Compatibility of Different Port Types at a Big IC

Piers Dawe

Rita Horner

John Petrilla

Avago Technologies

Contents

- Introduction
- Examples of ICs that support multiple port types or different medias
- Opportunity for even better modularity in P802.3ba
- Complexity implications for an IC of different port types
  - Transmitter
  - Receiver
  - Overall
- Conclusions

Introduction

- Many media types
- At least two likely classes of pluggable module
  - Big
  - Small
- Common MAC/PCS silicon, path to multi-port ICs
- This presentation shows that multiple media types can be supported with compatible specifications in the same MAC/PCS silicon
- Mix-and-match pluggability as in traditional Ethernet

Examples 1

- GBIC
  - Same socket can receive
    - Optical modules
      - for MMF
      - for SMF
    - Other?
  - XENPAK
    - As above, but too much electronics in the module

Examples 2: SFP/SFP+

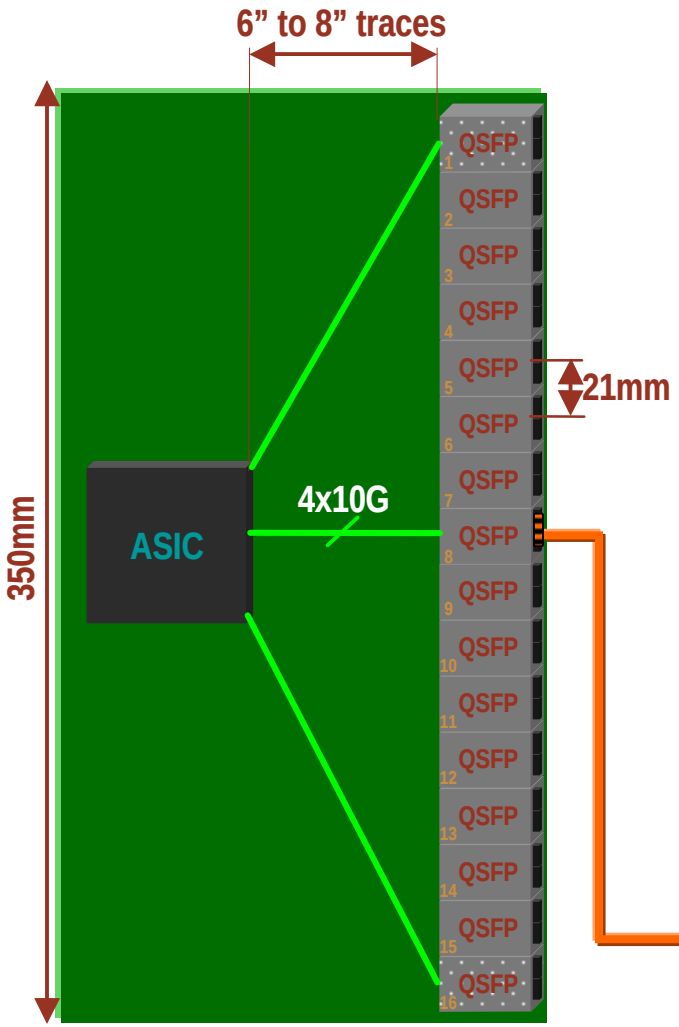
- Same socket can support
  - Optical modules
    - for MMF
    - for SMF
  - Passive electrical cable
  - Active electrical cable
  - Active optical cable
- Available for 10G Ethernet: "hard"
  - Intermediate PHY ICs used for challenging analog performance
  - Cost, heat, space disadvantages
  - It's not hard because it's unretimed and pluggable; it's hard because too much budget was given to the fibre
- Available for 8G Fibre Channel: "easy"
  - Believed to be implemented with multi-port "big ICs"
  - "Easy" because a generous proportion of the budget was given to the host, with enough for the module

We Can Do Better Than Those

- Small module
  - Same socket can support
    - Optical modules for 100 m MMF and 40G, 10 km SMF
      - Active optical cable if desired
    - Passive electrical cable: 2, 5 or 10 m
      - Active electrical cable if desired
    - These are the set of two port types per MAC rate for the Data Center, with one socket type and one big IC type
      - Backplane is physically different: same one big IC type
  - No need to make it "hard"!
- Big module
  - Same socket can receive
    - Optical modules for 10 km, 40 km SMF
    - Also as above (100 m, 40G, 10 km SMF and electrical cable) if desired
    - Assume will contain CDRs: not "hard"
      - Meaning not an unusual analog burden on the ASIC
    - The whole set of port types per MAC rate with one socket type and one big IC type
- Multiple modules
  - Four or ten, SFP+ or XFP modules
  - P802.3ba doesn't have to sweat to enable this; it can just happen

40G SerDes Port Side Connections to QSFP

Single ASIC – Maximum Case



ASIC SerDes power: 250mW per 10G

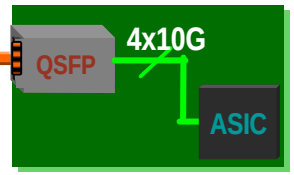
- Total power per ASIC due to SerDes
  4ch x 16QSFP x 0.25W / ch = 16W

