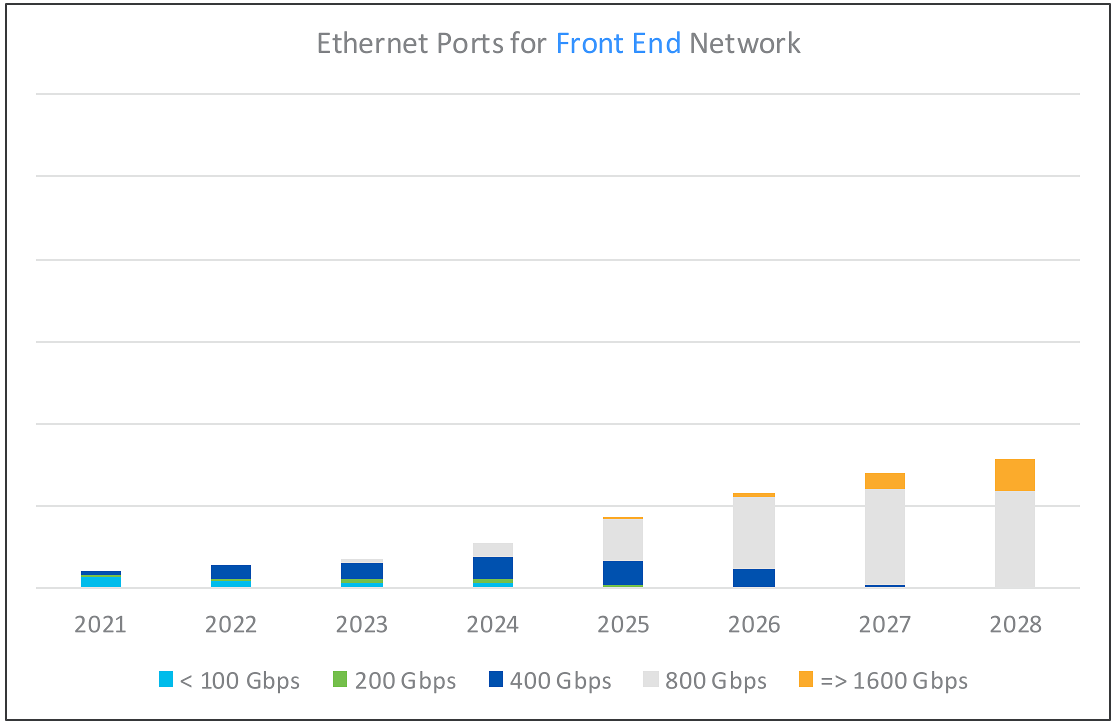
System and Rack Design considerations for 400G

