Trends in 400G Optics for the Data Center

Christian Urricariet
NANOG 75
San Francisco, February 2019
Due to the ongoing large increase in bandwidth demand, Data Center connections are expected to move from 25G/100G to 100G/400G.

- Within the Data Center Racks
  - 10GE still being deployed
  - 25GE starting to be deployed in volume
  - 100GE (or 50G) will follow

- Between Data Center Racks
  - 40GE still being deployed
  - 100GE starting to be deployed in volume
  - 400GE will follow at large Cloud Service Providers

- Long Spans/DCI & WAN
  - 10G DWDM/Tunable still being deployed
  - 100G/200G Coherent starting to be deployed
  - 400G Coherent will follow – Then 600G or 800G
Forecasted Data Center Ethernet Port Shipments

Source: Dell’Oro, 2018
Forecasted 400GE Shipments by Market Segment

Port Shipments in Millions

2018  2019  2020  2021  2022

- Top 4 Cloud SPs
- Rest of Cloud
- Large Enterprises

Source: Dell’Oro, 2018
## Mainstream 1RU Ethernet Switch Roadmap

<table>
<thead>
<tr>
<th>First Deployed</th>
<th>Electrical I/O [Gb/lane]</th>
<th>Switching Bandwidth</th>
<th>TOR/Leaf Data Center Switch Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>~2010</td>
<td>10G</td>
<td>1.28T</td>
<td>32xQSFP+ (40G)</td>
</tr>
<tr>
<td>~2015</td>
<td>25G</td>
<td>3.2T</td>
<td>32xQSFP28 (100G)</td>
</tr>
<tr>
<td>~2019</td>
<td>50G</td>
<td>6.4T</td>
<td>32 ports of 200G</td>
</tr>
<tr>
<td>~2020</td>
<td>50G</td>
<td>12.8T</td>
<td>32 ports of 400G</td>
</tr>
</tbody>
</table>

3.2Tb/s switches based on 100G QSFP28 modules being deployed in cloud data centers today.

Given the multiple switching ICs expected to be available, the market is likely to be fragmented in the future.

Large growth in bandwidth demand is pushing the industry to work on technologies and standards to support future **12.8T switches**.
400G and Next-Gen 100G Ethernet Optical Standardization

<table>
<thead>
<tr>
<th>Interface</th>
<th>Link Distance</th>
<th>Media type</th>
<th>Optical Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>400GBASE-SR16</td>
<td>100 m (OM4)</td>
<td>32f Parallel MMF</td>
<td>16x25G NRZ Parallel VCSEL</td>
</tr>
<tr>
<td>400GBASE-DR4</td>
<td>500 m</td>
<td>8f Parallel SMF</td>
<td>4x100G PAM4 Parallel (SiP)</td>
</tr>
<tr>
<td>400GBASE-FR8</td>
<td>2 km</td>
<td>2f Duplex SMF</td>
<td>8x50G PAM4 LAN-WDM (DML)</td>
</tr>
<tr>
<td>400GBASE-LR8</td>
<td>10 km</td>
<td>2f Duplex SMF</td>
<td>8x50G PAM4 LAN-WDM (DML)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Interface</th>
<th>Link Distance</th>
<th>Media type</th>
<th>Optical Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>400GBASE-SR16</td>
<td>100 m (OM4)</td>
<td>4f Parallel MMF</td>
<td>2x50G PAM4 850nm (VCSEL)</td>
</tr>
<tr>
<td>100GBASE-DR</td>
<td>500 m</td>
<td>2f Duplex SMF</td>
<td>100G PAM4 1310nm (EML)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Interface</th>
<th>Link Distance</th>
<th>Media type</th>
<th>Optical Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>400GBASE-SR8</td>
<td>100 m (OM4)</td>
<td>16f Parallel MMF</td>
<td>8x50G PAM4 850nm (VCSEL)</td>
</tr>
<tr>
<td>400GBASE-SR4.2</td>
<td>100 m (OM4)</td>
<td>8f Parallel MMF</td>
<td>8x50G PAM4 BiDi 850 / 910nm (VCSEL)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Interface</th>
<th>Link Distance</th>
<th>Media type</th>
<th>Optical Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>400G-FR4</td>
<td>2 km</td>
<td>2f Duplex SMF</td>
<td>4x100G PAM4 CWDM (EML)</td>
</tr>
<tr>
<td>100G-FR</td>
<td>2 km</td>
<td>2f Duplex SMF</td>
<td>100G PAM4 1310nm (EML)</td>
</tr>
<tr>
<td>100G LR</td>
<td>10 km</td>
<td>2f Duplex SMF</td>
<td>100G PAM4 1310nm (EML)</td>
</tr>
</tbody>
</table>