QSFP: 21mm pitch

➔ x16 QSFP – single row

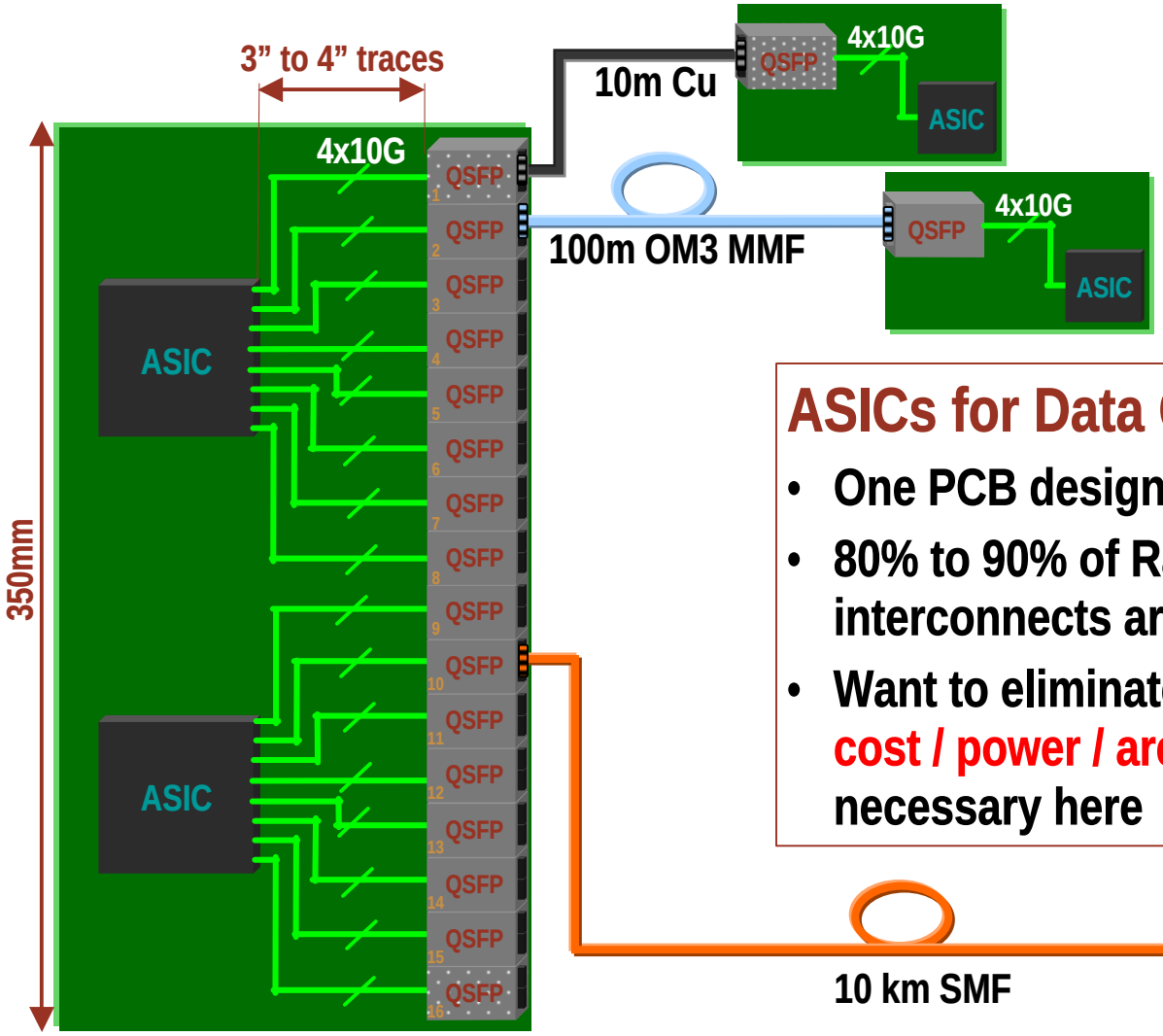
➔ x32 QSFP – stacked 2-rows

Max. Channel: 175mm ➔ > 6”



40G SerDes Port Side Connections to QSFP

Optics and Cu

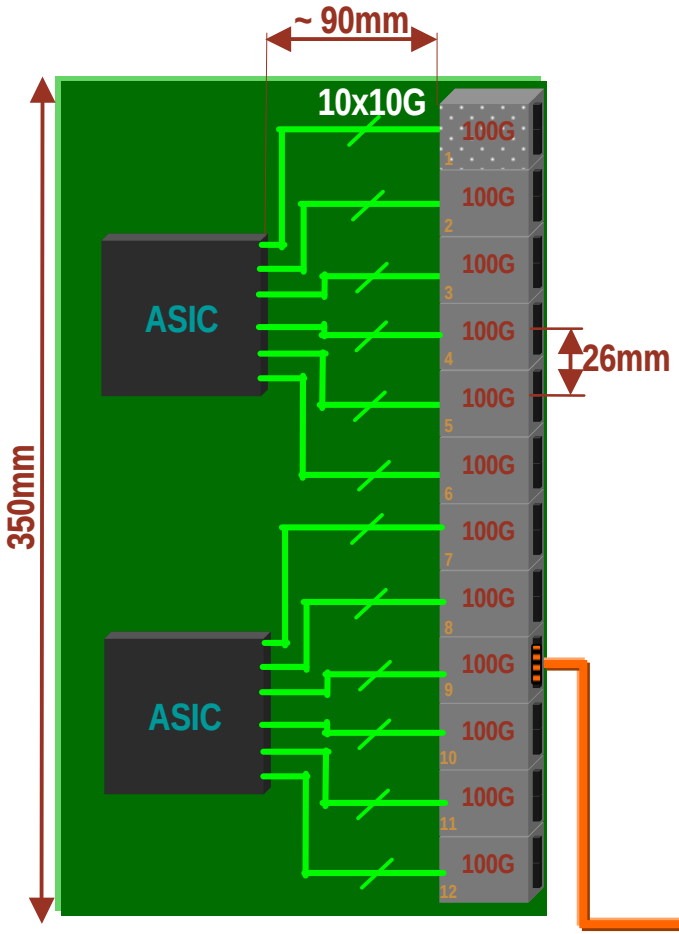


ASICs for Data Centers

- One PCB design for multiple platforms
- 80% to 90% of Rack-to-Rack stackable interconnects are Cu connections
- Want to eliminate the additional cost / power / area of glue logic, not necessary here

100G SerDes Port Side Connections

Two ASICs – Maximum Case



ASIC SerDes power: 250mW per 10G

- Total power per ASIC due to SerDes
  10ch x 12connectors x 0.25W / ch = 30W
- Requires at least two ASICs

100G connector: 26mm pitch

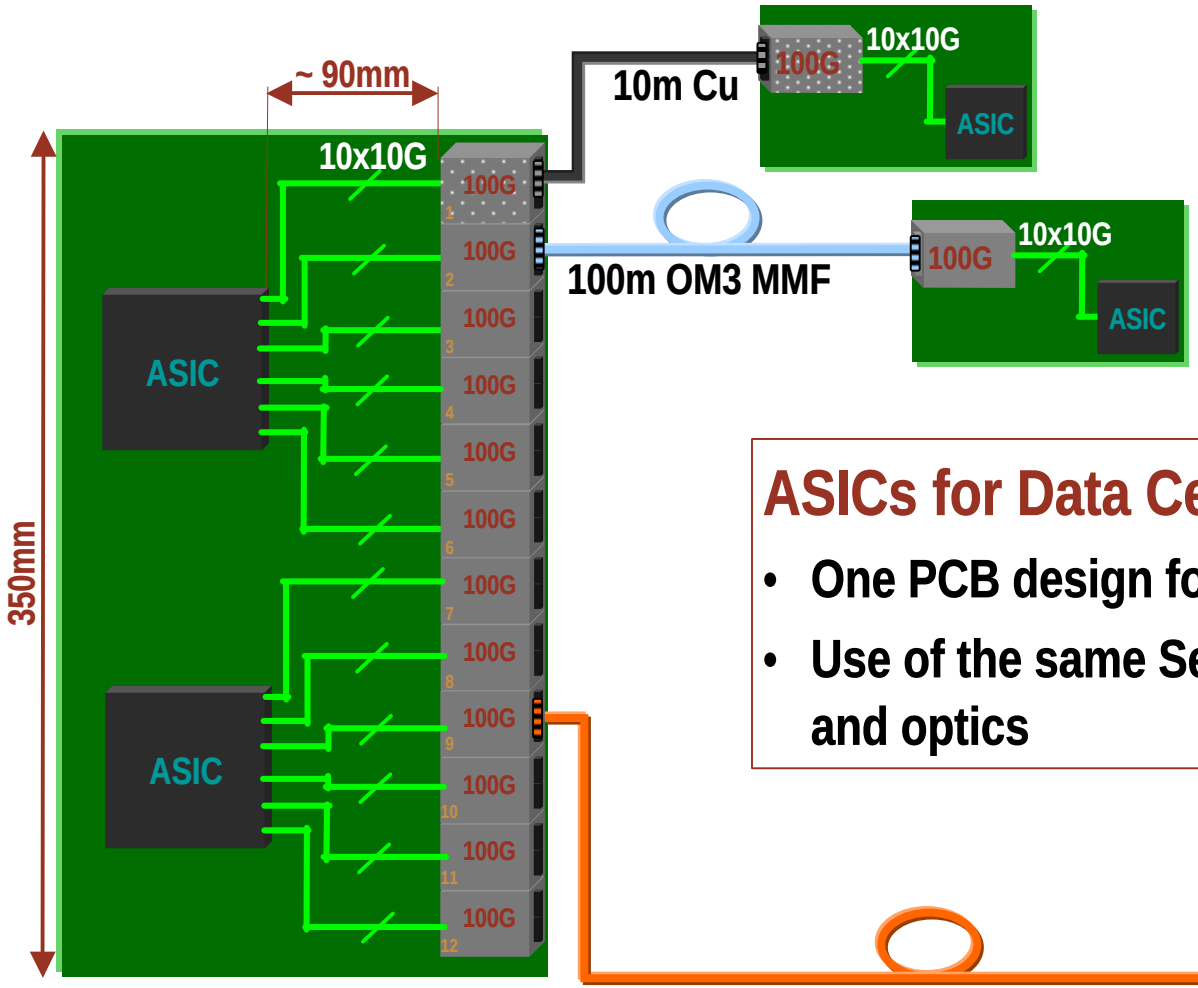
➔ x12 – single row

➔ x24 – stacked 2-rows

Max. Channel: 87.5mm ➔ < 4”