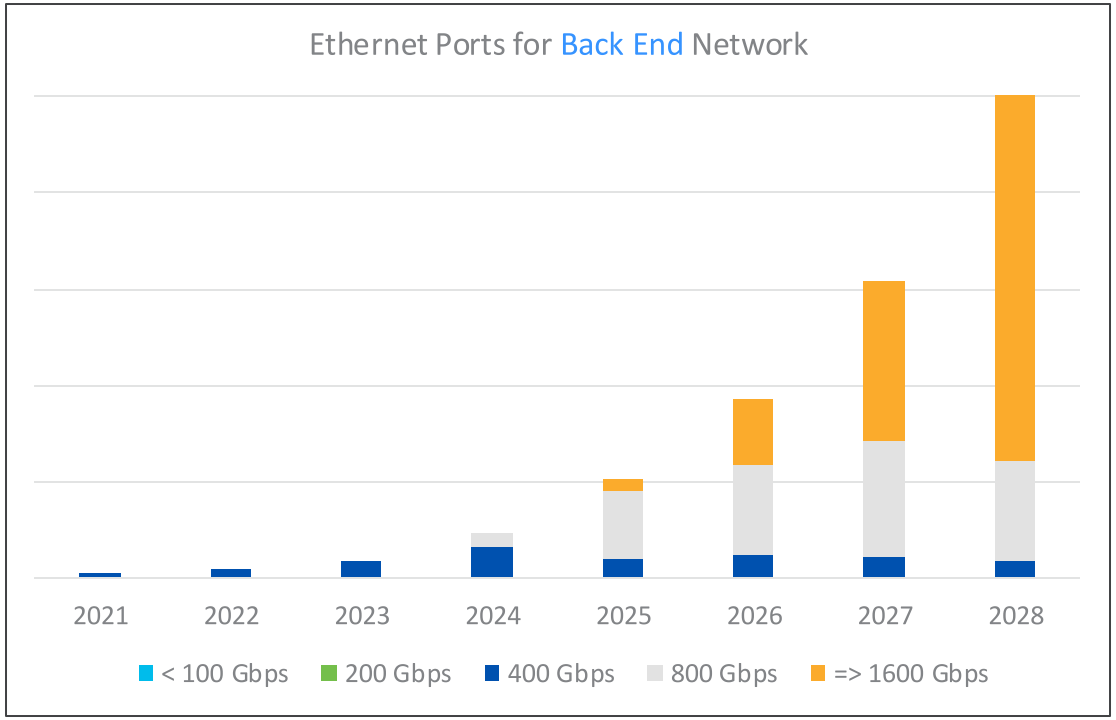
Mark Nowell, Fellow, Cisco

OIF 448G AI Workshop, April 15-16, 2025

Acknowledgements: This presentation would not exist without the inputs, expertise, and patience of many Cisco colleagues! Thanks to the Anthony Torza, Brian Welch and Cisco SI team including David Nozadze; Sayed Ashraf Mamun; Wenbin Ma; and Mike Sapozhnikov

Ethernet Speed Transitions in AI Networks





Majority of the switch ports in AI back-end Networks to be 800 Gbps in 2025 and 1600 Gbps in 2027, showing a very fast migration to the highest speeds available in the market.

Dell'Oro: 95N01_Advanced_Research_AI_Networks_For_AI_Workloads_Report_3Q24

Relentless Advancement

Switching Capacity	640G	1.28T	3.2T	6.4T	12.8T	25.6T	51.2T	102.4T	204.8T	409.6T
ASIC # SerDes	64	128		256		512			1024	
Serdes Gbps	10G NRZ		25G NRZ		50G NRZ		100G PAM4	200G PAM4		400G
Port Gbps	40G		100G		400G		800G	1.6T		3.2T
Ports/Silicon	16	32		64	32	64	64	64	128	

ASIC density continues to redefine how products are built.
Gates & GHz. SerDes & Interconnect. Optics & wavelengths.

Credit: adapted from slide by Rakesh Chopra & John Chapman
https://blogs.cisco.com/sp/co-packaged-optics-and-an-open-ecosystem

System Architectures evolve with technology

Fixed Box

Vertical Linecard System (VLC)

Modular & Rack System

Interconnects for an AI/ML world

AI/ML is a disruptive event for traditional networking

		Traditional Front-end DC	AI/ML Back-end DC
Rack Bandwidth (ToR/MoR)		3.2T-12.8T	>> 100T
Rack power		~10 kW	100 kW+
Packet Loss impact (reliability)		Low importance	Critical importance
Latency importance	Absolute	Low	Low
	Tail	-	High

Lots of interconnect
→ Speed matters
→ Power matters
→ Density matters

Massive rack density increase
→ Power matters
→ Copper cables matter
→ Density matters
→ Thermal solutions matter

Job completion time (JCT)
→Link BER performance critical.
→Tail latency most important

What to define and why?

AI Care-abouts

- Energy Efficiency
- Reliability
- Reach
- Scale Up
- Power Efficiency
- Training
- Scale Out
- Bandwidth
- Inference
- Latency
- Beachfront Density
- Ethernet
- Parallelization

400G

Equipment/ASIC Care-abouts

- SERDES
- Signal Integrity
- Power Density
- Cooling
- ASIC Packaging
- Copper
- Optics
- System Radix
- System Density
- Connectors
- Pluggable
- Co-packaged
- Ethernet
- Form Factors
- Channels

