50Gb/s Per Lane Electrical and Optical Technology: The Next Generation of Server I/O

SP-4: System Level Issues
The Open Server Summit
Santa Clara, CA
11 November 2014
Chris Cole
Outline

■ Caveat
■ Historical & Recent Rate Increases
■ 50G Standardization
■ 50G vs. 100G per lane debate
■ 50G Electrical Technology
■ 50G Optical Technology
■ Rate Roadmap
■ Final Caution
Caveat

- This presentation contains many predictions based on rigorous analysis pictured in the adjacent photo.

- Predictions are mostly linear extrapolations of the past.
- Innovation is rarely linear.
- Averaging many predictions is conventional wisdom.
- Innovation is rarely conventional wisdom.
- Most predictions go wrong after one technology generation.
- Listener beware!
Historical Rate Increases

“400 Gigabit Ethernet Call For Interest Consensus”, IEEE 802.3 Plenary Meeting, 19 March 2013, Orlando, FL
Historical & Recent Rate Increases

- Mainstream technology per lane rates Gb/s:
  \[ 1 \rightarrow 2.5 \rightarrow 5 \rightarrow 10 \rightarrow 25 \rightarrow 50 \]

- 802.3 Ethernet Router & Switch MAC rates Gb/s:
  \[ 1 \rightarrow 10 \rightarrow 100 \rightarrow 40 \rightarrow 400 \]

- 802.3 Ethernet Server & Switch MAC rates Gb/s:
  \[ 1 \rightarrow 10 \rightarrow 40 \rightarrow 25 \rightarrow 2.5 \rightarrow 5 \]

- 25 Gigabit Ethernet Consortium MAC rates Gb/s:
  \[ 25 \rightarrow 50 \]

- Observations:
  - Technology per lane rates increase in ~2x steps
  - Ethernet rates both drive and follow technology rates

- Conclusion: There will be 50 Gb/s Ethernet
50G Technology Standardization

- The OIF has been working for two years on 50Gb/s per lane electrical specs in the CEI-56G Project
  - Specify chip-to-module (VSR), chip-to-chip (MR), and other shorter reach electrical interfaces of 1 to N lanes
  - Specify per lane speeds from 39 to 56 Gb/s
- IEEE 802.3bs 400Gb/s Ethernet Task Force is investigating 8x50G electrical & optical interface specs
  - VSR & MR electrical interfaces
  - 500m, 2km, 10km SMF optical interfaces
  - Proposal to specify 50G electrical and optical lanes:
    - Support Nx50G optical and electrical lane configurations
      \( N = 1, 2, 4, 8, 16 \)
    - Operate down to 40G

“50Gb/s Per Lane Specification Considerations”, Chris Cole, IEEE 802.3 Plenary Meeting, 3 - 6 Nov. 2014, San Antonio, TX
50G Ethernet Standardization

- Straw Poll taken during the IEEE 802.3bs 400Gb/s Task Force meeting, 6 Nov, 2014, San Antonio, TX
  Should 802.3 standardize 50Gb/s Ethernet?
  - No: 0
  - Yes, as part of 802.3bs Project: 8
  - Yes, in a separate, new 802.3 Project: 86
  - No opinion at this time, don’t care, abstain: 20
  (Unapproved, unofficial results)

- As a result, 802.3bs will take into consideration writing 50G per lane specifications to support Nx50G configurations

- It is likely that a new 50G Ethernet Project will be started when 25G Ethernet Project starts winding down
## 50G vs. 100G Debate Relevance to Servers

Currently 100G per lane is only relevant to switch & router I/O

<table>
<thead>
<tr>
<th>Applications (** Areas of Debate)</th>
<th>50G per lane</th>
<th>100G per lane</th>
</tr>
</thead>
<tbody>
<tr>
<td>400G 2/10km &amp; High-Loss duplex SMF***</td>
<td>Yes</td>
<td>TBD</td>
</tr>
<tr>
<td>400G 500m PSM4***</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Next Gen 400G MMF (after 16x25G)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Next Gen 100G SMF (after 4x25G)***</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Next Gen 100G MMF (after 4x25G)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Next Gen Serial SMF &amp; MMF (after 25G)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Next Gen 40G SMF &amp; MMF (after 4x10G)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Next Gen Fibre Channel (after 32x)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Next Gen ASIC &amp; Module I/O (after 25G)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Next Gen Backplane &amp; Cu Cable (after 25G)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Next Gen BASE-T (after 25G &amp; 40G)</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>
50Gb/s NRZ SerDes RX Data: Electrical Alt1

<table>
<thead>
<tr>
<th>Speed</th>
<th>Pattern</th>
<th>Error Rate</th>
<th>Trace Length</th>
<th>Eye Margin</th>
</tr>
</thead>
<tbody>
<tr>
<td>50G</td>
<td>PRBS</td>
<td>&lt;10^{-15}</td>
<td>HCB-MCB (-10.38dB)</td>
<td>187.5mV</td>
</tr>
</tbody>
</table>

“56G NRZ Measured Test Results (in support of Chip to Module and Chip to Chip Interfaces)”, Haoli Qian (Credo), IEEE 802.3 Plenary Meeting, 3 - 6 Nov. 2014, TX
50Gb/s NRZ SerDes RX Data: Electrical Alt1

<table>
<thead>
<tr>
<th>Speed</th>
<th>Pattern</th>
<th>Error Rate</th>
<th>Trace Length</th>
<th>Eye Margin</th>
</tr>
</thead>
<tbody>
<tr>
<td>50G</td>
<td>PRBS31</td>
<td>~1e-10</td>
<td>13.1in (-31.41dB)</td>
<td>65.63mV</td>
</tr>
</tbody>
</table>