100G SerDes Port Side Connections

Optics and Cu

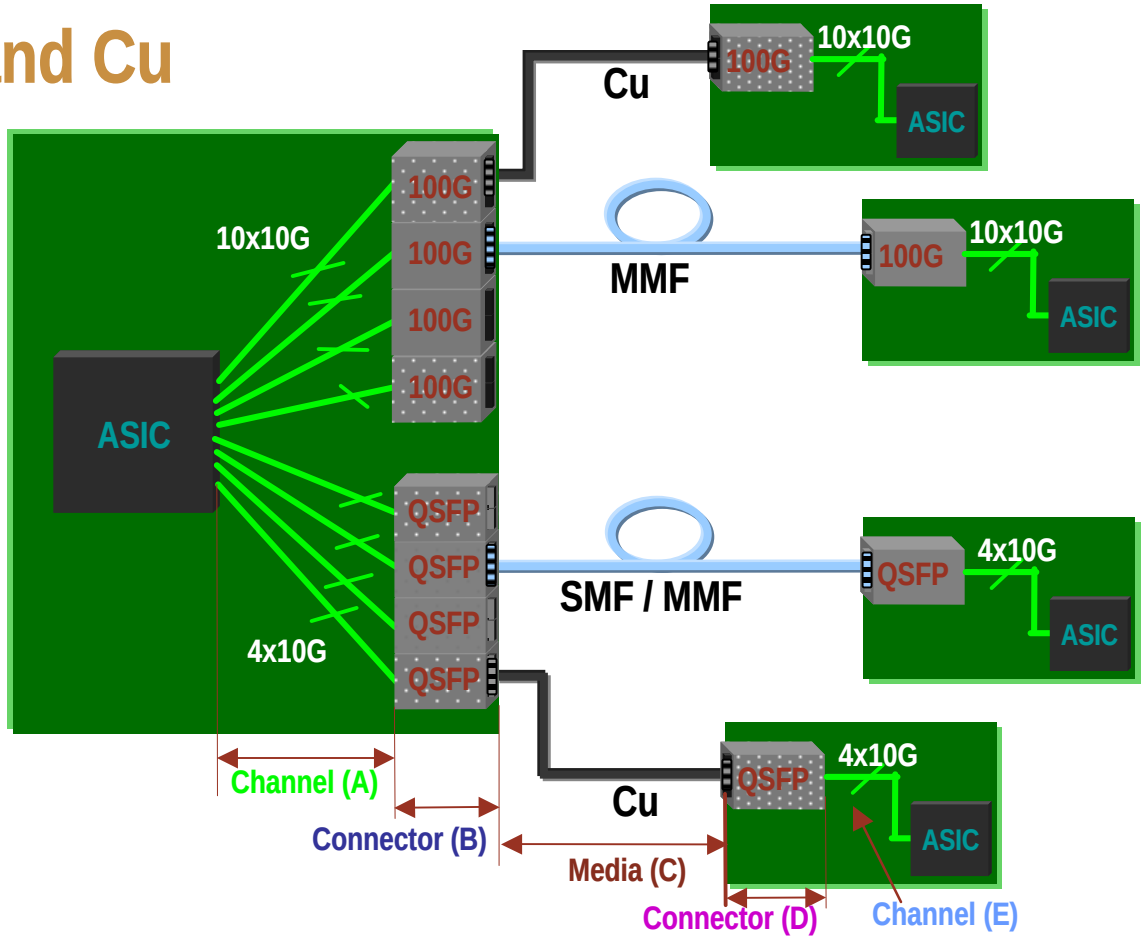


ASICs for Data Centers

- One PCB design for multiple platforms
- Use of the same SerDes to interface both Cu and optics

Need for Well Defined Port Side Interface Specification

Optics and Cu



- **Cu:** Channel Loss (A) + Connector Loss (B) + Interconnect Media Loss (C) + Connector Loss (D) + Channel Loss (E)
  - Have to trade off cable length vs. IC specs and PCB length wisely
- **Optical:** Channel Loss (A) + Connector Loss (B) + Impairments of Tx, fiber, Rx + Connector Loss (D) + Channel Loss (E)
  - Have to trade off optical jitter specs vs. IC specs and PCB length wisely

Examples Show High Density Pluggable Mix&Match HSE is Viable

- Committee has to strike a judicious balance between IC specs, Cu cable and PCB trace loss
- Glue logic chips can be eliminated for 100 m MMF, 40G 10 km SMF, and short copper cables
  - Save cost, power and space
- What does it take to do this?

Complexity Implications: Tx at a Big IC

- If PCB "channel" is easy, don't need Tx emphasis
  - e.g. GBIC or SFP at 1GE
- If harder, use Tx emphasis
  - Might be set blind e.g. XFP with short traces
- Harder still?
  - Either, set Tx emphasis per port
    - e.g. for FC-PI-4 (8GFC) with wide tolerances, 10GE SFP+ with narrow tolerances, can be according to PCB loss
  - Or, use per-link Tx emphasis
    - Needs some handshaking or auto-negotiation
      - e.g. 10GBASE-KR backplane
      - Adds complexity
      - Per lane or same emphasis for all lanes?
      - Typically cannot retune if channel evolves
        - » Not seen as a problem because Rx can retune
      - Is it necessary?
- Think in terms of clocked delay lines with 1, 2 or 3 taps
  - See chart later

Complexity Implications: Rx at a Big IC

- If it's easy, don't need any equaliser
  - e.g. GBIC or SFP at 1GE
- If harder, use an equaliser
  - Huge range from simple and low power to full strength LRM/SFP+ to long-haul maximum likelihood sequence detector (MLSD) equaliser...
  - Equaliser settings are obtained by the Rx
    - In normal operation
      - e.g. 10GBASE-LRM
      - Can retune if channel evolves
      - If FEC used, can use corrected errors to fine-tune Rx
    - In a training sequence
      - Can have an easy pattern to get started
        - » e.g. 10GBASE-T (and 10GBASE-KR?)
  - Think in terms of Feed Forward Equaliser (FFE) with *n* taps, and/or Decision Feedback Equaliser (DFE) with *m* taps
    - If no feedforward equaliser, *n*=1
    - If no feedback equaliser, *m*=1
    - See chart later

