40GBASE-KR4 backplane PHY proposal

Richard Mellitz & Ilango Ganga
Intel Corporation

Mar 18, 2008
Contributors & Supporters

- Andre Szczepanek, Texas Instruments
- Arthur Marris, Cadence Design Systems
- Pravin Patel, IBM
- Chris DiMinico, MC Communications
- Scott Kipp, Brocade
- Tom Palkert, Luxtera
- Jeff Cain, Cisco Systems
Key messages

- Proposal to adopt 10GBASE-KR as a baseline for specifying 40GBASE-KR4 with the following changes
  - Backplane layer diagram (Clause 69)
  - Leverage 10GBASE-KR PMD for operation over 4 lanes (Clause 72)
  - Auto-Negotiation (Clause 73)
  - Forward Error correction (Clause 74)
Considerations for 40G BPE PHY

- To be architecturally consistent with the Backplane Ethernet layer stack illustrated in Clause 69
- To interface to a 4-lane backplane medium with interconnect characteristics recommended in IEEE Std 802.3ap (Annex 69B)
  - Most generation 2 blade systems are built with 4-lanes (10Gbaud KR ready)
- Leverage 10GBASE-KR technology/specifications (Clause 72 and Annex 69A) to define 40GBASE-KR4 PHY:
  - 64B/66B block coding
  - Startup protocol (per lane)
  - Signaling speed 10.3125Gbd (per lane)
  - Electrical characteristics
  - Test methodology and procedures
- Optional FEC sublayer
  - PCS to interface to optional FEC sublayer consistent with Clause 74 specification
- Compatible with Backplane Ethernet Auto-Neg (Clause 73)
  - Enhancement to indicate 40GbE ability
Backplane Ethernet overview

- IEEE Std 802.3ap-2007 Backplane Ethernet defines 3 PHY types
  - 1000BASE-KX: 1-lane 1 Gb/s PHY (Clause 70)
  - 10GBASE-KX4: 4-lane 10Gb/s PHY (Clause 71)
  - 10GBASE-KR: 1-lane 10Gb/s PHY (Clause 72)
- Forward Error Correction (FEC) for 10GBASE-R (Clause 74) – optional
  - Optional FEC to increase link budget and BER performance
- Auto-negotiation (Clause 73)
  - Auto-Neg between 3 PHY types (AN is mandatory to implement)
  - Parallel detection for legacy PHY support
    - Automatic speed detection of legacy 1G/10G backplane SERDES devices
  - Negotiate FEC capability
- Clause 45 MDIO interface for management
- Channel
  - Controlled impedance (100 Ohm) traces on a PCB with 2 connectors and total length up to at least 1m.
  - Channel model is informative (Annex 69B)
- Interference tolerance testing (Annex 69A)
- Support a BER of $10^{-12}$ or better
Existing backplane architecture

Figure 69-1—Architectural positioning of Backplane Ethernet

AN = AUTO-Negotiation
GMI = GIGABIT MEDIA INDEPENDENT INTERFACE
MDI = MEDIUM DEPENDENT INTERFACE
PCS = PHYSICAL CODING SUBLAYER
FEC = FORWARD ERROR CORRECTION
PHY = PHYSICAL LAYER DEVICE
PMA = PHYSICAL MEDIUM ATTACHMENT
PMD = PHYSICAL MEDIUM DEPENDENT
XGMI = 10 GIGABIT MEDIA INDEPENDENT INTERFACE
Proposed backplane architecture with 40GbE

Figure 69-1 Architectural positioning of Backplane Ethernet
Proposed Auto-Neg changes

- IEEE Std 802.3ap defines Auto-Negotiation for backplane Ethernet PHYs
  - AN uses DME signaling with 48-bit base pages to exchange link partner abilities
  - AN is mandatory for 10GBASE-KR backplane PHY, negotiates FEC ability

- Proposal for 40GBASE-KR4 (Ability to negotiate with other 802.3ap PHYs)
  - Add a Technology Ability bit A3 to indicate 40GbE ability (A3 is currently reserved)
  - No changes to backplane AN protocol or management register format
  - No change to negotiate FEC ability, FEC when selected to be enabled on all 4 lanes
  - AN mandatory for 40GBASE-KR4, no parallel detect required for 40G

<table>
<thead>
<tr>
<th>Bit</th>
<th>Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>A0</td>
<td>1000BASE-KX</td>
</tr>
<tr>
<td>A1</td>
<td>10GBASE-KX4</td>
</tr>
<tr>
<td>A2</td>
<td>10GBASE-KR</td>
</tr>
<tr>
<td>A3</td>
<td>40GBASE-KR4</td>
</tr>
<tr>
<td>A4 to A24</td>
<td>Reserved for future technology</td>
</tr>
</tbody>
</table>

