Beyond 400 GbE Project Priorities for Data Center Networks

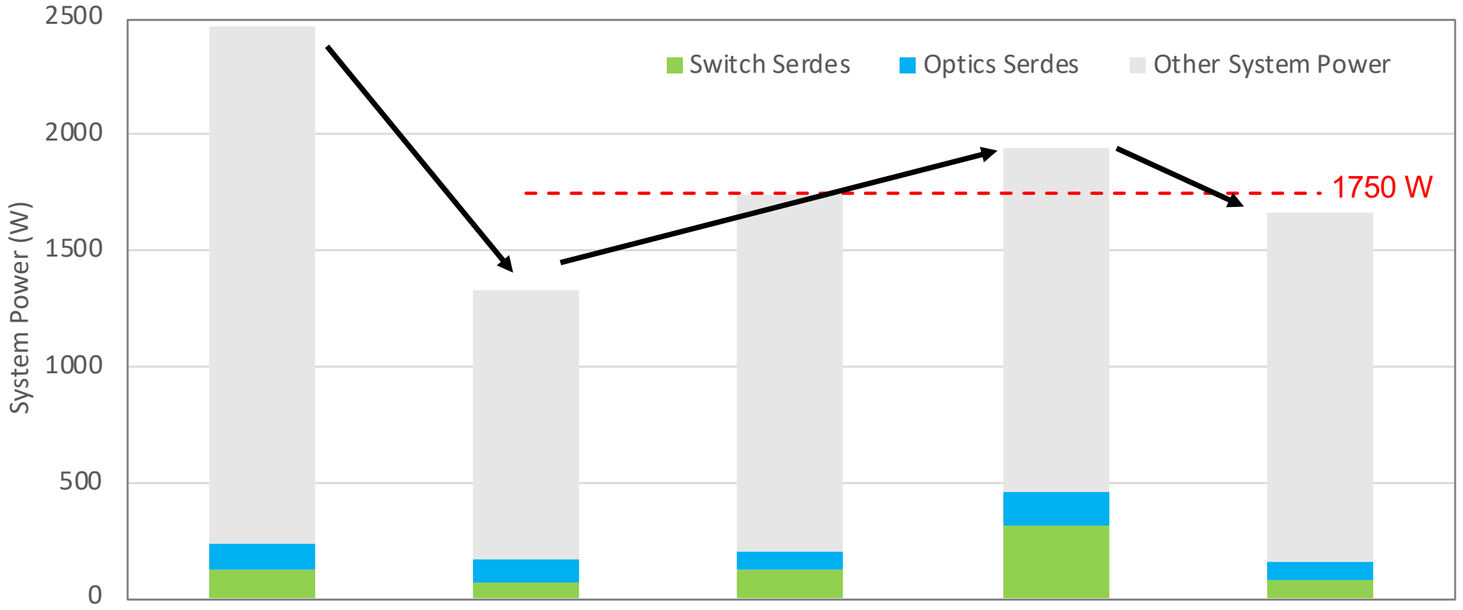
Rob Stone
Facebook

FACEBOOK Infrastructure

Data Center Challenges

- Limitation in data center network power budget
- Scaling (bandwidth demand from new workloads)
- Flexibility – operationally efficient backwards compatibility
  - i.e. ability to run existing devices in “downspeed” mode to connect to legacy equipment

Power Reduction Motivation



Bandwidth	12.8T	12.8T	25.6T	51.2T	51.2T
Codename	Backpack	Minipack	Minipack 2	Next-Gen	CPO Switch
System Architecture	12 Chip, LR	Single Chip, LR	Single Chip, LR	Single Chip LR	Single Chip, XSR
Optics Form Factor	128 x 100G QSFP-28	128 x 100G QSFP-28	128 x 200G QSFP-56	128 x 400G FPP	CPO

Front-Panel Pluggable Optics

Co-Packaged Optics

Backpack

Minipack

- Co packaged optics (CPO) provides the next big step in power reduction

Data Center Challenges: Implications for B400G

Limitation in data center network power budget

- Optimize for Efficiency
  - PCS / FEC Architectures (end – end vs by segment vs concatenated)
  - Introduction of Co-packaged Optics

Scaling (bandwidth demand from new workloads)

- New Speeds
  - MAC & PMD rates (800 GbE)
  - Lane Rate (switch ASIC wiring / escape)

Flexibility

- Backwards compatibility
  - Preserve ease of compatibility (e.g. retain prior generation optical wavelength grid)
- Preserve interoperability between implementations
  - Co-packaged and front-panel pluggable optics
  - Active / Passive Copper Cables (?)

