# The equivalent of XAUI and other architectural considerations in 40G/100G Ethernet

Piers Dawe

Avago Technologies

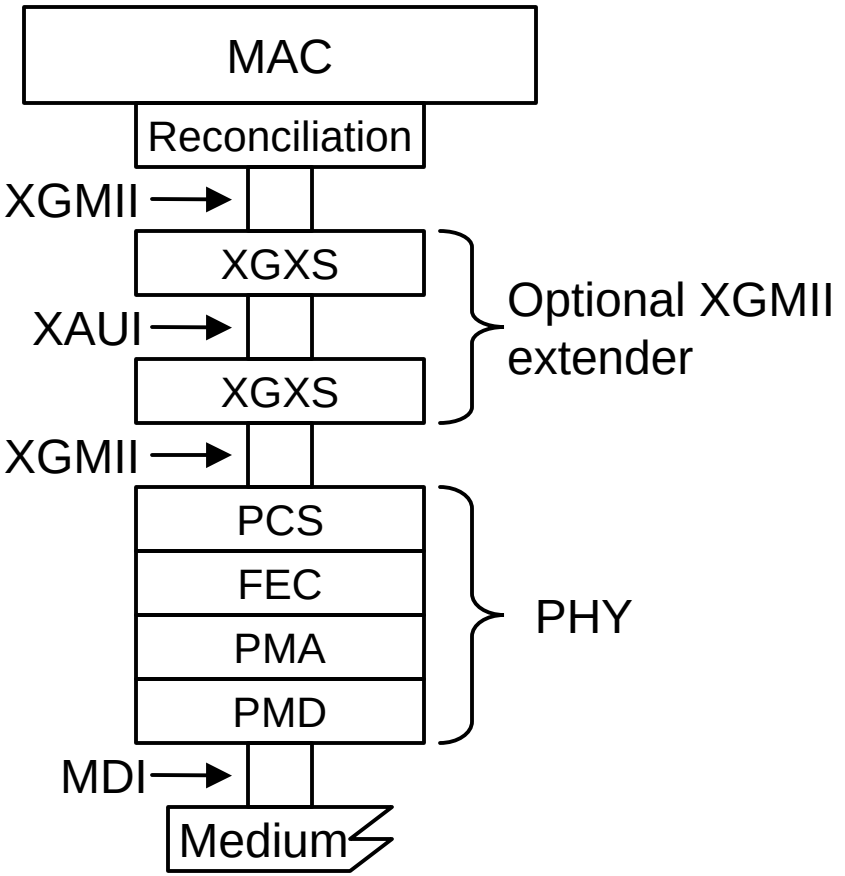
# Contents

- For those who like layer diagrams: 5 slides
- For those who like block diagrams: 4 slides including test points
  - Proposed mapping of function to sublayer names
- Use of compliance boards: 1 slide
- Compliance points and WDM: 2 slide
- Conclusions: 1 slide
- Backup

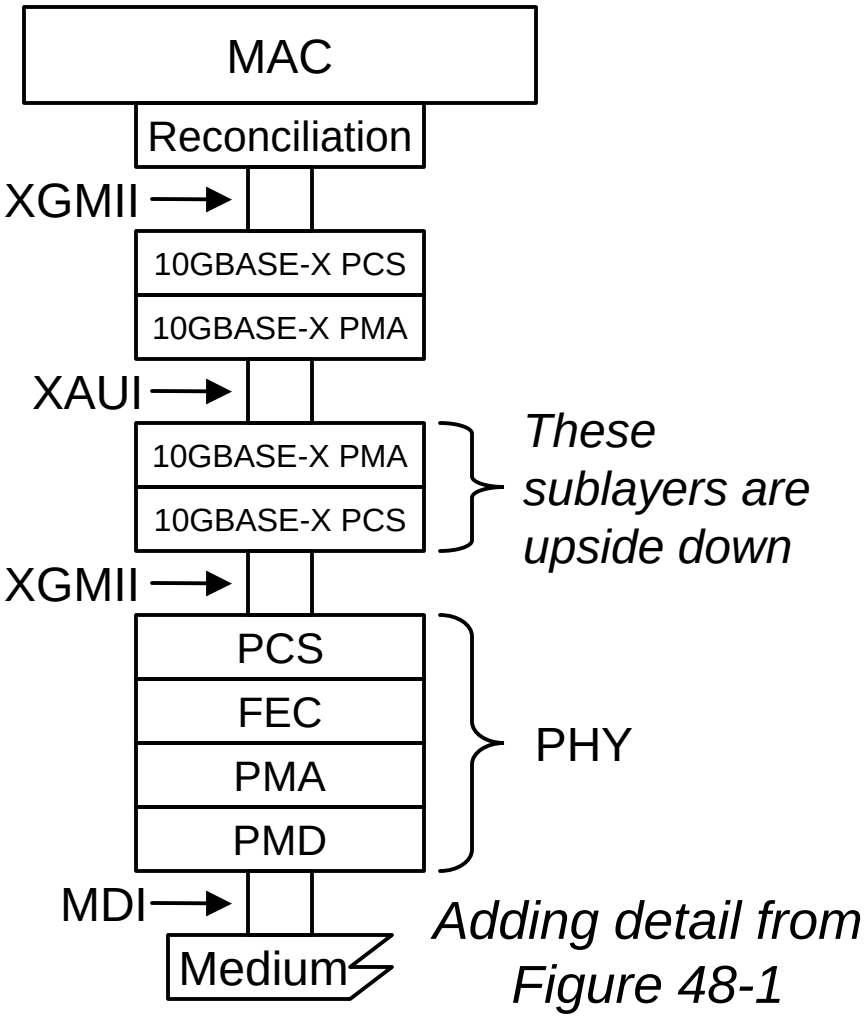
# General architecture objectives

- Architecture has to be a reasonable extraction of real-world implementation and partitioning
  - NOT force-fit reality to a pre-chosen documentation architecture
- Seek a flexible, long-lived architecture that will support evolution
  - No 10GBASE-SR/SFP+ painting into a corner!

