Intra Optical Data Center Interconnection:

Session 1: Component & Module Focus

Co-Organizer/Presider/Session Chair:

Dr. Ioannis Tomkos

Networks and Optical Communications group – NOC
Session 1 Speakers

- **Martin Zirngibl, Vice President and Technical Fellow, Finisar**
  - Before joining Finisar recently, Martin headed the IP routing research organization at Alcatel-Lucent’s Bell Labs, which he joined in 1990 as a researcher (initially performing pioneering work in fiber amplifiers and photonic integrated circuits). Since 1998, Martin has held various managerial positions at Bell Labs and drove early research and development in 100G that led to a significant industry disruption in WDM long-haul. Martin holds a PhD in applied physics from ETH Lausanne, Switzerland.

- **Brad Booth, Principal Engineer, Microsoft**
  - Brad Booth is heading up the 25/50G Ethernet Consortium and the Consortium for On-Board Optics (COBO). At Microsoft, he leads the development of hyperscale interconnect strategy for Microsoft’s cloud datacenters. He is also the founder and past Chairman of the Ethernet Alliance. Brad was previously a Distinguished Engineer in the Office of the CTO at Dell Networking and he has also held senior strategist and engineering positions at Applied Micro, Intel, and PMC-Sierra.

- **Toshiki Tanaka, Research Manager, Fujitsu Laboratories Ltd**
  - Toshiki joined Fujitsu in 1997 and has been engaged since then in research and development of dense WDM optical transmission systems and high capacity optical transceivers with higher order modulation. Since 2016, he has engaged in developing 100G and beyond 100G optical transceivers for short reach application at Fujitsu Optical Components Limited.
History behind ODCI event

- The optical networking scientific community recognized early the huge potential that optical technologies have to solve the scaling problems in data centre networks.
  - At OFC 2011 the ‘Datacom, Computercom and Short Range and Experimental Optical Networks’ subcommittee (now track DSN6 in the OFC 2016 technical programme) was formed
  - Since that time, the topic of optical communications for data centres has become mainstream and many other major conferences have adapted their coverage accordingly.

- Before the NGON 2015 event, there were a few sporadic talks about ODCI included in the NGON conference program (mostly focusing on inter-ODCI)

- In NGON 2015, we introduced for the first time a special workshop focusing on ODCI, with emphasis on intra-ODCI technologies/solutions

- Due to the success on the 2015 workshop and the tremendous market potential of ODCI solutions, IIR decided to expand the NGON conference by including a parallel track on ODCI
“Super-sizing the data centre”

- Warehouse-scale data centres are stretching the limits of current networking technologies
- The article described how advanced optics can remove performance bottlenecks to enable the data centres of the future → This is what we will discuss further today!
Outline

Session 1: Component & Module Focus

- Architectures and characteristics of intra-DC networks
- Transceiver solutions for optical interconnects in intra DC networks

Session 2: Debating Intra-DC solutions and Photonic Integration approaches

- Photonic integration approaches for intra-DC networks: InP vs. SiP
- Photonic integration approaches for intra-DC networks: Monolithic vs. hybrid
What is a (conventional) “Data Center”

- A **data center** is a facility used to house computer systems (aka servers) and associated components, such storage systems, that are networked together.

- The equipment may be used to store and process the (big) data providing relevant applications directly to the DC’s operator customers.
Intra-Data Center Interconnection Networks (intra-DCNs)

The Data Center Networks are usually based on a fat-tree topology using commodity switches

**Access Layer**: Top-of-Rack (ToR) switches are used to connect the servers in a rack (e.g. 1Gbps)

**Aggregate Layer**: Used to interconnect the ToR switches (e.g. 10Gbps)

**Core Layer**: Used to interconnect the access switches (e.g. 100Gbps)
Need for Big (Warehouse-scale) Data Centers

Computing and Storage are moving from PC-like clients to large Internet service providers that use the cloud.

There is the need for more powerful and efficient warehouse data centers to host the cloud computing applications.
# How Big are these Mega Data Centers?