Intra Data Center Network Evolution

FB Initial Deployment Year / Status	Port Speed (Gb/s)	Electrical Lane Speed (Gb/s)	Switch Silicon Bandwidth (Tb/s)	Switch System Configuration (Radix)	Major Optical PMD	Optical Lane Speed (Gb/s/λ)	Optical Module Type		
							Front Panel Pluggable Optics	On Board Optics	Co-packaged Optics
2016	40	10	1.28	128 x 40GbE	40GBASE-LR4	10	QSFP+	-	-
2018	100	50	12.8	128 x 100 GbE	100G-CWDM4 (OCP)	25	QSFP-28	Mini-Photon	-
2021	200	50	25.6	128 x 200 GbE	200G-FR4	50	QSFP-56	Next Gen OBO	-
Planning – 2023	400	100	51.2	128 x 400 GbE	400G-FR4	100	TBD	Next Gen OBO	CPO Gen 1 (XSR)
Exploration	800	200	102.4	128 x 800 GbE	800G-FR4	200	-	-	CPO Gen 2
Exploration	1600	??	204.8	128 x 1600 GbE	?				

- Goal: Preserve switch radix gen over gen while scaling port bandwidth
- Re-use existing fiber, power, cooling and physical infrastructure to enable “rolling upgrade” with minimal disruption

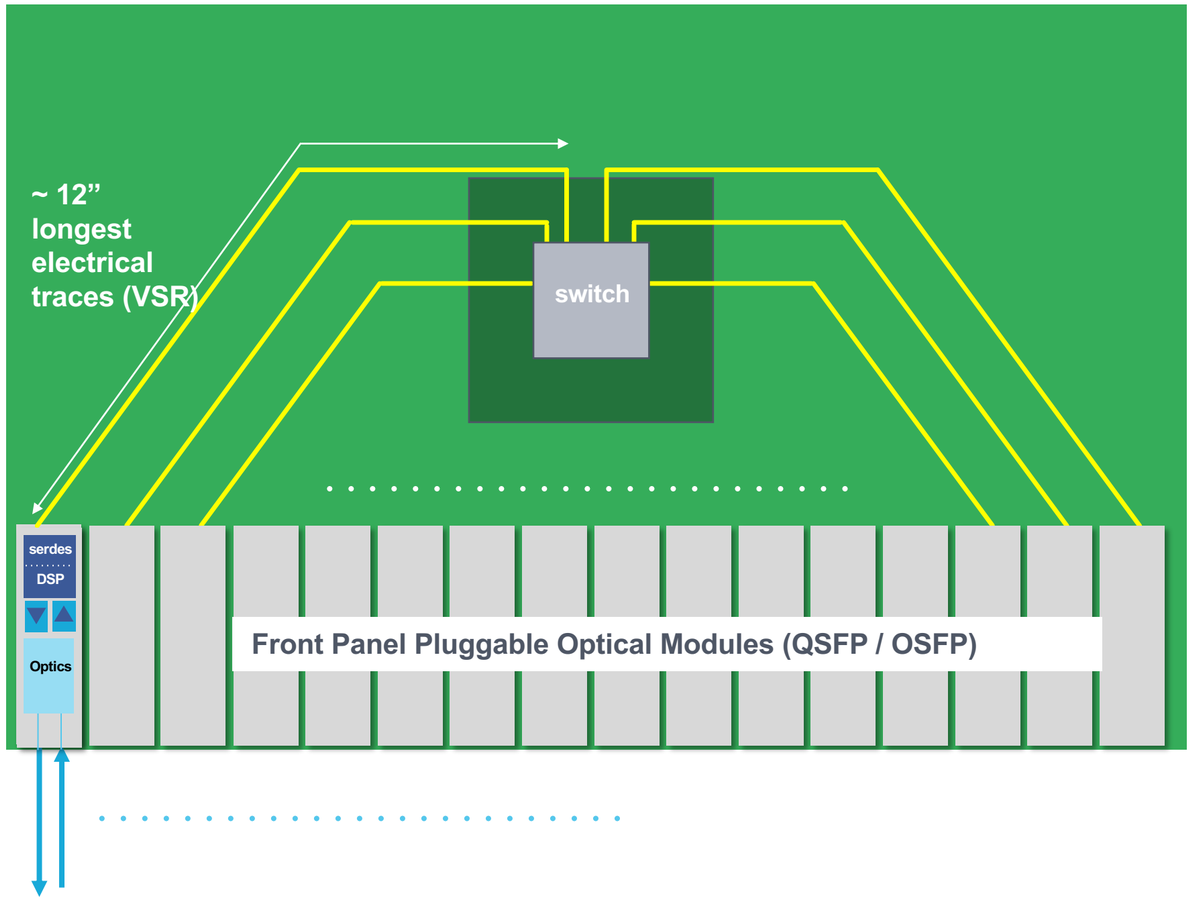
Backwards compatibility

- Not a typical IEEE objective, but required for operational reasons
- Tiers / regions of network may be upgraded at different times
- Example – operation of a 400GBASE-FR4 capable transceiver as 200GBASE-FR4 (400GAUI-4 → 200GAUI-4)