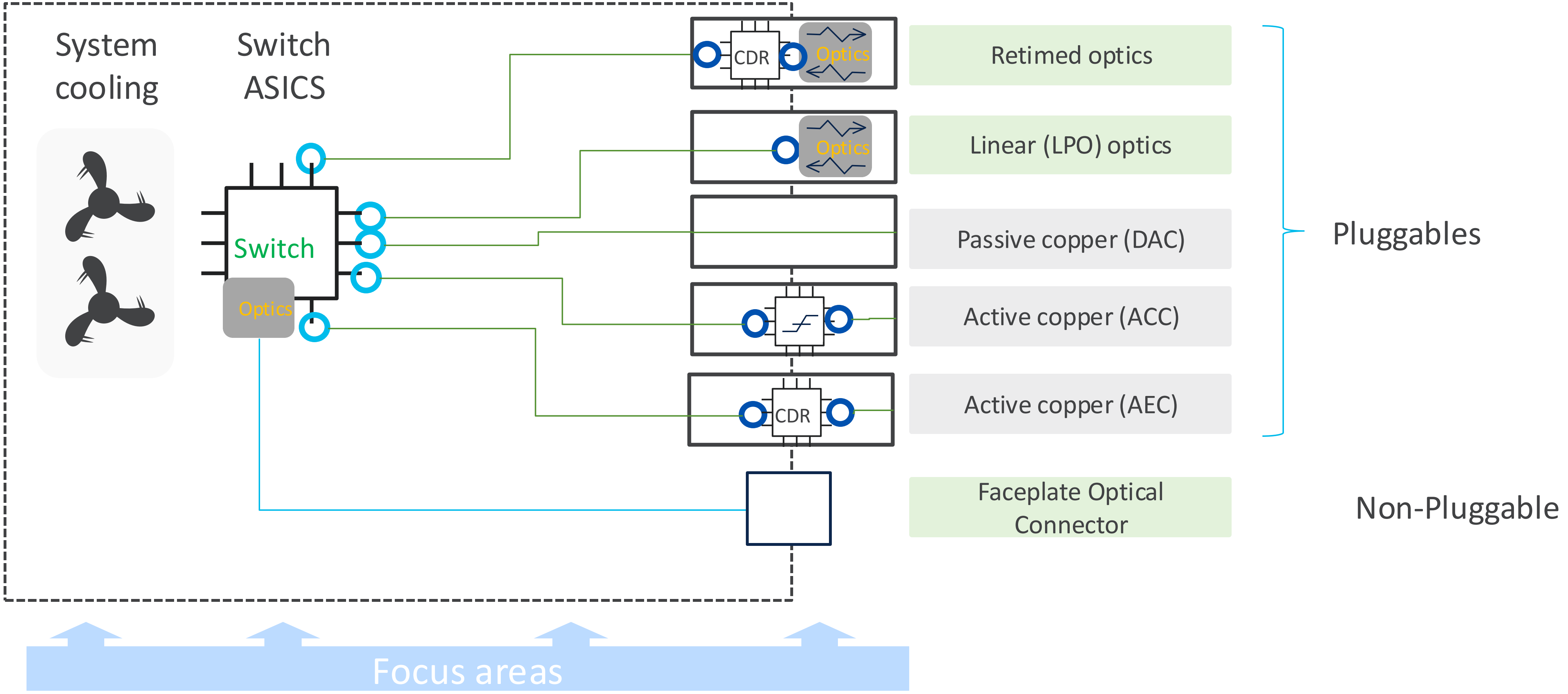
Key questions

- What specifically?
- Priorities?
- Feasibility?

System design considerations @ 400G

- 400G technology forces a complete end-to-end co-design mindset. Every detail matters
- Evolution from 50G→ 100G → 200G has made it clear that separate specialties are less and less able to be optimize independently
- 400G will be the hardest yet.
  - Margins are diminishing
  - Reliability requirements are heightened due to AI deployment scale
  - New technologies, materials, packaging, fabrication techniques, modulations, connectors all need to be brought to production
  - SNR impacts with greater sensitivity to channel impairments