Analog Challenge Implications

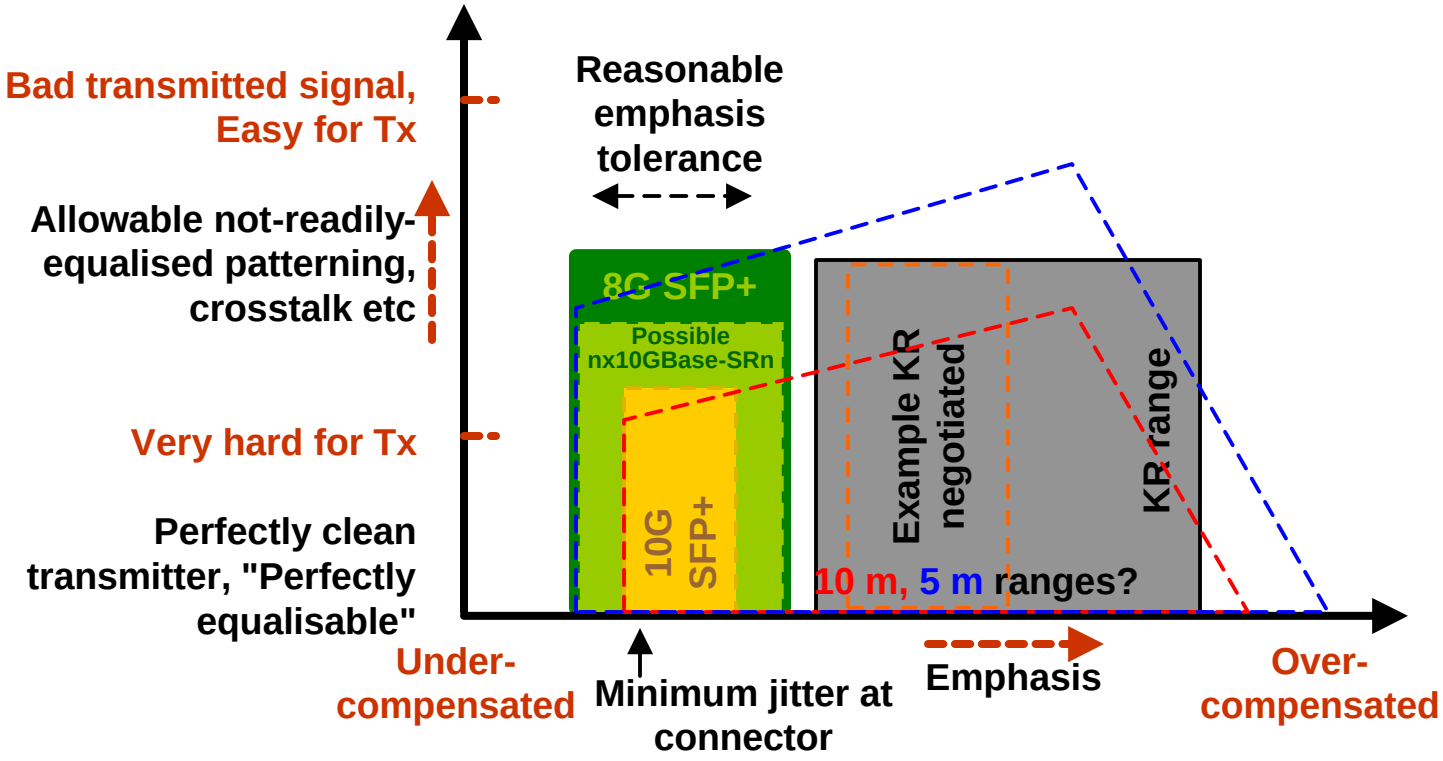
- This is crucial
- The actual spec numbers e.g. jitter, crosstalk, reflections make all the difference between high yield and unremarkable "it just works" and tolerance-intolerant, temperamental, applications-support-needing, pain
- With multiple lanes, unless we have a guarantee of correlation between the lanes, we have to be conservative to get all 4 or 10 working well every time, all the time

Comparison Chart - Tx

Design assumptions (real implementations can be different)	Transmit taps *	Transmit emphasis ^	Per-port Tx emphasis setting	Auto neg or hand-shaking	Thermal or complexity at big IC, Tx+Rx	Analog challenge for big IC, Tx+Rx	Difficulty ranking, Tx+Rx	Port type	Reference
Port/implementation type					(subjective)				
Traditional limiting e.g. GBIC, SFP at lower speeds	1 + 0	No	No	No	Lowest	Low	2	e.g. 1000BASE-SX	802.3 38.5 and SFF-8431 App. F
XFP	1 + 1? 1 + 2?	Yes	Yes	No, set by host	Low	Low	2		SFP ftp://ftp.seagate.com/sff/INF-8077.PDF
AUI (Ghiasi)	1 + 1? 1 + 2?	Yes	Yes?	No, set by host?	Medium? Low? ^^	Moderate	3		ghiasi_HSE_01_0608.pdf
5 m copper, .3ba (Horner)	1 + 1? 1 + 2?	Default?	Not needed?	Not needed?	Mid to low	Low-mid **	3		
FC-PI-4 optical limiting	1 + 1? 1 + 2?	No	Yes	No, set by host	Low	Low	4	e.g. 800-SN	FC-PI-4 Ch. 6 www.t11.org/ftp/t11/pub/fc/pi-4/08-138v1.pdf
FC-PI-4 electrical what?	1 + 1? 1 + 2?	No?	Yes	Required	Mid to low?	Low?	4	e.g. 800-DF-EL-S? 800-DF-EA-S?	FC-PI-4 Ch.9 www.t11.org/ftp/t11/pub/fc/pi-4/08-138v1.pdf
5 m copper, SFP+ (SFF-8461)	1 + 2	No	Not needed	Not needed	Mid to low	Low	4		
10GBASE-KR	1 + 2	Yes	Yes	Req'd; is it needed?	Mid to low?	Moderate?	5		802.3 72, 73
TP0/1/4/5 PMDSI (Petrilla)	1 + 1? 1 + 2?	No	Yes	No, set by host	Low	Low-mid **	5	40GBASE-SR10, 100GBASE-SR10	petrilla
SFP+ limiting	1 + 1 or 1 + 2	No	Yes	No, set by host	Moderate	High	7		ftp://ftp.seagate.com/sff/SFF-8431.PDF
10 m copper (Di Minico)	1 + 2	Will be needed	Yes?	Available; is it required?	Medium-high?	High?	8		
10 m copper, SFP+ (SFF-8431)	1 + 1 or 1 + 2	No		No, set by host	High	Very high	9		
SFP+ LRM	1 + 1 or 1 + 2	No	Yes	No, set by host	High	Very high	10	10GBASE-LRM	ftp://ftp.seagate.com/sff/SFF-8431.PDF
Notes									
Transmitter and receiver at a multi-port IC									
^ More emphasis than is needed to open the eye at the transmitter port									
** For 100 m or with FEC. Can always make it hard by objective creep.									
^^ Trade-off with PCB channel									

Too hard for multi-channel

Tolerancing Chart for Tx



- This is an attempt to show the Tx tolerancing graphically
- Can trade Tx number of taps against PCB length or loss
- No opinion ventured on the relative height of 8GFC and KR boxes

