Objectives

- To gauge the interest in starting a study group to investigate a 25 Gigabit Ethernet project

- Don’t need to:
  - Fully explore the problem
  - Debate strengths and weaknesses of solutions
  - Choose a solution
  - Create a PAR or 5 Criteria
  - Create a standard

- Anyone in the room may vote/speak
Agenda

• Overview
• MAC-PHY Mismatch
• Potential Use Cases
• Why Now?
• Straw Polls
Agenda

• Overview
• MAC-PHY Mismatch
• Potential Use Cases
• Why Now?
• Straw Polls
25G Ethernet Overview

• Provide a 25G media access control (MAC) that matches the single-lane 25G physical layer (PHY) technology
• In web-scale data centers, 25G Ethernet could provide an efficient server to top-of-rack (TOR) speed increase
  • Predominantly direct-attach copper (DAC) cable
• The speed of the PCIe host bus is not moving as fast as networking connectivity speeds
Existing 10G Topology

- Today’s volume topology for web-scale data centers
- 48 servers/TOR
- 3:1 oversubscription
- Uses low-cost, thin 4-wire SFP+ DAC cable

[Diagram showing TOR, 48-port 10G, 4-port 40G, and connected servers with 25G Ethernet CFI]
Existing 4x10G Topology

- Commonly used topology in web-scale data centers
- Permits non-blocking 10G mesh
- 40G ports used as 4x10G with QSFP+ to SFP+ break-out cable
- Same server network interface card (NIC) as 10G
40G Topology

- High-performance, low-volume topology
- Uses bulkier 16-wire QSFP+ DAC cable
- Max. 24 servers/TOR with 3:1 oversubscription
- Will transition to 100G
Agenda

• Overview
• MAC-PHY Mismatch
• Potential Use Cases
• Why Now?
• Straw Polls
MAC-PHY Mismatch History

• Single-lane PHY technology was able to keep pace with Ethernet up to 10 Gb/s

• 802.3ba 40G and 100G Ethernet technology relied on 10G as its foundation

• 25G was used only for long-reach optical interconnect

• 802.3bj and .3bm are building a foundation based upon 25G PHY technology
MAC-PHY Mismatch History Part 2

- Short reach 40G is 4 lanes of 10G
  - Led to the development of break-out cables both for copper and optical cables
  - Permits greater faceplate density on switches
- New 100G efforts are built on 4 lanes of 25G
  - Unlike at 40G, no ability to break-out to 25G as not a supported MAC data rate
  - 40G doesn’t map to 100G as easily as 25G
Learnings from 40G

• Four lanes is good
  • 4x 10G provides good TOR to server density
  • Provides the ability to use 4x 10G ports to build a single speed non-blocking mesh network
  • Increased faceplate density on TOR switch
• Web-scale data centers
  • 10G DAC in high volume for servers
  • Mates nicely with 40G or 10G mesh at TOR and above
100G Family

• Currently using 25G as its primary building block
• No means to take advantage of 25 Gb/s building block like industry did with 4x 10G – 40G family
• New 25G family could easily build upon existing 802.3 specifications
  • Permits a focused and timely standards project
  • Simplifies development of interoperable specification and systems
25G Industry Dynamics

• Technology re-use
  • Single-lane of 100G 4-lane PMD and CAUI-4 specifications
  • SFP28 being developed for 32G FC

• Areas of modification
  • PCS is based on 20 lanes and could support 5 lanes
  • Backplane FEC is striped over 4-lanes, but could operate over single-lane with increased latency

• Can support multiple data center refresh cycles
PCIe Gen3 Lanes Required per Ethernet Rate

<table>
<thead>
<tr>
<th>Ethernet rate</th>
<th>Single-port</th>
<th>Dual-port</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 Gb/s</td>
<td>16 lanes</td>
<td>32 lanes (uncommon)</td>
</tr>
<tr>
<td>40 Gb/s</td>
<td>8 lanes</td>
<td>16 lanes</td>
</tr>
<tr>
<td>25 Gb/s</td>
<td>4 lanes</td>
<td>8 lanes</td>
</tr>
<tr>
<td>10 Gb/s</td>
<td>2 lanes</td>
<td>4 lanes</td>
</tr>
</tbody>
</table>