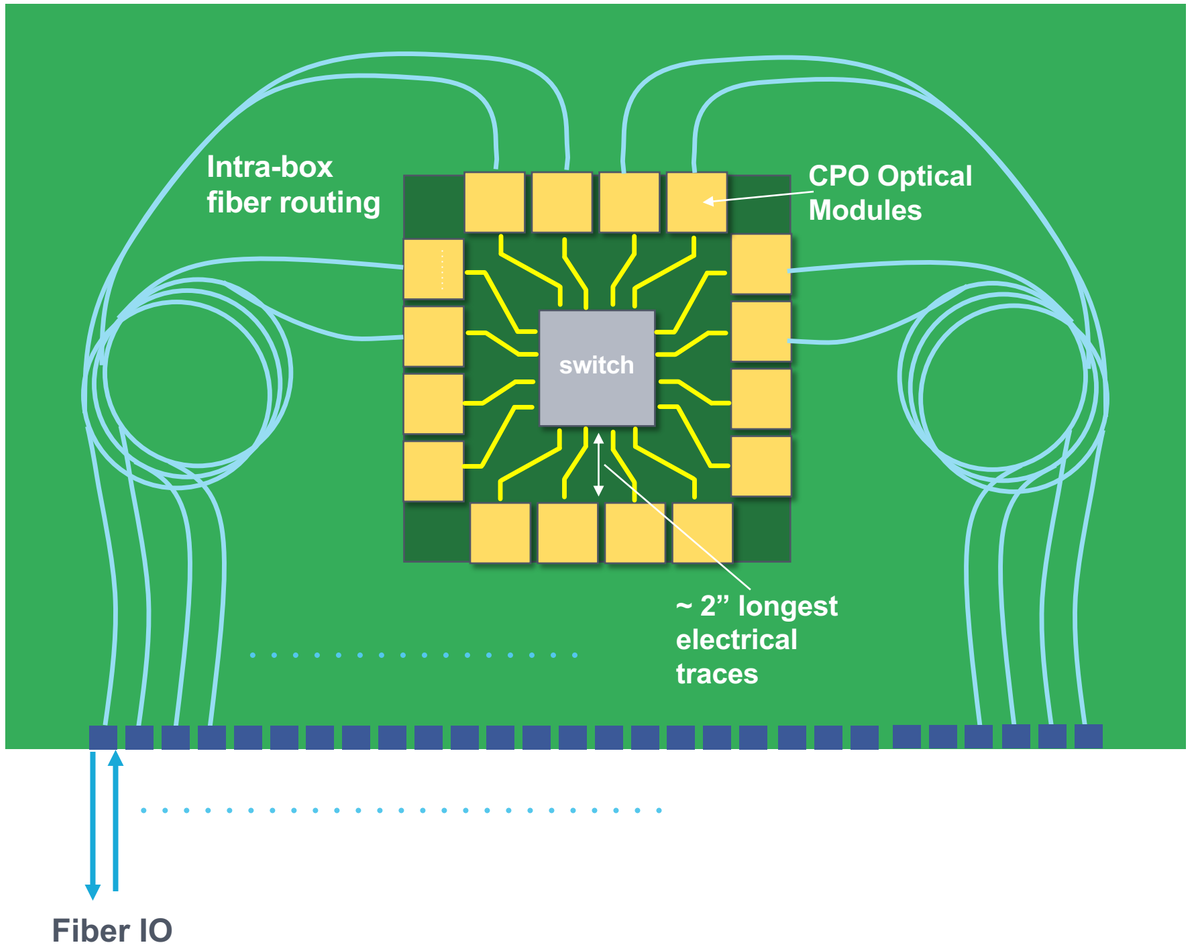
- Ideal Outcome:
  - Preserve optical link budgets and wavelength plan for B400G PMDs
  - Consider ease / relative cost of multi-rate serdes compatibility (i.e. PAM4 vs PAM6..)

