TF, 300 TB

12800 nodes
Cray XT5: Over 1,400 Components Packed Into Each Cabinet

Blade = 4 Nodes
8 processors
32 cores
4 Interconnect chips
16 (4 GB) memory chips = 64 GB
6 DC voltage converters

Node = 2 Processors
4 cores per processor
1 Interconnect chip
4 x (4 GB) memory chips = 16 GB

Processor = 4 Cores
2 memory chips

Cabinet = 24 Blades
768 cores
96 Interconnect chips
384 memory chips (1.5 TB)
144 voltage converters
+ power supply, liquid cooling, etc.
Power 480V, ~40,000 Watt per cabinet

Jaguar = 284 cabinets (XT5 and XT4), ~ 6.5 Megawatts
Interconnection Networks: Real HPC (Cray Jaguar)

- 1 chassis: 8 blades:

  - The basic building block is a single chassis
    - A chassis is 1 x 4 x 8
      - Dimensions are: X x Y x Z
    - Each node on a blade is connected in the Y dimension (mezzanine)
    - Each node in a chassis is connected in the Z dimension (backplane)
    - All X-dimension connections are cables
Interconnection Networks: Real HPC (Cray Jaguar)

1 rack (3 chassis)

1 blade (4 nodes)

1 chassis (8 blades)
Interconnection Networks: Real HPC (Cray Jaguar)

- 1 Rack: 3 chassis
Interconnection Networks: Real HPC (Blue Gene/Q)

- Blue Gene/Q:
  - 5D Torus, 131.072 nodes (system level)
Data Centers (DCs)

- Warehouse-scale computers
- Based on Clusters: Commodity (not high-end) hardware
- Wide variety of applications

- Large scale applications
  - webmail, websearch, facebook, youtube
- Cloud computing
  - Amazon EC2, Microsoft Azure

Facebook’s data centers store more than 40 billion photos, and users upload 40 million new photos each day, ~ 2,000 photos every second
Mega DC and modular units

- Mega Data Centers: 500,000+ servers
- Modular DC – quick deployment
  - Unit packaged (often) in standard shipping container formats (called *pods*)
  - Contains: ~2000 servers, storage, network
Data Centers architecture

- Fat tree logical topology:

- Implemented as folded clos:
Interconnection Networks: Data Center architecture/topology

- Most of the current data centers: based on commodity switches for the interconnection network.
- Fat-tree 2–Tier or 3–Tier architecture
  - Fault tolerant (e.g. a ToR switch is usually connected to 2 or more aggregate switches)
  - Drawbacks:
    - High power consumption of switches and high number of links required (bad scalability).
    - Latency (multiple store–and–forward processing).

In the front-end: route the request to the appropriate server.

Top Of Rack switch
1 Gbps links

Servers (up to 48) as blades
Full bisection bandwidth: can support N/2 simultaneous communication streams

Oversubscribed fat tree: offers less than full bisection bandwidth

- Reduced cost, complexity
- Compromise bandwidth
- Needs sophisticated task assignment – Communication locality

assigning a total committed information rate to a given port that is greater than that port's speed.
Traffic Profiles (HPC applications)

- Traffic patterns (locality, message size, inter-arrival times) play an important role in architecture/topology design
- **HPC applications**
  - comprise tasks that run on processors in a distributed/parallel manner and communicate through messages.
  - exhibit well defined communication patterns
  - **MPI** (Message Passing Interface Standard) has become the "industry standard" for writing message passing programs on HPC platforms.
- **Types of MPI messages:**
  - **Point-to-point (PTP) communication routines** (involve message passing between two, different MPI tasks).
  - **Collective communication routines** (involve all processes).
**Logical Communication Graphs**: A logical communication graph expresses the amount of data that is exchanged between processors throughout the application execution.
**Traffic Profiles (Data Centers)**

- **DataCenters**: multi-tenant environments, various applications, wide variations of requirements.
  - **Link Utilization**
    - Core > Edges > Aggregation
    - Many links are unutilized
  - **Losses**
    - Aggregation > Edges > Core
    - Core has relatively little loss but high utilization
    - Edge & Aggr have significantly higher losses
    - Few links experience loss
    - Loss may be avoided by utilizing all links (re-route traffic)
  - **Traffic adheres to ON-OFF traffic pattern**
  - **Arrival process is log normal**
Traffic Profiles (Data Centers)