Table 73-4—Technology Ability field encoding
Proposed 40GBASE-KR4 PMD

- Leverage 10GBASE-KR (Clause 72) to specify 40GBASE-KR4 with following changes for 4 lane operation
  - Change KR Link diagram for 4 lanes (similar to KX4)
  - Change KR PMD service interface to support 4 logical streams (similar to KX4)
  - Change PMD control variable mapping table to include management variables for 4 lanes
40GBASE-KR4 Link block diagram
Service Interfaces for KR4 PMD

- PMD Service Interface
  - Service interface definition as in Clause 72
  - Specify 4 logical streams of 64B/66B code groups from PMA
    - PMD_UNITDATA.request (txbit<0:3>)
    - PMD_UNITDATA.indication (rxbit<0:3>)
    - PMD_SIGNAL.indication (SIGNAL_DETECT<0:3>)
      - "While normally intended to be an indicator of signal presence, is used by 10GBASE-KR to indicate the successful completion of the start-up protocol". Enumerate for 4 lanes

- AN Service Interface (Same as Clause 73)
  - Support AN_LINK.indication primitive
  - Requires associated PCS to support this primitive
PMD MDIO function mapping (1)

- Support management variables for 4 lanes
- Include lane by lane Transmit disable

### Table 71-2: MDIO/PMD control variable mapping

<table>
<thead>
<tr>
<th>MDIO control variable</th>
<th>PMA/PMD register name</th>
<th>Register/ bit number</th>
<th>PMD control variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reset</td>
<td>Control register 1</td>
<td>1.0.15</td>
<td>PMD_reset</td>
</tr>
<tr>
<td>Global Transmit Disable</td>
<td>Transmit disable register</td>
<td>1.9.0</td>
<td>Global_PMD_transmit_disable</td>
</tr>
<tr>
<td>Transmit disable 3</td>
<td>Transmit disable register</td>
<td>1.9.4</td>
<td>PMD_transmit_disable_3</td>
</tr>
<tr>
<td>Transmit disable 2</td>
<td>Transmit disable register</td>
<td>1.9.2</td>
<td>PMD_transmit_disable_2</td>
</tr>
<tr>
<td>Transmit disable 1</td>
<td>Transmit disable register</td>
<td>1.9.2</td>
<td>PMD_transmit_disable_1</td>
</tr>
<tr>
<td>Transmit disable 0</td>
<td>Transmit disable register</td>
<td>1.9.1</td>
<td>PMD_transmit_disable_0</td>
</tr>
<tr>
<td>Restart training</td>
<td>10GBASE-KR PMD control register</td>
<td>1.150.0</td>
<td>mr_restart_training</td>
</tr>
<tr>
<td>Training enable</td>
<td>10GBASE-KR PMD control register</td>
<td>1.150.1</td>
<td>mr_training_enable</td>
</tr>
</tbody>
</table>
PMD MDIO function mapping (2)

- Support management variables for 4 lanes
  - Add lane by lane signal detect
  - Enumerate status indication per lane as appropriate

<table>
<thead>
<tr>
<th>MDIO status variable</th>
<th>PMA/PMD register name</th>
<th>Register/ bit number</th>
<th>PMD status variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fault</td>
<td>Status register 1</td>
<td>1.1.7</td>
<td>PMD_fault</td>
</tr>
<tr>
<td>Transmit fault</td>
<td>Status register 2</td>
<td>1.0.11</td>
<td>PMD_transmit_fault</td>
</tr>
<tr>
<td>Receive fault</td>
<td>Status register 2</td>
<td>1.8.16</td>
<td>PMD_receive_fault</td>
</tr>
<tr>
<td>Global PMD Receive signal detect</td>
<td>Receive signal detect register</td>
<td>1.10.0</td>
<td>Global_PMD_signal_detect</td>
</tr>
<tr>
<td>PMD signal detect 3</td>
<td>Receive signal detect register</td>
<td>1.10.4</td>
<td>PMD_signal_detect_3</td>
</tr>
<tr>
<td>PMD signal detect 2</td>
<td>Receive signal detect register</td>
<td>1.10.3</td>
<td>PMD_signal_detect_2</td>
</tr>
<tr>
<td>PMD signal detect 1</td>
<td>Receive signal detect register</td>
<td>1.10.2</td>
<td>PMD_signal_detect_1</td>
</tr>
<tr>
<td>PMD signal detect 0</td>
<td>Receive signal detect register</td>
<td>1.10.1</td>
<td>PMD_signal_detect_0</td>
</tr>
<tr>
<td>Receiver status</td>
<td>10GBASE-KR PMD status register</td>
<td>1.451.0</td>
<td>rx_trained</td>
</tr>
<tr>
<td>Frame lock</td>
<td>10GBASE-KR PMD status register</td>
<td>1.451.1</td>
<td>frame_lock</td>
</tr>
<tr>
<td>Start-up protocol status</td>
<td>10GBASE-KR PMD status register</td>
<td>1.451.2</td>
<td>training</td>
</tr>
<tr>
<td>Training failure</td>
<td>10GBASE-KR PMD status register</td>
<td>1.451.3</td>
<td>training_failure</td>
</tr>
</tbody>
</table>
KR4 PMD transmit & receive functions