Comparison chart - Rx

Design assumptions (real implementations can be different)	Receiver FFE taps *	Receiver DFE taps	Rx equaliser tracking?	Error bursts a concern?	Thermal or complexity at big IC, Tx+Rx	Analog challenge for big IC, Tx+Rx	Difficulty ranking, Tx+Rx	Port type
Port/implementation type					(subjective)			
Traditional limiting e.g. GBIC, SFP at lower speeds	1	0	No	No	Lowest	Low	2	e.g. 1000BASE-SX
XFP	1?	1? 2?	No	No	Low	Low	2	
AUI (Ghiasi)	1?	2 to 5?	Available	Minor?	Medium? Low? ^^	Moderate	3	
5 m copper, .3ba (Horner)	1	5 or less	Available	Minor?	Mid to low	Low-mid **	3	
FC-PI-4 optical limiting	1	2	Not required	No	Low	Low	4	e.g. 800-SN
FC-PI-4 electrical what?	1	2	?	Minor?	Mid to low?	Low?	4	e.g. 800-DF-EL-S? 800-DF-EA-S?
5 m copper, SFP+ (SFF-8461)	1	5	Available	Minor	Mid to low	Low	4	
10GBASE-KR	1	5	Available	Yes	Mid to low?	Moderate?	5	
TP0/1/4/5 PMDSI (Petrilla)	1?	2 or 3	Available	No	Low	Low-mid **	5	40GBASE-SR10, 100GBASE-SR10
SFP+ limiting	No spec (e.g. 1)	No spec (e.g. 2-3)	Yes (not required)	No	Moderate	High	7	
10 m copper (Di Minico)	1?	5 or more?	Available	Yes	Medium-high?	High?	8	
10 m copper, SFP+ (SFF-8431)	14	5	Available	Yes	High	Very high	9	
SFP+ LRM	14	5	Yes (required)	Minor	High	Very high	10	10GBASE-LRM
Notes								
Transmitter and receiver at a multi-port IC								
* There's always at least 1 tap even if no smarts or FFE								
** For 100 m or with FEC. Can always make it hard by objective creep.								
^^ Trade-off with PCB channel								

Too hard for multi-channel

Eye Chart

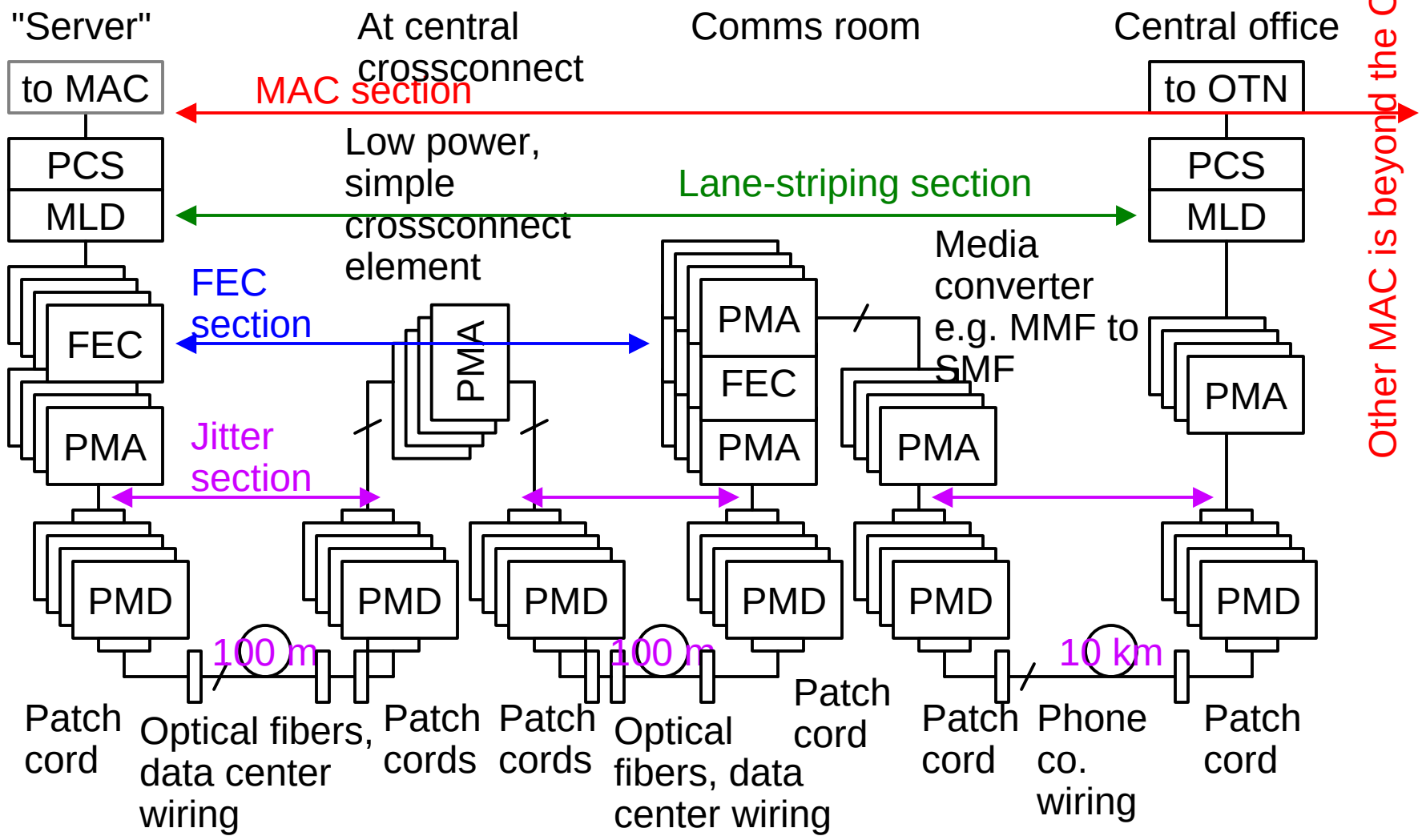
Design assumptions (real implementations can be different)	Transmit taps *	Transmit emphasis ^	Per-port Tx emphasis setting	Auto neg or hand-shaking	Receiver FFE taps *	Receiver DFE taps	Rx equaliser tracking?	Error bursts a concern?	Thermal or complexity at big IC, Tx+Rx	Analog challenge for big IC, Tx+Rx	Difficulty ranking, Tx+Rx	Port type	Reference
Port/implementation type									(subjective)				
Traditional limiting e.g. GBIC, SFP at lower speeds	1 + 0	No	No	No	1	0	No	No	Lowest	Low	2	e.g. 1000BASE-SX	802.3 38.5 and SFF-8431 App. F
XFP	1 + 1? 1 + 2?	Yes	Yes	No, set by host	1?	1? 2?	No	No	Low	Low	2		SFP ftp://ftp.seagate.com/sff/INF-8077.PDF
AUI (Ghiasi)	1 + 1? 1 + 2?	Yes	Yes?	No, set by host?	1?	2 to 5?	Available	Minor?	Medium? Low? ^^	Moderate	3		ghiasi_HSE_01_0608.pdf
5 m copper, .3ba (Horner)	1 + 1? 1 + 2?	Default?	Not needed?	Not needed?	1	5 or less	Available	Minor?	Mid to low	Low-mid **	3		
FC-PI-4 optical limiting	1 + 1? 1 + 2?	No	Yes	No, set by host	1	2	Not required	No	Low	Low	4	e.g. 800-SN	FC-PI-4 Ch. 6 www.t11.org/ftp/t11/pub/fc/pi-4/08-138v1.pdf
FC-PI-4 electrical what?	1 + 1? 1 + 2?	No?	Yes	Required	1	2	?	Minor?	Mid to low?	Low?	4	e.g. 800-DF-EL-S? 800-DF-EA-S?	FC-PI-4 Ch.9 www.t11.org/ftp/t11/pub/fc/pi-4/08-138v1.pdf
5 m copper, SFP+ (SFF-8461)	1 + 2	No	Not needed	Not needed	1	5	Available	Minor	Mid to low	Low	4		
10GBASE-KR	1 + 2	Yes	Yes	Req'd; is it needed?	1	5	Available	Yes	Mid to low?	Moderate?	5		802.3 72, 73
TP0/1/4/5 PMDSI (Petrilla)	1 + 1? 1 + 2?	No	Yes	No, set by host	1?	2 or 3	Available	No	Low	Low-mid **	5	40GBASE-SR10, 100GBASE-SR10	petrilla
SFP+ limiting	1 + 1 or 1 + 2	No	Yes	No, set by host	No spec (e.g. 1)	No spec (e.g. 2-3)	Yes (not required)	No	Moderate	High	7		ftp://ftp.seagate.com/sff/SFF-8431.PDF
10 m copper (Di Minico)	1 + 2	Will be needed	Yes?	Available; is it required?	1?	5 or more?	Available	Yes	Medium-high?	High?	8		
10 m copper, SFP+ (SFF-8431)	1 + 1 or 1 + 2	No		No, set by host	14	5	Available	Yes	High	Very high	9		
SFP+ LRM	1 + 1 or 1 + 2	No	Yes	No, set by host	14	5	Yes (required)	Minor	High	Very high	10	10GBASE-LRM	ftp://ftp.seagate.com/sff/SFF-8431.PDF
Notes													
Transmitter and receiver at a multi-port IC													
* There's always at least 1 tap even if no smarts or FFE													
^ More emphasis than is needed to open the eye at the transmitter port													
** For 100 m or with FEC. Can always make it hard by objective creep.													
^^ Trade-off with PCB channel													

