



Architecture Vision: Open Silicon Photonics for AI Systems

A Vision for Scaling AI Compute and Switching to Next-Generation Density Through Open Photonic Interconnect

1.0.6

Authors:

Bijan Nowroozi, Lightmatter

Michael Klempa, Qualcomm

Jared Wright, Qualcomm

Mohammad Alavirad, Dell Technologies

Jian Chen, Dell Technologies

Jiayin Ma, Flex

Date: JULY, 2026

Collaborators and Contributors

Aditya Raina, GUC	Michael Lane, Hyve Solutions
Ahmad El-Chayeb, Keysight	Michael Ransford, Corning
Boping Xie, Intel	Min-Jie Chong, Keysight
Dan Schaffer, Flex	Pete Tomaiuolo, Celestica
David Piehler, Dell	Ram Channappa, Keysight
Dennis S.C. Chen, Foxconn FIT	Randy Clark, Celestica
Devanshu Kant, Flex	Rob Nance, Jabil
Benoit Fleury, Corning	Sergio Nunes, Celestica
Igor Elkanovich, GUC	Steven C.W. Sher, Foxconn
Jason Wildt, Jabil	Tanuj Kalakuntla, Keysight
Kinny Chen, Wiwynn	Timothy Pak, Foxconn FIT
Larry Wikelius, Qualcomm	Winnie Lin, Hyve Solutions

Special Thanks

Brian Aspnes, Intel, OCP Server
Shawn Dube, Dell, OCP Server
Corey Harman, Dell, OCP Server
Jeff Kennedy, Dell, OCP Server
Greg Sellman, AMD, OCP Server

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Version History

Note, refer to the OCP Contribution Versions, Revisions and Errata best practices documentation. Versions and revisions are made to this document and logged here. Errata is a separate document such that the contribution specification document it refers to was not revised. Ex: Version 1, Errata E1

Date	Version #	Author	Description
25 FEB 26	0.1-0.8	Bijan Nowroozi	Community and LM comments, thoughts and edits added to illustrate the context space
26 FEB 26	0.8.1	Bijan Nowroozi	Integration issues resulted in additional/missing text or comments- fixed
03 MAR 26	0.8.2	Bijan Nowroozi	Added some discussion of latency
05 MAR 26	0.8.3	Bijan Nowroozi	Clarified the switching, added 3 high level diagrams per feedback
25 MAR 26 APR 26	0.9.1-3	Mohammad Alavirad/Bijan Nowroozi, et al.	Edits + Added the reference to the ecosystem steps, corrected organizations listed in bullets, added some detail about rear connectors.
4 MAY 26	0.9.43	Bijan Nowroozi	Added diagram in 10.2 as an overview to illustrate which specifications should apply. Appendix B diagram provides additional information.
8 MAY 26	0.9.6	Jian Chen	Added a new paragraph in section 8.1; and added a new section 8.4 on Diagnostic and Fault Isolation
15 MAY 26	0.10.0-0.10.8	Bijan Nowroozi	Added Channel Budget Reference, Placement Class Taxonomy, COM and Placement Selection,

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

			Envelope and Contract subsections. Reworked Executive Summary; harmonized reach numbers across document via Channel Budget Reference; relocated four-layer interoperability framework from Executive Summary to Envelope and Contract; added boundary transition pattern; added co-design recommendation in Adoption Sequencing; added cross-references and OCP RAS API to Standards Alignment. Elevated faceplate zone-modularity framing;refined Goal 5; highlighted worked on Appendix A's clarity. Index remapping
16 MAY 26	0.10.9-11	Bijan Nowroozi	Improve the mapping of and wording for sustainability and cooling requirements
17 MAY 26	0.10.11-18	Bijan Nowroozi	Scale-DSA visibility in Table F (Profiles A/B/C/D Primary Use); Scale-DSA mapping sentence after Table K; IOWN GF / Goal 3 expanded for protocol and traffic-class agnosticism including Scale-DSA; new Goal 5 inserted (passive optical backplane replacing copper midplane assemblies); old Goals 5/6/7/8 renumbered to 6/7/8/9; Goal 8 OCP-IOWN AI Computing Continuum paragraph added to Standards Alignment; inter-site scope clause added to Scope and Stance.
19 MAY 26	0.10.19	Mohammad Alavirad	Final review, and editions. Added sections 5.4 and 6.5
20 MAY 26	0.10.20	Bijan Nowroozi	Edits and review responses. Added Class V

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

			distributed rail-switch architecture, wavelength-scaled radix framing, multiple fabric-plane guidance, and document-wide cross-references.
21 MAY 26	0.10.21	Mohammad Alavirad	Edits and review responses to some comments, provide responses, fixing some terms throughout the text. Added some technical discussion to different sections
22 MAY 26	0.10.22	Mohammad Alavirad	final technical/editorial review, addressed collaborator questions with proposed response text, and added targeted revisions for terminology consistency, feasibility caveats, COM-to-RAS telemetry, and thermal/power validation.
26 MAY 26	0.10.23-5	Bijan Nowroozi	Clean ups on images and labeling. Clean up/rephrase some text in multiple sections.
29 MAY 26	0.10.24-6	Mohammad Alavirad	Cleaned up many cross-references. Revised some ambiguities. Also fixed some arithmetic calculations throughout the WP.
31 MAY 26	0.2	Bijan Nowroozi + Mohammad Alavirad	Clean ups and Followed the reviewer comments and ended up cleaning up much of the text in section 3 with rewrites (no new concepts), then moved sections around for better flow- hence the rebase and new v0.2
02 JUN 26	0.3	Bijan Nowroozi	Restructure of Chapter 6-9. Retained but realigned content for flow. Added new diagrams
02 JUN 26	0.3.2	Bijan Nowroozi	Added some clarity around IO escape with text and images. Cleaned up list of figures etc.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

08 JUN 26	0.3.3	Jiayin Ma	System Integrator Validation Matrix in §5.7 with a manufacturing-readiness dimension
09 JUN 26	0.3.4	Bijan Nowroozi + Mohammad Alavirad	Clean up edits. Added Diagram 5.18 and laser requirements in 5.10
09 JUN 26	0.3.5	Devanshu Kant	Added Laser Thermal Coupling and Reliability in 5.6
11 JUN 26	0.3.6-7	Bijan Nowroozi	Illustration clean up, 9.1.6 clean up Sorted out sync thoughts from a giant section and renumbered it.
17 JUN 26	0.3.8	Mohammad Alavirad	Reworked section 5.10 thoroughly
17 JUN 26	0.3.9	Bijan Nowroozi	Updated Diagrams in section 5
23 JUN 26	0.4.0	Bijan Nowroozi	Section 5, consolidated and updated the zone definitions, also Scale-DSA tweaked slightly. Added illustration
28 JUN 26	0.5.0	Bijan Nowroozi	Scale-DSA further improvements. Added Appendix E
30 JUN 26	0.6.0	Bijan Nowroozi	Added table OO to clarify the text around OpenCPX
02 JUL 26	0.6.1-7	Bijan Nowroozi	Clean ups/re-organization of section 9
05 JUL 26	0.7.0-1	Bijan Nowroozi	Styles fixes
08 JUL 26	1.0.1	Bijan Nowroozi	Fixed Broken Links
13 JUL 26	1.0.2	Bijan Nowroozi	Misc grammar fixes
17 JUL 26	1.0.3	Bijan Nowroozi	OCP Foundation feedback incorporated
28 JUL 26	1.0.4-5	Bijan Nowroozi	Collaborator correction. Fig E1/E2 were fuzzy-replaced.
30 JUL 26	1.0.6	Bijan Nowroozi	SC Feedback (Tradenames) incorporated

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Executive Summary

The Vision

For most of the history of computing, the scarce resource was information. We built capacity for data, then calculation, then ever-faster processors, then connected the world. That era is ending. We are awash in information.