Simple view



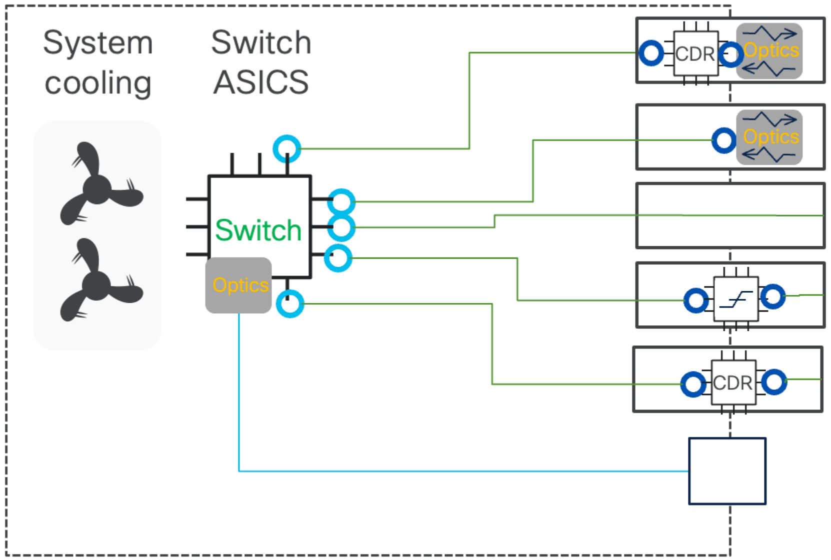
- Air vs liquid cooling

- Highest capacity
- Serdes capability

- Optimized channel

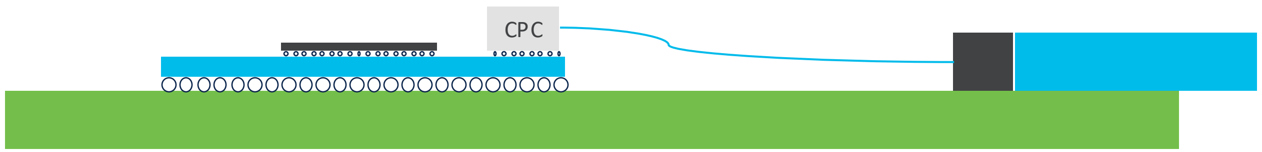
- Retimed optics
- Linear optics
- Copper cables

Simple view – key takeaway's



- Liquid cooling should be assumed. Reduces thermal challenges, but does not reduce the focus on power efficiency
- The duality of both copper and optics need to be supported to cover the breadth of requirements
  - Pluggable solution needed
- Some traditional (Legacy?) system design approaches are running out of runway. Need to understand what are acceptable channels to architect (and standardize) around.

Complex view:
At 400G co-consideration of every aspect is necessary



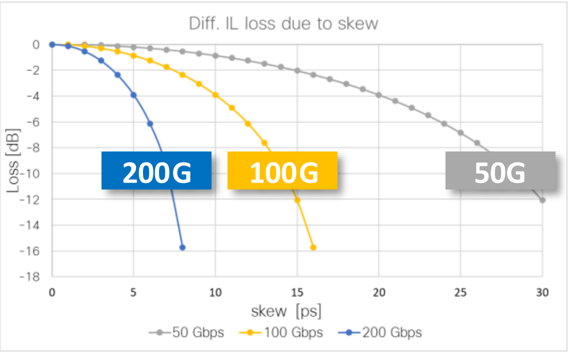
Even assuming a co-packaged copper (CPC) approach as the basis of designs numerous and inter-related challenges remain.

Higher speed signaling:
- New modulations for electrical (and optical?)
- Increased Loss
- Increased coupling
- Greater sensitivity to channel impairments, skew, nonlinearity
- New SerDes, DSPs, PLLs
- Improved FEC(?)
- Link Reliability

Complex packaging:
- Larger packages
- Increased Loss
- Manufacturing challenges
- Reliability

Next Gen pluggable:
- New connector approach
- New form factor
- Optics and Copper
- Reliability/FIT/Retention

SerDes Rate Increase: Enhanced Challenges



2x SerDes Rate = 4x – 6x more loss
Accelerating loss with increased skew

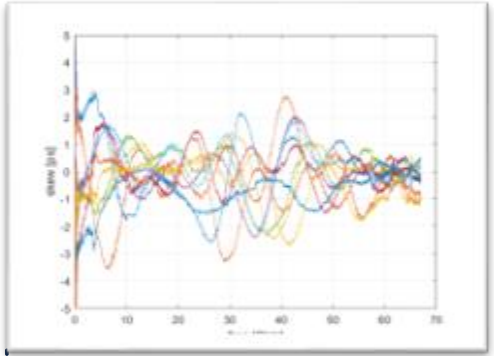
P/N Skew Budget

As the data rates increase, design and statistical skew will have a significant impact on loss
Skew becomes critical item to control

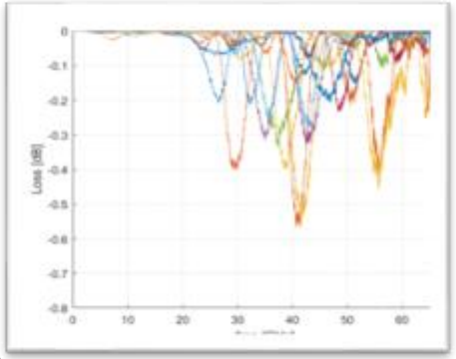
Channel Transitions

Channel transitions degrade link performance which is amplified by an SNR impact due to modulation. New packaging & fabrication techniques. New connectors will be key.