Too hard for multi-channel

Common socket and common IC implies coordinating test points

- **Next slides of multi-section links to set the scene, then focus on test points**

Example of link with multiple sections



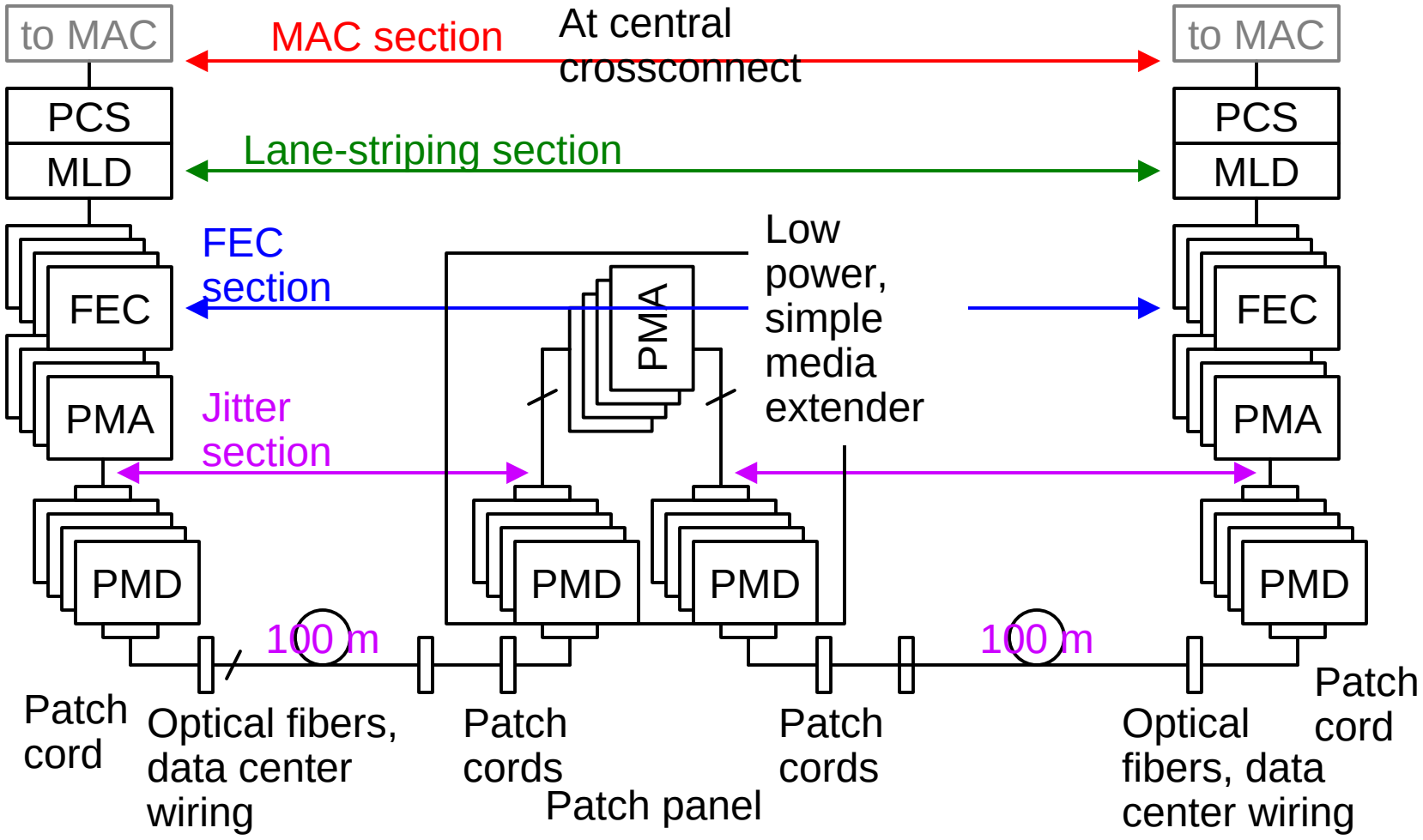
Sort of like SONET section, line and path – but four levels shown here