# Extender Sublayer,802.3ae way



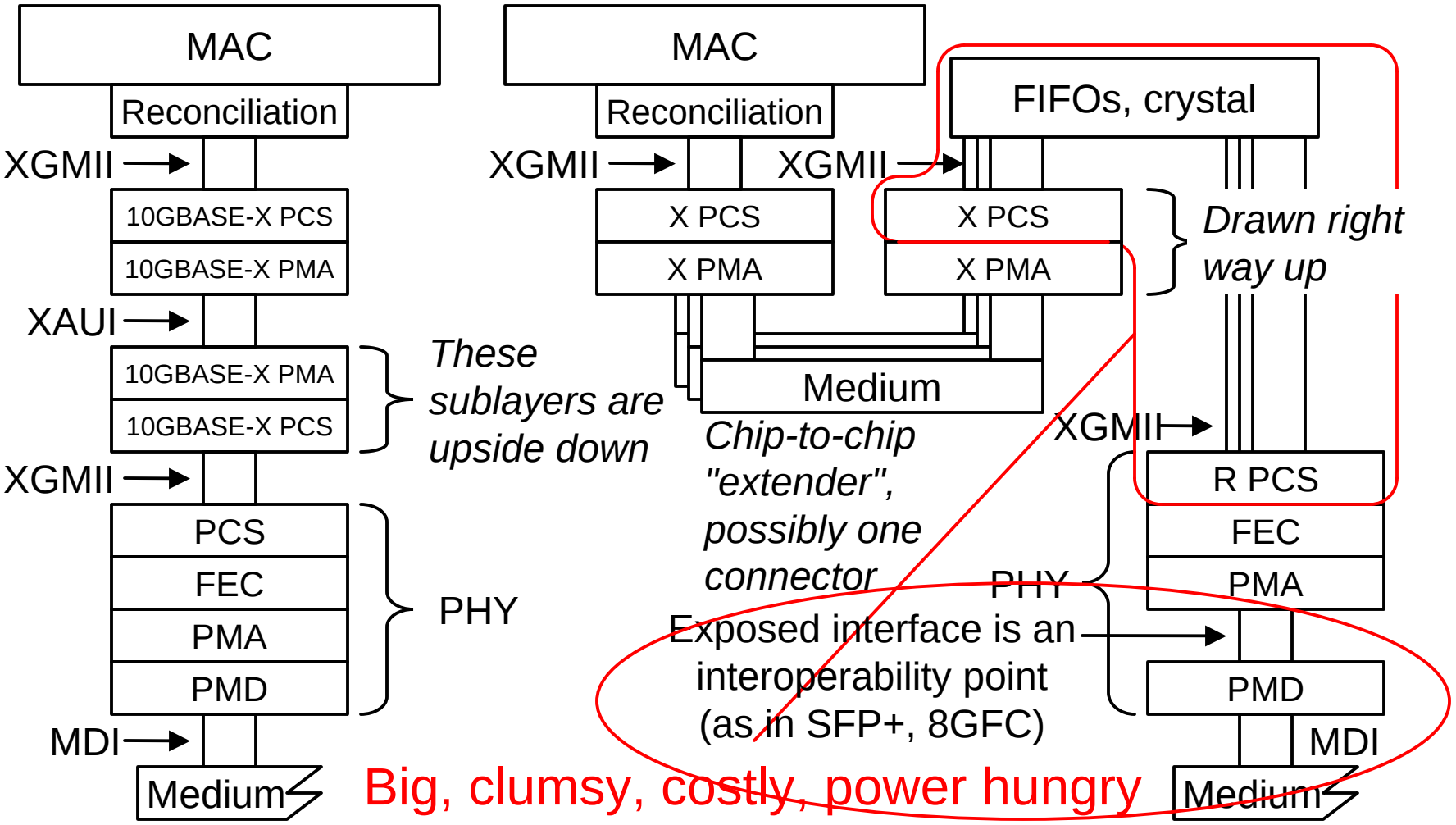

*From Figure 46-1 and Figure 74-1*



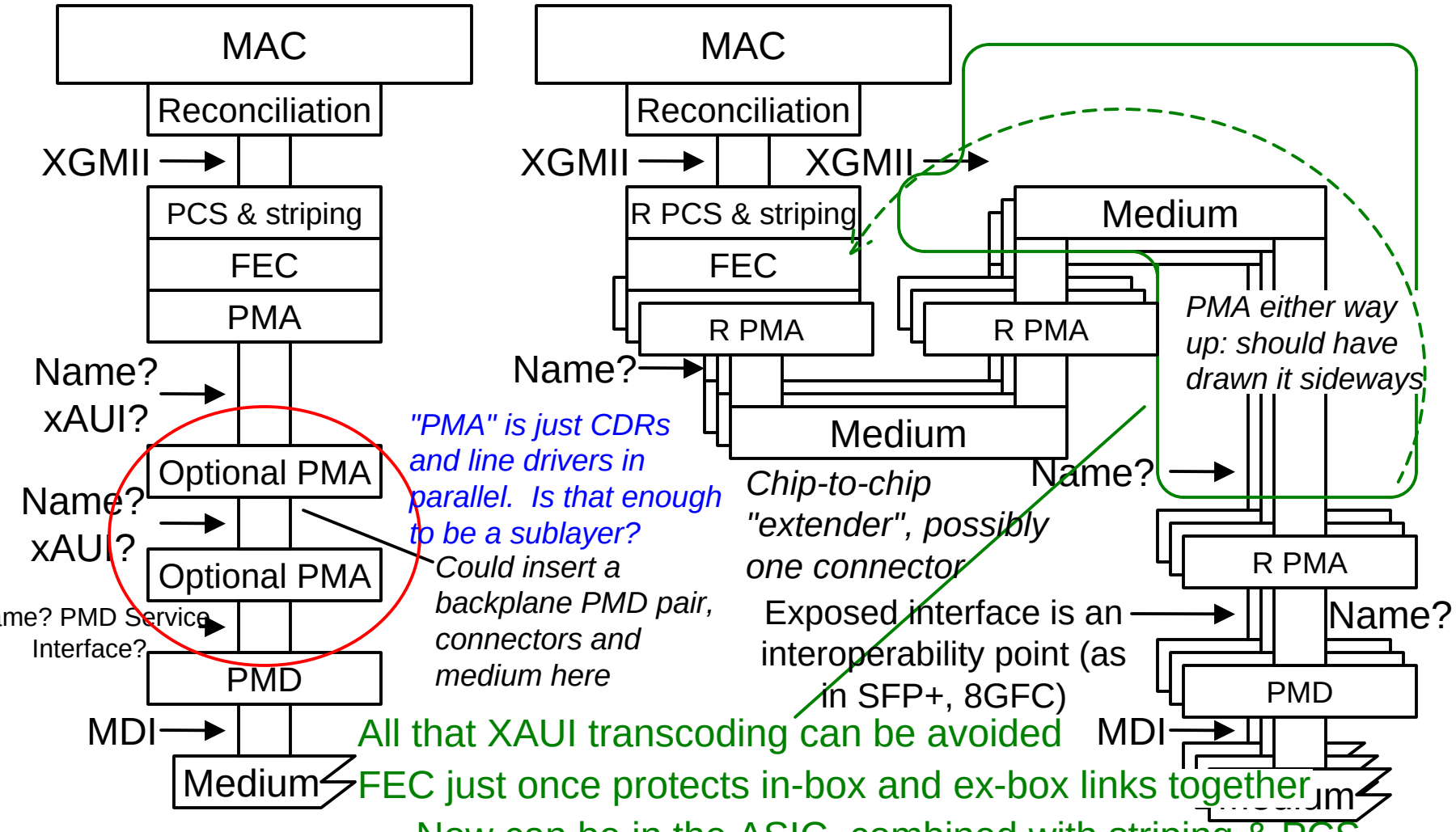

*Adding detail from Figure 48-1*

- XGXS/XAUI is 8B/10 encoded, 4 lanes
- PHY is typically 64B/66B encoded, one lane

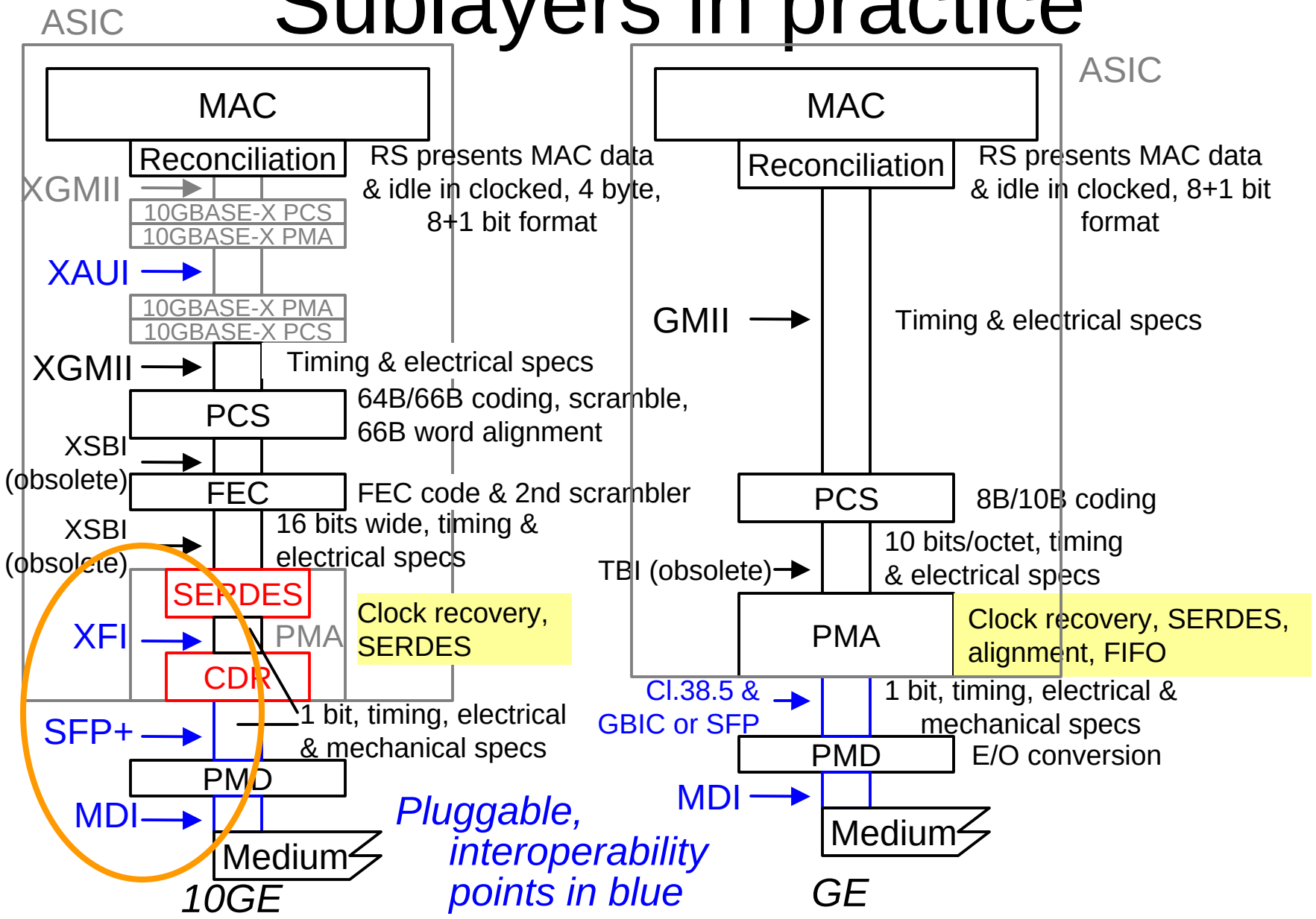
# More realistically...