- Traffic that remains within the data center: more than 70% of total traffic volume in DCs
- Traffic that flows from data center to data center
- Traffic that flows from the data center to end users through the Internet or IP WAN
**Traffic Profiles (Map Reduce)**

- **MapReduce**: Prominent traffic application in Data Centers
  - Originally proposed by Google
- **Apache Hadoop**: similar but open-source
  - used by many companies including Yahoo!, Facebook and Twitter
- MapReduce, Hadoop: use a divide-and-conquer approach.
  - data are divided ("mapped") into fixed size units/collections of key/value pairs, processed independently and in parallel by Map tasks, which are executed in a distributed manner across the nodes in the cluster.

![Diagram of MapReduce and Hadoop workflow](image-url)
Energy Consumption in HPCs

- **Power consumption and size**: main set of barriers in next-generation interconnection networks (Data Centers, High Performance Computing).
- Predictions that were made back in 2008-09 concluded that supercomputing machines of 2012 would require 5MWs of power and in 2020 will require a power of 20MWs.
- In 2012: The K-supercomputer has already reached the 10Pflops performance, requiring however approximately 10MW of power instead of the 5MW predictions four years ago!!
Energy Consumption of DCs

Energy Consumption of telecom & DC networks:

<table>
<thead>
<tr>
<th></th>
<th>2007 (In billion KWh)</th>
<th>2020 (In billion KWh)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Centers</td>
<td>330</td>
<td>1012</td>
</tr>
<tr>
<td>Telecom</td>
<td>293</td>
<td>951</td>
</tr>
<tr>
<td>Total</td>
<td>623</td>
<td>1963</td>
</tr>
</tbody>
</table>

For comparison: the total energy consumption of European Union in 2013 was 2798 billion kWh.

Sources:
Optical Interconnects

- Solution: optical interconnects
- Q: where to attach the optics?  
  A: Wherever possible. As close to as possible to the processor
- Critical issues: **Cost**, Reliability, Performance
Optical Interconnects

- **Currently**: communication via fibers between switches in the rack-to-rack level
- **SFP**: Small Pluggable Connector
- **XFP**: 10 Gigabit Small Form Factor Pluggable
- **AOC**: Active Optical Cables
**Optical Interconnects**

Currently: communication via fibers between switches in the rack-to-rack level
Optical Interconnects

- Devices that are widely used in optical networks:
  - **Splitter and combiner**: fiber optic splitter: passive device that can distribute the optical signal (power) from one fiber among two or more. A combiner: the opposite.
  - **Coupler**: passive device that is used to combine and split signals but can have multiple inputs and outputs.
  - **Arrayed-Waveguide Grating (AWG)**: AWGs are passive data-rate independent optical devices that route each wavelength of an input to a different output. They are used as demultiplexers to separate the individual wavelengths or as multiplexers to combine them.
  - **Wavelength Selective Switch (WSS)**: A WSS is a typical an 1xN optical component that can partition the incoming set of wavelengths to different ports (each wavelength can be assigned to be routed to different port). It can be considered as reconfigurable AWG and the reconfiguration time is a few milliseconds.
  - **Micro-Electro-Mechanical Systems Switches (MEMS switches)**: MEMS optical switches are mechanical devices that physically rotate mirror arrays redirecting the laser beam to establish a connection between the input and the output. The reconfiguration time is a few milliseconds.
  - **Semiconductor Optical Amplifier (SOA)**: Optical Amplifiers. Fast switching time, energy efficient.
  - **Tunable Wavelength Converters (TWC)**: A tunable wavelength converter generates a configurable wavelength for an incoming optical signal.
Optical Interconnects

Optical MEMS Spatial switching

Coupler

AWGR Wavelength switching

Wavelength and spatial switching

Wavelength Selectable Switch (WSS)
Rack-to-rack architectures

- Hybrid architectures (*Fat tree architecture is enhanced using Optical Circuit Switching*)
  - Easily implemented (commodity switches)
  - Slow switching time (MEMs). Good only for bulky traffic that lasts long
  - Not scalable (constraint by Optical switch ports)