FEC and lane-striping sections might be nested the other way

The things auto-negotiating may not have MACs. Need some autonomy within PHY

They may not see the aggregate stream, let alone parse packets

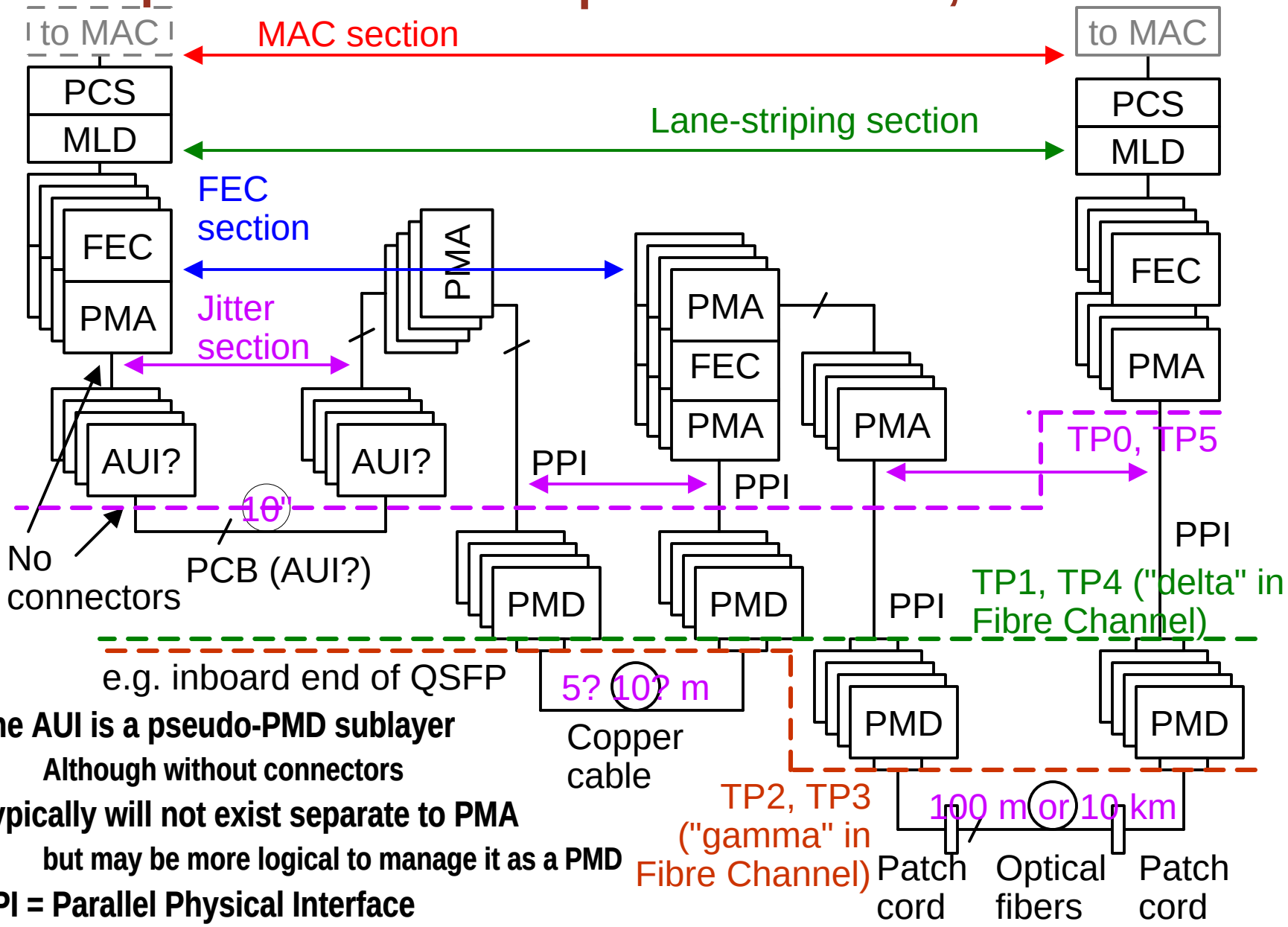
Example of link with multiple sections: media extender



For those with big data centers who must use MMF

Media extenders could be built into patch panel/crossconnect to save two connectors

Example of link with multiple sections: AUI, nBASE-CRm



The AUI is a pseudo-PMD sublayer

- **Although without connectors**

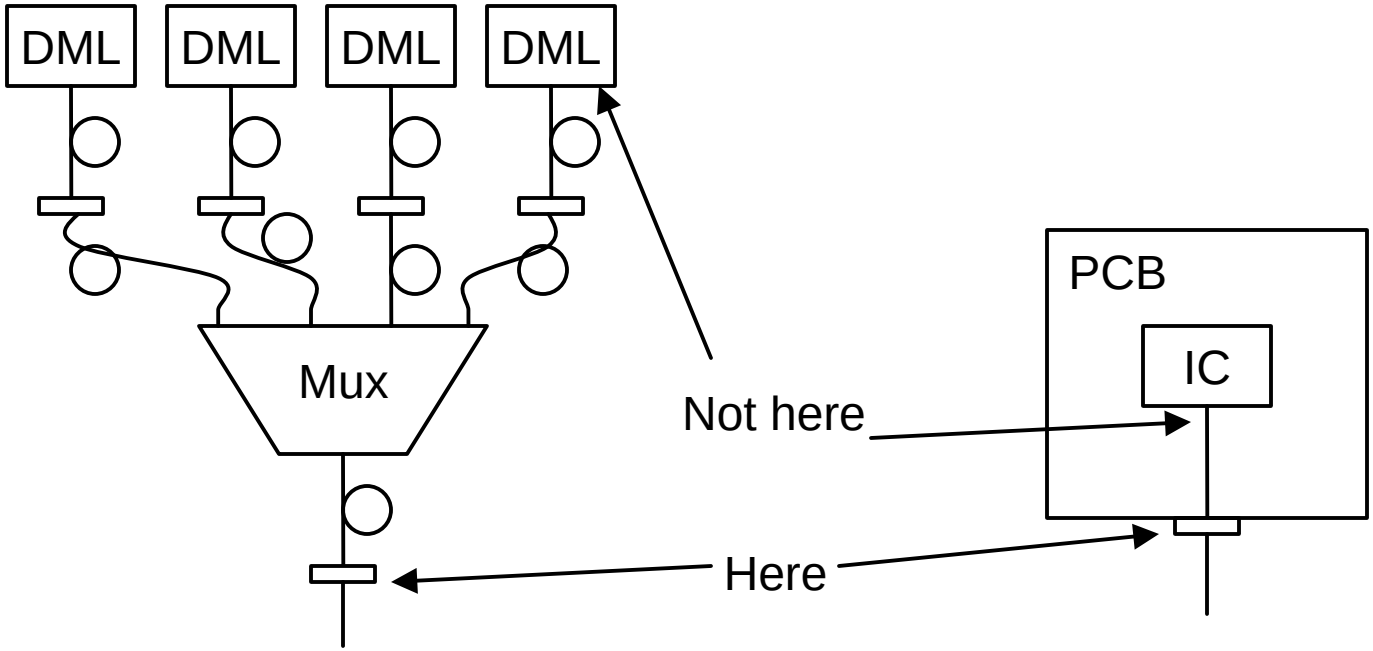
Typically will not exist separate to PMA