We are learning that information behaves like matter: it can be formed, transformed, and annealed into solutions we could not reach before. This is the dawn of a new kind of tool, one we call Artificial Intelligence. It began as conversation and became something larger.

The accelerators that train the largest AI models are no longer bound by how fast silicon can compute inside a reticle. The constraint is conversation: the relentless need to move data between accelerators at memory-semantic bandwidth, the speed and access pattern memory itself demands. Sustaining it across a rack is a rising tax in power, latency, and service complexity.

At the root, this is a problem of graphs. Every node needs to reach every neighbor at once, where the neighborhood is not set by distance but by how fast the two can talk.

Information, using electricity, has traveled over copper lines since the telegraph. This works as an electron pushes through a lattice of copper atoms that block and tug at it; the faster it travels, the more it crowds onto the surface and fades. Drive each pulse faster and it reaches less far.

Light behaves differently. A photon glides down its waveguide and can be driven at high rates and still cross the whole rack having paid only for the passage. The electron

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

makes the same trip only by being stopped and rebuilt, again and again, each repair spending power and time. The electron is not slow. Past a certain reach, it needs help.

The question is not whether our machines will use light, but how, and whether we build that capability in the open, together, or fracture it into a hundred incompatible chassis.

This white paper proposes an answer: an openly developed, vendor-neutral photonic interconnect architecture for AI infrastructure, agnostic to the optical technology that implements it. It extends a tray's scale-up neighborhood from within a rack to across racks and rows, and it builds on existing OCP community work already proven and sized for hyperscale-class AI training. The aim is to unleash an ecosystem for next generation photonic AI systems designed at hyperscale and open to all.

The thesis is one line: copper makes topology follow proximity; photonics lets topology follow the workload. Copper is reaching the end of its economic life at the tiers that now bind, and the reason is physics, not effort.

This effort does not remove topology constraints; it reimagines the composition. Reach and retiming give way to radix plateaus, the connector budget at high lane rates, serviceability, and link telemetry, while fiber escape buys back physical space. Separating the passive optical plant from the optical engine (OE) that lights it lets the OCP community build hyperscale interconnect as an open platform capability, not a proprietary chassis. The goal of this paper is to seed an OCP workstream, and from it a buildable base specification, for Open Silicon Photonics for AI Systems.

§ Proposed Deliverable Structure: Informative Vision, Normative Contracts, and Validation Profiles frames this hierarchy in detail. § Ecosystem Collaboration and Next Steps lists the specific actions this paper requests of OCP working groups.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Figure 1.1 places this work in the wider ecosystem: the external standards bodies it references and the other OCP efforts it builds on. The architecture is designed to compose with that work, not stand apart from it.

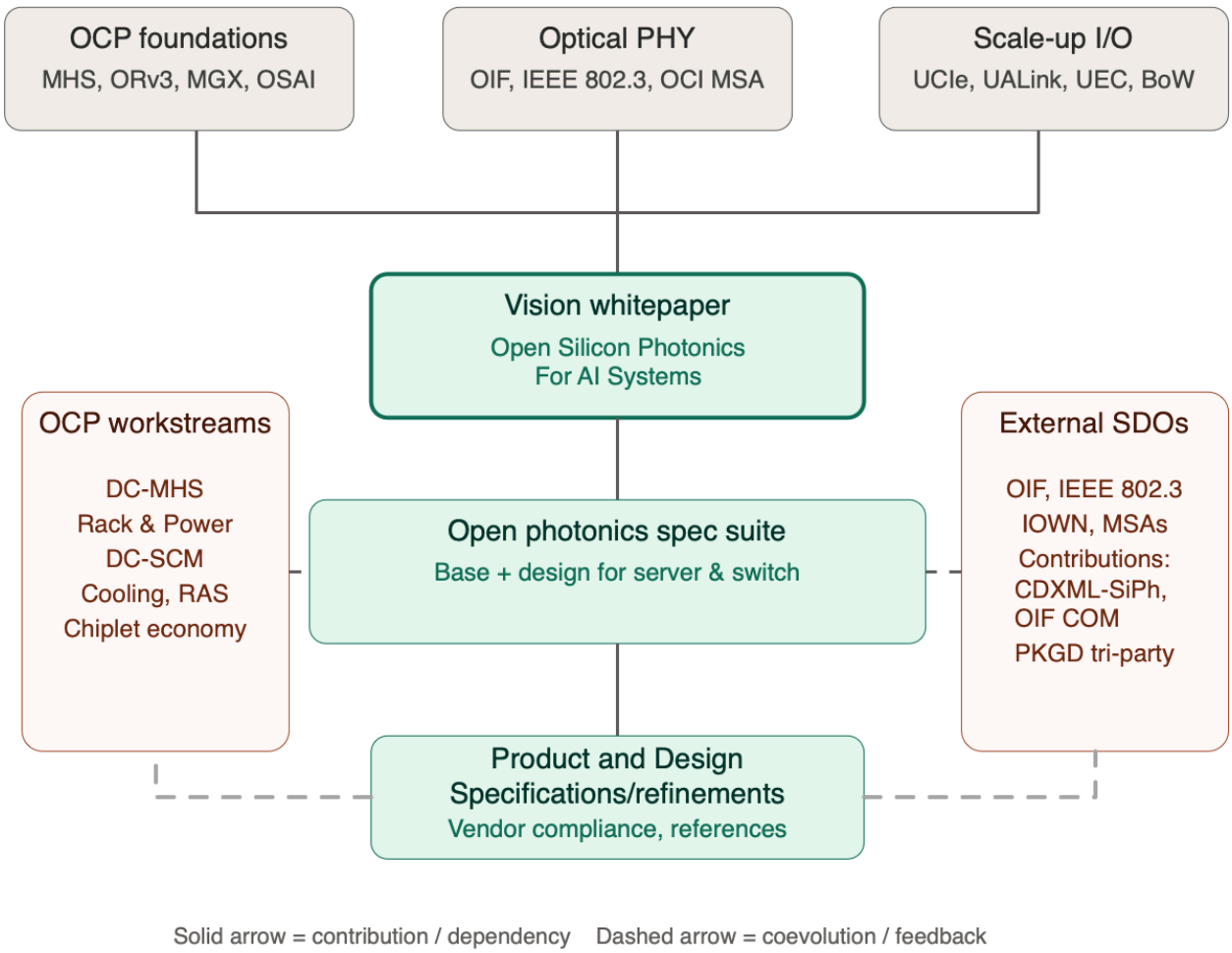


Figure 1.1. *Architecture Vision: Open Silicon Photonics for AI Systems Contribution Graph.*