Farrington et al (SIGCOMM 2011)

Wang et al (SIGCOMM 2010)
Rack-to-rack architectures

- Optical switch architectures with high radices in order to lead to more flat DC architectures (less tiers) by replacing electrical switches in the upper tiers of the fat-tree

Alternatives to fat trees

Ye et al (SANCS 2010)

Farrington et al (OFCC 2013)

Gripp et al (OFCC 2010)  Xia et al (TR 2010)

Nephele Project
Eg Petabit switch fabric: three–stage Clos network and each stage consists of an array of AGWRs that are used for the passive routing of packets.

In the first stage, the tunable lasers are used to route the packets through the AWGRs, while in the second and in the third stage TWC are used to convert the wavelength and route accordingly the packets to destination port.
(General) Structure of Switch
OPCB (Optical Printed Circuit Boards) building blocks

• All previous architectures target mainly rack-to-rack communication.
• However photonic technology building blocks are being manufactured for board-to-board, module-to-module (on-board) and on-chip communication.

- Light source
  - Edge/Surface emitting 850, 1300, 1550nm
- Detector
  - Si, InGaAs, InP
- Interconnect medium
  - Fibres (single/ multimode), polymer, silicon, glass like dielectrics,
- Drive and receive circuits

• Need decisions where integration is essential
  - Eg wavelength, waveguide, light source,
**OPCB building blocks**

Multilayer Waveguide Arrays

- **2-layer waveguide array** (by mask exposure)
  - W=250μm  W=50μm
  - 4x12 waveguides

- **4-layer waveguide array** (by laser writing)
  - 1 mm Polymer
  - Layer thickness control better than +/- 5%!
    - (e.g. vertical pitch: 250 +/- 5 μm)

---

Optical fibers across the boards

~2011

Optical waveguides in/on boards

~2015

Optical interconnects integrated with the processor

~2020
OPCB building blocks
OPCB building blocks

Elements of an optical card-to-card link

Card-to-card optical interconnect scheme

Card-to-card optical interconnect prototype

Optical printed circuit board panel realization

Embedded waveguides

Optical printed circuit board

Optical connectors

Opto-electronic transceivers
E.g. Opto–electronic router chip
- 168 optical channels: 168 Tx (VCSELS) & 168 Rx (PhotoDiodes) elements
- 8 Gbps/channel
- Packet switching
- O/E/O conversion of packets for processing

Already incorporated in routers
- Next step: integrate on OPCBs (waveguide–based, not cable–based connections)
Modules:
- multi- and single-mode optical PCBs (various materials: glass, polymers, silicon photonics, ...)
- Optochips
- chip- and board-to-board connectors
Differences of optical and electrical on-PCB interconnects

- X–Y routing
  - the method of choice for high complexity, high performance designs with many high pin count parts

- River routing
  - via reduction and often fewer layers
**Differences of optical and electrical on-PCB interconnects**

- **Thompson model** (X-Y model for electrical interconnects topologies)
  - Nodes laid-out in a 2D grid: only column- or row-wise comm. (X-Y routing)
  - 2 routing layers: 1 for vertical wiring & 1 for horizontal wiring
  - At each layer no crossing is allowed (inter-layer connectors/vias used)

- **Extensions**: >2 wiring layers (to reduce area)
  - Nodes laid-out in a 2D grid exhibiting only column- or row-wise communication
  - Multilayer 2D grid model: only 1 layer contains nodes
  - Multilayer 3D grid model: more than one layers contain nodes
Differences of optical and electrical on-PCB interconnects

- X–Y routing
- River routing
Differences of optical and electrical on–PCB interconnects

Waveguided communication (differences to electrical):
1. Waveguide bends require a (non-sharp) bending radius $r$

2. Crossings possible at the same layer (at various crossing angles), but need to account for the induced losses and crosstalk

$90^\circ$  $45^\circ$
3. WDM (Wavelength Division Multiplexing): many logical links over a physical link

WDM and crossings in the same layer allow in-principle denser integration
Usual power budget (=transmission power – detector sensitivity): **15 dB**

**Losses**: mainly insertion losses for coupling (from chip-to-board and board-to-board) and waveguide modules (bends, crossings,...)