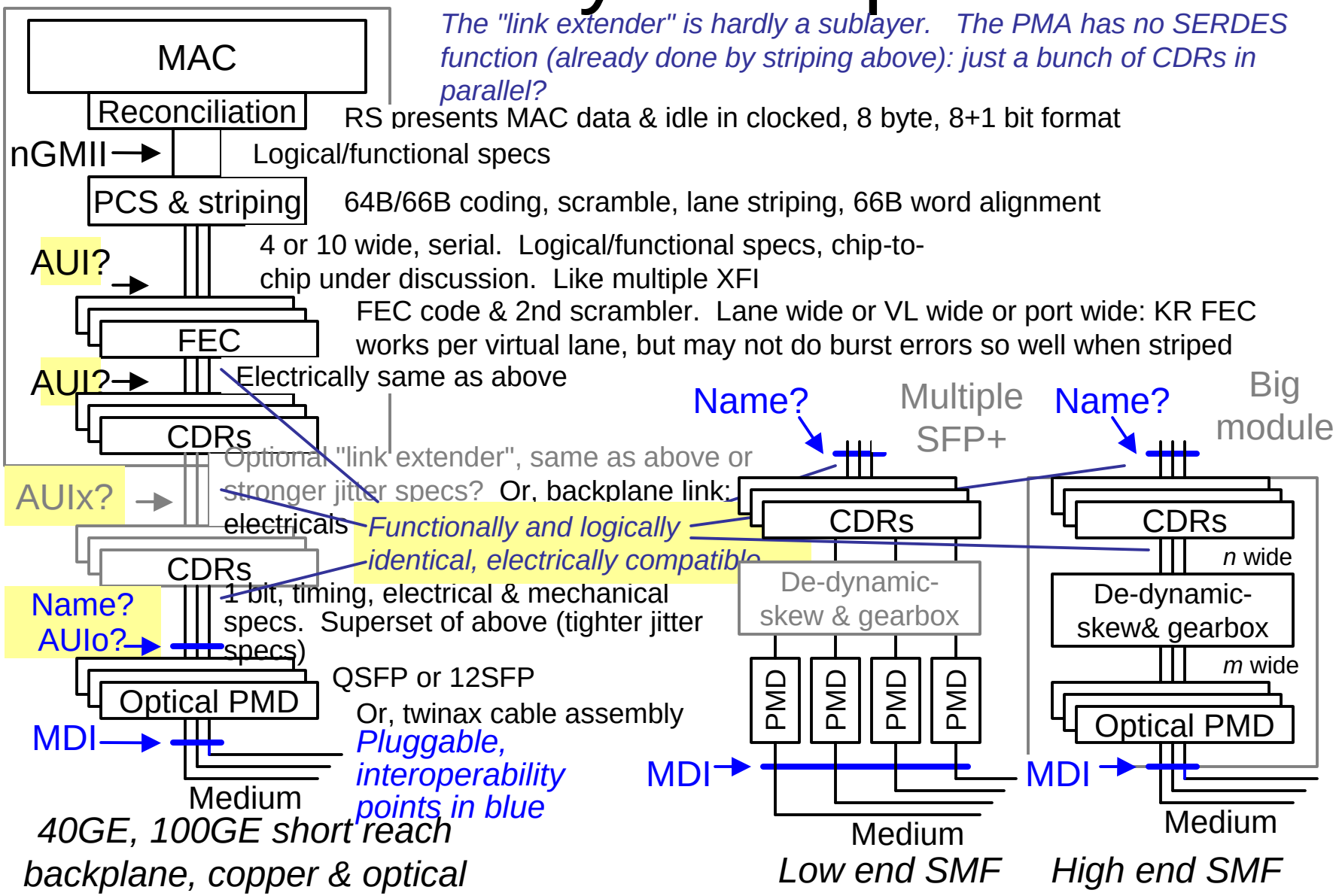
# Much simpler in .3ba



# Sublayers in practice



# HSE sublayers in practice




*The "link extender" is hardly a sublayer. The PMA has no SERDES function (already done by striping above): just a bunch of CDRs in parallel?*

# What's a PMA?

- In Gigabit Ethernet
  - Clock recovery, SERDES, alignment, FIFO
  - May? contain clock multiplier
- In 10G Ethernet
  - Clock recovery, SERDES
  - May? contain clock multiplier
- Proposed for 40G/100G Ethernet
  - Usually no SERDES function, already done by striping function associated with PCS
  - 4 versions
    - 4 wide, 10.3125 GBd: lanes can have independent timing
    - 10 wide, 10.3125 GBd: lanes can have independent timing
    - 4 wide, 41.25 GBd: lanes can have independent timing
    - 10:4 gearbox: has to remove "dynamic" skew.  use any lane for clock recovery?
    - For the future: 4:1 or 10:2 gearboxes
    - Each with 4 grades of electrical spec: see another slide
  - Generally doesn't need clock multiplier: "line timed", clock recovered from signal
  - If feasible, should allow for up to 6 line-timed PMAs in series (10G XFP scenario has a minimum of 2)
  - Does the 20:10 or 20:4 mapping of virtual lanes to physical lanes count as a PMA?  Or a PMA like function within the PCS?
- See Gustlin for a presentation on PMA functionality

# What's an AUI?

- per Clause 7, it is the interface above a PMA
- Previous generations have had MII, GMII, XGMII denoting speed as well as type
- But have not done the same with PMA, PMD: no "XGPMA"
- Propose avoid the clutter and call them all "AUI"
  - Might add suffixes e.g. AUIx for link extender type, AUIo for unretimed optical TP1/TP4,
    - AUIb for backplane, AUIc for electrical 10m cable
    - But the last two are already called "medium" already?
- Similarly, just have "MII" rather than "XL/CGMII"?

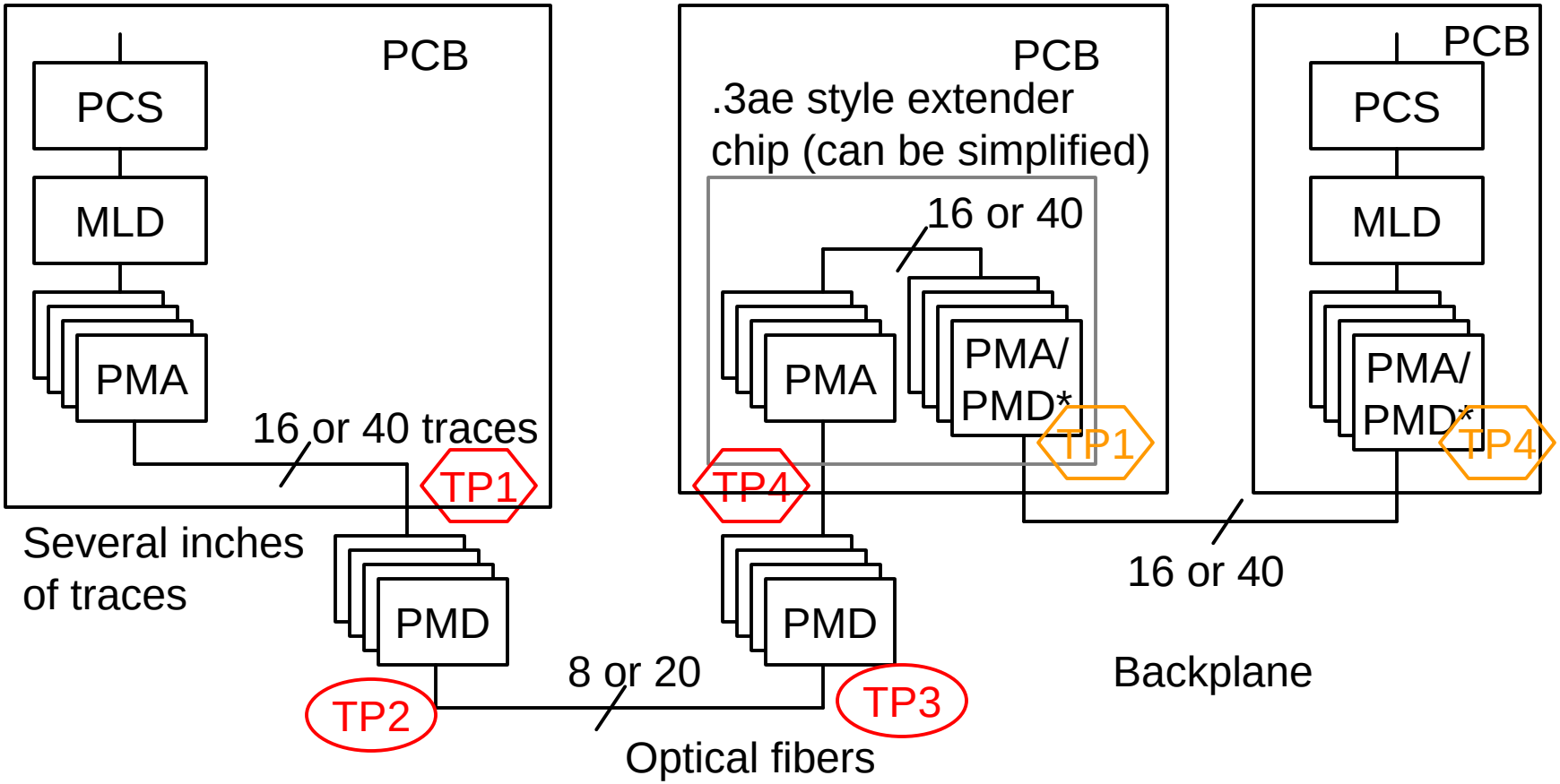
# Summary of slides 4-10

- In-box links and most external links (electrical cable, 100 m optical, 4WDM for 40G SMF) have the same number of physical lanes everywhere
  - 4 lanes for 40G, 10 lanes for 100G
    - (the 4:10 gearbox for 100G SMF is associated with/is inside the 100G optical module)
  - All accessible electrical lanes are 10G (but provide for 25G 2nd generation)
  - Most HSE PMAs, PMDs and media are groups of several PMAs, PMDs and media in parallel
- The scenarios that used XAUI at 10G can be addressed MUCH more cleanly in HSE
  - No transcoding
  - No FIFOs, no crystal, no SERDES function
  - No intermediate deskew
  - No bit manipulation at all!
- In HSE, the PMA doesn't have to do SERDES function: already done in PCS/striping sublayer above
  - A PMA is a group of CDRs in parallel, with appropriate simple EDC e.g. line drivers
  - Sublayer naming using PMA and AUI proposed
- FEC done just once can protect all the way from one ASIC to another
  - FEC is now in the right place, in the digital ASIC not near the optics
- The PMD service interface, being pluggable, is an interoperability point that must be fully (analog) specified

# Block diagrams

- The next few slides and five more in the backup show similar material but in block diagram style
  - First a diagram showing a parallel-optics link with backplane extension with test points, then a more realistic representation of the extender chip showing vendor domains and connectors, then TP0, TP5