This document (focal, thicker outline) draws from OCP foundations, optical PHY standards, and scale-up interconnect (top); produces the Open Photonics spec suite that OCP workstreams refine into vendor products (bottom); and coevolves with parallel OCP workstreams and external SDOs (left, right). The mid-density variant appears in §Proposed Deliverable Structure; full-detail enumeration in Appendix B.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

The Interconnect Bottleneck

Conversation became the constraint, for a physical reason that worsens each generation. At scale, interconnect bandwidth, latency, and reliability now decide how much of an accelerator's compute is realized. The cause is the shrinking reach of the scale-up interconnect, with memory bandwidth pressing from the other side. As SerDes rates climb from 224 toward 448 Gb/s/lane, viable copper reach compresses roughly 2–4× under iso-modulation, at fixed channel and equalization assumptions.

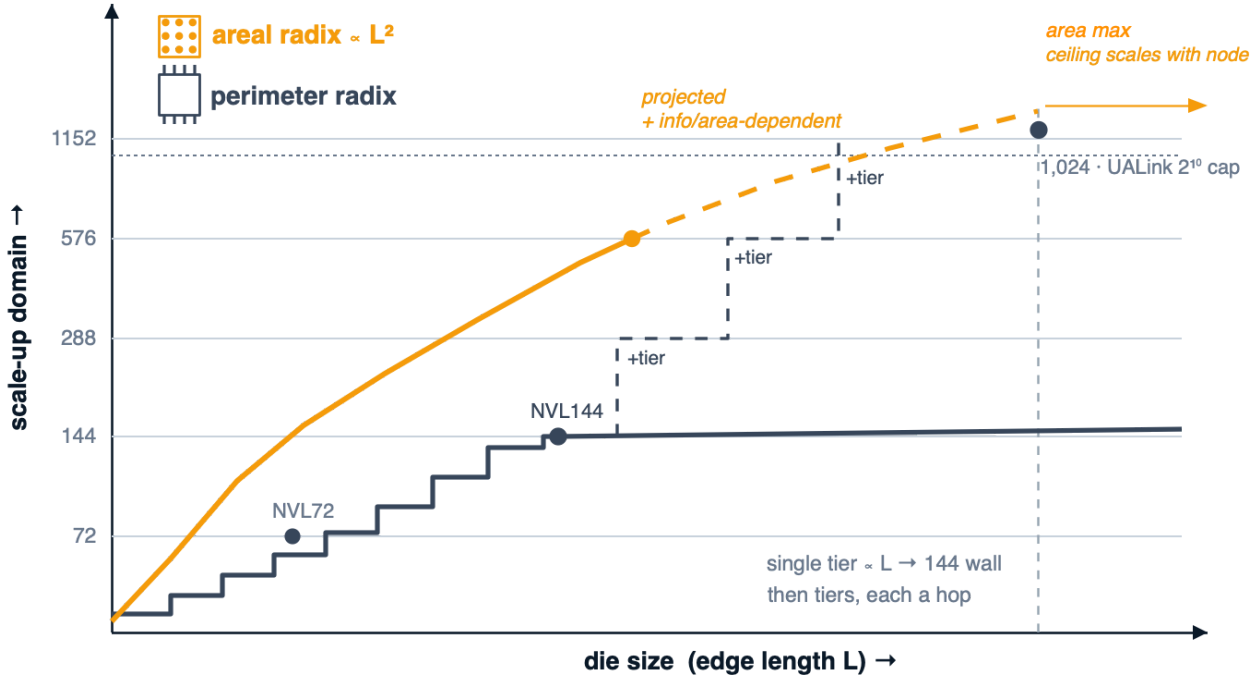


Figure 1.2. *Domain follows die size through escape.*

Die size L is the independent variable; the known scale-up domains sit on the y-axis. Perimeter radix rises as a single tier ∝ L, in chip-I/O steps, to the 144-port radix wall, then composes one switch tier at a time, each a latency hop. Areal radix (APO) rises with L², solid where grounded in demonstrated escape density and dashed where it extrapolates beyond. The area-max ceiling is not fixed. More wavelengths and higher-order modulation raise it; process and materials scaling raises it durably, packing more devices, drivers, and channels per face. The 1,024 line is UALink™'s 2¹⁰ addressing cap, not a radix step. Derivation in §4.

The usual remedies postpone this physics rather than escape it. Retimers and heavier conductors restore reach, but each retiming stage adds latency that accumulates along the path, a structural delay no topology can route around. Copper also runs out of

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

shoreline: the I/O it can physically escape from a package edge is perimeter-bound, a ceiling reach extension cannot lift. The cost surfaces as idle silicon. Meta's Llama 3 405B pretraining reported BF16 Model Flops Utilization (MFU) of 38–43%, with the largest losses in communication and synchronization. Remarkably, this shows over half of an extraordinary machine's throughput went unrealized.

The Open Photonic Solution

The way through is to let light do what copper cannot. Where copper reach, physical IO connectivity limits, retimer power, midplane complexity, or serviceability becomes the limit, optical interconnect becomes the preferred option, subject to link-budget, power, thermal, cost, and serviceability closure. That much is increasingly agreed. The harder question is how to make the transition without fragmenting the ecosystem into incompatible islands.

This architecture answers with an interface-and-contract method: define the interfaces every component should fit and the contract it honors, then let many technologies meet inside them. It crowns no single winner. Silicon photonics, Vertical-Cavity Surface-Emitting Laser (VCSELs), micro-Light Emitting Diodes (LEDs), and coherent optics share one passive plant and speak across common contracts. No one is locked out, and no one is locked in. That plant stays protocol-agnostic, indifferent to what travels over it, and carries scale-up, scale-out, and cross-architecture Scale-Domain Specific Acceleration (Scale-DSA) bridging traffic on the same fiber.