Skew



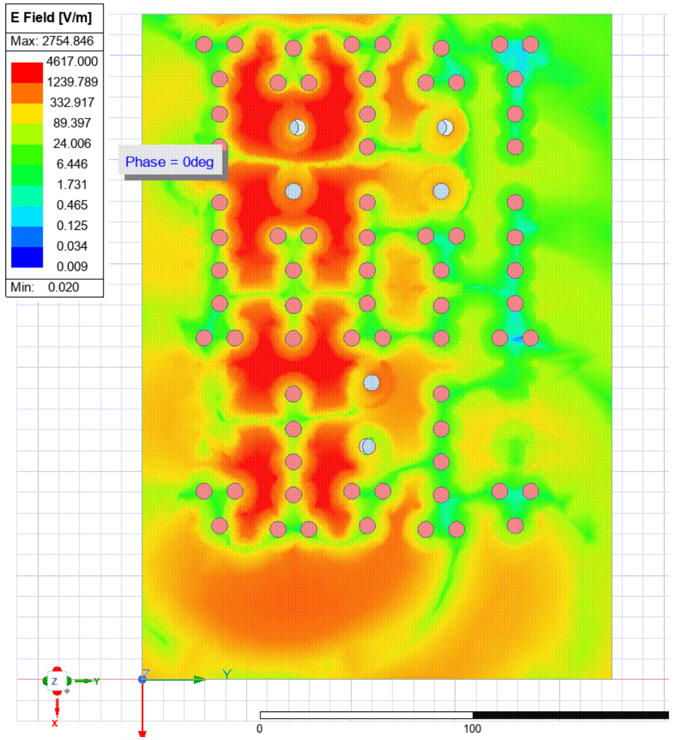
Loss



Twinax cables aren't immune

Multiple Cable Skew Signatures Exist
Key to understand the Signature & Impact on Performance.

SerDes Rate Increase: Old methods break down

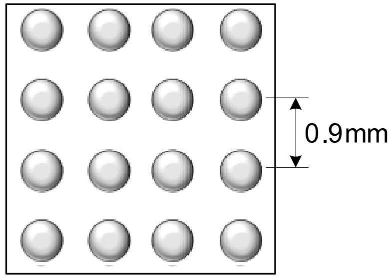


EM Field analysis showing effectiveness of ground via isolation

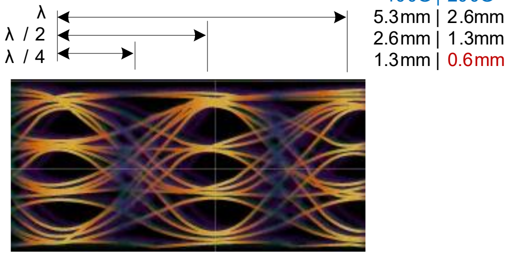
Return Loss & Cross Talk

200G PAM4 (56G) ¼ wavelength (0.6mm) is smaller than ASIC ball pitch (0.9mm)
Traditional use of ground vias to isolate signals fail
Return loss & cross-talk require new methods for optimization

Ball Pitch



Wave Length



Rule of thumb design limits:

λ/2 ; cut off by GND via fence resonances
λ /4 ; signal via localization
λ /8 ; GND via fence shielding

