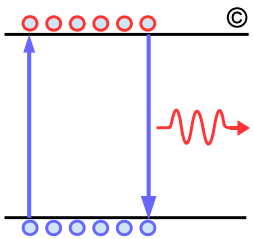


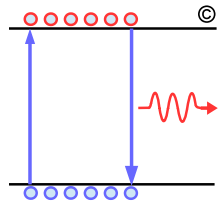
Alternate Path to Consensus on 400 GbE PMDs

Ali Ghiasi
Ghiasi Quantum LLC

IEEE 802.3bs Task Force

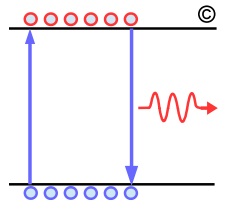
November 2014

Current 802.3bs Objective Per Dallas Meeting



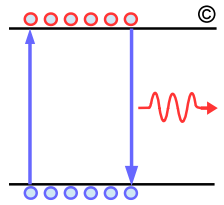
- **Provide physical layer specifications which support link distances of at least 100 m over MMF**
- **Provide physical layer specifications which support link distances of at least 500 m over SMF**
- **Provide physical layer specifications which support link distances of 2 km on SMF**
  - Where we have seen significant contentions
- **Provide physical layer specifications which support link distances of at least 10 km over SMF**
- **Key questions where consensus need to be developed are:**
  - Do we define in .bs more efficient PMDs?
  - Do we define higher bit rate narrower CDAUI in .bs?

The Reality



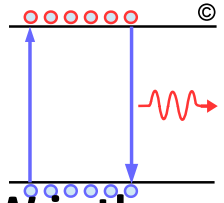
- **The 400 GbE market likely will follow 100 GbE market where even after 4 years of shipment the volume in 2014 will reach ~50K units**
  - The largest switch silicon announced to date has 3.2 Tb of capacity
    - Due to radix limitation an 8 ports 400 GbE switch is not a data center solution
  - The primary applications of 400 GbE in the next 3-4 years will be Routers and Router-OTN
  - What is needed in first phase deployment of 400 GbE is something short to connects racks (SR-16 or AOC) and 10 km
  - The world largest carrier has presented to this body the need for 10 km PMD
    - http://www.ieee802.org/3/bs/public/14_07/huang_3bs_01_0714.pdf
- **Then what is the reason for such high level of interest in 400 GbE PMDs including several competing proposals!**
  - Ladies and gentleman you are seeing replay of the movie called 802.3bm
  - Under existing PAR if 802.3bs can backfill where 802.3bm could not reach consensus it will be a service to humanity as long we don't delay 400 GbE!

Reaching Consensus on PMD Selection without Common Signaling



- **Common signaling based on 50 Gb/s PAM4 as proposed could have created as set of unified PMDs with common SerDes**
  - http://www.ieee802.org/3/bs/public/14_09/ghiasi_3bs_01a_0914.pdf
  - But there is no consensus to define a common signaling for electrical PMDs and optical PMDs
- **The catalyst for common signaling were**
  - 50 Gb/s MMF PMD to support 100 m reach
  - Operation of CADUI-8 at 50 Gb/s over current CAUI-4 channel
- **The reality is that the industry is not ready for 50 Gb/s MMF PMD yet**
  - 50 Gb/s/lane MMF likely to be the solution with introduction of 50 GbE
- **Does 802.3bs need to standardize 50 Gb/s/lane electrical now?**
  - OIF dodge the NRZ vs PAM4 down selection process by standardizing both NRZ and PAM4 for 56G-VSR
  - Even if we can reach consensus on the electrical signaling it will add 12-18 month to the project!
  - Completion of 802.3bs is not dependent on having 50 Gb/s/lane electrical signaling
  - The 50 Gb/s electrical signaling could be better engineered as part of a solution set when 50 GbE and 50 Gb/s/lane MMF are standardized.

