40GBASE-KR4 backplane PHY proposal and Next Steps

Richard Mellitz & Ilango Ganga
Intel Corporation

May 13, 2008

Supporters

- Andre Szczepanek, Texas Instruments
- Arne Alping, Ericsson
- Arthur Marris, Cadence Design Systems
- Brad Booth, AMCC
- David Koenen, HP
- Frank Chang, Vitesse
- Gourgen Oganessyan, Quellan
- Jeff Lynch, IBM
- Scott Kipp, Brocade
- Tom Palkert, Luxtera

Supporters for mellitz_01_0308

- Jeff Cain, Cisco Systems
- Chris DiMinico, MC Communications
- Pravin Patel, IBM

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 Backplane Ethernet 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

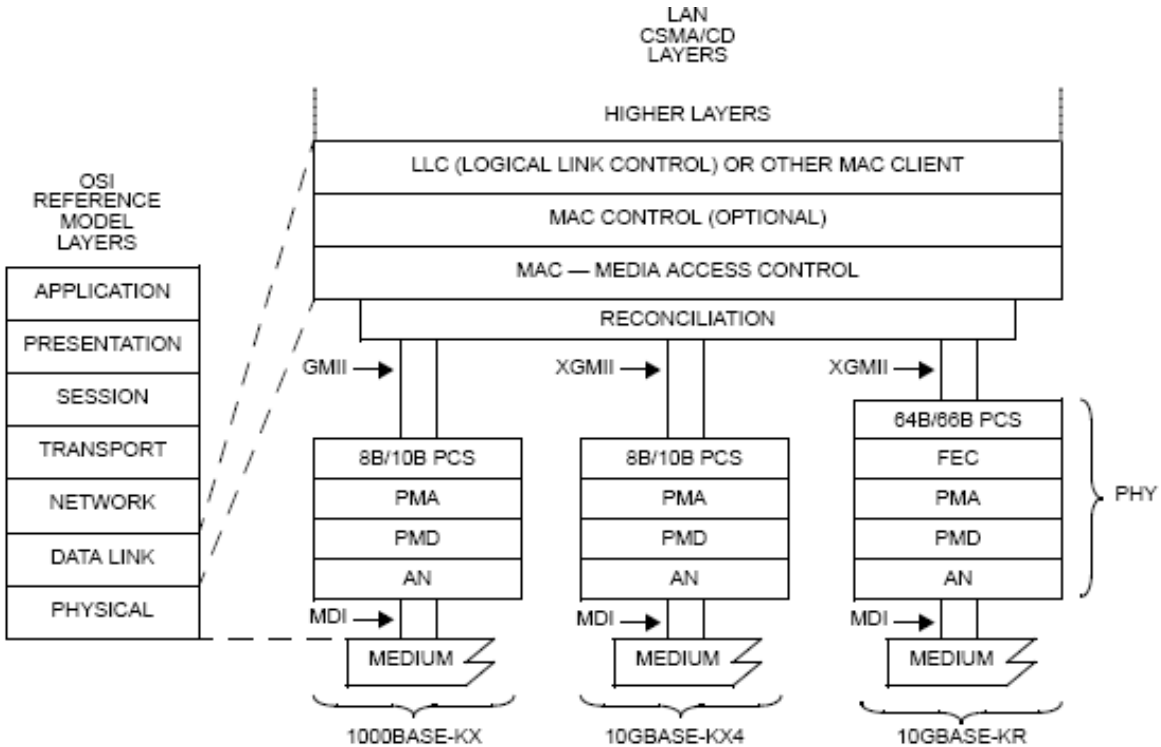
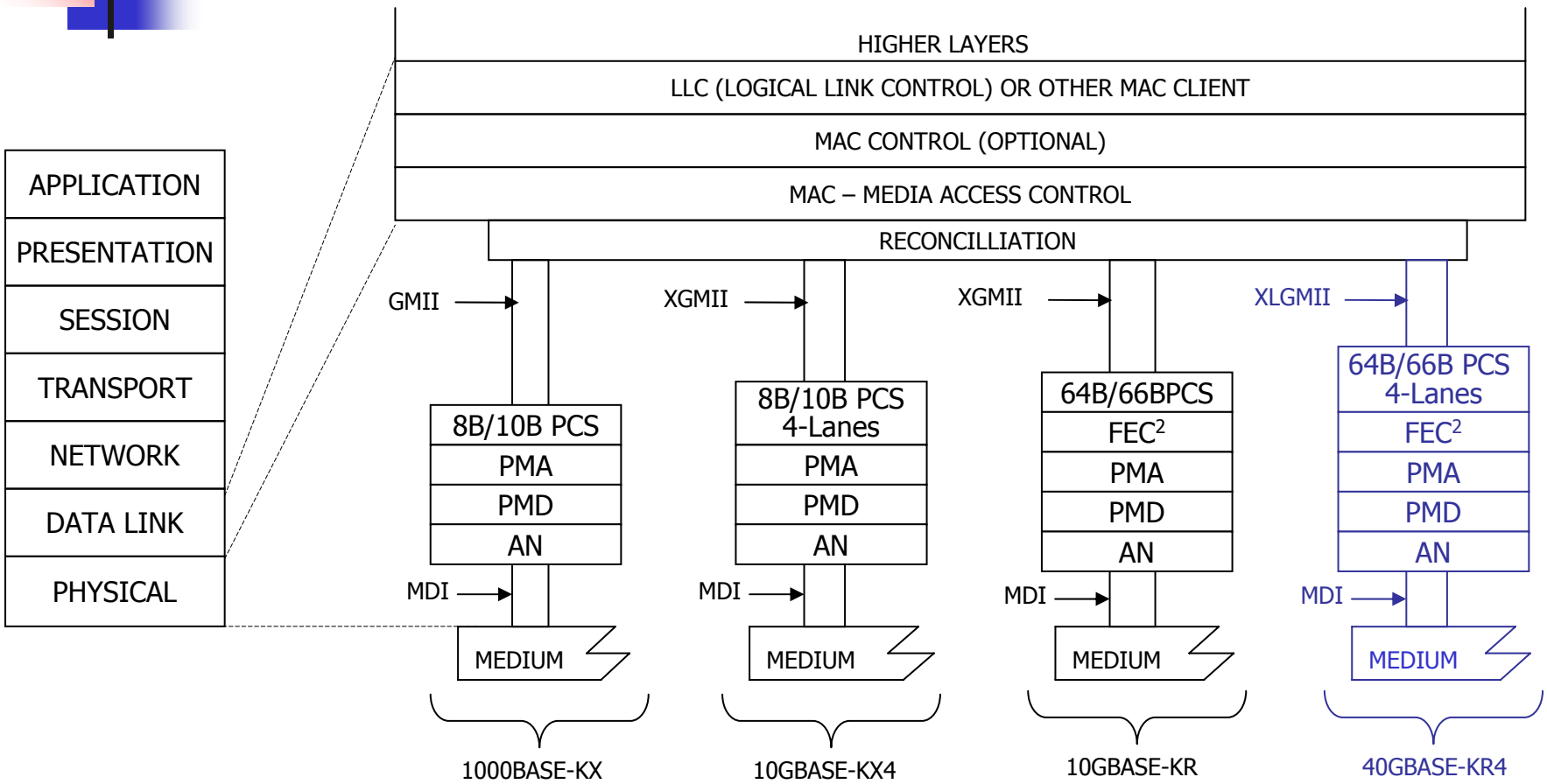