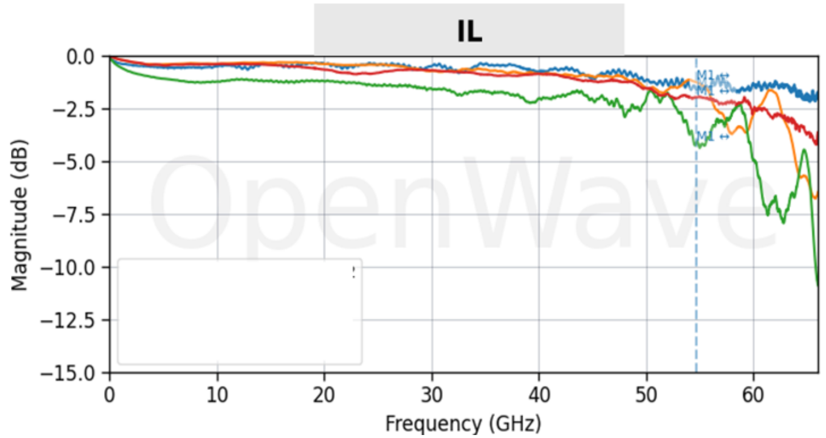
SerDes Rate Increase

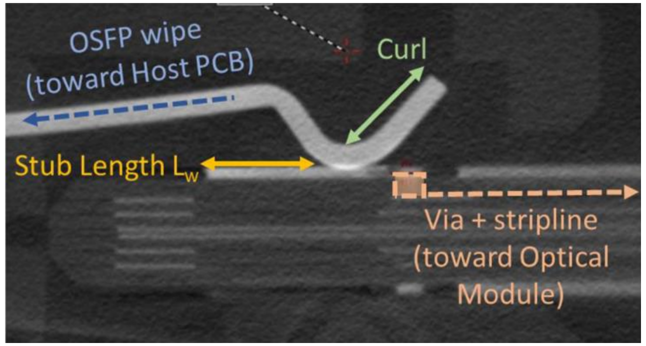
- PAM4 (electrical) system channels do not appear feasible.
  - PAM4 (optical) generally preferred.
- PAM6/8 – under investigation
  - SNR impact
  - Signal:
    - Mitigate channel loss with improved techniques such as CPC
  - Noise:
    - Skew, Coupling, Reflections, Transitions, Non-linearities all have heightened impact
    - Packaging, connectors, channels
- FEC becomes a trade-off between link reliability/performance and latency. Increased FEC overhead can have diminishing return due to device f_T limitations

SerDes Rate Increase: Existing OSFP connectors

Current card edge connector approaches look limited beyond 200G

- Return Loss is very high
- Near-end crosstalk is very high
- Far-end crosstalk looks manageable
- EM simulations @ 400G are not encouraging & BW limited.





	RL [dB] (0-53GHz)	IL [dB] ~53GHz	PSNEXT [dB] ~53GHz	PSFEXT [dB] ~53GHz
Industry Ranges for 1X1 OSFP Connectors	-6 to -9dB	-1.5 to -4dB	-35 to -44dB	-25 to-30dB