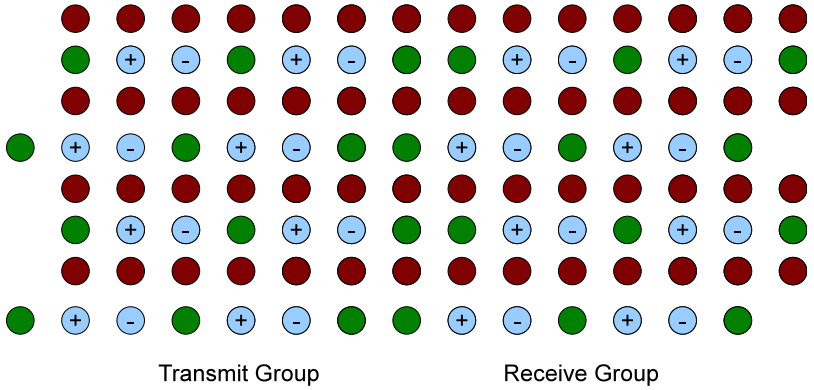
How Soon Do We Need 50G I/O on ASIC's

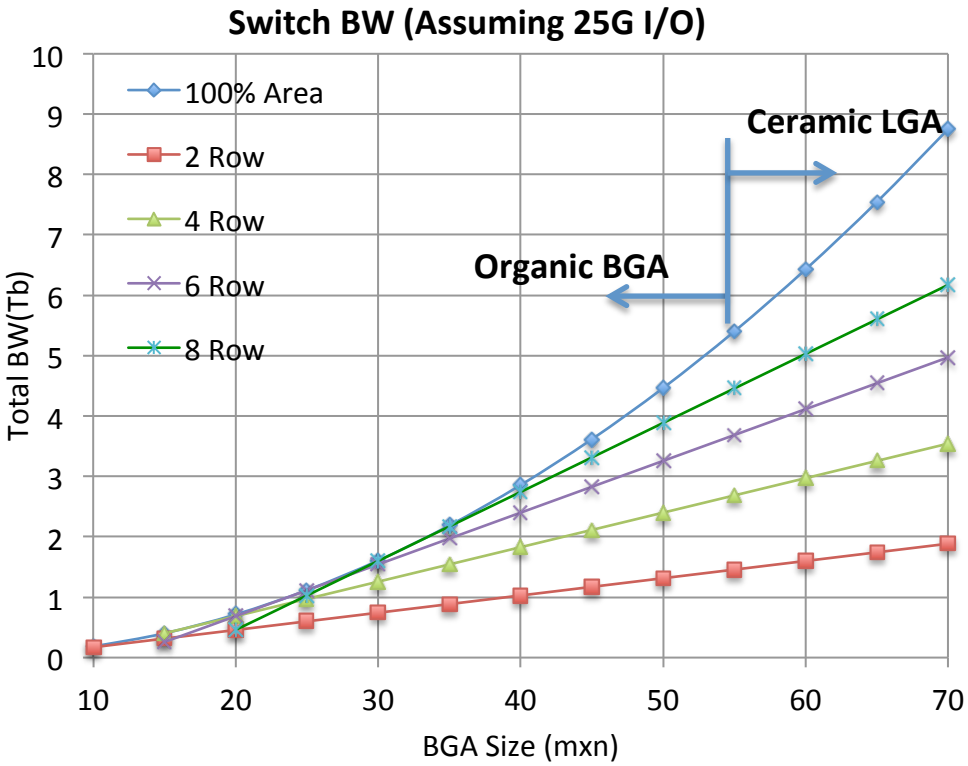


- For a given package where R is number of high speed rows and BW is the BW density for given ball field assuming 25G I/O, then total BW is calculated as (assuming 25Gb/s I/O) *
- A ceramic LGA package can support 6+ Tb/s BW with just 25G I/O

$$Total\ BW = 2 \times R \times \left((m+n) \times 2 - R \times 8 \right) \times BW$$