Target Outcomes

By trading the extreme engineering and manufacturing required for high-complexity multi-layer PCBs for a generation-invariant passive fiber plant, this architecture targets the measurable outcomes below:

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table: Target Outcomes

Metric	Current State	Target Outcome
Scale-up bandwidth fraction of memory bandwidth (α), one hop, untapered	α ≈ 1/6–1/4 across shipping XPUs; copper edge escape holds the band down	α ≥ 1/2, yielding flat domains of 1,024–2,048 XPUs bounded by crossbar silicon, not escape — area I/O (APO) retires the escape limit (Figure 1.2)
Realized Model FLOPS Utilization	38–43% (large model training)	55–65% (as communication overhead and stalls reduce)
Interconnect energy	15–20 pJ/bit (pluggable optics)	~10 pJ/bit (FEI-class) ~5-7 pJ/bit (short-reach / CPO-class demonstrated) ~3 pJ/bit (aspirational, next-gen EEI target)
Network power ratio	≈31% of system power (Class H switched, Blackwell-class)	≈15% of system power (Class V switchless, Rubin-class)
Link reliability	0.55M hour MTBF for pluggable optics[1]	2.6M+ hour MTBF for CPO-class implementations[2]
Fiber infrastructure lifespan	Redesign per SerDes generation	Generation-invariant passive fiber plant
Extreme Complexity PCBs	Required for high-density copper fabrics	Eliminated or reduced to passive fiber distribution

[1] https://credosemi.com/blogs/the-path-to-zero-flap-reinventing-optical-reliability-for-scalable-ai-clusters/
[2] subject to normalization by link count, operating temperature, connector count, laser architecture, and service model. Appendix D consolidates these and the document's other load-bearing quantitative claims with evidence status (§ Appendix D: Architectural Anchors and Sensitivities, Table AR).

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Metric	Current State	Target Outcome
		where photonics links replace copper reach extension
Scale-up link latency	40–120 ns accumulated overhead in multi-stage retimed copper paths	Propagation-dominated, ~5 ns/m on non-retimed optical paths; removes the retimer-stage accumulation OIF CEI [3]; ESUN Base Spec [12]

Good fences make good neighbors. In open systems, the fences are interfaces: shared requirements that make the boundary explicit and leave everything behind it to the vendor.

The depth that follows is deliberate. We begin with no assumptions beyond the laws of physics. Each section takes one constraint to its floor, from escape geometry through power to the time axis, and derives the boundary it sets. Those boundaries then assemble into a practical vision: photonics integrated into open AI platforms with modularity, serviceability, and multi-vendor choice intact. The next step is to turn that vision into buildable base specifications, developed in the open with the OCP community.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table of Contents

Collaborators and Contributors **1**
Special Thanks **1**
Version History **2**
Executive Summary **6**
The Vision 6
The Interconnect Bottleneck 9
The Open Photonic Solution 10
Target Outcomes 10
Table of Contents **13**
Table of Figures **16**
Table of Tables **18**
1. Mission **20**
Purpose and Audience 20
Document Contract 21
Scope and Stance 21
2. Compliance with Open Compute Project Tenets **22**
Openness 22
Efficiency 23
Impact 23
Scale 24
Sustainability 24
3. Introduction **26**
The Interconnect Problem 27
3.2. Channel Budget Reference 30
3.3. Optical Link Composition 32
3.4. The Density-Bandwidth Dilemma 36
3.5. Case Study: The Inter-Canister Tapering Problem 38
3.6. The Midplane PCB as Single Point of Complexity 40
4. The Photonic Interconnect System **43**
5. Architectural Vision **49**
5.1. Design Goals 50
5.2. Scale-DSA: The Third Scaling Axis 53
5.3. Optical Engine Classes 60
5.4. Fiber Attach Envelope 61
5.5. Placement Class Taxonomy 65
5.6. Envelope and Contract 72

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

5.7. System Integrator Validation Matrix 79
5.8. Reference Platform Scope 81
5.9. Integration by Reference 110
5.10. Passive Optical Infrastructure Components 111
6. Optical Engine Selection and Economics 134
6.1. Channel Operating Margin and Placement Selection 135
6.2. Effective pJ/bit and Effective pJ/token 144
6.3. Heterogeneous Optical Engine Integration 144
6.4. Adoption Sequencing 145
7. Link Profile Framework 146
7.1. Candidate Link Profiles 146
7.2. Profile-Invariant Infrastructure 153
7.3. Protocol Accommodation 153
7.4. Time and Synchronization 154
8. The Scale Boundary 159
8.1. Premise: Two Trajectories, One Wall 159
8.2. How to Read These Tables 161
8.3. Class-H (MHS/MGX Trays): Reach Scenarios 162
8.4. Class-V (Reach): Interconnect Reach Binds 165
8.5. Class-V (Throughput): Escape Density Binds 168
8.6. The Wall 172
9. Platform Form Factor Evolution 176
9.1. Tray Orientation Classes 176
9.2. Optical Integration Continuum at the Switching Silicon 180
9.3. Density on the Tray 180
9.4. Three-Phase Platform Evolution 181
10. Reliability, Serviceability, and Economics 183
10.1. Reliability Improvements 183
10.2. Serviceability Through Architecture 185
10.3. Economic Benefits 186
10.4. Diagnostic and Fault Isolation: COM-to-RAS Runtime Closure 187
11. System-Level Impact on AI Workloads 188
11.1. MFU Improvement Path 188
11.2. Power Efficiency at Scale 189
11.3. Photonic HBD: Extending the High Bandwidth Domain 189
11.4. Data-Center Roll-Up at the Scale Boundary 191
12. OCP Platform Integration 197
12.1. Proposed Modifications to ORv3/MHS 200

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

12.2. Standards Alignment 203
13. Proposed Deliverable Structure: Informative Vision, Normative Contracts, and Validation Profiles 208
13.1. Compliance as an Ecosystem Accelerator 211
14. Ecosystem Collaboration and Next Steps 212
14.1. Proposed Working Group Cooperation 213
14.2. Invitation to Collaborate 215
15. Conclusion 218
Appendix A: Lower PHY Architecture 220
Fiber Count Derivation 226
Appendix B: Photonic Ecosystem Reference 228
Appendix C: Photonic Link Budget Reference 231
C.1 Scope and Method 231
C.2 Physical Layer Losses 234
C.3 Per Topology Insertion Loss Accounting 239
C.4 Transmitter Output and Receiver Sensitivity Envelopes 249
C.5 SNR Closure by Modulation Column 256
C.6 Reference Cases by Lane Generation (DR and OCI) 261
C.7 Tolerance Envelope Partitioning 265
C.8 Open Items and Vendor-Data Boundaries 269
Appendix D: Architectural Anchors and Sensitivities 272
D.1 Perimeter versus area scaling 272
D.2 Single-reticle perimeter ceiling 274
D.3 Retimer latency accumulation 275
D.4 Rack-level interconnect energy savings 276
D.5 Escape across tiers and SerDes generations — the CPO/APO frontier 277
D.6 Key claims and evidence 279
Appendix E: Time-Axis Anchors and Sensitivities 282
E.1 The memory-hierarchy access-time cliff 282
E.2 Pooling the reachable endpoint set 284
16. Glossary 286
17. References 290
18. License 292
18.1. Creative Commons 292
19. About Open Compute Foundation 293