- VCSEL technology to be used <100m
- Silicon Photonics to be used <1km
- DML/EML technology to be used <40km

SWDM to enable 400GE over Duplex MMF in the future
CFP8 is the 1st-generation 400GE module form factor, to be used in core routers and DWDM transport client interfaces.

Module dimensions are slightly smaller than CFP2
Supports either CDAUI-16 (16x25G NRZ) or CDAUI-8 (8x50G PAM4) electrical I/O

QSFP-DD and OSFP modules being developed as 2nd-generation 400GE, for high port-density data center switches.

Enable 12.8Tb/s in 1RU via 32 x 400GE ports
Support CDAUI-8 (8x50G PAM4) electrical I/O only
QSFP-DD host is backwards compatible with QSFP28
## Potential Types of 400G Transceivers in the Market

<table>
<thead>
<tr>
<th></th>
<th>PARALLEL (MPO)</th>
<th>DUPLEX (LC)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MULTIMODE</strong></td>
<td><strong>SR8 &amp; 8x50G-SR (breakouts)</strong></td>
<td><strong>TBD / TBD</strong></td>
</tr>
<tr>
<td></td>
<td>70/100/100m</td>
<td><strong>Future support for existing enterprise duplex fiber infrastructure</strong></td>
</tr>
<tr>
<td></td>
<td><strong>SR4.2</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>70/100/150m</td>
<td></td>
</tr>
<tr>
<td><strong>SINGLE MODE</strong></td>
<td><strong>DR4</strong></td>
<td><strong>FR8 / FR4 / CWDM8</strong></td>
</tr>
<tr>
<td></td>
<td>500m</td>
<td>2km</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>LR8 / LR4</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10km</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>ER8</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>40km</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>ZR</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>80km</td>
</tr>
</tbody>
</table>

Multimode distances refer to OM3/OM4/OM5 MMF; Single mode distances refer to SMF.

**BLACK** = Standard IEEE interfaces

**RED** = MSA and Proprietary interfaces
# 400G Ethernet Is Taking Shape in the Cloud Data Center

**Metro DCI (< 80km)**
- Deploying 100G/200G Coherent
- Roadmap is 400GE LR8/ER8, ZR Coherent

**Tier 2 Switch to Tier 1 Switch links**
- Deploying 100GE CWDM4/PSM4
- Roadmap is 400GE FR8/FR4/DR4

**Tier 1 Switch to TOR Switch links**
- Deploying 100GE SR4/AOC
- Roadmap is 400GE DR4/SR4.2/SR8/AOC

**TOR Switch to Server links**
- Deploying 25GE SR/AOC/DAC(3m)
- Roadmap is 50GE/100GE SR/AOC

**Roadmap**
- Deploying 100G/200G Coherent
- Roadmap is 400GE LR8/ER8, ZR Coherent

**Tier 2 Switch**
- 12.8Tb Switches
- 3.2Tb Switches

**Tier 1 Switch**
- Tier 1 Switch

**TOR Switch**
- Tier 2 Switch

**RACK**
- 25G Servers
- 50G/100G Servers
General Trends in Data Center Optical Interconnects

- Continuous increase in bandwidth density
  - On-board optics vs. pluggable optics discussion

