Optics in computers servers and computers: challenges and opportunities

H.J.S. Dorren

Eindhoven University of Technology

H.J.S.Dorren@tue.nl
Acknowledgements

Teams at

TE Connectivity in the Netherlands for various components (VCSEL and PD arrays and cmos ICs)

IPtronics Denmark (soon to be Mellanox) for CMOS driver and receiver ICs and wafers

Draka/Comteq Eindhoven Netherlands for OM4+ MMF fiber samples

Cleanroom at COBRA/Eindhoven for help in processing

At TU/e: Oded Raz, Pinxiang Duan, Karel van den Plassche
Data-center networks/architecture

DC networks employ fat-tree networks for cost reasons
Tree topologies introduce bandwidth bottlenecks
Datacenter networks

- 300,000 servers
- Each server: 1 150 Gflop/sec Intel Sandy bridge CPU
- Each processor 4 memory controllers for DDR3-1333 (4x 120 Gbps raw bandwidth)
- Multilayer fat-tree network
- 40 servers per rack
- Each rack: 1 tor switch (48 port GE switch; 1:5 bandwidth to higher layer; cost U$15/port)
- With in rack Cat5 based communications (0.2$ per cable)
- From layer 0 tor to layer 1: optical communications using SFPs
- Layer 1 same bandwidth deviation as layer 1 but entirely SFP based
- Layer 2 aggregation in 10 Gbps lanes (SFP+)
- Eventually in 40 Gbps (QSFP)
Data center network: bandwidth

Total network bandwidth $\sim 0.2$ Petabit/sec
(excl. on chip communication and CPU-DIMM interconnect bandwidth)
Evolution data center network

Red: factor 1000 per 10 years
Blue: factor 100 per 10 years
Green: factor 10 per 10 Years

Computations that follow based on fat-tree (folded clos) architecture

Amdahl’s law: in balanced parallel systems: 1 byte/sec communications per flop/sec performance
Bandwidth to each server

Red: factor 1000 per 10 years
Blue: factor 100 per 10 years

By 2022: 100 - 400 Gbps to each server  !!!
Evolution power data center network

Red: factor 1000 per 10 years
Blue: factor 100 per 10 years

Computation based on “typical” publicly available data in combination with ITRS roadmaps

Includes switch power, transceivers, etc
Switch power dominates !!!
Evolution **average** energy per bit to stay at 2013 level

48 x 400 Gbps switch can dissipate a maximum of 60 Watts of power

- **Blue:** factor 100 per 10 years
- **Red:** factor 1000 per 10 years

Includes switch power, transceivers, etc

Switch power dominates
Evolution of **average $** per Gbps for DC network to stay at 2013 level

A 48 port rack interconnect operating at 400 Gbps per channel can cost U$ 400

Blue: factor 100 per 10 years  
Red: factor 1000 per 10 years
Evolution of number of OE interfaces in data-center

Number of E-O conversions

Year

2012 2015 2017 2019 2022

25 Gbps links
DC network scenarios

- DC applications require more communications between servers (more east-west traffic)
- Scaling of bandwidth between servers is unclear but some algorithms applied today scale as NP-hard problems
- Networking scenarios allowing for more east-west traffic require more interconnectivity, power, cost
Roadmap on-board photonics
need for ultra-compact transceivers!

Currently implemented

Implemented in high-end machines

Optical engine mounted on package or on switch engine
Prior-art on 3D stacked transceivers (2)

850nm top-emitting VCSEL and photodiode array
Flip-chip bonding with “Holey” Optochip

F.E. Doany, Electronic Components and Technology Conference, 247-255, 2010 (IBM)
Prior-art on 3D stacked transceivers (3)

~ 1000nm back-emitting VCSEL and photodiode array
Flip-chip bonding on the analog IC – 168 VCSELs and 168 PDs @ 8 Gbps

Hasharoni, Kobi, OFC 2013, paper: OTu3H.2 (Compass EOS)
Motivation for our work

- **Low cost**
  - *Industry standard devices*
  - *Wafer scale manufacturing, such as lithographic metallization*
  - *Simple process flow*

- **High bandwidth density**
  - *Compact size transceiver (Gbps/mm²)*
  - *Three dimensional ICs*

- **High date rate per channel**
  - *Impedance matched connections*
  - *Co-planar waveguides instead of wire bonds*
Low cost: wafer-scale manufacturing

- Process on a wafer-scale and not on individual dies
- Pick and place process (fast pick and place machines with sufficient accuracy exist)
- Lithographic metallization
- Suitable for low cost packaging (passive alignment)
Photoresist ramp-schematic structure

- 220µm step, high viscous PR is needed
- Appropriate thickness of photoresist
- Multilayer spinning to decrease edge bead effect
- Above soft-point is reshaping
- Surface tension determines final shape
- Slowly ramp up and down temperature to reduce cracks
Process flow (1) - die to wafer bonding

VCSEL array
Thickness: 200um

bonding layer

VCSEL driver
Process flow (2)
photoresist pattern after lithography

VCSEL array
Photoresist pattern
VCSEL driver
Process flow (2)
photoresist ramp after reflow

VCSEL array
Photoresist ramp
VCSEL driver
Lens exploration

- Photoresist ramp
- VCSEL array
- Micro-lens
Process flow (3)- metal paths by plating

- VCSEL array
- Metal path
- Photoresist ramp
- VCSEL driver pad
Module testing
3D stacked receiver IC
2D (side-by-side) receiver ICs
Uniform performance

Biased with 3 mA current
5 mA amplitude of the modulated signal
(a pair of differential signals at 10 Gb/s with $2^{31}-1$ NRZ PRBS format)
Present - waferscale manufacturing

After VCSEL array placement
After photoresist patterns
After reflow process
After plating
Conclusions (1) - DC networks

Over the next 10 years:

- Bandwidth to each server increases with 2 orders of magnitude

- Bandwidth per interface on each switch grows from 1 Gbps to 25 Gbps and number interfaces on each grows with factor 4 (density issue)

- Power efficiency has to improve with 3 orders of magnitude to stay at 2012 level

- Cost have to go down with 3 orders of magnitude to stay at 2012 level
Conclusions (2)

• Motivation 3D die-stacked transceiver ICs
  Bandwidth density
  Cost
• Advantages
  Industry standard devices directly being used
  Wafer scale metallization, Impedance matched design
  Process compatible with normal CMOS process technology
• Experimental results
  Transmitter (incl. transmission 500 m OM4+ MMF), Receiver
  Power consumption is depends on the CMOS IC
  Thermal issues controllable