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table of Figures

Figure 1.1. Architecture Vision: Open Silicon Photonics for AI Systems Contribution Graph.....8
Figure 1.2. Domain follows die size through escape.....9
Figure 3.1. Moving forward from Today's Constraints.....28
Figure 3.2. Escaping The Copper Bottleneck.....30
Figure 3.3. Per-inch insertion loss at the 448G Nyquist (112 GHz).....31
Figure 3.4. One link, one grammar: loss scales with connector count (N).....32
Figure 3.5. Connector-count convergence.....36
Figure 3.6. NVIDIA Kyber (public, GTC): vertical compute trays; the in-rack scale-up domain at the leading edge of density.....38
Figure 3.7. NVIDIA Kyber NVL1152.....39
Figure 3.8. Copper connectors and midplane (GTC, March 2025) — high-density copper connectors (left) and the assembled midplane they populate (right).....42
Figure 4.1. Photonic Interconnect: Same architecture as copper; different economics.....43
Figure 4.2. Example Concept CPO Switch with ELSFP Shown at ECOC 2025.....44
Figure 4.3. The two axes of scale-up interconnect.....46
Figure 5.1. Reaching for Ideal Topology and Eliminating Constraints.....49
Figure 5.2. Conceptually, three scaling axes of the AI fabric, scale-up, scale-out, and scale-DSA — over a unified passive fiber plant.....54
Figure 5.3. The three traffic classes on coupling scope versus latency envelope, split by the electrical reach limit: scale-up within it, scale-out and Scale-DSA beyond it over the shared multi-plane optical fabric (cf. Table B, Three Scaling Axes Compared).....55
Figure 5.4. Example of Fiber Attachment Unit Connectivity.....62
Figure 5.5. Co-Packaged Optics and Fiber Attach Units.....66
Figure 5.6. Near Package Optics and Fiber Attach Units.....67
Figure 5.7. Socketed Optics and Copper High-Speed Interfaces.....68
Figure 5.8. Mezzanine-Socketed Optics and Copper High-Speed Interfaces.....69
Figure 5.9. XPO Optics with Copper High-Speed Interfaces.....70
Figure 5.10. External Laser Small Form Factor Pluggable (ELSFP).....71
Figure 5.11. OIF EEI AI Compute aggregates Pods with Scale-out links.....77
Figure 5.12. The interconnect stack, from Copper to Photonics.....82
Figure 5.13. Class-H and Class-V trays composing rack architecture.....84
Figure 5.14. Example Compute Tray (Class-H) with 4 XPU.....85
Figure 5.15. Example 2-ASIC (Rail Type Switch) Configuration with XPO.....86
Figure 5.16. Example 8-XPU MHS Configuration.....89
Figure 5.17. Class V canister depth-budget view.....101
Figure 5.18. Passive optical infrastructure components — Optical Faceplate, ODF, Fiber

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Raceway, Rear Interface Zone, and Front-Panel Module Thermal Envelope....112
Figure 5.19. Example of Modular and Configurable Optical Faceplate for Compute Tray showing ELSFP and VSFF (Very Small Form Factor) Connectors....113
Figure 5.20. Tray Depth Zones and Placement Coordinates....114
Figure 5.21. Example of Optical Distribution Frame....117
Figure 5.22. Example of ELS Cage positioning in Compute Tray....121
Figure 5.23. Example of ELS Cage (Liquid Cooled)....122
Figure 5.24. Laser distribution and recovery domains....125
Figure 5.25. Example of Optical Distribution Frame showing main fiber routes....126
Figure 5.26. Example of Expanded Beam Optical Connectors....128
Figure 5.27. Example of Front and Rear I/O Access in Compute Tray....129
Figure 6.1. Example ingress and egress I/O FRU Placement....138
Figure 7.1. Representative Example: Optical 16-λ DWDM Bi-Di Link Architecture....148
Figure 7.2. One reference, two media, one boundary: a frequency carrying an embedded phase marker, distributed over copper and embedded onto the optical plane at the board edge....155
Figure 7.3. What needs to be synchronized, and what provides it....157
Figure 8.1. Demand versus the seating element, converging on the copper-reach wall....161
Figure 8.2. Compute sits below the copper shoreline; the switch is first to exhaust it....165
Figure 8.3. Reach as the binding constraint: copper reach collapsing with lane rate, the optical rollup flat over distance....168
Figure 8.4. Throughput: the perimeter-versus-area escape gap, the connector bill by rollup, and the tray-height surface....172
Figure 9.1. (Above) ORv3 Rack with Class H trays....178
Figure 9.2. (Above) ORv3 Rack with Class V trays for compute and Class H shown for switching in front only placement....178
Figure 9.3. Class V Showing rear-switch mode vs. distributed rail-switch mode; dimensions are illustrative, not normative....179
Figure 9.4. Three-Phase Platform Evolution....181
Figure 11.1. Binding-constraints across the four scale anchors....194
Figure 12.1. SDO Governance Map across the hyperscaler build flow....204
Figure 13.1. Relationship between this white paper, future efforts, and external groups....209
Figure B.1. Photonic ecosystem build-flow governance reference (full-detail variant)....229
Figure B.2. Relationship between this effort, the OCP, and outside SDOs and future efforts...230
Figure E.1. The memory-hierarchy load-to-use ladder: the roughly seventyfold gap from the sub-microsecond pooled tier to bulk storage, and the resulting average-access-time reduction at h ≈ 0.9 and capture ≈ 0.9....284
Figure E.2. Mean remote-access wait gating a 512-way collective....285

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table of Tables

Table: Target Outcomes..... 11
Table A: Optical Link Topology Composition (Architectural Framework)..... 33
Table B: Three Scaling Axes Compared..... 57
Table C: Optical Engine (OE) Classes..... 60
Table D: Contract Dimensions..... 73
Table E: MHS 1OU Common Tray and Chassis Envelope..... 81
Table F: The Radix Ladder at 200G and 400G per Lane..... 94
Table G: The Lane-Rate Axis (radix held at 1,024)..... 97
Table H: Tray Placement Positions..... 115
Table I: 224 Gb/s PAM4 per lane..... 140
Table J: 448 Gb/s PAM4 per lane (PAM4 baseline; see Forward-looking note on modulations beyond PAM4)..... 141
Table K: Network Profiles..... 149
Table L: Latency Characteristics..... 152
Table M: Class-H Reference Provisioning — Reach Scenarios..... 163
Table N: Class-V Reference Provisioning — Reach-Bound Case..... 166
Table O: Perimeter-versus-Area Escape at the Switch..... 169
Table P: Class-V Throughput — Connector and Laser Bill by Rollup..... 170
Table Q: Tray-Height Lever — 204.8T Switch Trays per Tier..... 172
Table R: Socketed near-package escape — convergence for 204.8T scale-up..... 173
Table S: Platform Form Factors..... 176
Table T: Serviceability..... 185
Table U: Photonic HBD..... 189
Table V: Four-anchor scaling rollup..... 192
Table W: Platform Integration Requirements Coverage..... 197
Table X: Deliverables..... 209
Table Y: Bandwidth Matrix..... 220
Table Z: Per-Tray Optical BW and ELS Sizing by Configuration Case..... 224
Table AA: Fibers per ASIC by Profile (Worked Example)..... 226
Table AB: OE-Internal Loss Mechanism Summary..... 238
Table AC: OE-Internal IL Comparison at OCI 4λ Density..... 241
Table AD: Per-Topology Fiber Path Composition..... 242
Table AE: Total Link IL by Topology Class..... 245
Table AF: C.3 Reference Anchors and Public-Literature Citations..... 247
Table AG: Transmitter Envelope Components (Laser Power and Modulator IL)..... 250