Co-packaged Optical Switch – Quick Recap

Conventional Fixed Box Switch System

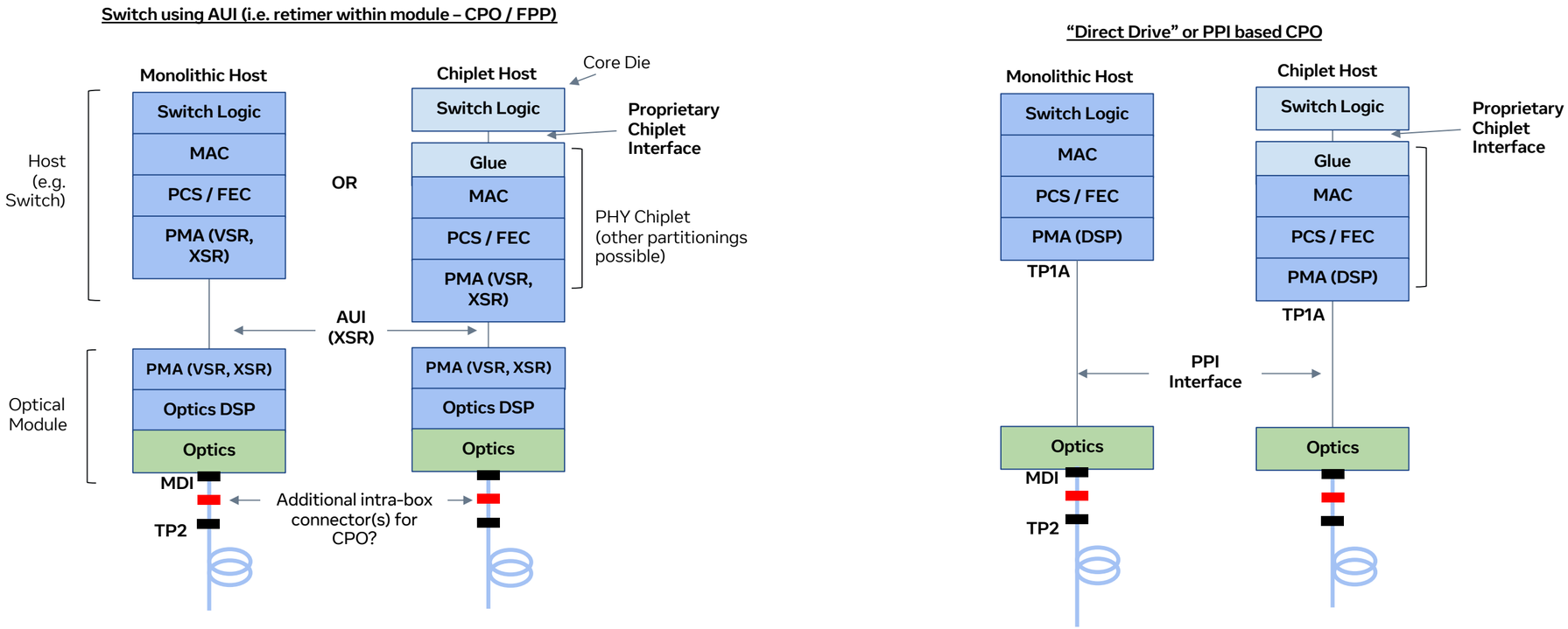


CPO Switch System



- Move optics closer to switch → shorter electrical channels, less serdes power

Co-packaged Optics – Architectural Considerations



- PMD (and by extension PCS) interoperability at TP2 is required for CPO and FPP modules
  - Need to account for additional intra-box connector(s) (as compared to FPP based optics)
  - Compatibility between CPO and FPP based systems is required
- Interoperability of PPI approaches (optics, serdes) is required for deployment at scale
  - Need multi-vendor interoperability to preserve supply chain robustness
  - Specifications will need to be developed for PPI channel, serdes and optics (see Annex 86A for example)
- *Note: end-end FEC shown here for simplicity, but many options to be considered see wang_b400g_01_210208.pdf for example*

Summary

- We need to optimize for efficiency!
  - Both power, and operational
- 200G Electrical signaling looks to be required to support efficient 102.4T generation switch systems (wiring / ASIC escape limitations)
- 800GBASE-FR4 is the next required major optical PMD for Facebook applications
- New architectures such as CPO are emerging to support higher efficiency designs
  - These need to be compatible with current approaches (i.e. pluggable modules)
  - Serdes interoperability will continue to be critical
    - Electrical Serdes (XSR, VSR, MR, LR)
    - Optical (i.e. serdes capable of supporting “direct drive” / PPI)

Thank you!