- Increasing adoption of optics in Server-to-TOR Switch links

- Low-latency optics for certain niche cognitive-computing applications

- Maturity of key technologies
  - High-speed VCSELs
  - Silicon photonics

- Arrival of coherent optics for data center interconnects
The Market Demands Continuous Improvement in Bandwidth Density

<table>
<thead>
<tr>
<th>Module Type</th>
<th># of I/O lanes</th>
<th>Electrical I/O</th>
<th>I/O Baud Rate</th>
<th>Module BW</th>
<th>Width (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SFP+</td>
<td>1</td>
<td>10Gb/s-NRZ</td>
<td>10G</td>
<td>10Gb/s</td>
<td>13</td>
</tr>
<tr>
<td>QSFP+</td>
<td>4</td>
<td>10Gb/s-NRZ</td>
<td>10G</td>
<td>40Gb/s</td>
<td>18</td>
</tr>
<tr>
<td>QSFP28</td>
<td>4</td>
<td>25Gb/s-NRZ</td>
<td>25G</td>
<td>100Gb/s</td>
<td>18</td>
</tr>
<tr>
<td>QSFP56</td>
<td>4</td>
<td>50Gb/s-PAM4</td>
<td>25G</td>
<td>200Gb/s</td>
<td>18</td>
</tr>
<tr>
<td>QSFP-DD / OSFP</td>
<td>8</td>
<td>50Gb/s-PAM4</td>
<td>25G</td>
<td>400Gb/s</td>
<td>18</td>
</tr>
<tr>
<td>Form factor?</td>
<td>8</td>
<td>100Gb/s-PAM4</td>
<td>50G</td>
<td>800Gb/s</td>
<td>?</td>
</tr>
</tbody>
</table>

256 x 25G Switch System – 2 RU (64 x QSFP28 interfaces)
Is Pluggability Still a Requirement for Optics?

- Some optics are not pluggable; they are mounted directly on the system host PCB.
  - BOAs have been used for several years on core routers (inter-chassis) and supercomputers.
  - Very short host PCB traces enable low power dissipation and high port density.

- Higher bandwidth density can be achieved by:
  - More channels, e.g., up to 16 Tx and 16 Rx channels in a module.
  - Higher data rate per channel: 10G/ch and 25G/ch variants deployed today, 50G/ch in the future.

- The industry view however is that **pluggable optics** will be preferred for 400G Ethernet.
  - Facilitates maintenance and pay-as-you-grow model.
Optical Technologies for Next-Generation Data Centers

- Short Reach (0 to 100 meters)
  - Higher bandwidth VCSELs
  - VCSELs with sparing capability
  - VCSELs with low RIN

- Intermediate Reach (500 meters to 2 km)
  - Silicon photonics

- Long Reach (10 km and beyond)
  - DML/EMLs
  - Low-power coherent optics
For short reach links (<100 m), multimode optics have better cost, power, density than silicon photonics

**VCSEL Tx components:**
- High speed 850nm VCSEL, 150µm x 150µm die, ~8 µm diameter
- Coupling optics with 10 µm’s tolerance
- VCSEL driver, <10 mA
- Non-hermetic packaging
- MM optical alignments: 1

**Si Photonics Tx components:**
- High power CW 1310nm DFB, >500 µm long
- Optical isolator
- External modulator
- Coupling optics with ~1 µm tolerance
- Si Photonic PIC + IC driver, several tens of mA
- Hermetic box or enclosure
- SM optical alignments: 2
Today’s application space is divided by fiber type and reach:

- **0-100m MM**: MM VCSEL today (OM3/OM4 fiber types)
- **0-300m MM**: MM VCSEL (OM4/OM5 fiber type for 4x25G)
- **0-2 km SM**: DML, Si Photonics *New application space since 100G*
- **0-10 km SM**: EML, DML

