Optical Interconnects for Data Center Systems

Xuezhe Zheng
Photonics Director
Netra Systems and Networking Group
Oracle
Challenges in Data Centre

- Version / Releases
- Patching
- Instances
- Utilization (max, not used)
- Performance / Response
- Scalability
- Availability / Resilience
- DC Floor space / Power
- Mixed Workloads

- Test, Dev Environments
- SLA, Change Control
- Configuration
- Infrastructure Refresh
- Planned/Unplanned Maint
- Backup & Recovery
- Provisioning
- Responding to business
- Maintenance Costs
Oracle’s Engineered Systems

- Engineered systems provide extreme performance, simpler management, one-stop support, expedited time to value, lower cost of ownership, and scalability

---

**Example: Oracle’s Exalogic Elastic Cloud for the Enterprise**

<table>
<thead>
<tr>
<th>Rack Type</th>
<th>Number of Nodes</th>
<th>Memory 1</th>
<th>Memory 2</th>
<th>Storage 1</th>
<th>Storage 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eighth Rack</td>
<td>4</td>
<td>384 GB RAM</td>
<td>800 GB SSD</td>
<td>60 TB NAS</td>
<td></td>
</tr>
<tr>
<td>Quarter Rack</td>
<td>8</td>
<td>768 GB RAM</td>
<td>1.6 TB SSD</td>
<td>60 TB NAS</td>
<td></td>
</tr>
<tr>
<td>Half Rack</td>
<td>16</td>
<td>1.5 TB RAM</td>
<td>3.2 TB SSD</td>
<td>60 TB NAS</td>
<td></td>
</tr>
<tr>
<td>Full Rack</td>
<td>30</td>
<td>2.8 TB RAM</td>
<td>6 TB SSD</td>
<td>60 TB NAS</td>
<td></td>
</tr>
<tr>
<td>Multi-Rack</td>
<td>240+</td>
<td>23+ TB RAM</td>
<td>48+ TB SSD</td>
<td>60+ TB NAS</td>
<td></td>
</tr>
</tbody>
</table>

---

**Example: Oracle’s Big Data Platform**

- Acquire
- Organize
- Analyze
- Decide

---

**Oracle’s Engineered Systems**

- Exadata Database Machine
- Exalogic Elastic Cloud
- Database Appliance
- Exalytics
- Big Data Appliance
- SPARC SuperCluster
- Virtual Compute Appliance
Engineered System Interconnect

- Software & Hardware Engineered to Work Together
  - Highly compact and efficient system design
    - More aggregation at all levels: package; box; and rack; instead of vanilla boxes
  - Best overall performance & efficiency

- High-Bandwidth System Interconnect
  - Standardized pluggable modules not necessary and most likely not sufficient
  - Bandwidth-density is the challenge

~1.4x every 2yrs
Evolution of system interconnect

Yesterday

- Long Reach Serdes
  - BW*density limited
  - BW*distance limited
  - <5-10Gbps

- Active Optical Cables
  - <15Gbps

- Long/Short Reach Serdes
  - BW*density limited
  - Mid-Board Optical Modules
  - 25+ Gbps

- Very Short Reach Serdes
  - CPU/DRAM + EO/OE MCM
  - 40+ Gbps

Today

- 40" trace
- 25" trace
- 15" trace
- <15m
- <5m
- <7" <20-30m

- Density & Total BW increase as we get closer to the ASIC

Next?

- **New solutions to address**
  - Bandwidth distance limitation
  - Bandwidth density limitation at both front panel and package levels
  - Energy efficiency
  - Future scalability
Industry efforts: 100G Optical Multi-Source Agreements

- Four recently announced TxR MSAs target 100G via 4x25.78Gb/s (or higher)
  - Three WDM MSAs specify a QSFP form factor TxR w/ 2km reach using single-mode fiber
  - One parallel MSA specifies a 500m reach transceiver w/ single-mode fiber ribbon
  - OpenOptics, CLR4, PSM4 have silicon photonic implementations

<table>
<thead>
<tr>
<th>MSA</th>
<th>Optical I/O</th>
<th>Wavelength band</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenOptics MSA (Mellanox, Ranovus)</td>
<td>LC duplex/WDM</td>
<td>1550nm C-band on ITU-T grid</td>
</tr>
<tr>
<td>100G CLR4 Alliance “Coarse LR4”</td>
<td>LC duplex/WDM</td>
<td>1310nm coarse WDM O-band</td>
</tr>
<tr>
<td>(Intel, Arista, others)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CWDM4 MSA “Coarse WDM 4” (Finisar,</td>
<td>LC duplex/WDM</td>
<td>1310nm coarse WDM O-band</td>
</tr>
<tr>
<td>others)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSM4 MSA “Parallel SMF 4” (Luxtera</td>
<td>MPO/Parallel</td>
<td>1310nm O-band</td>
</tr>
<tr>
<td>,Avago, MSFT, Juniper,...)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Motivation for Optics “In the Box”

**MBOM**: mid-board optical IO solution with optical transceivers near ASIC

- Eliminate the front panel bandwidth limit
- Improve signaling across the system
- VCSEL based parallel optics for 25Gbps +

8x panel density improvement compared with QSFP
Optics to the package eventually

Multi-chip Package with Optical IOs

- Addresses bandwidth limitation at both front panel and ASIC package
- Better energy efficiency by minimizing the reach of electrical links
- Single mode WDM SiP interconnects further extends the system reach with better scalability