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table AH: Modulator Output OMA Envelope per Channel....250
Table AI: Receiver Sensitivity Envelope Components....251
Table AJ: Per-Channel Sensitivity Envelope at Photodetector....252
Table AK: Available Budget at Spec-Grade vs Envelope Range....253
Table AL: SNR Closure Visibility Map by Topology and Modulation....254
Table AM: Residual SNR Margin per Topology × OE Class....258
Table AN: Worked Reference Cases by Lane Generation (DR and OCI)....262
Table AO: Tolerance Envelope Contributors and Physical Origins....265
Table AP: Escape Ladder — Copper, CPO, and Area Across the Switch Tier Ladder....278
Table AQ: Connector and Laser Bill Across the Ladder — Where Discrete Terminates....278
Table AR: Key Claims and Evidence....279

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

1. Mission

Purpose and Audience

This white paper builds on OCP's Open Systems for AI: Blueprint for Scalable Infrastructure [6], which frames AI infrastructure as a cross-project systems problem requiring common, evolvable building blocks rather than vertically isolated implementations. The ecosystem converged on a common modular platform, now the de facto standard for AI server infrastructure. The premise here is to evolve it for density, not replace it. Within that broader blueprint, this contribution focuses on the integration of the photonic high-speed I/O layer, with the placement classes, passive optical infrastructure, link profiles, validation contracts, and platform extensions needed to make open AI-scale photonic interconnect accessible and deployable through OCP. It serves three primary audiences, each with a distinct stake in the architecture:

For hyperscalers and cloud service providers: open specifications for extending scale-up domains from rack-bounded systems to row-scale and larger deployments, while preserving operator choice across vendors, MSAs, and standards tracks. The goal is interconnect that does not limit the accelerator capacity it serves.

For ODMs and system integrators: an envelope-and-contract method for qualifying multiple OE technologies, SerDes generations, and HPM configurations on shared passive optical infrastructure without redesigning the platform for each generation.

For silicon photonic suppliers and the open ecosystem: vendor-neutral interface and qualification contracts that define where components interoperate, without binding the architecture to one MSA or implementation family.

The architecture is informative and directional rather than prescriptive. It identifies what should be exposed for interoperability and recommends how systems may converge, without dictating each implementation. Its main contribution is the open interoperability contract: the observable semantics, interfaces, and behaviors that different

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

implementations will support. The design and specifications behind that contract remain the ecosystem's work.

The targets are the three concurrent scaling axes that define modern AI integration: scale-up within homogeneous compute domains, scale-out across fabric boundaries, and Scale-Domain Specific Acceleration (DSA) leveraging heterogeneous compute architectures.

Document Contract

This white paper is an informative architecture vision, not a normative specification. It describes the problem space, proposed architectural direction, candidate interface boundaries, and areas where open collaboration can produce interoperable photonic AI infrastructure. Any requirement-style language in this document, including uses of "should," "must," or similar terms, should be read as proposed direction for future specification work rather than as binding compliance language. Normative requirements, conformance criteria, validation methods, and pass/fail obligations are intentionally deferred to future OCP Base Specifications and related derivative specifications developed through the appropriate OCP workstreams.

Scope and Stance

This document focuses on and addresses the system tray and the rack interfaces that touch it, down to the network cabling. It does not develop interconnect beyond that scope, nor does it claim one modality is universally superior. Copper and photonics each have physical regimes where they are the rational engineering choice. The boundary between those regimes is generation-dependent, not absolute.

The concern here is narrower. Silicon photonics is entering AI compute platforms because interconnect bottlenecks reduce MFU, copper reach compression raises scale-up cost, and facility-scale power limits constrain each SerDes generation. These

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

pressures do not decide every placement. They make the electron-to-photon boundary an explicit integration decision.

The architecture therefore defines the open foundation needed for vendor-neutral photonic integration across SerDes generations.

Where the document analyzes copper reach, retimer cost, or channel margin, it treats those factors as integration-cost inputs to placement-class decisions, not modality verdicts. The integrator owns the boundary decision for a specific system. The document provides the framework for making that decision transparent.

Inter-site and multi-datacenter photonic infrastructure lies beyond this work; it falls to the OIF, OCP–IOWN AI Computing Continuum, and other relevant efforts, covered in Standards Alignment.

2. Compliance with Open Compute Project Tenets

A full explanation of the OCP core tenets can be seen here.

Openness

The architecture's contribution to openness is the envelope-and-contract methodology developed in § Envelope and Contract: every interface that joins two ecosystem participants is published as a vendor-neutral contract that any qualifying implementation can satisfy. Openness extends in two directions. Inward, the architecture integrates proven OCP work being perfected for AI systems (MHS-IAI, OSAI, OCP RAS API, Open Rack v3) rather than re-deriving foundations the broader ecosystem already maintains. Outward, it coordinates with external SDOs and MSAs (OIF, IEEE, JEDEC, DMTF, and the photonic-relevant MSAs) so work originated in OCP composes with work originated elsewhere. Contribution through the OCP open contribution model means rights flow

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

outward to all implementers rather than inward to a closed group; ecosystem-wide adoption is the architectural-success criterion rather than ecosystem-wide compliance to a single vendor's reference.

Efficiency

The architecture's contribution to efficiency operates at the system level rather than the component level. The integrating metric is effective pJ/bit (§ COM and Placement Selection): direct energy-per-bit divided by effective goodput, capturing both the nominal interconnect energy cost and the energy effectively lost to error-driven workload disruption. Photonic placement classes correctly chosen against the channel-evaluation function deliver lower effective pJ/bit than perimeter-limited copper at every relevant SerDes generation, with the gap widening as line rates advance. Efficiency improves further through complexity removal: each retimer eliminated subtracts both its power draw and its structural latency (tens to ~120 ns per traversal); each midplane PCB layer reduced or eliminated subtracts the materials and manufacturing complexity of the highest-loss segment of the channel. The architecture's discipline is to make placement-class selection grounded in effective pJ/bit and complexity cost rather than in inherited packaging convention.

Impact

On the consumer side, the architecture expands viable scale-up domains from rack-bounded to row-scale and beyond; at AI training scale, the HBD size determines which model topologies, parallelism strategies, and FLOPS utilization regimes are achievable, with interconnect as the binding constraint. The architecture provides the open foundation on which row-scale and pod-scale HBDs are buildable across vendors and across SerDes generations. On the supplier side, photonic interconnect at hyperscale is still pre-deployment relative to its addressable market; the photonic ecosystem currently spans more than fifty companies — component vendors, integrators, MSAs, foundries — whose specializations are not yet composed into open

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

AI-system infrastructure. The architecture's open contract layer is the substrate on which that composition becomes practical, mobilizing the existing photonic ecosystem into AI-system participation rather than waiting for vendor-by-vendor MSA negotiation.