<table>
<thead>
<tr>
<th>Data Center Site</th>
<th>Sq ft</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook (Lulea, Sweden)</td>
<td>290,000</td>
</tr>
<tr>
<td>Facebook (Altoona, IA)</td>
<td>476,000</td>
</tr>
<tr>
<td>Google (Mayes County, OK)</td>
<td>~500,000</td>
</tr>
<tr>
<td>Apple (Maiden, NC)</td>
<td>505,000</td>
</tr>
<tr>
<td>Switch (SuperNap, Las Vegas-7)</td>
<td>515,000</td>
</tr>
<tr>
<td>Microsoft (Northlake, IL)</td>
<td>700,000</td>
</tr>
</tbody>
</table>

Wembley Stadium: **172,000 square ft**
Traffic Characteristics of Data Centers

(Source: Cisco global cloud index: 2014-2019)

- IP traffic over data center networks has reached almost 5 Zettabytes in 2015
- Cloud DC traffic is growing 33% CAGR
- Most traffic remains within the DC → BW>73% of total network traffic
Data Centers Power Consumption

- Data centers consumed **330 Billion KWh** in 2007 and is expected to reach **1012 Billion KWh** in 2020

2007 electricity consumption. Billion kWh

- **US**: 5436
- **China**: 3023
- **Russia**: 1023
- **Japan**: 925
- **Cloud computing**: 623
- **India**: 588
- **Germany**: 547
- **Canada**: 536
- **France**: 447
- **Brazil**: 404
- **UK**: 345

[Source: How Clean is Your Data Center, Greenpeace, 2012]
Current Data Center Networks:

What the business need (Terabit Data Centers):

- Current data center networks cannot affordably satisfy the bandwidth requirements of upcoming applications
  - Bandwidth intensive applications
  - Data Centers based on low cost commodity hardware

[Source: “Scaling Networks in Large Data Centers, Facebook”]
Facebook’s fabric based data center

- For building-wide connectivity, fb created four independent “planes” of spine switches, each scalable up to 48 independent devices within a plane. Each fabric switch of each pod connects to each spine switch within its local plane. Together, pods and planes form a modular network topology.

To connect all these switches an enormous number of (low-cost & low-power) transceivers are required!
Data Growing Faster than Technology

- Typical silicon devices (electronics) cannot follow the growth of the data => Optical interconnects based on new paradigms are needed!
Energy and cost targets for interconnects

<table>
<thead>
<tr>
<th></th>
<th>Inter DC</th>
<th>Rack to Rack</th>
<th>Board to Board</th>
<th>Module to Module</th>
<th>Chip to Chip</th>
<th>Core to Core on chip</th>
</tr>
</thead>
<tbody>
<tr>
<td>DISTANCE TARGET</td>
<td>few km up to multi km</td>
<td>1-300m</td>
<td>0.1-1m</td>
<td>5-30cm</td>
<td>1-5cm</td>
<td>&lt;2cm</td>
</tr>
<tr>
<td>ENERGY TARGET</td>
<td>&lt;10PJ/b</td>
<td>&lt;1PJ/b</td>
<td>&lt;1PJ/b</td>
<td>&lt;0.5PJ/b</td>
<td>&lt;0.1PJ/b</td>
<td>0.01PJ/b</td>
</tr>
<tr>
<td>COST TARGET</td>
<td>&lt;$10.000</td>
<td>~$100</td>
<td>~$10</td>
<td>~$5</td>
<td>~$1</td>
<td>~$0.01</td>
</tr>
</tbody>
</table>

- Electronic interconnects suffer from many limitations, while Photonic based interconnects show much better promise.
Optical interconnect links

- Optical Links can provide Tbps Data rates
  - But still we are going to need E/O, O/E, transceivers and high performance switch fabrics…

[Source: Intel, The 50Gbps Si Photonic Link]
Options for realizing different 40 Tb/s interconnect scenarios - I

- The need to handle the increasing volumes of data traffic drives the need for ever-increasing communication port densities.

- This increased density leads to smaller surface areas available to dissipate the heat generated and therefore requires decreased power consumption per port.