Figure 69–1—Architectural positioning of Backplane Ethernet

Proposed backplane architecture with 40GbE



Note: 2. Optional

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
  - Lane 0 of the MDI is used for Auto-Negotiation, of single or multi-lane PHYs
- 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 73–4—Technology Ability field encoding

Bit	Technology
A0	1000BASE-KX
A1	10GBASE-KX4
A2	10GBASE-KR
~~A3 through A24~~	~~Reserved for future technology~~
A3	**40GBASE-KR4**
A4 through A24	**Reserved for future technology**

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

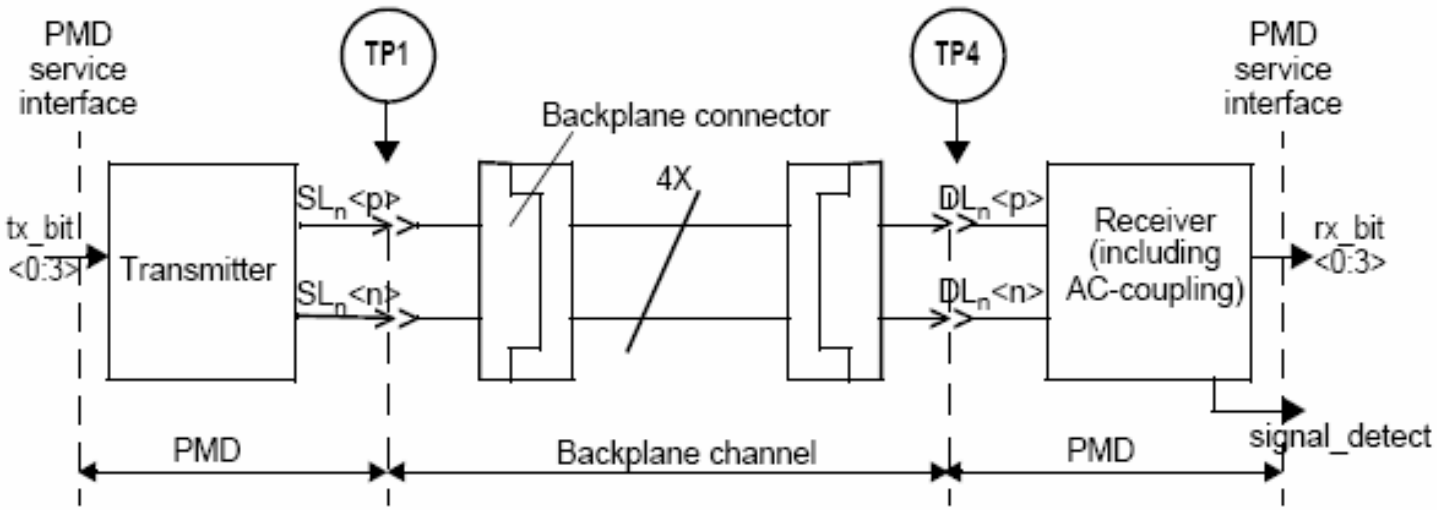


Figure ~~71–1~~—Link block diagram

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

MDIO control variable	PMA/PMD register name	Register/ bit number	PMD control variable
Reset	Control register 1	~~1.0.15~~	PMD_reset
Global Transmit Disable	Transmit disable register	~~1.9.0~~	Global_PMD_transmit_disable
Transmit disable 3	Transmit disable register	~~1.9.4~~	PMD_transmit_disable_3
Transmit disable 2	Transmit disable register	~~1.9.3~~	PMD_transmit_disable_2
Transmit disable 1	Transmit disable register	~~1.9.2~~	PMD_transmit_disable_1
Transmit disable 0	Transmit disable register	~~1.9.1~~	PMD_transmit_disable_0
Restart training	~~10GBASE-KR~~ PMD control register	~~1.150.0~~	mr_restart_training
Training enable	~~10GBASE-KR~~ PMD control register	~~1.150.1~~	mr_training_enable

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 ~~71–3~~—MDIO/PMD status variable mapping

MDIO status variable	PMA/PMD register name	Register/ bit number	PMD status variable
Fault	Status register 1	~~1.1.7~~	PMD_fault
Transmit fault	Status register 2	~~1.8.11~~	PMD_transmit_fault
Receive fault	Status register 2	~~1.8.10~~	PMD_receive_fault
Global PMD Receive signal detect	Receive signal detect register	~~1.10.0~~	Global_PMD_signal_detect
PMD signal detect 3	Receive signal detect register	~~1.10.4~~	PMD_signal_detect_3
PMD signal detect 2	Receive signal detect register	~~1.10.3~~	PMD_signal_detect_2
PMD signal detect 1	Receive signal detect register	~~1.10.2~~	PMD_signal_detect_1
PMD signal detect 0	Receive signal detect register	~~1.10.1~~	PMD_signal_detect_0
Receiver status	~~10GBASE-KR~~ PMD status register	~~1.151.0~~	rx_trained
Frame lock	~~10GBASE-KR~~ PMD status register	~~1.151.1~~	frame_lock
Start-up protocol status	~~10GBASE-KR~~ PMD status register	~~1.151.2~~	training
Training failure	~~10GBASE-KR~~ PMD status register	~~1.151.3~~	training_failure

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 signal detect 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)
  - Enumerate the management registers for coefficient update field and status report field for 4 lanes

