Considerations for 25GbE

John D’Ambrosia, Dell
Jon Lewis, Dell
Kapil Shrikhande, Dell

IEEE 802.3 25GbE Study Group
IEEE 802.3 Sept 2014 Interim
Kanata, Canada
Ethernet is Evolving

• Changing environment
  – “10x the Performance @ 3x the cost”
  – “The Ethernet of Everywhere” – being used everywhere for everything
    - Pick one – “Web Scale Data Center, Enterprise Data Center,, Enterprise, Campus, Client Side Connections, Etc”
    - Architectures Top-of-Rack, End-of-Row, Middle-of-Row
  – IEEE 802.3 Ethernet Bandwidth Assessment – 32% to 95% CAGR
  – Connections from ≈0m to 40km
  – Market Timings
  – PoE Certification discussions

• Per the 25GbE CFI Consensus Presentation:
  – Web-scale data centers and cloud based services need
    - Servers with >10GbE capability
    - Cost sensitive for nearer-term deployment

• Remember that Ethernet products designed for this space will move into other applications!
Things to Consider for Objectives

- Cu Cable Reach
- Need for an MMF Objective?
- Need for electrical interfaces?
A Few Words First…

- Lane Rate / Maximizing Switch Efficiency / Breakout to lower rates driving new issues

- Examples –
  - Success of 40GbE or 10GbE?
    - Breakout from QSFP has been a noted success.
    - Challenges in quantifying application volumes
    - “Fixed Ports” – Form factor – yes, Media / Rate – no
  - On-going debate in IEEE P802.3bs 400GbE in relation to 100GbE breakout
  - Formation of IEEE 802.3 25GbE Study Group

- From 25GbE Consensus Presentation

- 25Gb/s technology standardized, developed, productized for 100GbE can be leveraged now!
<table>
<thead>
<tr>
<th>Technology</th>
<th>Nomenclature</th>
<th>Description</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backplanes</td>
<td>100GBASE-KR4</td>
<td>4 x 25 Gb/s (NRZ)</td>
<td>IEEE Std 802.3bj™-2014 Ratified</td>
</tr>
<tr>
<td></td>
<td>100GBASE-KP4</td>
<td>4 x 25 Gb/s (PAM-4)</td>
<td></td>
</tr>
<tr>
<td>Cu Twin-Axial</td>
<td>100GBASE-CR4</td>
<td>4 x 25 Gb/s</td>
<td></td>
</tr>
<tr>
<td>Chip-to-Chip</td>
<td>CAUI-4</td>
<td>4 x 25 Gb/s</td>
<td>IEEE P802.3bm in Sponsor Ballot</td>
</tr>
<tr>
<td>Chip-to-Interface</td>
<td>CAUI-4</td>
<td>4 x 25 Gb/s</td>
<td></td>
</tr>
<tr>
<td>Module Form Factor</td>
<td>SFP28</td>
<td>1 x 28 Gb/s</td>
<td>Summary Document SFF-8402</td>
</tr>
<tr>
<td></td>
<td>QSFP28</td>
<td>4 x 25 Gb/s</td>
<td>Style 1 - MDI for 100GBASE-CR4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Summary Document SFF-8665</td>
</tr>
<tr>
<td></td>
<td>CFP2</td>
<td>4 x 25 Gb/s</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CFP4</td>
<td>4 x 25 Gb/s</td>
<td>Style 2 MDI for 100GBASE-CR4</td>
</tr>
</tbody>
</table>
Cloud Adoption of 25 GE in Stand-Alone Servers

Percent of Server Shipments


2014 Dell'Oro Group
Big Driver: Total Cost of Ownership

Consider Today’s Cloud Scale Data Centers

<table>
<thead>
<tr>
<th>Server I/O</th>
<th>Oversubscription</th>
<th>Servers</th>
<th>100G Uplinks</th>
<th>Throughput (Tb/s) per ToR Switch</th>
<th>Utilization (%)</th>
<th># TORs for a 100K Server Data Center</th>
</tr>
</thead>
<tbody>
<tr>
<td>40Gbe (4x10G)</td>
<td>2.8:1</td>
<td>28</td>
<td>4</td>
<td>1.52</td>
<td>47.5</td>
<td>3572</td>
</tr>
<tr>
<td>40Gbe (2x20G)</td>
<td>2.4:1</td>
<td>48</td>
<td>8</td>
<td>2.72</td>
<td>85</td>
<td>2084</td>
</tr>
<tr>
<td>25Gbe Single Lane</td>
<td>3:1</td>
<td>96</td>
<td>8</td>
<td>3.2</td>
<td>100</td>
<td>1042</td>
</tr>
</tbody>
</table>