**Technology**
- MM VCSEL
- MM VCSEL
- DML, Si Photonics
- EML, DML

**Reach**
- 0m
- 100m
- 300m
- 500m
- 1000m
- 10 km

**Loss budget**
- ~3 dB
- ~3 dB
- ~3-4 dB
- 6.5 dB

**Average Optical Link Length Distributions**
- **2016?**
- **2014?**
- **2018?**
Implementation of 100G-PSM4 With Silicon Photonics

- Over 84% of the functions are integrated with high yield at wafer level in a Si fab
  - SiP integrates 4 modulators, 4 pin diodes, 4 x drivers, 4 x TIAs, CDR in a 3-D stacked self-hermetic, uncooled chip.
  - A similar architecture is being implemented for 400G-DR4 (“400G PSM4”), driven by the need for longer reach, future-proof singlemode cabling, and integration in optics.

Traditional

With silicon photonics

© Finisar Corporation
Several New Interface Types and Form Factors to be Deployed

- Enabled by high I/O count and by 400G-DR4 to 100G-DR breakout interoperability, high-density 100G implementations will thrive in Leaf-Spine topologies.

- Large I/O Line Cards will have QSFP-DD or OSFP sockets. OSFP slots may use QSFP adapters.

© Finisar Corporation
400G, 200G & 100G PAM4 Transceiver Demos at OFC/ECOC 2018

- 400G QSFP-DD LR8/FR8 (10km)
- 400G QSFP-DD AOC (70m)
- 400G QSFP-DD eLR8 (30km)
- 400G QSFP-DD DR4 (500m)
- 200G SFP56 FR4 (2km)
- 200G QSFP56 eFR4 (10km)
- 100G QSFP28 DR/FR (2km)

Additionally, several interoperability demos were done by the MSAs
- Coherent interfaces are likely to capture the 80km market at 400Gb/s and higher rates.
- For 40km and shorter reaches, direct detection may be lower power and cost than coherent for the next few years. Example: 8x50Gb/s (PAM4) ER8 and eLR8 modules.
- Currently coherent technology is about 2x higher power and cost relative to 100Gb/lane direct detection.
- Standardization work by OIF 400ZR IA and IEEE P802.3cn/ct Task Forces (400G ER8 and ZR).
- Aggressive innovation will be required to maintain long-term trends to support 1.6 Tb/s ~2024.
Coherent Transmission for DCI Applications

- 100G/200G links require a transponder box to convert to coherent optical transmission in order to support 80~100km and beyond.

- Several system OEMs provide a 1RU transponder box for DCI applications, most of which use pluggable Coherent CFP2-ACO optical transceivers.

- Expected coherent transceiver evolution is driven by improvements in optical packaging and DSP power dissipation:

  \[ 200\text{G CFP2-ACO} \rightarrow 400\text{G CFP2-DCO} \rightarrow 400\text{G QSFP-DD DCO} \]

\[ \text{400G DCO transceivers are expected to be plugged directly into switches and routers} \]
Coming Next: What Shape Will 800G Ethernet Take?

- **Switch 25Tb**
  - 800G / 1.6T
  - DCI <80km?
  - Coherent DWDM

- **DATA CENTER 2**
  - 32x800G Optical Modules
  - **Pluggable: Is it Feasible?**
    - No
    - Yes

- **METRO/REGIONAL**
  - BOA/OBO
  - Form Factor?
  - Electrical I/O Fly-over Cables? 2RU Switch?

- **FABRIC/TIERx**
  - 32x800G 2 km

- **FABRIC/TIER1**
  - 32x800G 500 m

- **TOR**
  - 100G DAC / AOC

- **SERVERS**
  - 100G

---

100G PAM4 electrical I/O being standardized in IEEE P802.3ck
Thank You

Christian Urricariet
christian.urricariet@finisar.com