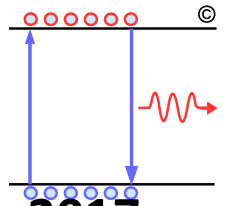
A Suitable 25G I/O Ball Map



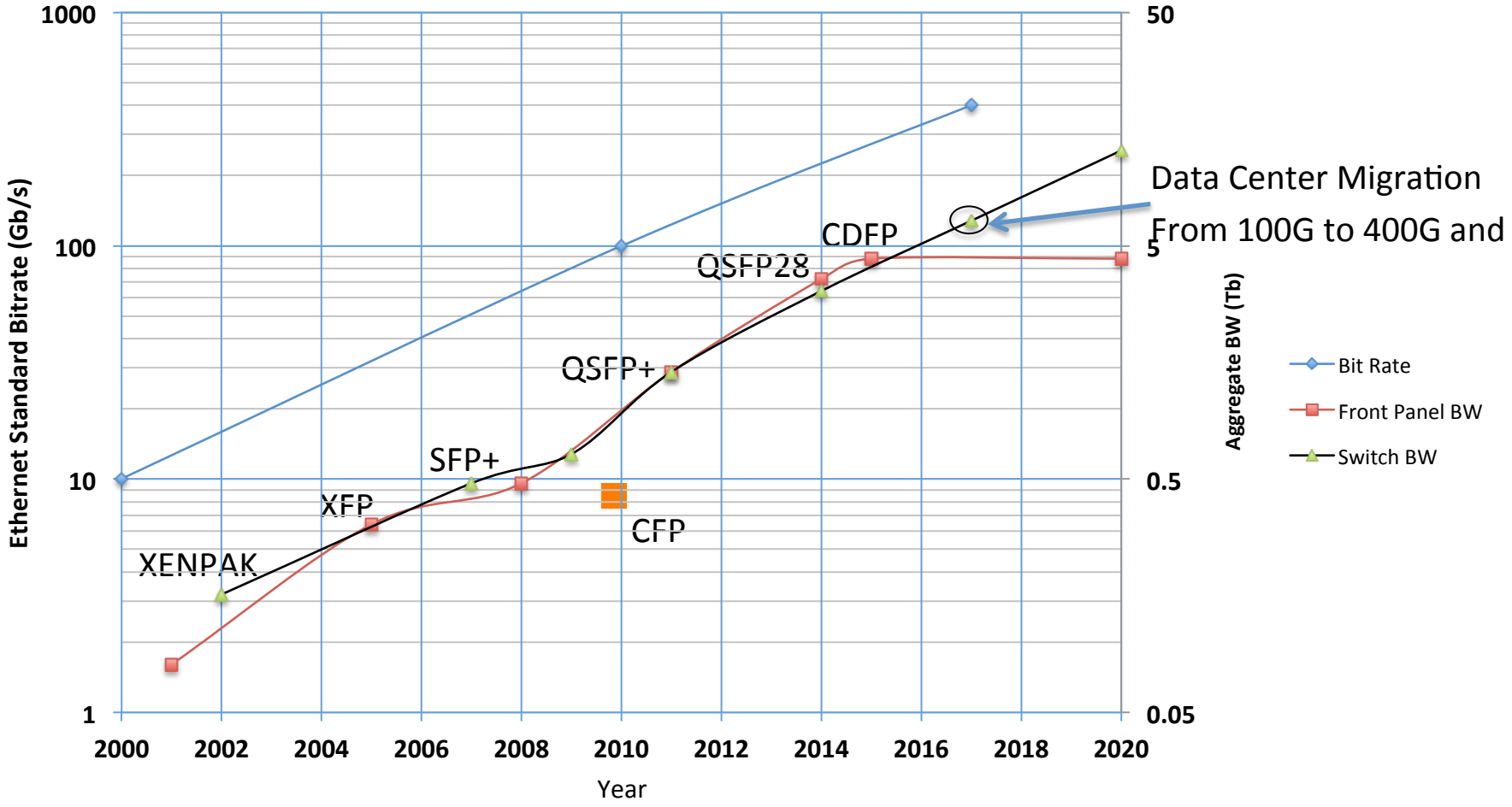


* Invited Paper: A. Ghiasi, Is There a Need for on-Chip Photonic Integration for Large Data Warehouse Switches, IEEE Photonics Group IV, 2012.

Ethernet Speed and Switch BW Growth

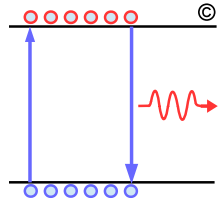


- **Updated with addition of CDFP to original published paper***
- **With addition of CDFP now 50G electrical signaling is needed in 2017 coinciding with 6.4 Tb switches**

