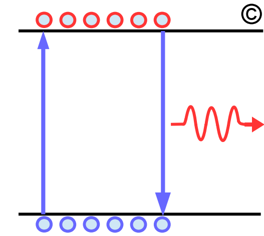


LPPI "The Ultra Low Power C2C Interface"

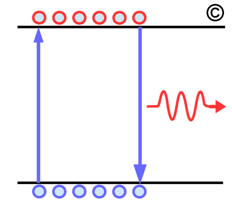
Ali Ghiasi

Ghiasi Quantum LLC

IEEE 802.3bs Plenary Meeting

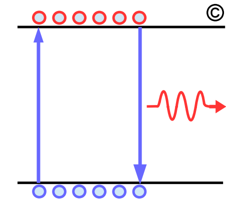
March 11, 2015

List of Supporters and Contributors



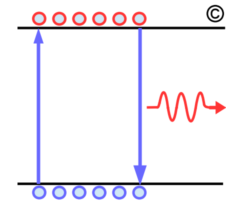
- **Graeme Boyd – PMC**
- **Tom Palkert – Molex**
- **Scott Irwin – MoSys**
- **Jeff Twombly – Credo**
- **Haoli Qian – Credo**
- **Dan Dove – Dove Networking**

Addressing Big Ticketed Items - LPPI (Page 12)



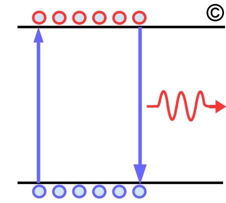
- **Actions:**
  - C2C and C2M interaction across the 3 reaches
  - Refine proposal and build consensus

Overview

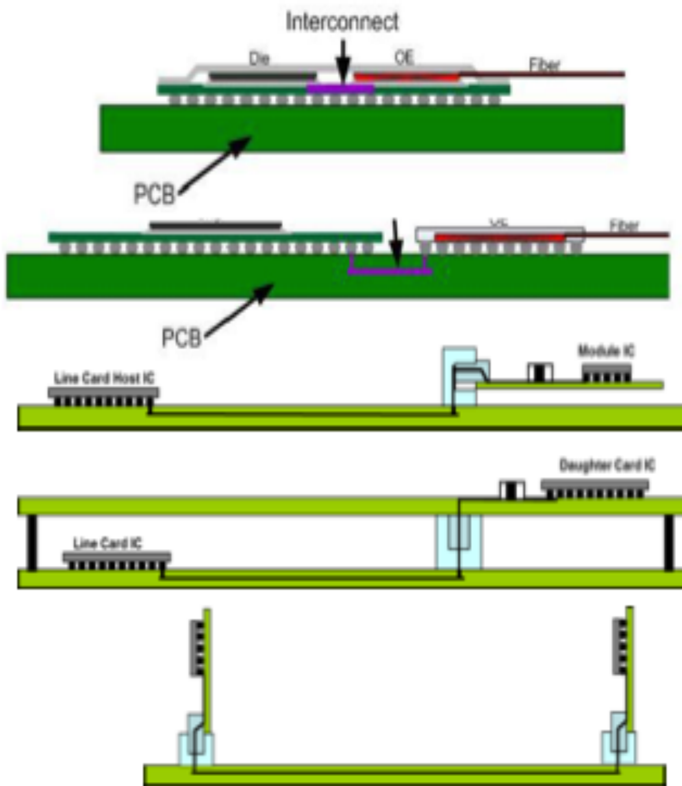


- **Does an OIF XSR like "C2C with 5-7 cm without a connector" interface has distinct identity for Ethernet applications?**
  - A key distinction of XSR interface is that the clock forwarded, DC coupled, and ISI is not a driving factor
  - For details OIF-56G-XSR please see http://www.ieee802.org/3/bs/public/15_03/dawe_3bs_01_0315.pdf
  - Traditionally xPPI designation has been associated with unretime interfaces
  - To avoid confusion with legacy xPPI, suggest to use C2S "chip to shim" naming if 802.3bs decide to standardize this interface
- **High capacity switching ASIC are pushing die area and power dissipation**
  - To integrate 400G-KR8/200G-KR4 on a large ASIC may not be feasible due to area and/or power dissipation
    - An alternative approach is to use an external shim connected with ultra lower power SerDes interface to the ASIC
  - Front panel BW could be limited as result of optical modules
    - An on board PIC connected with ultra low power SerDes interface could provide significantly more BW and lower system PD.

50G Interconnect Space



- **OIF is defining USR, XSR, VSR, MR, and LR but IEEE only defining CDAUI-8 Chip-to-chip and chip to module**
  - Should IEEE define equivalent of XSR?