# Parallel optics and backplane, with test points

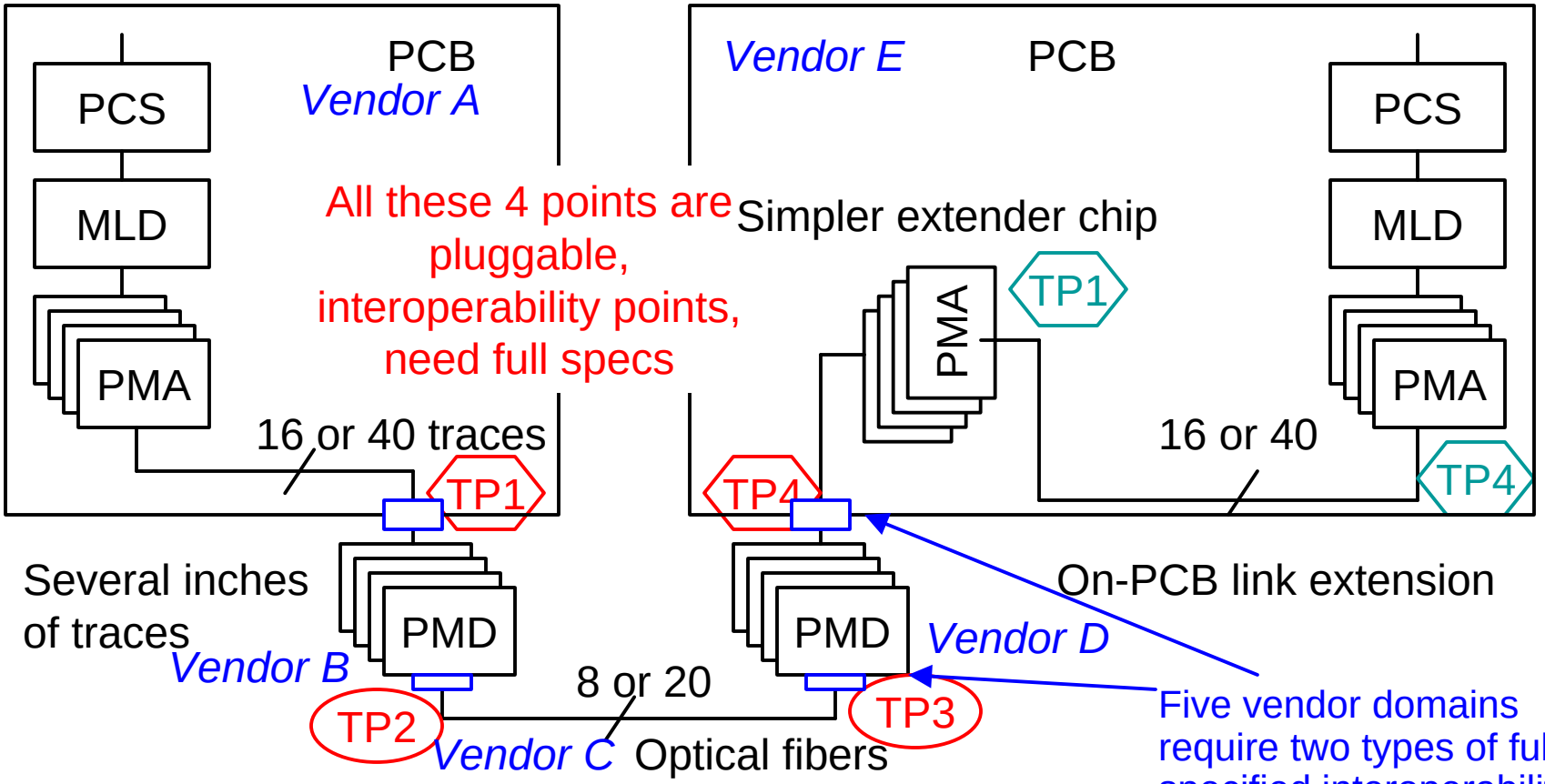



For optics, PMA is a set of CDRs, possibly with simple EDC: does not mux or demux

* In 10GBASE-KR, the PMD performs a similar function to an optics PMA

- Connectors not shown – see later

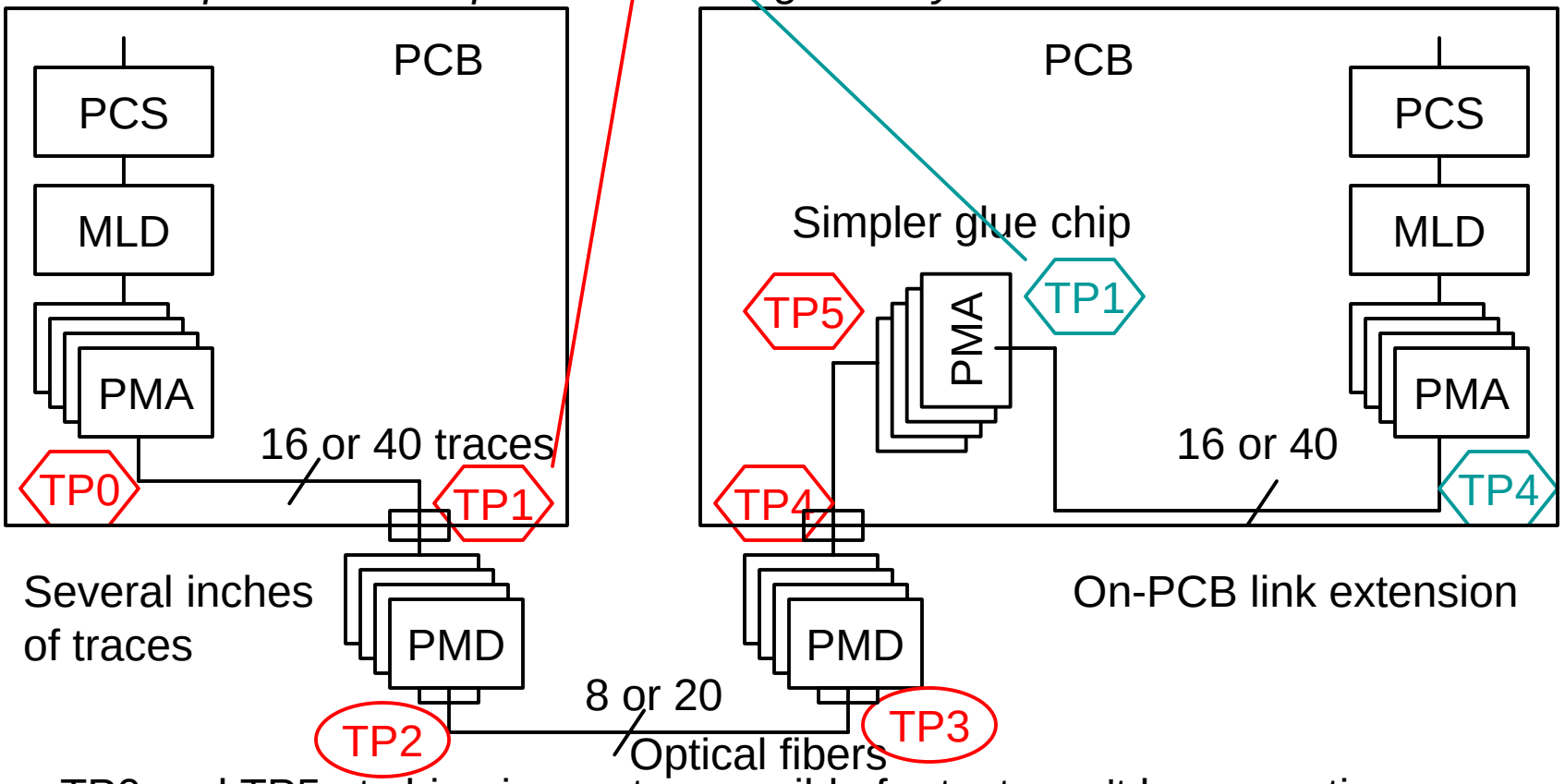
# Parallel optics and in-box link extenders ("AUI" or "XLAUI") real way - connectors




- TP1 and TP4 are at the module electrical connector
- TP2 is 2 m after the module optical transmitter
- TP3 is the output of the optical fiber
- For link extension, propose TP1 and TP4 relate to SMA IC pads

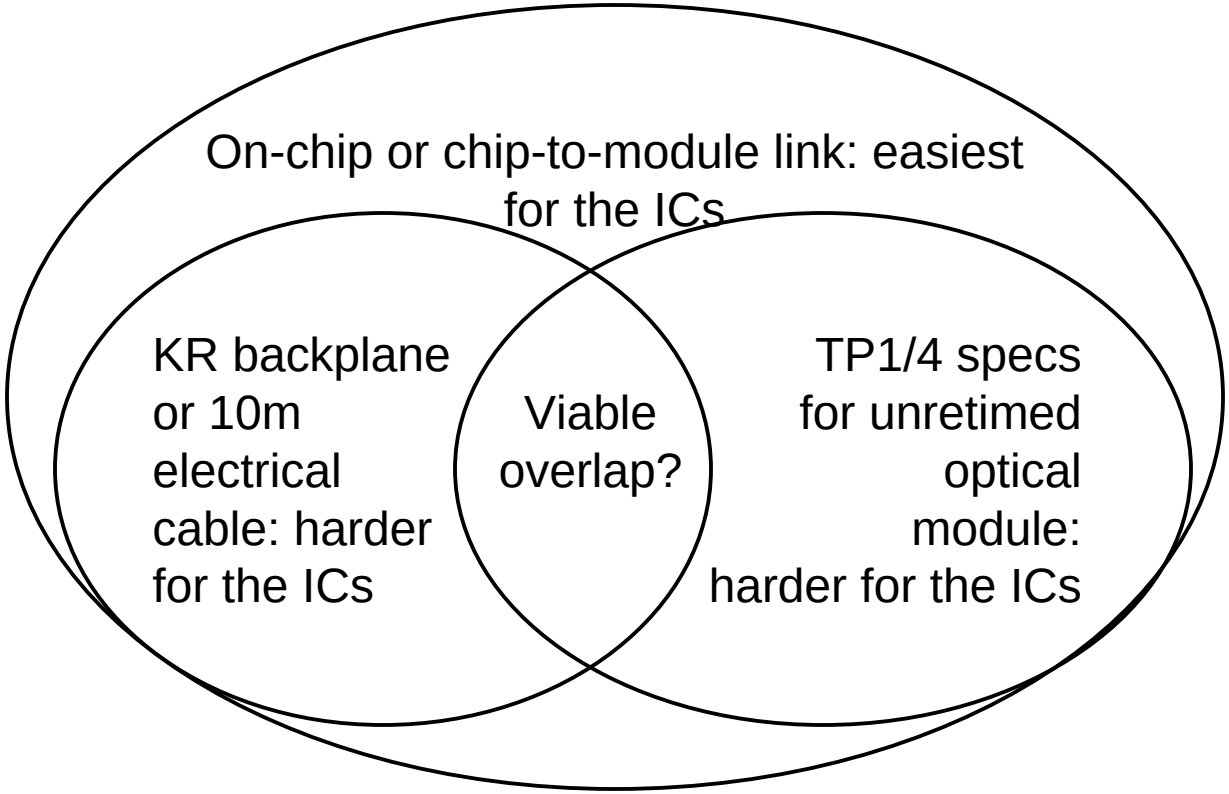
# TP0 and TP5

*These points must have different electrical specs because different noise environment*
*Compliance with optical TP1/TP4 generally harder than on-PCB TP1/TP4*