“56G NRZ Measured Test Results (in support of Chip to Module and Chip to Chip Interfaces)”, Haoli Qian (Credo), IEEE 802.3 Plenary Meeting, 3 - 6 Nov. 2014, San Antonio, TX
50Gb/s PAM-4 SerDes TX Data: Electrical Alt2

“CDAUI-8 Chip-to-Chip and Chip-to-Module Interfaces using PAM4”, Adam Healey (Avago) et. al, IEEE 802.3 Plenary Meeting, 3 - 6 Nov. 2014, San Antonio, TX
50Gb/s PAM-4 SerDes RX Data: Electrical Alt2

PRBS31
25 GBd

Packaged
PAM4 TX

Eval.
board

Cables

zQSFP+
MCB/HCB

Cables

Eval.
board

Packaged
PAM4 RX

Differential insertion loss ~ 11.7 dB at 12.5 GHz

RX internal eye plot

BER ~ 1.64E-11

“CDAUI-8 Chip-to-Chip and Chip-to-Module Interfaces using PAM4”, Adam Healey (Avago) et. al, IEEE 802.3 Plenary Meeting, 3 - 6 Nov. 2014, San Antonio, TX
4x50G NRZ Transceiver (200G): Optical Alt1

- 4x50G (200G) Transceiver (ex. for TOR): CFP4 or QSFP
- 1x50G (single lane) Transceiver (ex. for NIC): SFP
Functionality in dashed lines on the previous page

Finisar 2x50G hybrid SiP PIC fabricated in ST Microelectronics BiCOMS
50G NRZ SiP PIC TX Data: Optical Alt1

40Gb/s, PRBS9 TX optical eye diagram at $\pi/2$ bias:
- Measurement data,
- Simulation

56Gb/s, PRBS9 TX optical eye diagram at $\pi/2$ bias:
- Measurement data,
- Simulation
50G PAM-4 SiP PIC TX Analysis: Optical Alt2

56Gb/s (28GBaud) simulated TX optical eye diagram
Ex. Forecast: Cloud Server Speed Increase

“25GBASE-T Call For Interest Consensus”, David Chalupksy (Intel) et. al, IEEE 802.3 Plenary Meeting, 3 - 6 Nov. 2014, San Antonio, TX
Ex. Forecast: Enterprise Server Speed Increase

“25GBASE-T Call For Interest Consensus”, David Chalupksy (Intel) et. al, IEEE 802.3 Plenary Meeting, 3 - 6 Nov. 2014, San Antonio, TX
40G vs. 50G Deployment

- 40G initial limited server deployment as 4x10G
- 40G may have limited deployment as proprietary 2x20G
- 50G initial limited deployment as proprietary 2x25G
- 40G & 50G per lane (Serial) technology will be defined together (40G as reduced speed 50G)
- 40G & 50G Serial will have similar cost, i.e. 50G Serial will offer 25% more bandwidth for the same cost
- 50G Serial volume will quickly surpass combined 40G Serial and 40G 4x10G volume
- 40G combined volume will quickly plateau and decline (sorry Del’Oro 😊)
What Happened?

- 100G Ethernet was first specified for core networking applications using forward looking 25G technology.
- 40G Ethernet was then added for cost sensitive Switch and Router applications using existing 10G technology.
- 25G single lane technology is now more cost effective than 40G 4x10G technology.
- There would be no 25G & 50G Ethernet if 80G using 20G technology was standardized instead of 100G:
  10 → 80 → 40 → 20
- Lesson learned:
  - No more 10x rate increases, i.e. no 1T Ethernet.
  - Predicted future MAC rates Gb/s:
    100 → 200 → 400 → 800 → 1600
# Mainstream Server Rate Roadmap

<table>
<thead>
<tr>
<th>Rate Gb/s</th>
<th>25</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethernet Standard Status</td>
<td>In Definition</td>
<td>To be defined</td>
<td>Done</td>
<td>To be defined</td>
<td>In Definition</td>
</tr>
<tr>
<td>Single lane technology</td>
<td>Exists</td>
<td>In R&amp;D</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Deployment if 2x rate / 2 years I/O increase</td>
<td>2015</td>
<td>2017</td>
<td>2019</td>
<td>2022</td>
<td>2025</td>
</tr>
<tr>
<td>Deployment if 2x rate / 3 years I/O increase</td>
<td>2015</td>
<td>2018</td>
<td>2021</td>
<td>2024</td>
<td>2027</td>
</tr>
</tbody>
</table>

Please review caveat on page 2
Rate Increase Prediction Caution

- I/O 2x rate increases are predicted to replace 10x & 4x
- Flex Ethernet (also FlexMAC) is a cautionary example:
  - Efficient Lane bonding (LAG evolution)
    - LAG is ~80% efficient (traffic dependent)
    - LAG is difficult to manage
  - Generalized channelization & sub-rate aggregation (MLG evolution)
    - MLG is only defined for 10GbE & 40GbE channels
    - MLG has to be re-defined for each new PMD
- Today Flex Ethernet is only focused on connecting Routers, directly or over a Transport Network
- May find broader use tomorrow

“FlexEthernet”, David Offelt (Juniper) et. al, 4 Nov. 2014, Flash Mob meeting coincident w/ IEEE 802.3 Plenary Meeting, San Antonio, TX
50Gb/s Per Lane: Next Gen. of Server I/O

Thank you