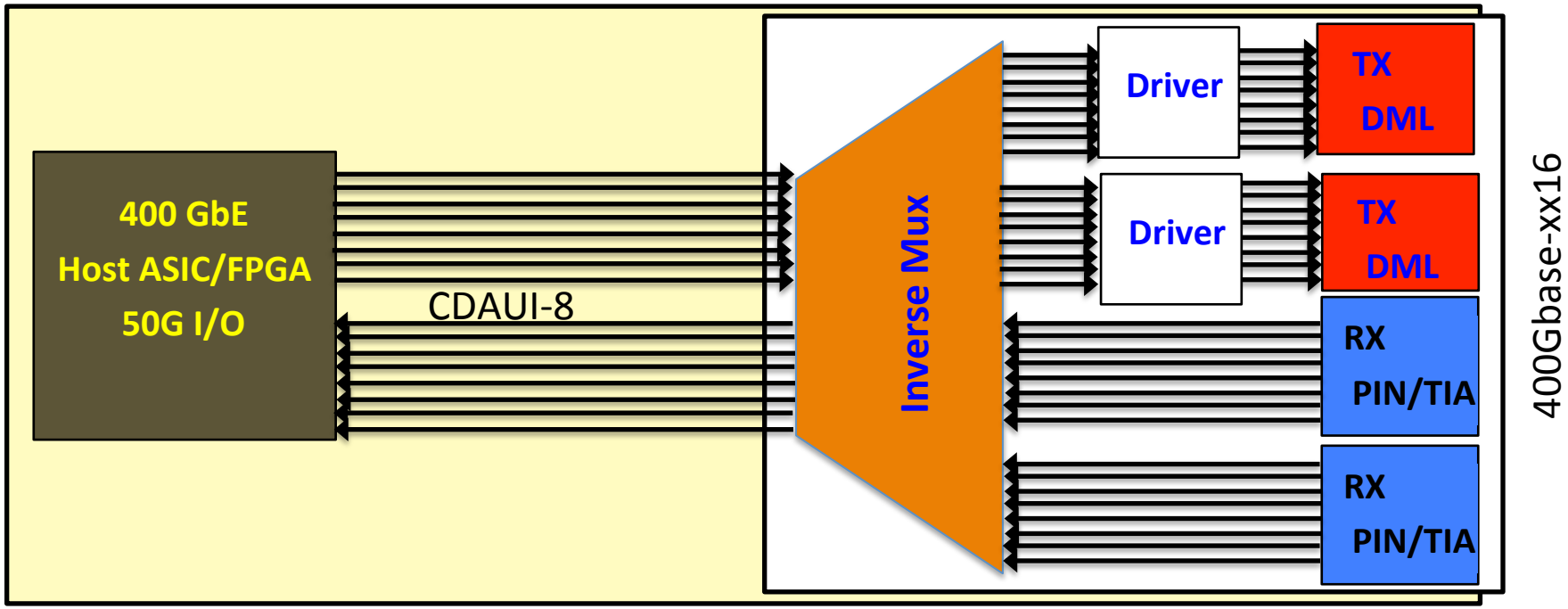


*** Invited Paper: A. Ghiasi, Is There a Need for on-Chip Photonic Integration for Large Data Warehouse Switches, IEEE Photonics Group IV, 2012.**

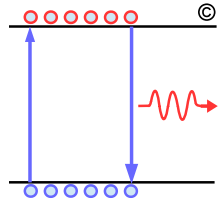
Why CDAUI-16 Will Remain the Electrical Interface of Choice



- **The lowest cost PMD the SR16 will require an in verse Mux on every port**
- **ASIC/FPGA's for many years to come will need to support 25 GbE**
- **25G-KR4 is a unified SerDes based on NRZ with moderate PD and size that can support backplane, C2C, and C2M applications**
  - No equivalent at 50G over the horizon!



To Meet the Project Timeline Need to Focus on Minimum PMD Set



- **PMDs in green best address eco system of 400 GbE and 100 GbE breakout**

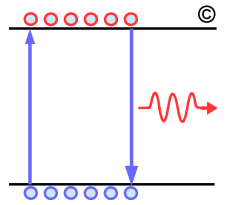
PMD	Remark	25G NRZ	50G NRZ	50G PAM4	100G PAM4	100G DMT
CADUI-16	Needed for ASIC Currently in Design	✔				
400G-SR16	Single port of 400G, 100G Breakout	✔				
CADUI-8 C2M	Linecard I/O BW in 2016		✔	✔		
CADUI-8 C2C	ASIC I/O BW in 2016		No	✔		
400G-SR8	Require CADUI-8 to Avoid Inverse-Mux To Support 100 m NRZ not an Option		✔ ~50 m	✔ 100 m		
400G-PSM4	Single Port of 400G, 100G Breakout Needs Single λ, 500 m reach				✔	
400G-LR8/4	Addressing 10 km Routers Needed now		✔	✔	*	*
400G-FR8/4	Addressing 2 km Datacenters Needed in 2018		✔	✔	*	*