Waveguide losses greatly depend on launch conditions in the waveguide

**In general:**
- **crossings** with 90 degree crossing angle are “cheap” (e.g. for polymers at 850μm wavelength: ~ 0.01 – 0.006 dB/crossing)
  - Crossing angle ≠ 90 degree: more losses and potential crosstalk problems
- **90 degree bends** are more “expensive” (for polymers and 50μm×50μm waveguide widths ~ 1dB/bend for 90 degree bend with radius=10mm)
  - S-bends are somewhat “cheaper”
- **Combiners and Splitters** even more expensive (1.5–3dB)

---

**Bend losses**

![Bend losses graph](image)

*Fig. 7. Power transmission of waveguide bends for three widths w = 50 μm, 75 μm and 100 μm*

**Crossings losses**

![Crossings losses graph](image)

*Fig. 11. Insertion loss of waveguide crossings for SMF and 50-μm MMF inputs.*
Board-to-board & On-board architectures

Pitwon et al (JLT 2012)

Dou et al (OPTO 2010)

Beals et al (AP 2009)

Bamiedakis et al (JLT 2014)
Board-to-board & On-board architectures

E.g. regenerative optical bus architecture

Bamiedakis et al (JLT 2014)
Networks On Chip (NOC) architectures

- (IBM–Columbia University): 3D stacking, lots of data on chip. Circuit Switching.
Networks On Chip (NOC) architectures

Step 1: Path SETUP request

Step 2: Path ACK

Step 3: Transmit Data

Meanwhile: Path Contention

Step 4: Path TEARDOWN
Networks On Chip (NOC) architectures

- Photonic layer: PSE (Photonic Switching element) based on silicon ring resonator

  ![Diagram of ring resonators](image)

- Alternatives to ring resonators: MZI (Mach–Zehnder Interferometer)

  ![Diagram of MZI](image)
Networks On Chip (NOC) architectures

- High Order Switch Designs
Networks On Chip (NOC) architectures

- G: gateways, locations on each node where a host can initiate or receive data transmissions.
- X: 4x4 non-blocking photonic switches
- Torus requires an additional access network. ‘I’ (injection) and ‘E’ (ejection) to facilitate entering and exiting the main network.
Networks On Chip (NOC) architectures

Crossbar
- Kirman, MICRO'06
- Vantrease, ISCA'08

Torus
- Shacham, TOC'08

Quasi-Butterfly
- Batten, IEEE Micro'09
- Morris, JSTQE'10

Fat-Tree
- Li, DAC'09
- Gu, DATE'09

Mesh
- Cianchetti, ISCA'09

Quasi-Clos
- Pan, ISCA'09

Bus
- Beamer, ISCA'10

Clos
- Joshi, NOCS'09
- Koka, ISCA'10

Multiple Buses
- Kurian, PACT'10
- Pan, HPCA'10

Quasi-Butterfly
- Xue, ISCA'10

Source: Batten, (Workshop on the Interaction between Nanophotonic Devices and Systems December 2010)
Optical interconnects are a promising solution for tackling the power and bandwidth requirements of HPC systems and Data Centers.

Architectural issues:
- on-board, on-backplane and system level topologies
- number of routers on board, number of boards per backplane
- number of channels/waveguides for chip-to-router and router-to-router communication
- topology lay-outs on PCBs
- switching paradigms (packet vs. circuit)
- benefits of WDM

All the above need to be re-visited, re-addressed, and re-evaluated.
Outline

- **Interconnection Networks**
  - Terminology
  - Topology basics
  - Examples of interconnects for
    - Real HPC systems (Cray Jaguar, IBM’s Blue Gene/Q)
    - Data Centers (DC)
  - Traffic profiles of HPC and DC

- **Optical Interconnects**
  - Motivation
  - Building blocks
  - Architecture examples for all packaging hierarchy levels:
    - Rack-to-rack
    - On-board and board-to-board
    - On-Chip
  - Sum-up – issues