- TP0 and TP5 at chip pins: not accessible for test, can't be normative
- Jitter budget for these three things differ:
  - On-PCB link extension
  - Backplane link (might be same/similar as 10 m electrical cable)
  - Link containing optics

# Jitter/electrical specs




- Digitally and functionally, all these interfaces are the same
- See Petrilla, Ghiasi and Di Minico for proposed electrical and cable specs

# Summary of slides 13-16

- As before, and
- TP1/TP4 for an optical module must have different electrical specs to TP1/TP4 for a chip-to-chip link ("AUI" or "XLAUI", "CAUI") because different noise environment
  - Compliance with optical TP1/TP4 generally harder than on-PCB TP1/TP4
- Backplane specs also harder than chip-to-chip specs
  - Backplane PMD seems to be functionally equivalent to .3ba PMA for optical, although different analog specs

# Four grades of AUI, or PMD service interface, in words

- In decreasing order of need of standardization

A. Unretimed optical PMD attach (TP1/TP4)
  - Pluggable multi-vendor interoperability point from day 1
    - Unlike 10GE when it came later and we didn't plan well enough for it
    - Specs in between FC-PI-4 ("easy") and SFP+ ("difficult")
    - More demanding than version D because optical link has significant random noise

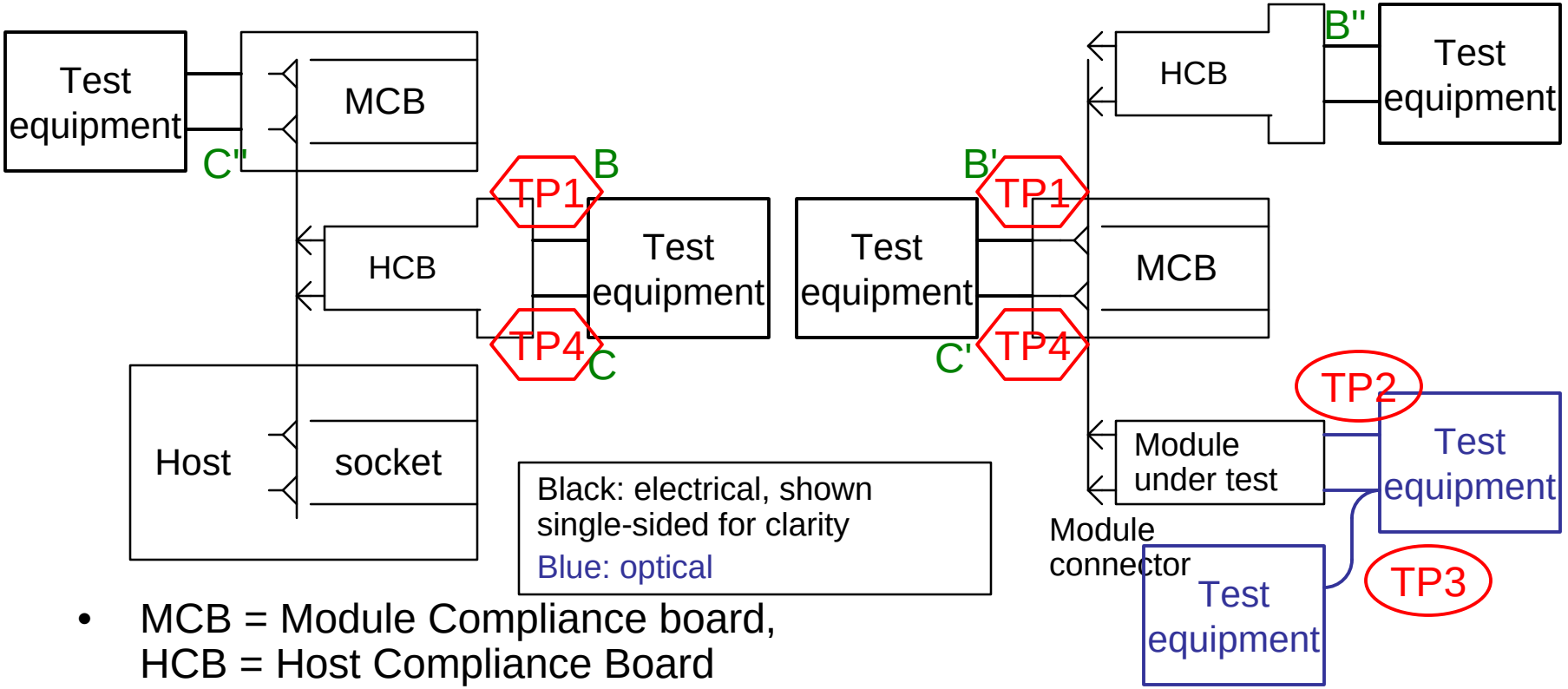
B. (Optical TP2, TP3
  - Ethernet always does these)

C. Electrical cable attach (TP1,2,3,4?)
  - Pluggable multi-vendor interoperability point from day 1
  - Physically the same point as A above in many implementations
    - Specs can learn from Backplane Ethernet, FC-PI-4 and SFP+
    - More demanding than version D because substantial frequency-dependent loss
    - **Same connector as A**
    - Little point re-using CX4 connector as an old CX4 cable won't have the specs for HSE

D. Basic "chip-to-chip"
  - Probably to include one module connector to connect to a retiming module
    - Chip to chip: common spec for IC procurement, less necessary for standardization
    - But chip to module is a pluggable multi-vendor interoperability point
    - Least demanding analog/jitter specs

E. Backplane
  - It turns out that the test points are placed so as not to ensure multi-vendor interoperability but to allow a common spec for IC procurement
    - Hence less necessary for standardization
    - More demanding than D because frequency-dependent loss, reflections and crosstalk

# 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

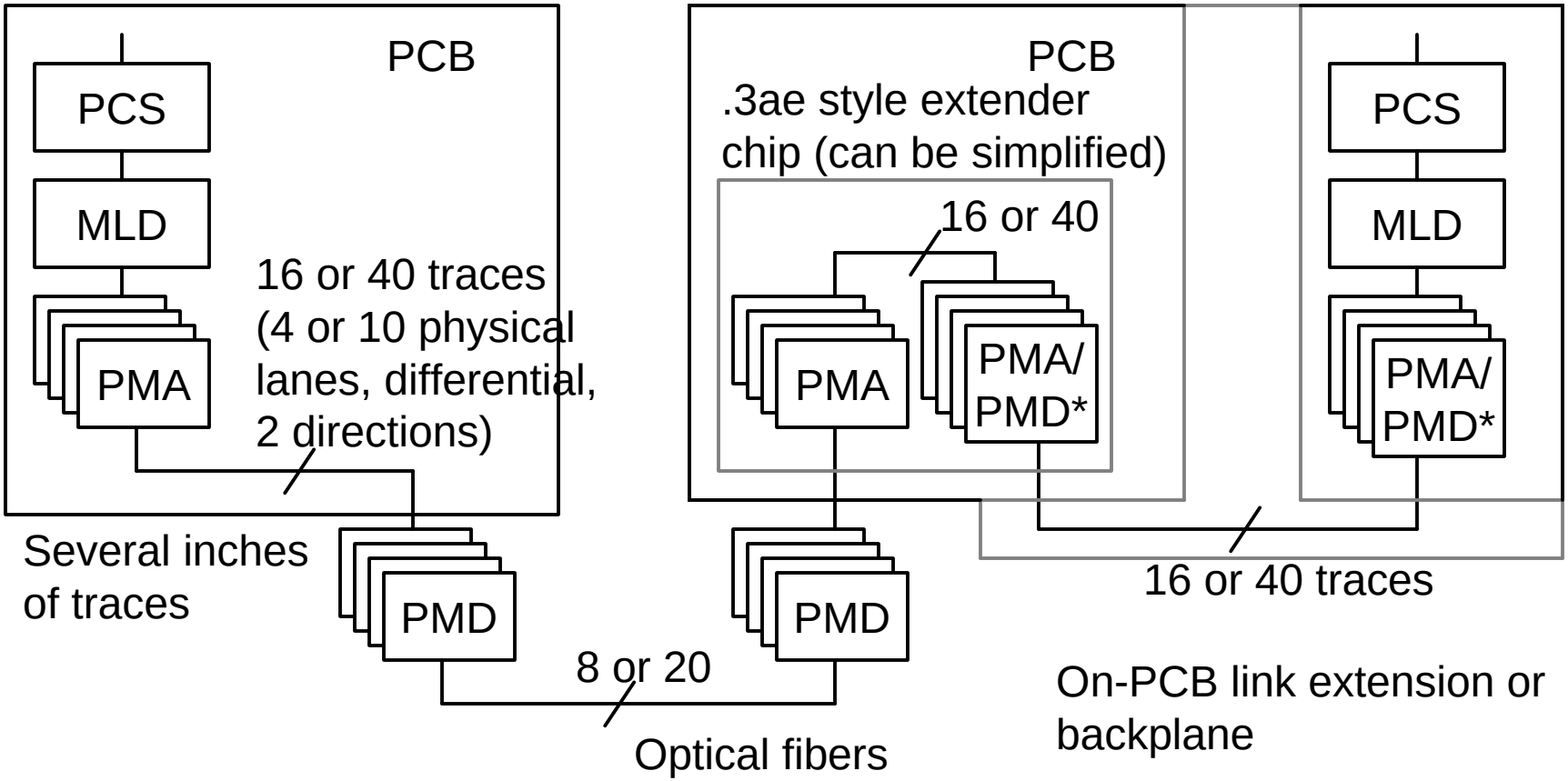
# Conclusions

- A well-designed striping scheme allows huge improvements in avoiding big, costly, power hungry XAUI-to-64B66B transcoding with its FIFOs and crystals
- Line coding can be done just once for multiple hops
- The replacement for all that XAUI stuff is a bunch of CDRs
  - Each lane is independent: no deskew or any bit manipulation needed at intermediate retimers
- FEC just once can protect the multiple hops, should be used everywhere
- An identical logic/management/digital port in a MAC ASIC can support ALL port types
- The optical module or electrical cable electrical connector is a pluggable interoperability point that must be fully (analog) specified