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 69A

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

FEC MDIO variable mapping

Table 74–2—MDIO/FEC variable mapping

MDIO variable	PMA/PMD register name	Register/bit number	FEC variable
10GBASE-R FEC ability	10GBASE-R FEC ability register	1.170.0	FEC_ability
10GBASE-R FEC Error Indication ability	10GBASE-R FEC ability register	1.170.1	FEC_Error_Indication_ability
FEC Enable	10GBASE-R FEC control register	1.171.0	FEC_Enable
FEC Enable Error Indication	10GBASE-R FEC control register	1.171.1	FEC_Enable_Error_to_PCS
FEC corrected blocks	10GBASE-R FEC corrected blocks counter register	1.172, 1.173	FEC_corrected_blocks_counter
FEC uncorrected blocks	10GBASE-R FEC uncorrected blocks counter register	1.174, 1.175	FEC_uncorrected_blocks_counter

- Enumerate the following counters for 4 lanes
  - FEC_corrected_blocks_counter
  - FEC_uncorrected blocks_counter
  - Possibly use indexed addressing to conserve MDIO address space

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

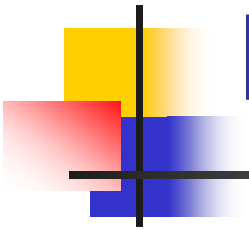
Summary

- 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)