Other Area of Dispute with OIF Selecting Both PAM4 and NRZ

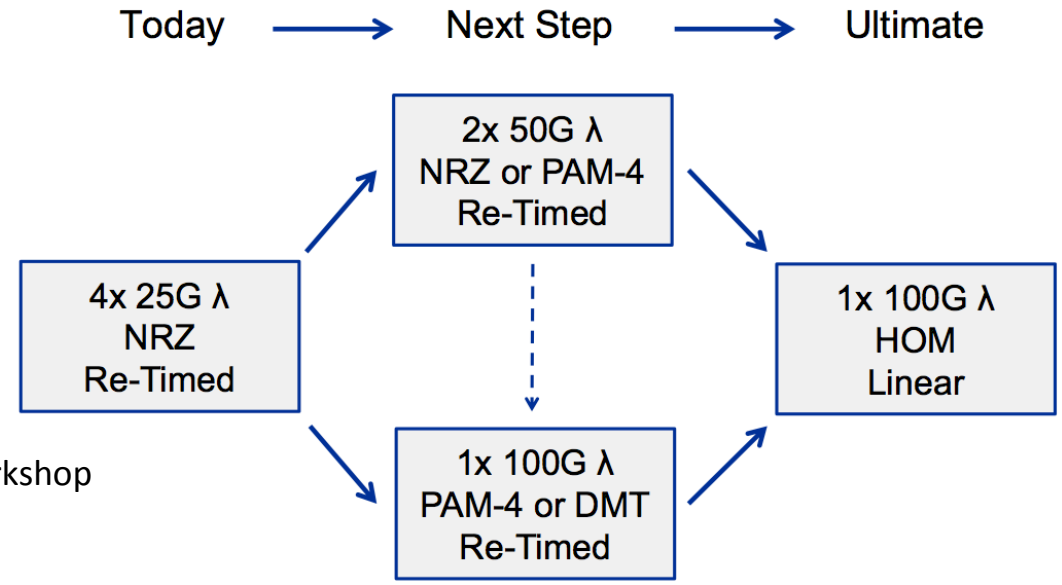
Area of Dispute

*** Insufficient link budget and/or requirement for 9+ dB FEC based on results presented as of Sept 2014.**

What Should We Aim for?

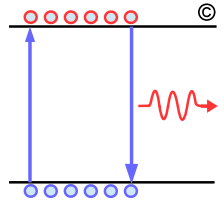


- **The challenge for aiming for ultimate PMD 1x100G λ**
  - With introduction of 100 GbE switches supporting 25G/lane overnight 25 GbE has became an IEEE project
  - Following the same logic the IEEE will have a 50 GbE project in the next 3 years where ultimate solution for some applications would be 8x50G!
    - Not to mention 850 nm MMF speed will <100G
  - Some applications can be better be served with 1x100G λ if it is technically feasible