# Backup

- More scenarios that architecture covers

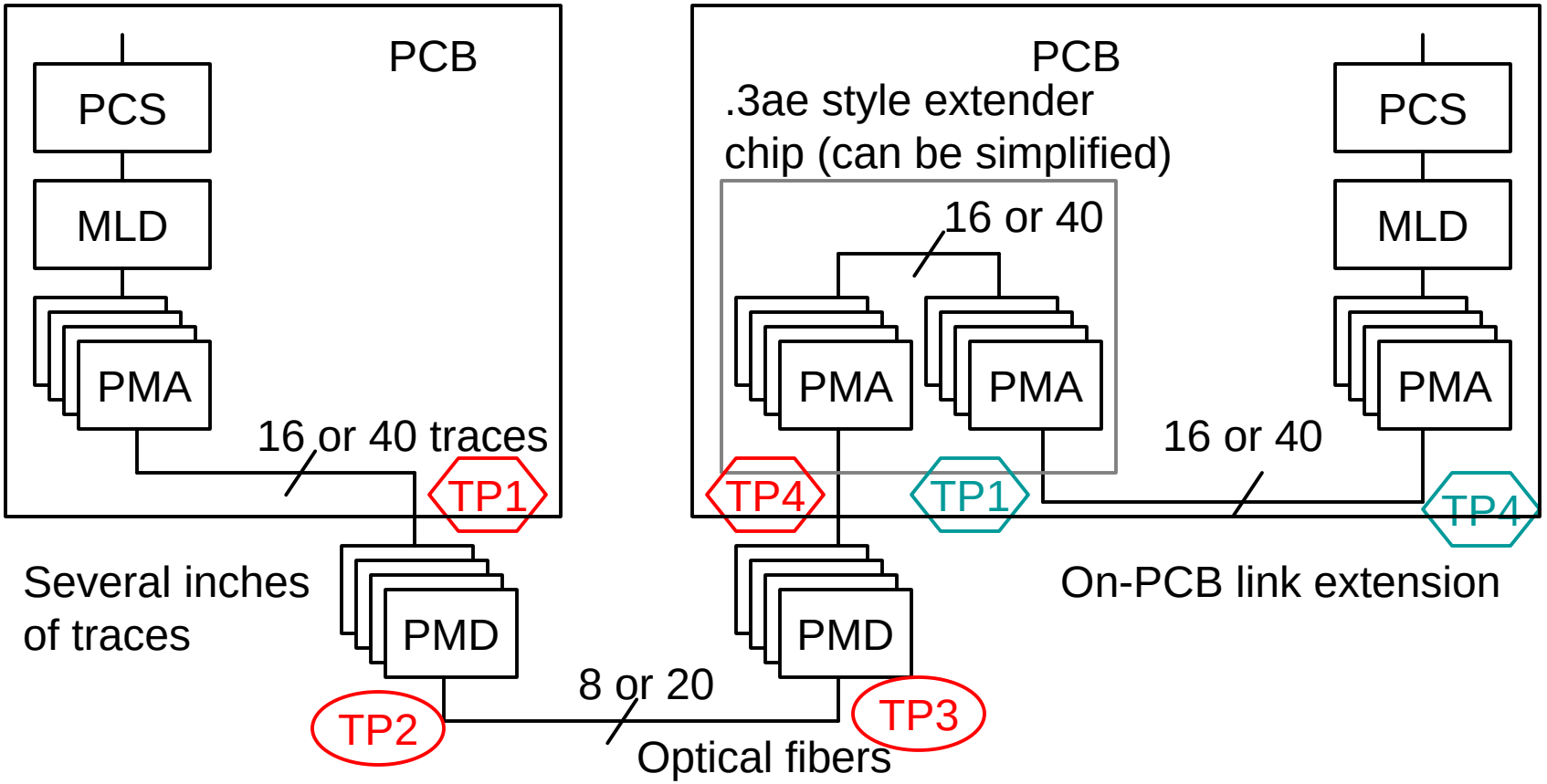
# Parallel optics compatible with in-box or backplane link extenders, block diagram




For optics, PMA is a set of CDRs, possibly with simple EDC: does not mux or demux

* In 10GBASE-KR, the PMD performs a similar function to an optics PMA

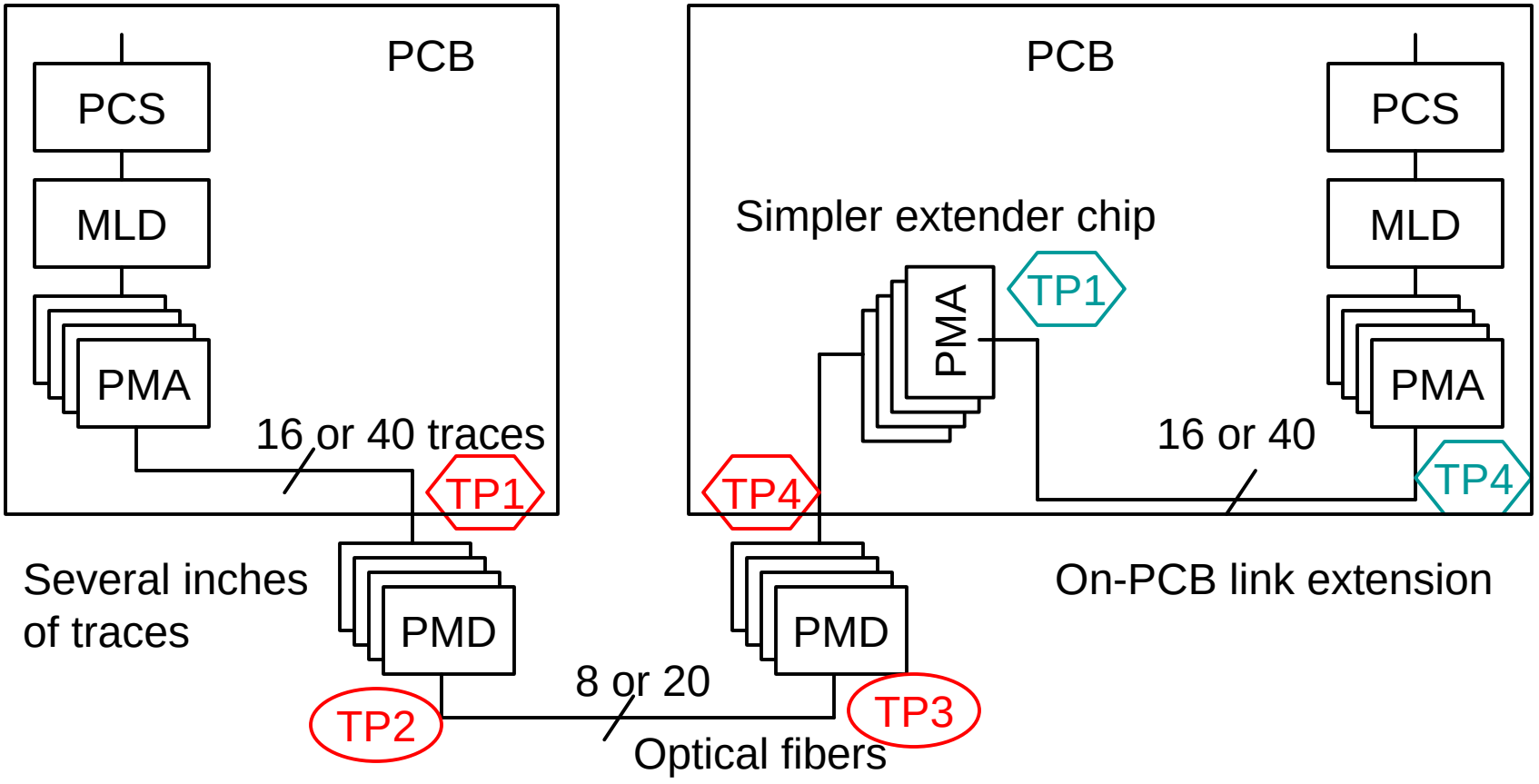
# Parallel optics and in-box link extenders ("XLAUI") clumsy way




For optics, PMA is a set of CDRs, possibly with simple EDC: does not mux or demux