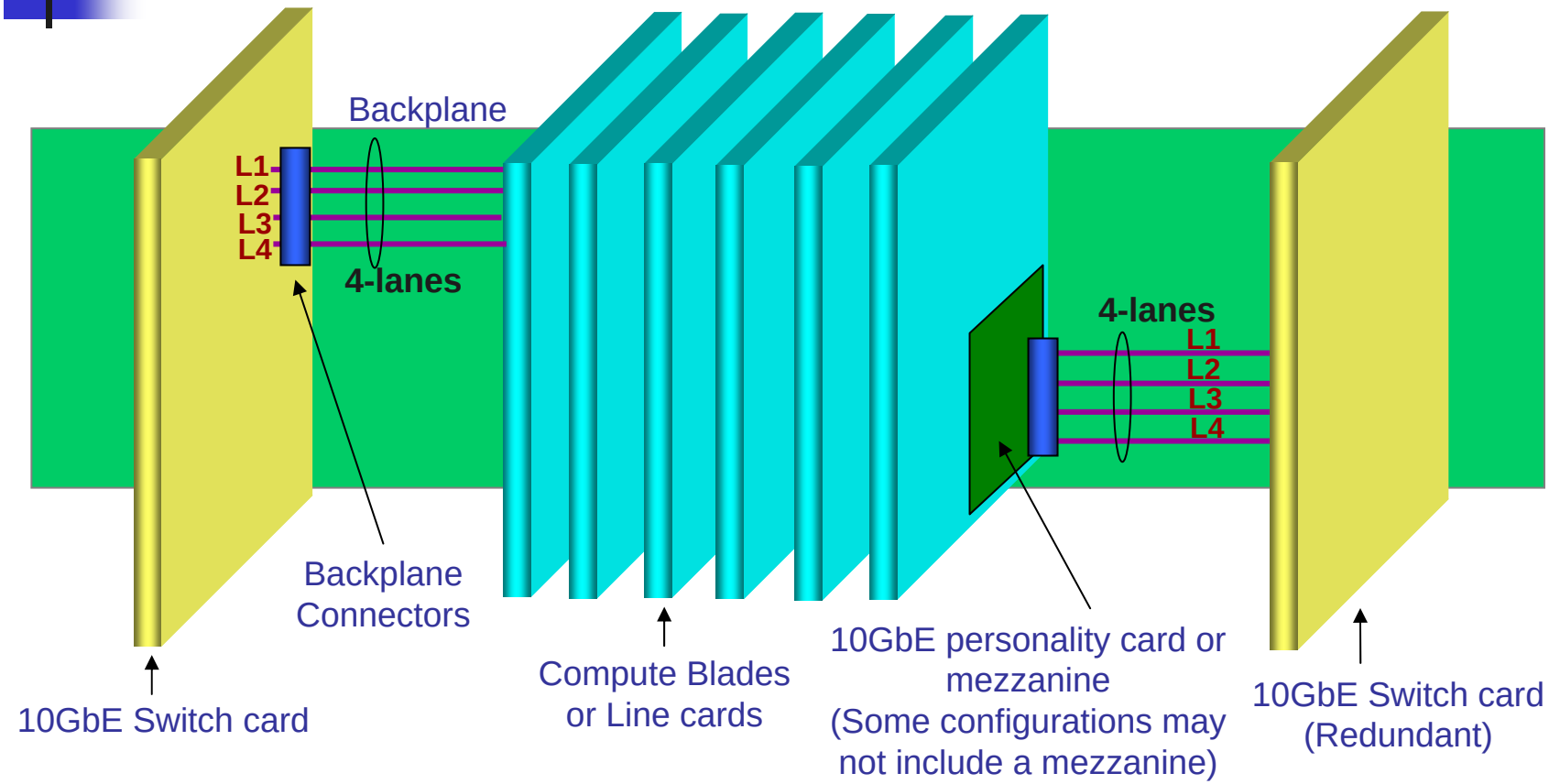
Next Steps

- Make a second generation blade channel model (IEEE Std 802.3ap KR compatible) available to the P802.3ba task force by July '08
- Simulations showing technical feasibility of 40GBASE-KR4 over 40G ready IEEE Std 802.3ap compatible 4 lane backplane system with compliant receivers



Backup

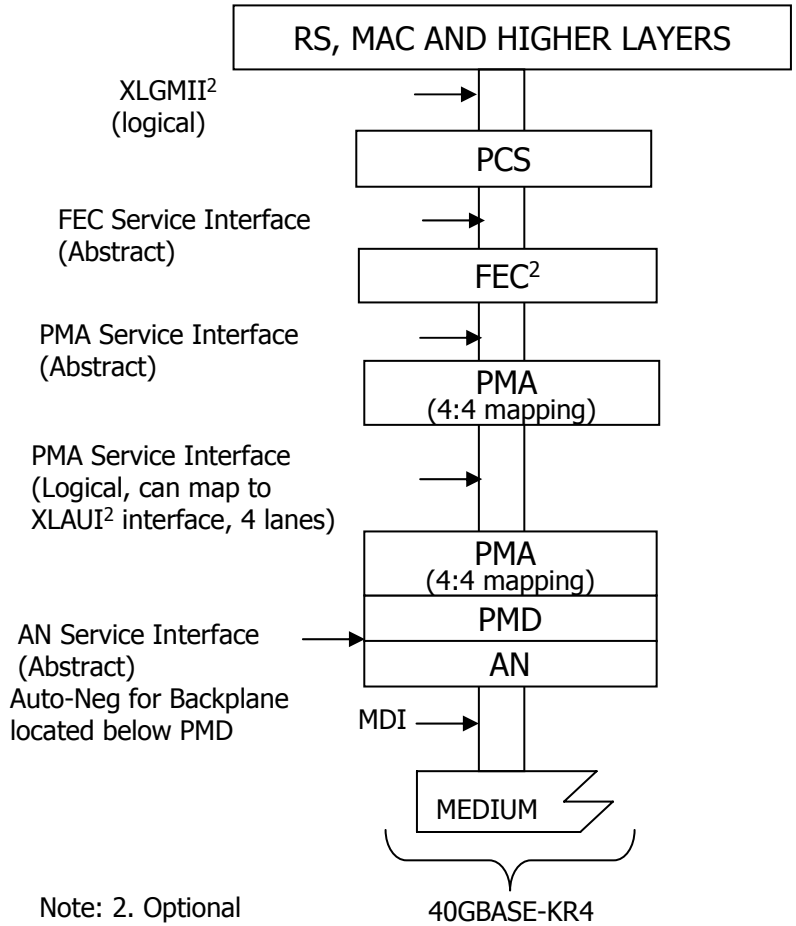
Typical backplane system illustration



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 architecture

- XLGMII (intra-chip)
  - Logical, define data/control, clock, no electrical specification
- PCS
  - 64B/66B encoding
  - Lane distribution and alignment
- XLAUI (chip-to-chip)
  - 10.3125 GBaud electrical interface
  - 4 lanes, short reach
- FEC service interface
  - Abstract, can map to XLAUI electrical interface
- PMA Service interface
  - Logical n lanes, can map to XLAUI electrical interface
- PMD Service interface
  - Logical



See ganga_01_0508 for 40/100G architecture and interfaces

Possible implementation examples