Scale

The architecture's contribution to scale is generation-invariant passive infrastructure paired with generation-aware active components. The modular tray geometry supports increases in photonic performance such as link architectures without the need for replacement; active OEs refresh on their own cadence; the placement-class taxonomy accommodates inward migration of the integration boundary as channel margin tightens at each SerDes generation. Scale is spatial (1024+ XPU HBD horizons), temporal (multi-generation infrastructure lifecycles), and transitional. The Link Profile framework (§ Link Profile Framework) extends infrastructure discipline across the scale-up, scale-out, and inter-pod boundaries so that profiles are protocol-agnostic and infrastructure-shared, reducing the structural thrash that typically accompanies tier transitions and the architectural rework that historically follows each generation step.

Sustainability

The architecture's contribution to sustainability operates across three timescales. Operationally, photonic placement classes cut interconnect energy and network power share toward the targets in § Target Outcomes; at hyperscale, each joule eliminated at the interconnect is also a joule eliminated at the facility cooling plant.
In embodied-material terms, the generation-invariant passive fiber plant decouples infrastructure lifecycle from the SerDes generation cycle: the same fiber, raceway, ODF, and faceplate support both 224 Gb/s/lane and 448 Gb/s/lane engines without replacement; 40+ layer midplane PCBs are eliminated or reduced to passive fiber distribution; and area-domain integration scales bandwidth per unit reticle area with L^2 rather than L (§ Placement Class Taxonomy), shrinking the silicon required per Tb/s.

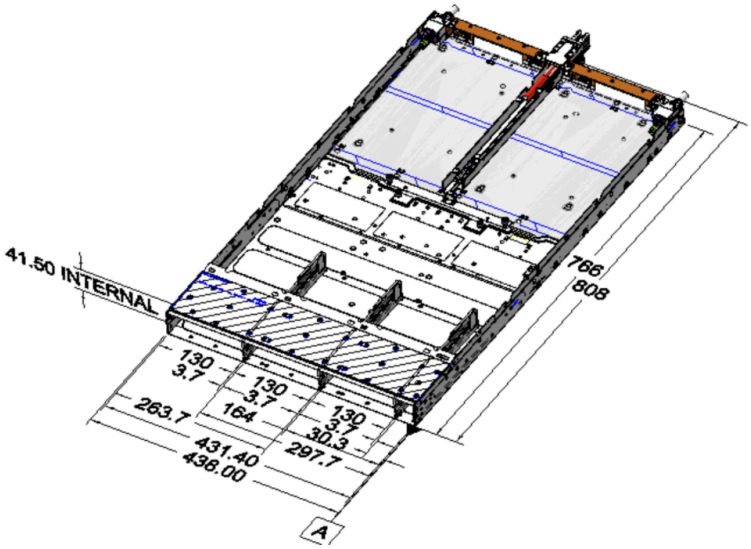
This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Across the component lifecycle, FRU-level serviceability (§ Serviceability Through Architecture) combined with secure component-instance identity (§ Ecosystem Collaboration and Next Steps) supports repair, refurbishment, and second-life pathways that deeply integrated assemblies foreclose. Modular boundary placement is therefore not only a reliability and economics decision but a circularity decision.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

3. Introduction

An architecture is always an inheritance before it is an invention. We have made the case for light; what remains is how to build it in the open, so that what one team proves, the whole ecosystem can build on. That discipline already has a home and a history. The Open Compute Project's Modular Hardware System (MHS) Subproject defines interoperability among key elements of data-center, edge, and enterprise infrastructure, supplying consistent interfaces and common form factors across modular building blocks. The first specifications, contributed in the open in 2022, now underpin much of today's dense AI training and inference infrastructure. The OCP MGX Accelerated Computing Rack and Trays Specification, contributed by NVIDIA in 2024, extends those building blocks into the high-density rack and tray envelopes that accelerated computing now requires [16]. Together, MHS and MGX provide an open foundation for the AI hardware ecosystem.



This whitepaper proposes the next extension of that foundation: integrating silicon photonics as the high-speed interconnect layer for next-generation AI systems. The extension is therefore not only integrative but compositional: existing MHS and MGX form factors are reused in new rack configurations to support density classes beyond what current platforms accommodate. What is added is the platform infrastructure required to host photonic interconnect at the tray boundary: the Optical Faceplate,

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Optical Distribution Frame (ODF), Raceway, optional ELS positions, an optional Rear Optical Interface Zone, and the associated manageability hooks. The workstream therefore extends the platform grammar as well as the interconnect medium. This work defines the technical scope and architectural vision for the proposed MHS Interconnect for AI (MHS-IAI) workstream.

The Interconnect Problem

Every successful technology eventually meets the physics that limits it, and copper has reached that point. The obstacle is not a single wall but five, closing in together. Copper is being overtaken on five axes: the information a medium can carry, the channels a processor can release from its edge, the paths threaded across a board, the distance a signal travels before it dissolves into noise, and the energy spent per bit. On each, the constraint tightens with every doubling of rate rather than easing.

Photonics does more than extend the same channel. In its richer integration forms, it relaxes three of these limits at once: wavelength and bidirectionality can add information dimensions, area-domain optics can escape through the chip face rather than only its edge, and fiber routes through space with far less distance-dependent penalty than copper while disrupting copper's bandwidth at distance with power lock. One difficulty stays common to both: the connector, the necessary seam where one path joins another. Fiber needs very special consideration to manage operational performance. Electrons were easy but now copper needs this same level of attention.

The following subsections take the axes in turn, resting on the OCP, the OIF plus other SDOs and MSAs work where the fundamentals are settled and arguing only where they are not.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