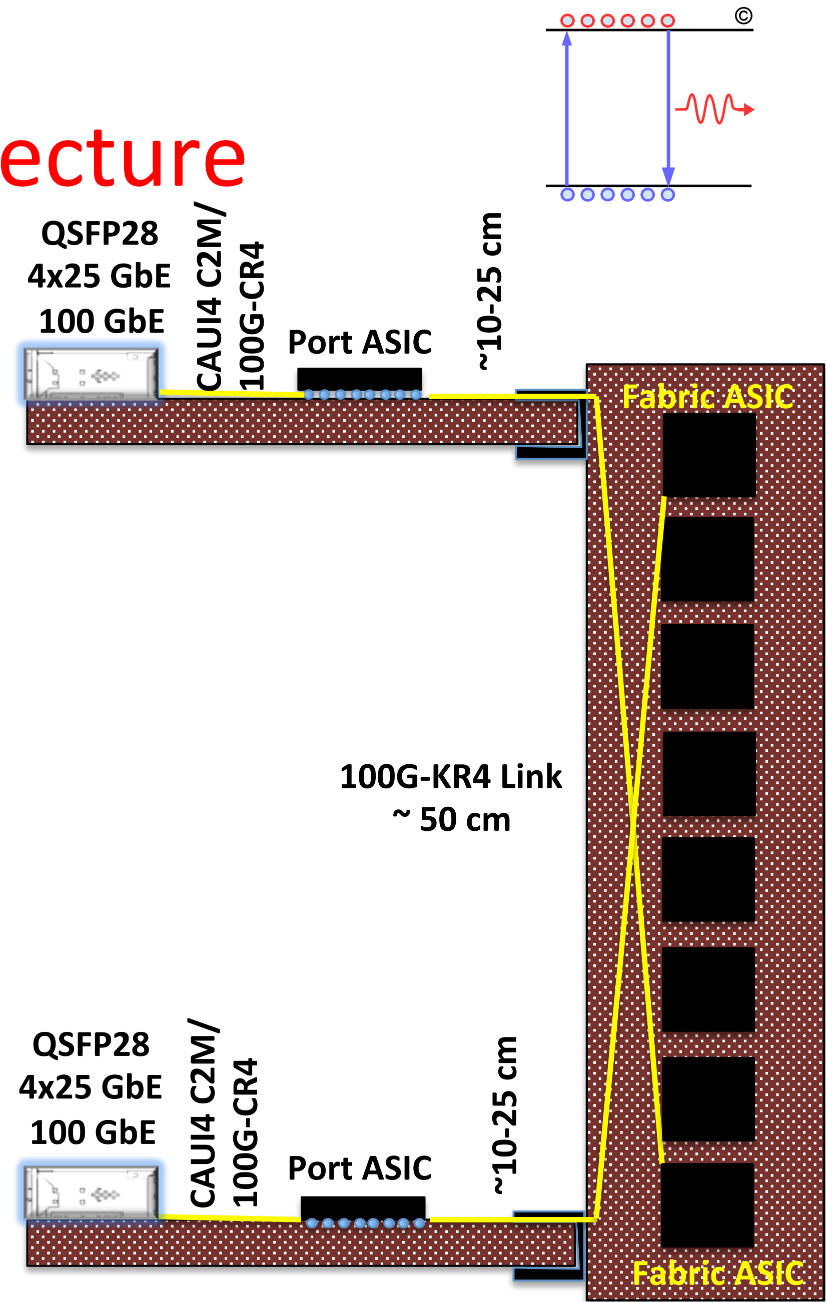


Source: Frlan Designcon 2015

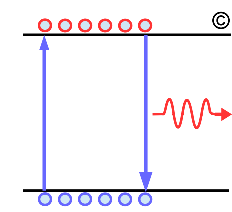
Application	Interface OIF/IEEE	Interconnect	Maximum Loss	Modulation Scheme Under Investigation
Chip-to-OE	CEI-56G-USR	< 1 cm (no connector or package)	1-2 dB @28 GHz	NRZ
Chip-to-nearby OE	CEI-56G-XSR	< 5 cm (no connector)	5-10 dB @28 GHz	NRZ and PAM4
Chip-to-Module	CEI-56G-VSR/ CDAUI-8	<10 cm (one connector)	7-10 dB @ 14 GHz 10-20 dB @ 28 GHz	NRZ and PAM4
Chip-to-chip	CEI-56G-MR/ CDAUI-8	<50 cm (one connector)	15-25 dB @ 14 GHz 20-50 dB @ 28 GHz	NRZ and PAM4
Backplane	CEI-56G-LR/ TBD	<100 cm (2 connectors)	25-40 dB @ 14 GHz	PAM4