- Total Cost of Ownership - Minimize cost per bit!
  - CAPEX – Top of Rack Switches, Interconnect Structure
  - OPEX – Power / Cooling

Represents 100% port utilization and no stranded ports
Cu Cable Distribution

• Data obtained from
  - Two product groups within Dell (past 1 to 1.5 years)
  - 10GbE based products (servers & switches)
  - 40GbE based products (servers & switches)

<table>
<thead>
<tr>
<th>Total (Cu Cable)</th>
<th>Division A</th>
<th>Division B</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;=3m</td>
<td>79%</td>
<td>63%</td>
</tr>
<tr>
<td>5m</td>
<td>21%</td>
<td>28%</td>
</tr>
<tr>
<td>&gt;5m</td>
<td>0%</td>
<td>8%</td>
</tr>
</tbody>
</table>

• Data obtained from
  - Two cabling companies (Molex, Tyco)

<table>
<thead>
<tr>
<th>Total (Cu Cable)</th>
<th>Company A MDI1/MDI2</th>
<th>Company B</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;=3m</td>
<td>62% / 69%</td>
<td>80%</td>
</tr>
<tr>
<td>5m</td>
<td>30% / 24%</td>
<td>15%</td>
</tr>
<tr>
<td>&gt;5m (Passive)</td>
<td>1%</td>
<td>5%</td>
</tr>
<tr>
<td>&gt;7m (Active)</td>
<td>7% / 6%</td>
<td>-</td>
</tr>
</tbody>
</table>
Thoughts Related to Cu Cable Objective

- Different applications easy to envision –
  - Lower density based on 25GbE / SFP28 from server to switch
  - Higher density based on breakout from 100GBASE-CR4 / QSFP28 on switch to SFP28 on server

- Channel budget?
  - Switch – must be constrained to meet 100GBASE-CR4 budget
  - Server – different options –
    - Different server port types to support budget for lower / higher density switch applications?
    - Add budget to server from cable?
    - Reduced budget for server NIC, 3m no FEC?
Beyond Top of Rack

I/O per 10G Server** ports

<table>
<thead>
<tr>
<th>SFP+ (PHY Type)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SR Optics</td>
<td>31.5%</td>
</tr>
<tr>
<td>CR (all passive, no active cables)</td>
<td>Some portion of 68.5%</td>
</tr>
<tr>
<td>Unknown*</td>
<td>Some portion of 68.5%</td>
</tr>
</tbody>
</table>

- in some instances dual ports are used for physical redundancy, but one port may not be populated

** - data gathered from Dell general purpose server family

IEEE 802.3 25GbE Study Group
Sept 2014 Interim, Kanata, Canada
Comparison Between Options

• Option #1 – Reduce cable reach to 3m / assign budget to server
  – High density passive Cu switching applications limited to intra-rack / higher density server form factors
  – Potential for stranded ports on high density switches increases
    – More switches – CAPEX / OPEX
  – Forces use of active cable assemblies / optics for reaches beyond 3m
  – Limits broad market potential to intra-rack applications

• Option #2 – Choose 5m reach objective and TF can specify 3m cable with no FEC
  – Asymmetrical budget, NIC needs less loss than budgeted for host. Leave switch budget alone, 3m cable. No FEC
  – Reduces latency (for those applications)

• Option #3 –
  – Choose objective targeting 3m intra-rack applications
  – Choose objective targeting 5m inter-rack applications
Chip-to-Module (C2M) Interfaces

• Chip-to-module (C2M) interfaces will happen
  – SFP28 for 25GbE connections anticipated
  – QSFP28 for 4x25 GbE connections anticipated
  – Recommend SFP28 / QSFP28 for MDI

• C2M channel budget details need to be consolidated with host trace portion of Cu cable channel budget

• Chip-to-chip interface should be defined

• Leverage IEEE 802.3bm work

• Recommend adopting objectives for chip-to-chip and chip-to-module electrical interfaces
Summary

- **Recommendations**
  - Adopt a Cu cable objective for a 5m reach only (inter-rack)
  - Adopt a Cu cable objective for a 3m reach only (intra-rack)
  - Adopt a MMF Objective targeting xx m
    - Data on reach not available at this time
  - Consider SFP28 / QSFP28 for 25GbE MDIs
  - Adopt objectives for chip-to-chip and chip-to-module electrical interfaces