- **Server ports**
  - Trend is towards single-port servers due to cost
  - Volume servers are typically deployed with x4 PCIe
- **25G Ethernet is an easier upgrade path from 10G**
  - Requires half the number of lanes compared to 40G (x4 instead of x8 PCIe lanes)
  - Better PCIe bandwidth utilization (25/32=78% vs. 40/64=62.5%) with lower power impact
- **PCIe Gen4**
  - Work is in progress
  - Lane reduction by a factor of 2 with same utilization/power impact considerations
  - PCIe Gen3 will be a sizable part of the market for some time

* Source: Adee Ran, Intel
† Reflects nominal data rates, not traffic patterns
PCIe to Ethernet Throughput Matching

* Source: Adee Ran, Intel
† Reflects nominal data rates, not traffic patterns
Agenda

- Overview
- MAC-PHY Mismatch
- Potential Use Cases
- Why Now?
- Straw Polls
25G Direct Connect

- Same topology as 10G
- 48 servers/TOR
- 3:1 oversubscription w/ 100G uplinks, non-blocking w/ 400G
- Uses 4-wire SFP28 DAC cable
4x25G Breakout

- Same topology as 4x10G
- Permits non-blocking 25G mesh
- 100G ports used as 4x25G with QSFP28 to SFP28 break-out cable
- Same server network interface card (NIC) as 25G direct connect
High Density 25G

- Increased port switch port density
  - 64 servers in non-blocking architecture
  - 96 servers in a 3:1 oversubscription
- 24-port 400G TOR
- 192 servers in non-blocking architecture
Cost Dynamics

• 25G Ethernet
  • Single SERDES
  • Can use SFP28
  • Requires only 4-wire DAC cabling (similar to 10G)

• 40G Ethernet
  • Four SERDES (using 4x10G) & 16-wire DAC
  • No spec for 2x20G; different module?
Agenda

• Overview
• MAC-PHY Mismatch
• Potential Use Cases
• Why Now?
• Straw Polls
Creating 25G Ethernet

• While the 25G SERDES technology exists within IEEE 802.3, there is no supporting 25G MAC
  • Specifying a 25G MAC is extremely simple
  • Draw from 802.3bj and .3bm to create PHYs to expedite the standards development

• Creates useful breakout functionality
  • 100G ports can support 4x 25G
  • 400G ports could support 16x 25G
Why Start Now?

• 802.3bj is coming to a close
  • Experienced and knowledgeable folks will be ready for a new project
• 25G Ethernet would draw heavily on their capabilities
• Market readiness
  • 100G switch silicon using 4x25G will hit the market in the next 12-18 months
  • Project could capitalize on 100G (4x25G) market adoption
Study Group Considerations

- Timeliness of effort vs. breadth of PMDs
- Development of future SERDES technology (i.e. 40 Gb/s, 56 Gb/s)
- BASE-T technologies
- Optical PMDs... SR only? Should LR or ER be considered?
- Impact to ITU
Contributors

• Adee Ran, Intel
Supporters

• Nathan Tracy, TE Connectivity
• Tom Palkert, Molex
• Scott Sommers, Molex
• Mark Bugg, Molex
• Tom Issenhuth, Microsoft
• Bernie Hammond, TE Connectivity
• Greg McSorley, Amphenol
• Theodore Brillhart, Fluke Networks
• Andy Moorwood, Infinera
• Nathan Farrington, Packet Counter
• Rich Mellitz, Intel
• Adee Ran, Intel
• Shamim Akhtar, Comcast
• Martin Carroll, Verizon
• Andy Bechtolsheim, Arista
• Kent Lusted, Intel
• Mike Li, Altera
• Mark Gustlin, Xilinx
• Arlon Martin, Mellanox
Straw Polls
Call-for-Interest Consensus

• Should a study group be formed for “25 Gigabit Ethernet”?

Y: N: A:
Participation

• I would participate in a “25G Ethernet” study group in IEEE 802.3?
  • Tally:

• My company would support participation in a “25G Ethernet” study group?
  • Tally:
Future Work

• Ask 802.3 at Thursday’s closing session to form a 25 Gigabit Ethernet study group

• If approved:
  • 802 EC informed on Friday of formation of the study group
  • First study group meeting would be during May 2014 IEEE 802.3 interim meeting
Thank you!