2015 Switch Architecture

- 100GBASE-KR4/CR4/CAUI4 integrated into the Port ASIC
  - Dual mode
- Port ASIC allow seamless support of following interfaces with QSFP28
  - 100GBASE-CR4/KR4
  - 100GBASE-SR4
  - 100GBASE-LR4
  - 100G CWDM4/CLR4/PSM4
  - 4x25GBASE-CR/KR
  - 4x25GBASE-SR
- These high capacity systems operate at the limit of cooling capacity
  - System with 8 line cards could deliver 256 port of 100G with QSFP28
  - System with 16 line cards could deliver 512 ports of 100G with QSFP28
- The implementation is made possible as result of shim-less design by integrating 100GBASE-KR4 SerDes into the ASIC.



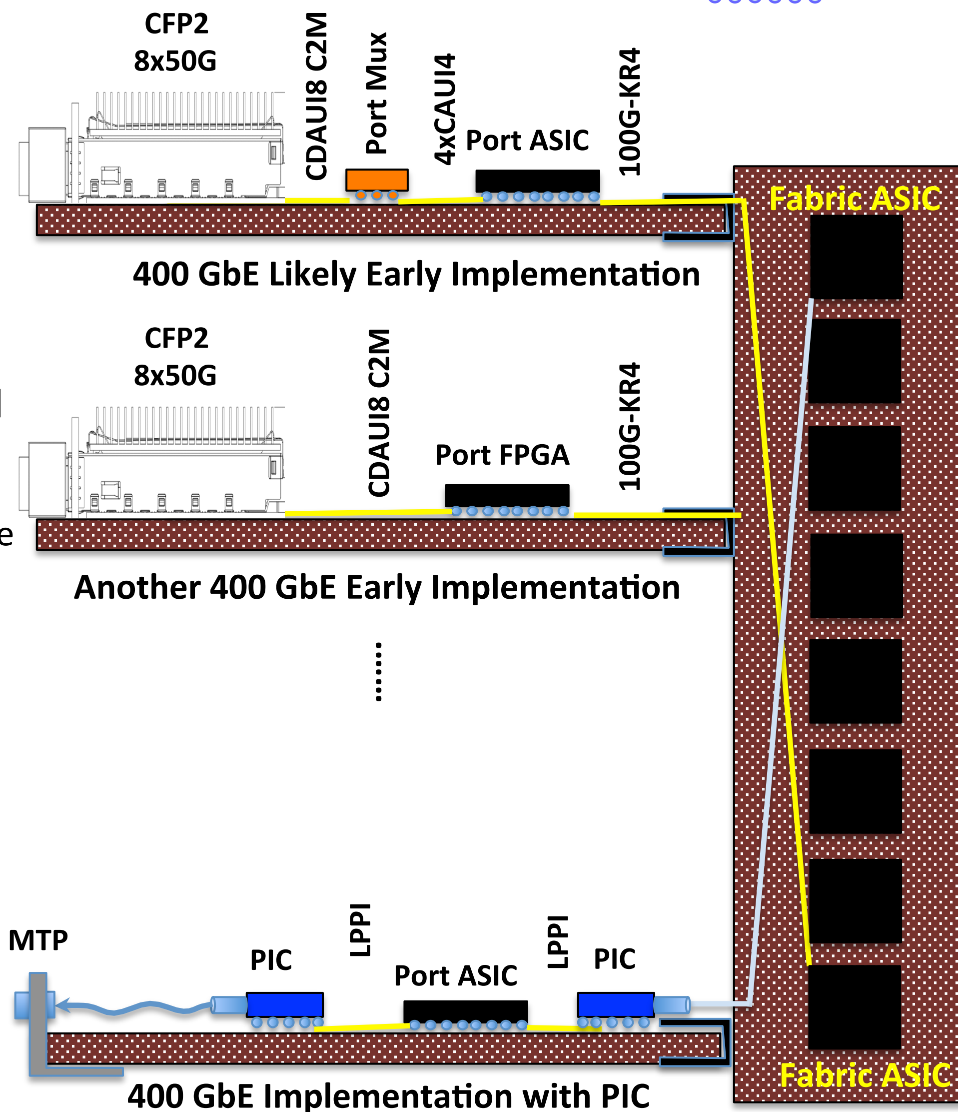
Early 400 GbE Capable Systems



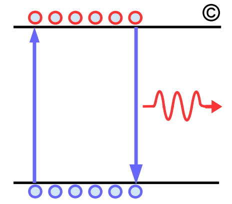
- **The 400 GbE systems expect to deploy by ~2017 based on CFP2 (8x50G) will deliver the same capacity as QSFP28 linecards delivering in 2015 (3.2-3.6 Tb/s)**