- Connectors not shown – see later

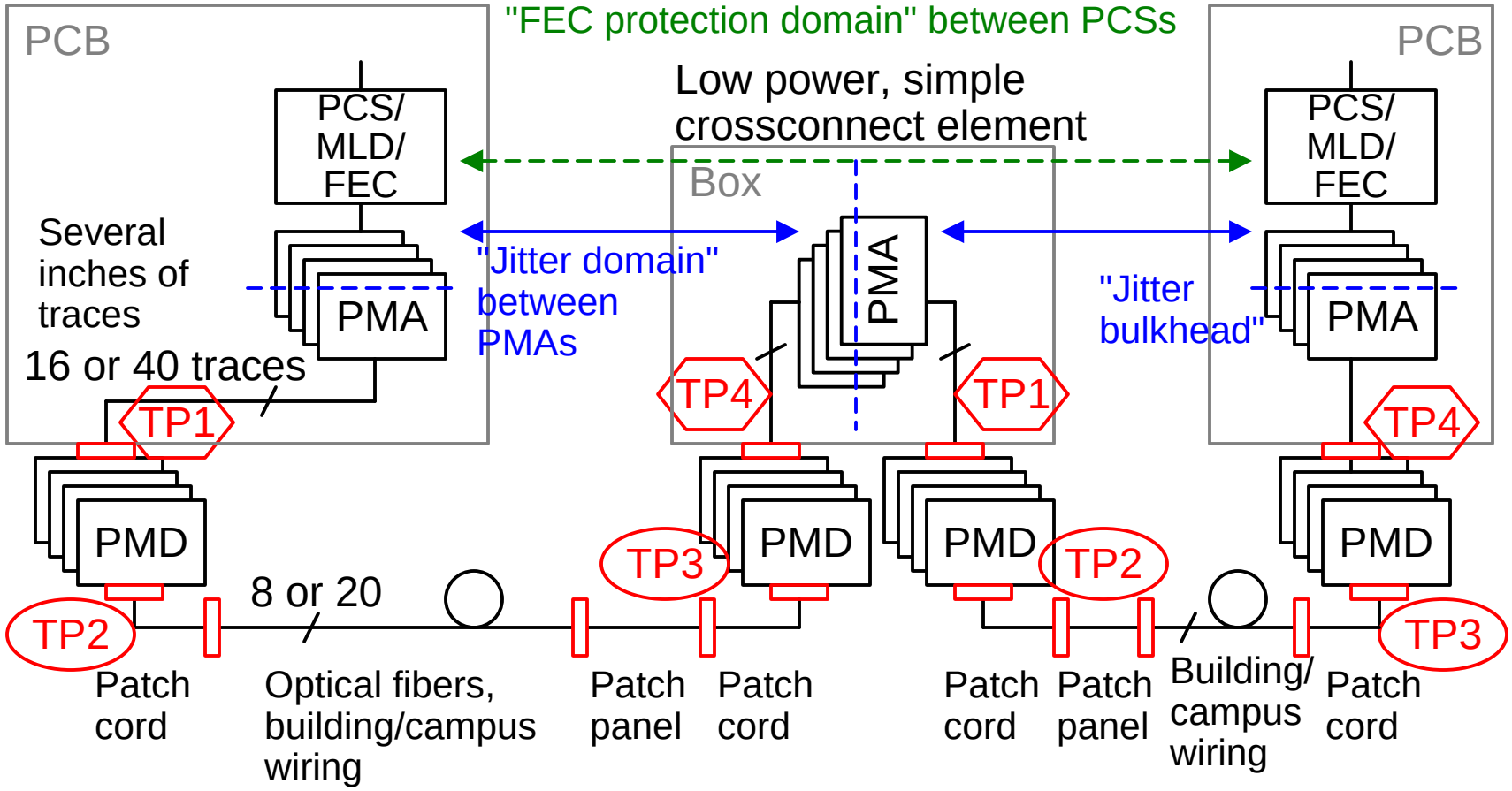
# Parallel optics and in-box link extenders ("AUI" or "XLAUI") real way




For optics, PMA is a set of CDRs, possibly with simple EDC: does not mux or demux