- **but may be more logical to manage it as a PMD**

PPI = Parallel Physical Interface

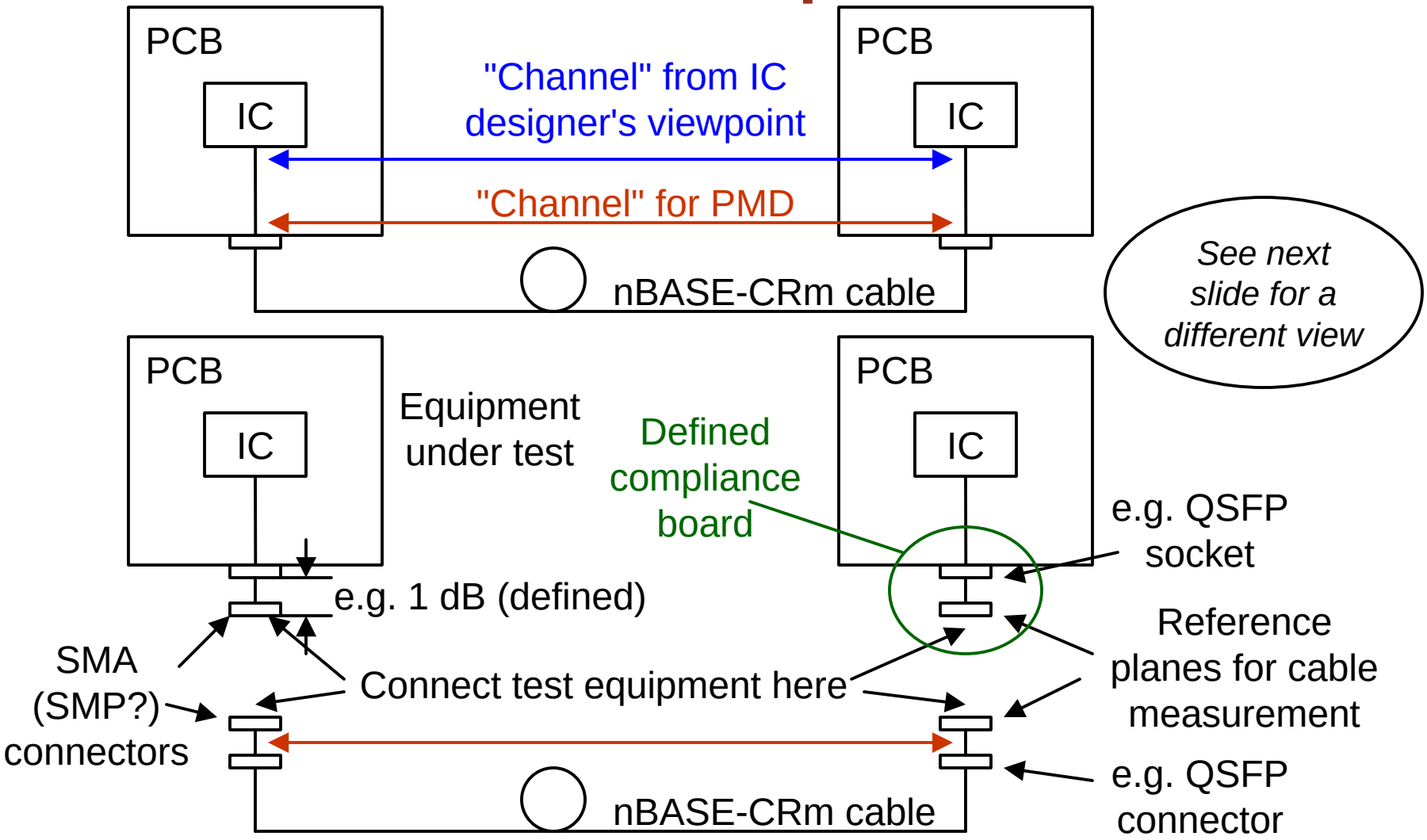
- **Same as un-retimed PMD service interface**
- **Two flavours, optical and electrical?**

Where is the end of the PMD?



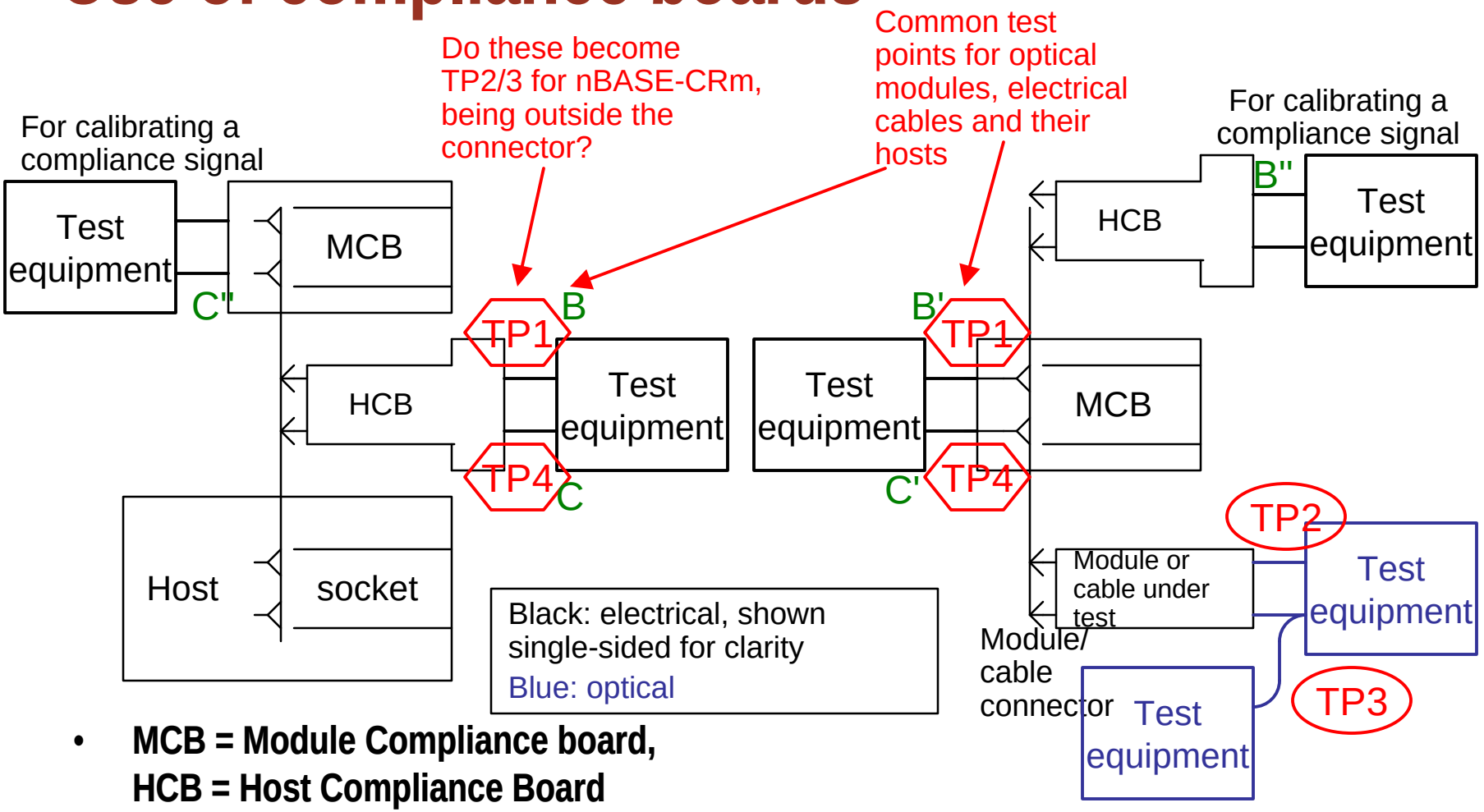
- **A PMD is not a module**
- **A PMD is not an IC**
- **A PMD extends to a connector where a signal can be measured**
  - **TP2 (optical) is 2 m beyond the connector**
  - **Electrical test points should be e.g. 1 dB before/after the connector**
    - **See next two slides**
- **PMDs are always normative**

Ends of the PMD and compliance boards



- **Need to consider both cable and PCB when choosing specs**
- **BUT the PCB is internal to the PMD, specs apply at connectors**

Use of compliance boards



- **MCB = Module Compliance board, HCB = Host Compliance Board**
- **SFP+ test points shown in green. A and D are at an ASIC/SERDES (informative)**
- **For link extension compliance points (IC testing), propose similar methodology: measure at SMA connections a defined distance from the IC**
  - **See SFP+ Appendix C.1.3**

Conclusions

- On the Tx side, 1+1 or 1+2 taps is normal
  - Do we really need to negotiate Tx emphasis with Rx?
- On Rx side, 2 to 5 tap DFE (or equivalent performance by a different implementation) is commonplace
  - Can support SMF, MMF, copper cable and backplane with the same KR class circuitry
  - Can support copper cable, 100 m MMF and 40G 10 km SMF, with same physical socket ("small module")
  - Can support all SMF, MMF, copper cable with the same physical socket ("big module")
  - Common socket goes with common test points
- FEC makes everything more robust
  - e.g. backplane and copper cable: mitigates bursts of errors
  - e.g. 40/100GBASE-SR10: mitigates random noise & jitter