  - 400 GbE address critical needs of the OTN and router applications
  - Likely based on redesign of current line card for 400 GbE support
- **A 400 GbE implementation based on PIC could double the line card capacity to 6.4 Tb/s in 2017**
  - Ultra-low power LPPI might be the right interface for host ASIC when defining 400G-KRx is not on the table in the bs project
- **With CFP2 delivering only 3.2 Tb/s only an on board PIC could deliver higher capacity**
  - CFP2 belly to belly is not viable for blade systems
  - QSFP56 could deliver higher capacity but supports only 200G
- **We are seeing a diverging needs**

  Carriers – Require highest capacity in fiber

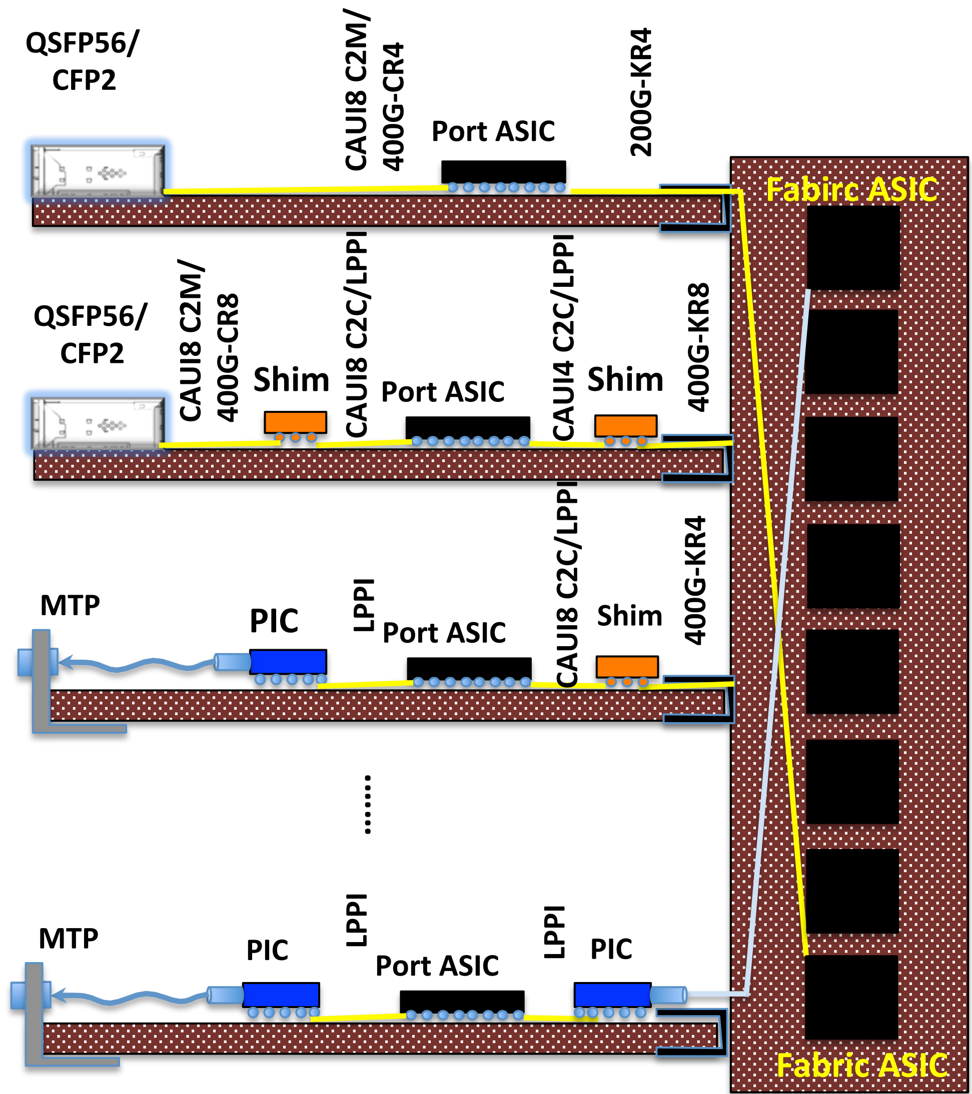
  Cloud data center – highest capacity for given radix