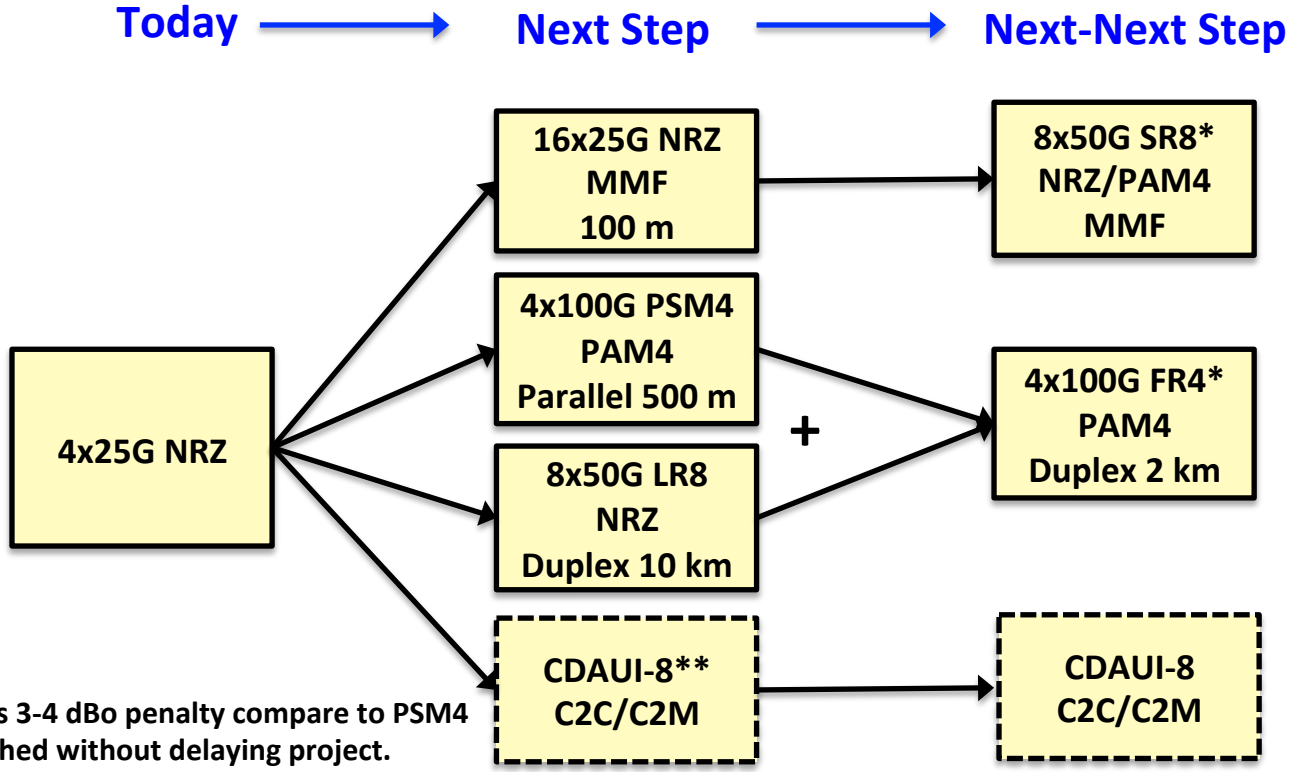


Cole OIDA/EA Workshop
May 2014

What Might Actually PMD Evolution Look Like

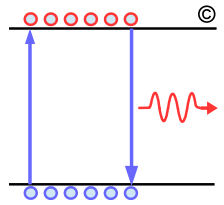


- **Defining 4x100G single λ 2 km duplex now pushes feasibility limit and require PMD specific high gain FEC that needs to be supported for ever!**
  - With components improvement duplex 2 km is better defined as part of next-next PMD development cycle and may use moderate gain host FEC



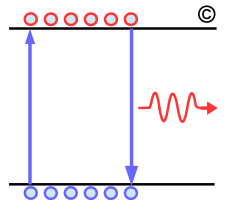
*** Optical Mux/de-mux adds 3-4 dBo penalty compare to PSM4**
**** If consensus can be reached without delaying project.**

The Logic Behind the Proposed PMD Set



- **CDAUI-16**
  - Compatibility with current and next generation ASICs/FPGAs, 25 GbE, 100 GbE, QSFP28, and CDFP
- **SR-16**
  - Address 400 GbE rack-rack applications
  - Compatibility with current and next generation ASICs/FPGAs, 25 GbE, 100 GbE, QSFP28, and CDFP
- **400G-PSM4**
  - Addresses 100 GbE data centers and 400 GbE co-located racks
  - Form factor compatibility QSFP28 and CDFP
- **400G-LR8**
  - Addresses the 400 GbE early deployments of routers and OTN equipment's that require 10 km
  - Form factor potentially CDFP and TBD.

Summary



- **Initial 400 GbE applications will be routers and OTN**
  - The primary PMD used on Routers and OTNs are 10 km duplex
- **Other consideration on the PMD selection are 100 GbE break out and 25GbE compatibility**
- **The following set of PMDs best addresses the 400 GbE as well as ecosystems**
  - CADUI-16
  - 400G-SR16
  - 400G-PSM4 based on 100 Gb/s PAM4 with 500 m reach
  - 400G-LR8 based on 50 Gb/s NRZ with 10 km reach
- **CDAUI-8 potentially will delay this project by at least a year and forces the lowest cost PMD SR16 to use inverse mux**
  - Early CFP2 implementation leveraged OIF 28G-VSR prior to CAUI-4 definition
  - Early implementation of 400 GbE based on 50G/lane could also use OIF-56G-VSR/MR.