- Connectors not shown – see later

# Jitter and FEC domains with link extender example

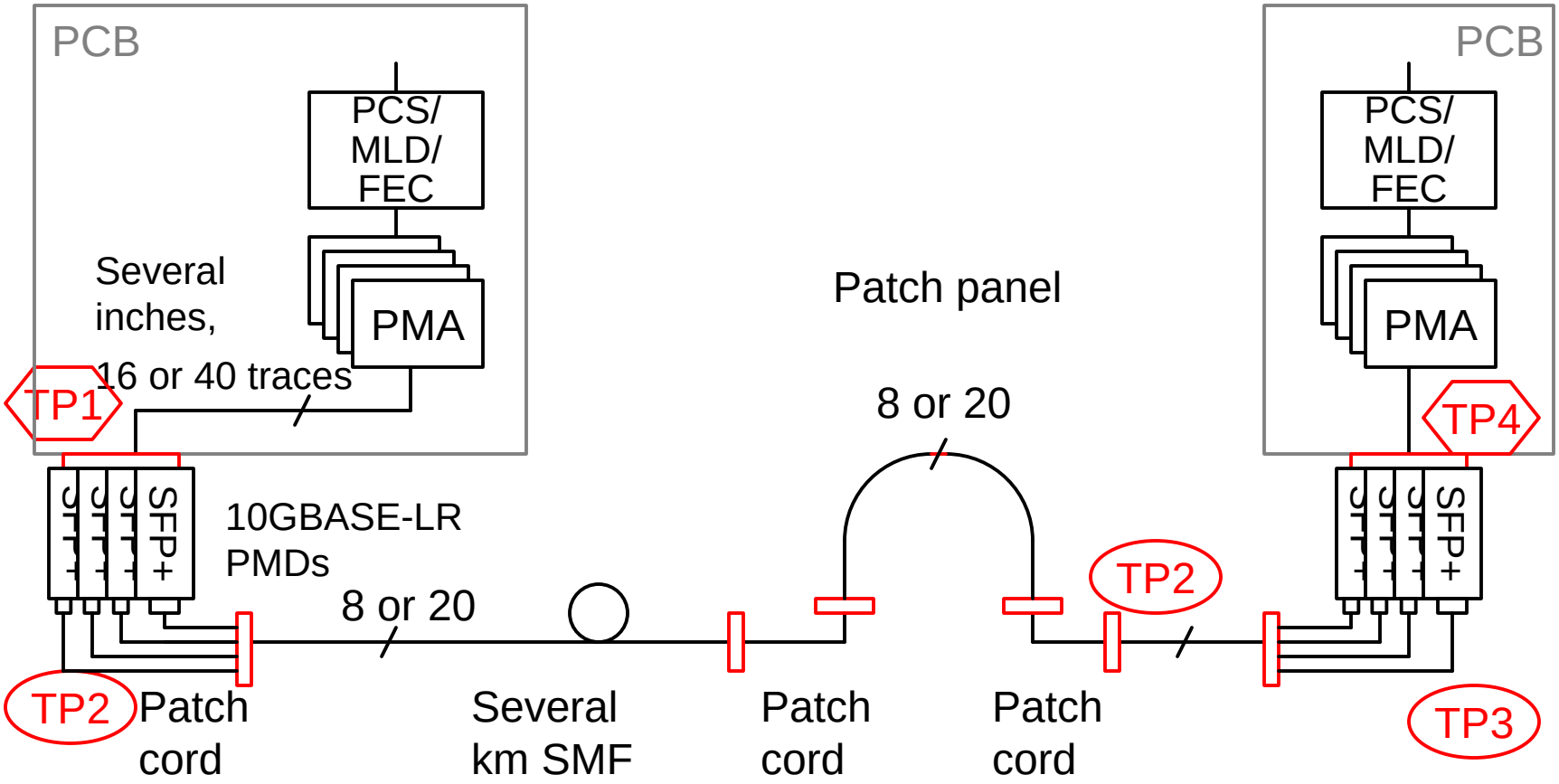



Pluggable connectors and test points shown in red

Cabling doesn't have to be HSE-aware, just not too mismatched in length

PMA in centre is a bunch of CDRs: "jitter bulkhead" resets jitter budget

# Time-to-market SMF if enough fibre available

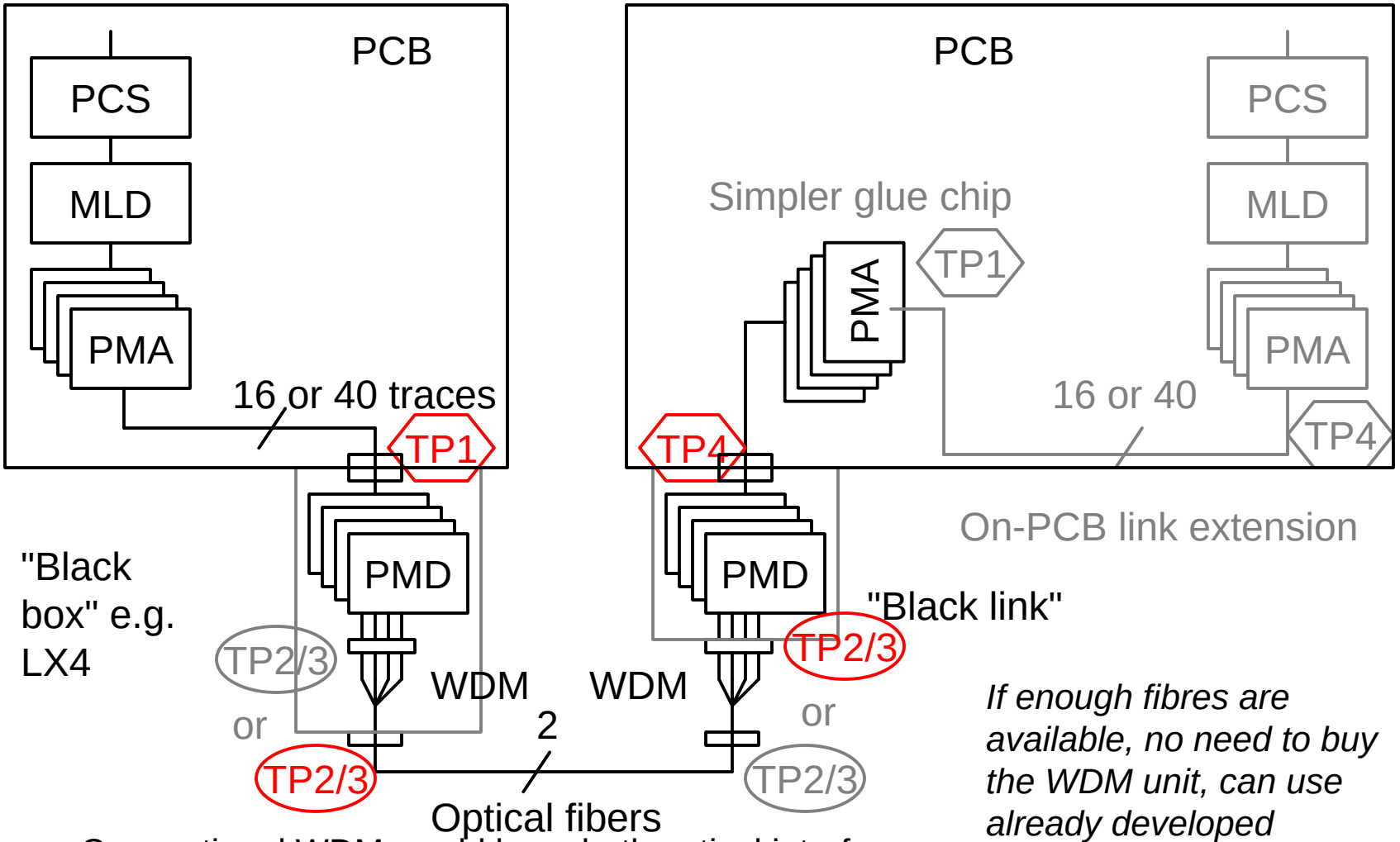



Pluggable connectors and test points shown in red

Cabling doesn't have to be HSE-aware, just not too mismatched in length

Alternatively could use WDM dongles and "coloured" SFP+ or XFP modules

# WDM optics: two scenarios



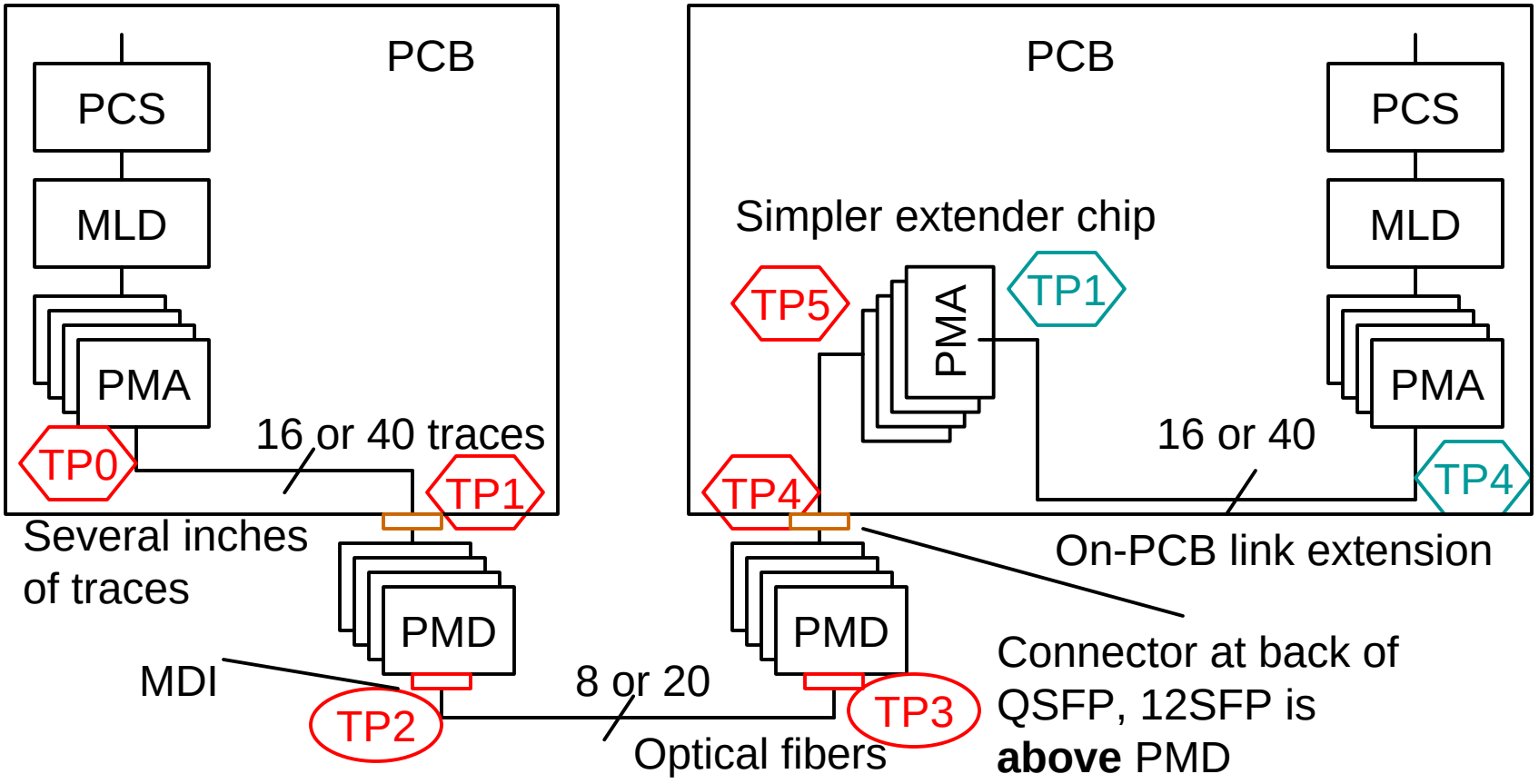

- Conventional WDM would have both optical interfaces
- WDM unit is small, can be inside a patch cord

*If enough fibres are available, no need to buy the WDM unit, can use already developed 10GBASE-LR modules*

# Black box or black link?

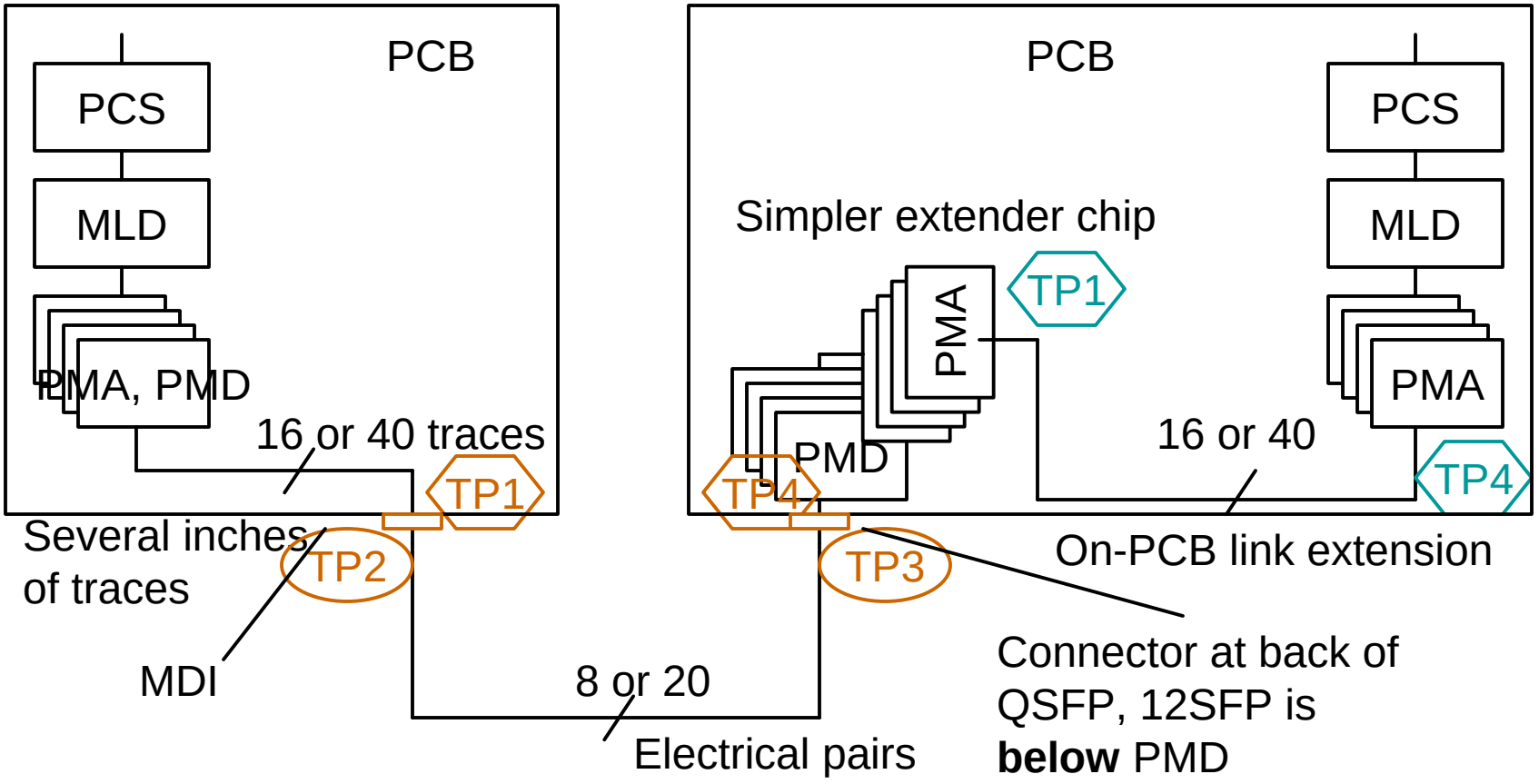
- Black link allows "pay as you grow"
- Defining both test points (either side of WDM) allows each end to be of each hardware partitioning type
  - Downside is that WDM mux/demux is specified with no trade-off with other components
- A generic, long-lived architecture should not exclude either

# Parallel optics and in-box link extenders




PMA is a set of CDRs, possibly with simple EDC: does not mux or demux

# Parallel electrical and in-box link extenders




PMA is a set of CDRs, possibly with simple EDC: does not mux or demux