- As an example, in the Table, we present the required data rate per fibre and the number of links for different 40 Tb/s interconnect scenarios.

```
<table>
<thead>
<tr>
<th>Effective Speed Scenarios</th>
<th>Bits Per Symbol</th>
<th>Symbol Rate (GSp/s)</th>
<th>Wavel. per Fiber</th>
<th>Data Rate per Fiber (Gbps)</th>
<th>No. of Links (Fibers)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1x10G</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>10</td>
<td>8000</td>
</tr>
<tr>
<td>1x25G</td>
<td>1</td>
<td>25</td>
<td>1</td>
<td>25</td>
<td>3200</td>
</tr>
<tr>
<td>1x40G</td>
<td>1</td>
<td>40</td>
<td>1</td>
<td>40</td>
<td>2000</td>
</tr>
<tr>
<td>1x50G</td>
<td>1</td>
<td>50</td>
<td>1</td>
<td>50</td>
<td>1600</td>
</tr>
<tr>
<td>4x25G</td>
<td>1</td>
<td>25</td>
<td>4</td>
<td>100</td>
<td>800</td>
</tr>
<tr>
<td>2x50G</td>
<td>2</td>
<td>50</td>
<td>1</td>
<td>100</td>
<td>800</td>
</tr>
<tr>
<td>8x25G</td>
<td>2</td>
<td>25</td>
<td>8</td>
<td>400</td>
<td>200</td>
</tr>
<tr>
<td>16x25G</td>
<td>2</td>
<td>25</td>
<td>4</td>
<td>800</td>
<td>400</td>
</tr>
<tr>
<td>10x40G</td>
<td>1</td>
<td>40</td>
<td>10</td>
<td>400</td>
<td>200</td>
</tr>
<tr>
<td>8x50G</td>
<td>2</td>
<td>50</td>
<td>10</td>
<td>400</td>
<td>200</td>
</tr>
<tr>
<td>20x40G</td>
<td>2</td>
<td>40</td>
<td>10</td>
<td>800</td>
<td>100</td>
</tr>
<tr>
<td>16x50G</td>
<td>1</td>
<td>50</td>
<td>16</td>
<td>800</td>
<td>100</td>
</tr>
<tr>
<td>40x40G</td>
<td>4</td>
<td>40</td>
<td>10</td>
<td>1600</td>
<td>50</td>
</tr>
<tr>
<td>32x50G</td>
<td>4</td>
<td>50</td>
<td>8</td>
<td>1600</td>
<td>50</td>
</tr>
</tbody>
</table>
```

“OIF Next Generation Interconnect Framework,” April 2013
Options for realizing different 40 Tb/s interconnect scenarios - II

- Given that a bundle of a hundred or less optical fibres could be considered acceptable for an optical conduit, it is easy to understand that in order to achieve 40Tb/s interconnect capacity, a combination of parallel optics, higher baud rates, increased bits per symbol, WDM or/and polarization multiplexing, space-division multiplexing using MCFs, etc. should/could be implemented.

- The use of parallel optics, advanced modulation formats and WDM are regarded as likely choices to increase the interconnect bit rate and density.
- On the other hand, orthogonal modulation in the optical domain and polarization multiplexing are not currently considered due to cost and complexity, but potentially, with the emergence of more advanced SiPh, some of these schemes could become more feasible.
- MCFs could be used to increase port densities by a significant factor and could be an alternative to VCSEL/MMF technology.
- However, at some point the bandwidth limit either in the EO/OE or on the system host chip is reached, where the only other option to further increase the bandwidth is with optical switching.
Future Data Center Networks will require high performance optical switches & WDM

- We need **high-radix, scalable, energy efficient** Data Centers that can sustain the exponential increase of the network traffic

[Source: Kachris, Bergman, Tomkos, “Optical Interconnects for future Data Center Networks”, Elsevier publication]
Roadmap of Ethernet speeds

- Ethernet roadmap, where the year in which the standard was/will be completed is shown

Source: Ethernet Forum
Relative size and power consumption of optical interfaces today and tomorrow

Source: Ethernet Forum
Some possible potential advanced modulation formats that could be utilized up to 400 GbE for intra DCI connectivity are:

- NRZ OOK
- Duobinary modulation
- Pulse amplitude modulation with 4 levels (PAM-4)
- Discrete multi-tone (DMT) modulation (OFDM)
- Others?
PAM-4 based transceiver

- The 4-levels PAM signal is generated by a PAM-4 encoder and is fed into the digital to analog converter (DAC). Then the signal is electrically amplified by a linear driver (DRV). As a final step, at the transmitter’s site, the output of the driver is modulated using an electro-absorption modulated laser (EML) or a directly modulated laser (DML).
- At the receiver’s side, the signal is directly detected with a simple PIN diode and then it is amplified by an electrical amplifier. Finally, the received electrical signal may be processed by digital signal processing (DSP).
### PAM versus DMT

<table>
<thead>
<tr>
<th></th>
<th>PAM4</th>
<th>DMT</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equalizer</td>
<td>FFE, DFE</td>
<td>IFFT,FFT</td>
<td></td>
</tr>
<tr>
<td>Complexity (Multiply-Acc)</td>
<td>x</td>
<td>10x</td>
<td>Dominated by FFT/IFFT complex mult</td>
</tr>
<tr>
<td>DAC resolution</td>
<td>2 bits</td>
<td>8 bits</td>
<td>ENOB, DMT Clipping issues</td>
</tr>
<tr>
<td>ADC Power</td>
<td>x</td>
<td>1.5 - 2x</td>
<td>ENOB, Peak to avg. ratio</td>
</tr>
<tr>
<td>FEC</td>
<td>x</td>
<td>x</td>
<td>DMT higher coding gain?</td>
</tr>
<tr>
<td>Non Linear Canceller</td>
<td>n/a</td>
<td>Volterra</td>
<td>DMT is more sensitive to non-linearity</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>30G</td>
<td>20G</td>
<td>PAM requires higher BW</td>
</tr>
<tr>
<td>Serdes</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td><strong>Total IC Power</strong></td>
<td>Y</td>
<td>2.5Y</td>
<td></td>
</tr>
</tbody>
</table>

Source: Ethernet Forum

- **Power consumption assuming 28nm CMOS technology:**
  - PAM4: 1.5W
  - DMT: 3.5W
Issues to be discussed/debated

- Material system: InP vs. SiP vs. ???
- Integration approaches: monolithic vs. hybrid?
- Packaging approaches?
- Wavelength of operation: 850nm vs. 1310nm vs. 1550nm?
- Laser type: VCSELs vs. DFBs vs. ???
- Direct vs. external modulation?
- The road to 400G and then to 800G/1600G?
- Modulation formats: PAM vs. DMT vs. QAM?
- Direct vs. coherent detection?
- Extend of use of DSP?
- Optical switching in the DC?
- Others???
Session 1 Speakers

- **Martin Zirngibl, Vice President and Technical Fellow, Finisar**
  - Before joining Finisar recently, Martin headed the IP routing research organization at Alcatel-Lucent’s Bell Labs, which he joined in 1990 as a researcher (initially performing pioneering work in fiber amplifiers and photonic integrated circuits). Since 1998, Martin has held various managerial positions at Bell Labs and drove early research and development in 100G that led to a significant industry disruption in WDM long-haul. Martin holds a PhD in applied physics from ETH Lausanne, Switzerland.

- **Brad Booth, Principal Engineer, Microsoft**
  - Brad Booth is heading up the 25/50G Ethernet Consortium and the Consortium for On-Board Optics (COBO). At Microsoft, he leads the development of hyperscale interconnect strategy for Microsoft’s cloud datacenters. He is also the founder and past Chairman of the Ethernet Alliance. Brad was previously a Distinguished Engineer in the Office of the CTO at Dell Networking and he has also held senior strategist and engineering positions at Applied Micro, Intel, and PMC-Sierra.

- **Toshiki Tanaka, Research Manager, Fujitsu Laboratories Ltd**
  - Toshiki joined Fujitsu in 1997 and has been engaged since then in research and development of dense WDM optical transmission systems and high capacity optical transceivers with higher order modulation. Since 2016, he has engaged in developing 100G and beyond 100G optical transceivers for short reach application at Fujitsu Optical Components Limited.