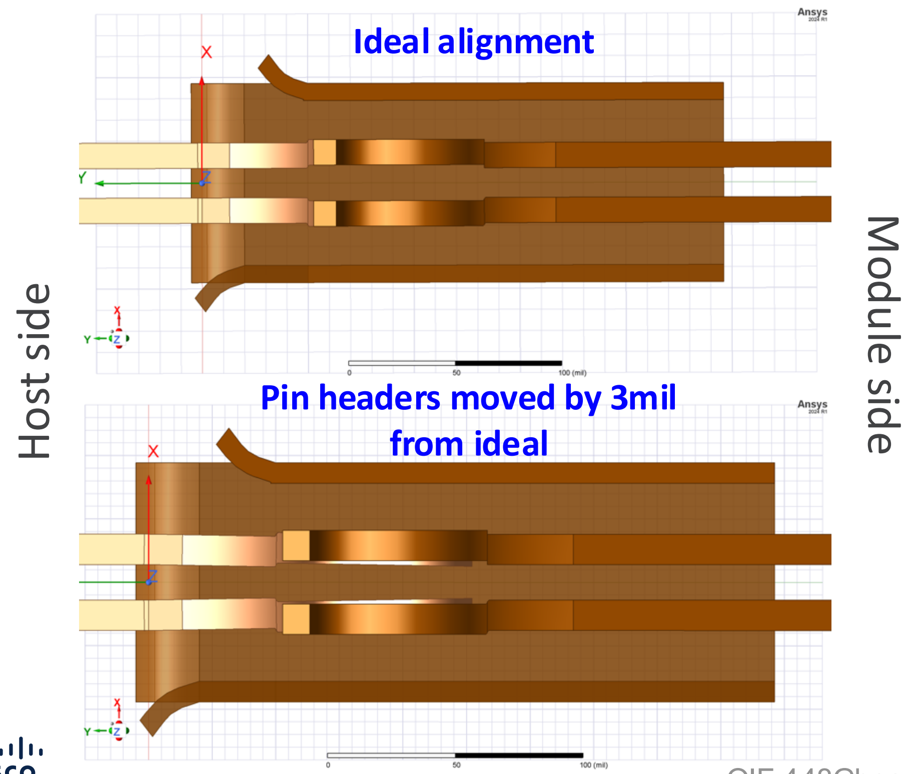
*Measured and de-embedded

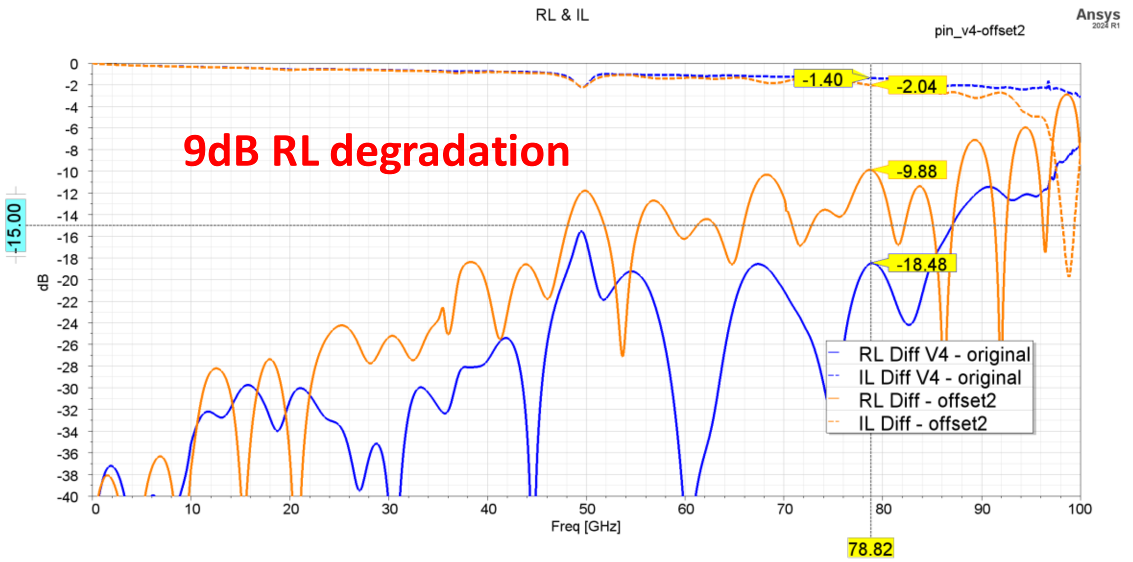
Connector Mating Consistency

Preliminary EM analysis of various connector structures (one shown below)

- Show viable performance without mating & manufacturing variations
- Very sensitive to minor offset – 8-10dB degradation in RL is observed
    - (Suggests a 2-piece smaller connector would have better mating consistency and tolerances to forces)

Top view of a p/n mating pair within a connector





400G Optics

- Optimized modulation for optics being different from electrical implies that pluggables will be retimed.
  - Initial priority would be an optimized interface for short reaches. Market forecasts suggest strong need for defining a common interface to support volumes.
  - Radix priorities push towards parallel fiber interfaces
  - Allows some evolution of interfaces to happen for different (longer) reaches as market becomes clear
- Co-packaged Optics
  - Will co-exist with pluggable, so common optical interface definition will be needed.
  - Power reductions due to co-packaging will drive adoption.

400G: Where do we stand?

- The details really matter!
- Establishing the system and network (initial) requirements is imperative to allow these inter-dependencies to be analyzed and traded off.
- Need to restrict the scope to what are “reasonable” implementations
- Key will be to define an initial starting point for industry specs. There will be time to build out the full portfolio of specs
  - Electrical interface capable of supporting pluggable (Is there an acceptable DAC reach of tbd?)
  - Active copper to extend reach
  - Short optical interface capable of supporting intra-DC, high radix, compatible with pluggable/CPO.

The bridge to possible