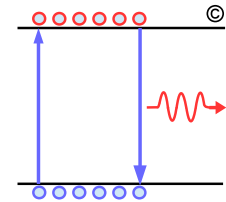
2nd Generation 50G/200G/400G Systems



- **Natural evolution of current systems to double the system capacity in 2-3 years**
  - Supports 50G/200G/400G CR/CR4/CR8
  - Supports 50G/200/400G optical PMDs
  - Double backplane capacity at least to 200G
  - Double the module capacity with QSFP56
  - Requires integration of 50G/200G/400G-KRx into the Port ASIC
  - Lanes will be triple speeds 10/25/50G
- **If integration of 50/200/400G-KRx SerDes into large ASIC is not feasible due to PD and/or area then the options are**
  - Use External shim
    - LPPI – between the ASIC and shim
    - Use PAM4 between shim and module
  - Use on board PIC
    - LPPI – between the ASIC and PIC.



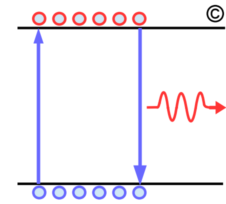
Next Generation 50/200/400G ASIC I/O Requirements



- **Scaling current system by 2x depends on feasibility of 50G/200G/400G-KRx core integration into the port ASICs**
  - Alternatively use onboard PIC or external shim IC
- **ASIC IO decision depends on**
  - If 50G/200G/400G KRx/CRx ports can be integrated meeting ASIC PD and area constrain then signaling obviously would be PAM4
  - If external shim or PIC are needed then ASIC IO is better optimized with NRZ (LPPI)
- **With expected CR/KR support in future project then PAM4 would be the natural signaling in conjunction with QSFP56 either driven from ASIC directly or from shim IC.**

50/200/400GbE Implementation	ASIC IO		Module VSR/C2M	C2C	50/200G -CRx	50/200G -KRx
KR Core on ASIC	PAM4	=	PAM4	PAM4	PAM4	PAM4
ASIC + KR Shim	NRZ (LPPI)	≠	PAM4	PAM4	PAM4	PAM4
VSR ASIC + active Cu	PAM4/NRZ	=	PAM4/NRZ	NA	Active Cu	NA
ASIC + PIC	NRZ (LPPI)	NA	NA	NA	NA	NA

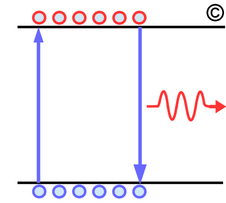
ASIC and Module I/O Evolution



- **802.3bs is trying to decide the signaling on the host ASIC and module without consideration for CR or backplane support**
  - As illustrated below going from Gen1, Gen2, to Gen3 the ASIC IO can vary between each generation
  - LPPI could be an important interface both for PIC and shim IC
  - As long as CRx/KRx core can be integrated on larger ASIC then I/O on the ASIC will be PAM4, but if CRx/KRx can not be integrated then will see bifurcation of implementations.

Implementation	ASIC IO	External Shim	VSR/C2M	Module
Gen 1 400 GbE	16x25G NRZ	NA	NRZ	CDFP
	16x25G NRZ	16x25G NRZ 8x50G	NRZ	CFP2
Gen 2 400 GbE	8x50G NRZ or PAM4	NA	NRZ or PAM4	CFP2
Gen 3 400 GbE with PIC	8x50G NRZ (LPPI)	NA	NA	PIC

Summary



- **Initial 400 GbE applications will be routers and OTN**
  - These system are not doubling BW capacity
  - Initial ASIC will be based on 16x25G directly interfacing to CDFP or use external shim to interface with CFP2
  - 50G signaling is required on next generation high capacity data center systems which expect to support 50/200/400 GbE
- **CDAUI C2M/VSR is a module interface**
  - If Cu DAC is supported then C2M interface at 50 Gb/s/lane will be based on PAM4
- **The decision what core to integrate into ASIC depends 1st on the feasibility of integrating 50 Gb/s/lane KR/CR core**
  - The interface to module would remain PAM4 as long passive Cu support is required
  - If external shim is used to drive CRx/KRx then LPPI could be a better choice for ASIC I/O
  - If an external PIC is used the optimum interface would be LPPI
- **An ASIC implementing LPPI on every port require using a shim IC or a PIC to drive external port or backplane**
- **As illustrated there is sufficient distinct identity to define a chip to shim interface with 5-7 cm PCB trace "C2S" without any connector**
  - Taskforce need to balance the need to define LPPI "C2S" vs potential delay to the project.