- PMD transmit function (enumerate for 4 lanes)
  - Converts 4 logical streams from PMD service interface into 4 separate electrical streams delivered to MDI
  - Separate lane by lane TX disable function in addition to Global TX disable function

- PMD receive function (enumerate for 4 lanes)
  - Converts 4 separate electrical streams from MDI into 4 logical streams to PMD service interface
  - Separate lane by lane signal detect function in addition to Global TX disable function

- Same electrical specifications as defined in Clause 72 for 10GBASE-KR PMD
  - Receiver Compliance defined in Annex 69A (Interference Tolerance Test) and referenced in Clause 72
PMD Control function

Startup & Training

- Reuse Clause 72 control function for KR4 PMD (Startup & Training)
  - Used for tuning equalizer settings for optimum backplane performance
  - Use Clause 72 training frame structure
  - Use same PRBS 11 pattern, with randomness between lanes
- Same Control channel spec as in Clause 72, enumerated per lane
  - All 4 lanes are independently trained
  - Report Global Training complete only when all 4 lanes are trained
  - Same Frame lock state diagram (Fig 72-4)
  - Same Training state diagram with enumeration of variables corresponding to 4 lanes (Fig 72-5)
Electrical characteristics

- **40GBASE-KR4 Transmit electrical characteristics**
  - Same as 10GBASE-KR TX characteristics and waveforms as specified in Clause 72
  - Same test fixture setup as in Clause 72

- **40GBASE-KR4 Receiver electrical characteristics**
  - Same as 10GBASE-KR RX characteristics specified in Clause 72 and Annex 69 A
**Receiver Interference tolerance test**

- Test procedure specified in Annex 69A
- Receiver interference tolerance parameters for 40GBASE-KR4 PMD
  - Same as Receiver interference tolerance test parameters as in Clause 72
  - No change to broadband noise amplitude for KR4
Forward Error Correction

- Reuse FEC specification for 10GBASE-R (Clause 74)
  - The FEC sublayer transparently passes 64B/66B code blocks
  - Change to accommodate FEC sync for 4 lanes
    - Same state diagram for FEC block lock
    - Report Global Sync achieved only if all lanes are locked
    - Possibly add a FEC frame marker signal that could be used for lane alignment
Interconnect Characteristics

- Interconnect characteristics (informative) for backplane is defined in Annex 69B
  - No proposed changes
- 40GBASE-KR4 PHY to interface to the 4 lane backplane medium to take advantage of 802.3ap KR ready blade systems in deployment
Summary

- 40GbE backplane PHY to be architecturally consistent with IEEE Std 802.3ap layer stack
- Adopt 10GBASE-KR as baseline to Specify 40GBASE-KR4 PHY with appropriate changes proposed in this document
- Interface to 4 lane backplane medium to take advantage of 802.3ap KR ready blade systems in deployment
- Appropriate changes to add EEE feature, when adopted by 802.3az for KR
- PCS proposals and interface definitions to accommodate backplane Ethernet architecture (including FEC and AN)
Backup
Typical backplane system illustration

Note: The switch cards are shown at the chassis edge for simplicity.
In real systems there could be multiple fabrics located at the center, edge, or rear of the chassis.
Proposed 40GbE interfaces

- **Inter-sublayer interfaces** common to 40/100G where possible
  - XLGMII (intra-chip)
    - Logical; define data, clock, no electricals
  - PCS
    - 64B/66B encoding
    - Fixed 4-lane distributor based on 64B/66B header for alignment
    - Used for backplane & XLAUI
  - XLAUI (chip to chip)
    - Define electricals
  - FEC service interface
    - Abstract, can optionally use XLAUI
  - PMA Service interface
    - Define electricals, same as XLAUI
  - PMD Service interface
    - Abstract, or logical

- Optional overhead can be used with Optical PHYs
- Auto-Neg for Backplane located below PMD

- RS, MAC AND HIGHER LAYERS
- XLGMII (logical)
- PCS 64B/66B
- 4 Lane Distributor
- Optional overhead can be used with Optical PHYs
- FEC Service Interface (Abstract or XLAUI)
- XLAUI
- Define electricals
- PMA Service Interface (electrical, XLAUI)
- PMD Service Interface (abstract or logical)
- MDI
- MEDIUM
- “40GBASE-R4” or 100GBASE-???
Possible implementation examples

40GbE separate Backplane PHY

MAC  PCS 64B/66B

FEC  x4  PMA  x4  PMD Serial 4xKR  AN

XLAUI

MAC-PHY interface (Chip to Chip)

40GbE integrated backplane PHY

MAC  PCS 64B/66B

FEC  x4  PMA  x4  PMD Serial 4xKR  AN

XLGMII

MAC-PHY interface (intra-chip)