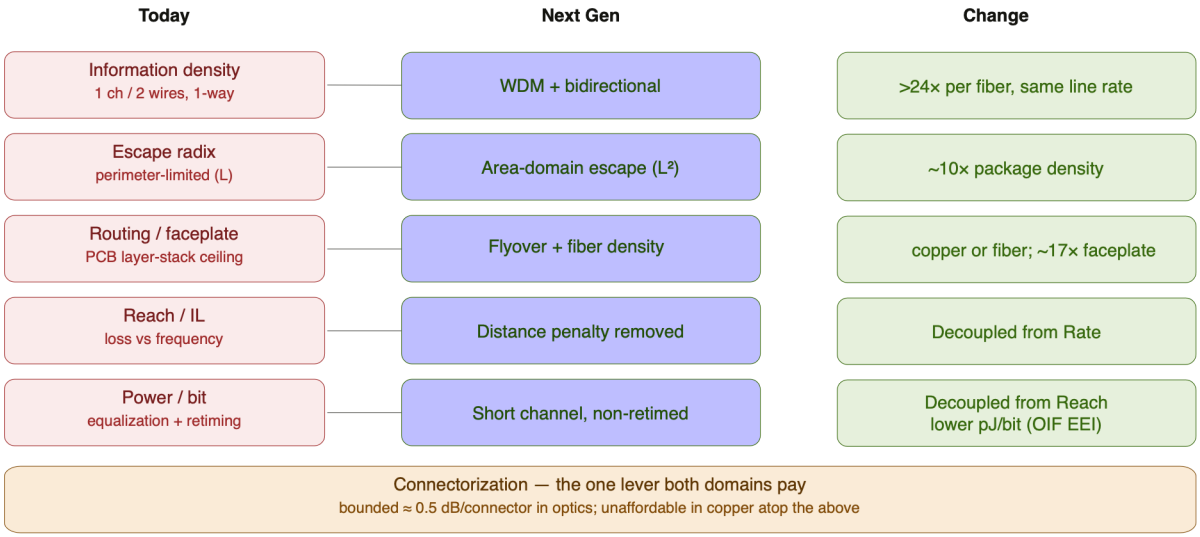


Figure 3.1. *Moving forward from Today's Constraints*

Copper limit → photonic improvement vectors across five axes: information density (>24×, WDM × bidirectionality), escape radix (~10×, area-domain), routing / faceplate density (~17×), reach & insertion loss, power per bit — with connectorization as the one shared lever (~0.5 dB / connector).

3.1.1. The Copper Reach Constraint

Copper's resistance to high frequency is a property of the medium, not a defect in it, and at the rates frontier AI now demands the channel spends its margin well before it covers any useful distance. At the ~112 GHz Nyquist of the 448G generation, the loss that finally closes the channel is often dominated by connector and reflection penalties; bulk dielectric attenuation (which crosses three decibels an inch on ordinary board material and holds near one only on flyover twinax [Figure 3.3]) is the secondary contributor. On standard 32 AWG twinax the workable reach falls from about a meter at 224 Gb/s/lane to roughly a quarter of that at 448 Gb/s/lane. The trajectory is settled: the OIF's long-reach work has tracked it for a decade (three meters at 56G, two at 112G, about one at 224G, and 448G still open [OIF CEI-LR[3]]), so the section cites the slope rather than re-deriving it. Reach falls with each doubling of rate; demand moves the other way.

[3] https://www.oiforum.com/technical-work/hot-topics/common-electrical-i-o-cei-224g/

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

The industry continues to work towards point solutions, but at this stage they are postponements rather than escapes. This group includes retimers, lower-loss laminates such as Megtron 9, heavier-gauge conductors or waveguides, and deeper PCB stackups. Ignoring the literal economics, each restores some reach and charges for it elsewhere, in power, latency, or board area, and each returns less as the rate climbs [Figure 3.1]. Retimers expose the core tradeoff. Reach and power are not separate problems; they are the same constraint viewed from opposite ends. Each added stage helps close the electrical channel, but it also adds equalization energy and conversion latency. The channel that finally meets budget is therefore the one that burns the most energy per bit and takes the longest to cross.

OIF's energy-efficient-interfaces work makes this coupling visible by plotting transceiver power against electrical loss between the XPU and OE [OIF EEI-COI]. The implication is direct: copper buys reach with both power and time. At these rates, those costs arrive together.

The effect compounds with scale. Each generation the radius a processor reaches grows, as the models we train demand wider domains, while the ultimate reach a signal can sustain contracts. The failure is not that any one link breaks but that the reachable domain retreats from the demand placed on it. Photonics does not merely postpone this; it changes the scaling law. Over the short distances inside a rack the loss of a fiber is negligible, so reach largely decouples from rate over the rack scale distances considered here; the electrical path shrinks to the short hop from die to OE, and the retiming that taxed both power and time can be eased or removed — leaving little but the propagation itself, on the order of five nanoseconds a meter. Here, on the axes of reach and power, the gap between copper and light stops being a difference of degree and becomes a difference of kind. How much of the recovered margin survives the journey is set by the connector — the subject of the next two subsections.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

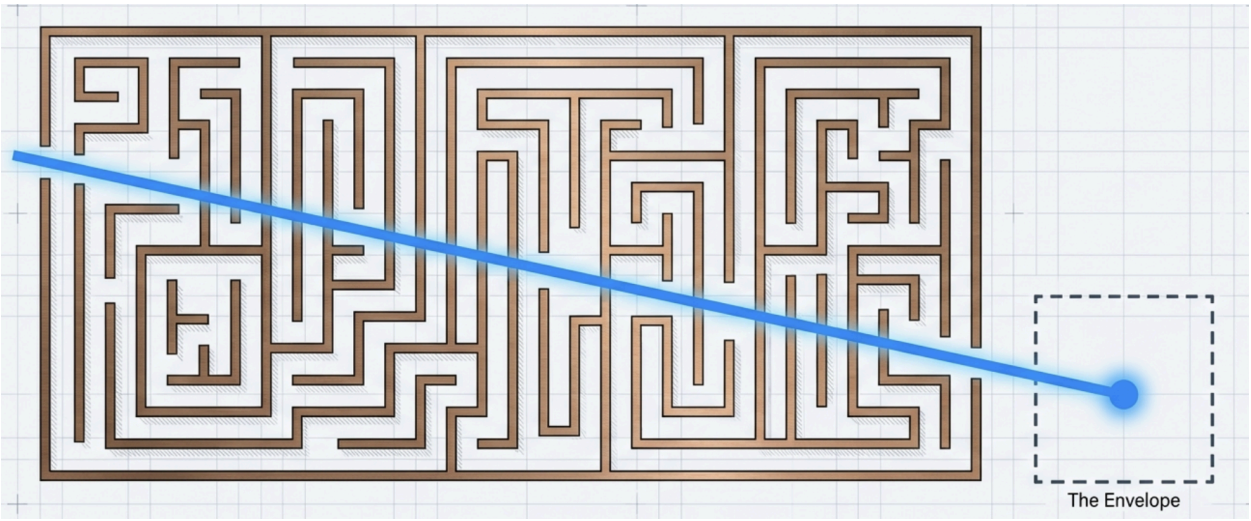


Figure 3.2. *Escaping The Copper Bottleneck*

The consequence is that the number of accelerators reachable within the copper radius shrinks at each generation, precisely when AI model scaling demands larger domains. The industry response has been to add retimers, use heavier gauge conductors, and engineer increasingly complex PCB stackups. These are mitigations, not solutions; each adds power, latency, cost, and board area for diminishing returns.

The architectural framework for choosing between extending copper reach and migrating the integration boundary inward is developed in § Placement Class Taxonomy and § Channel Operating Margin (COM) and Placement Selection.

3.2. Channel Budget Reference

Channel reach is not a standalone number. It is an operating point in an electromagnetic channel budget. The same physical distance can pass or fail depending on line rate, placement class, retiming model, host-PCB allowance, transmission-line structure, and the impairment profile. That profile captures loss, dispersion, impedance discontinuities and reflections, crosstalk, mode conversion, jitter, noise, and equalization limits. A reach claim without those assumptions is incomplete. This architecture records them once, in

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

a single reference, so later decisions cite the same channel basis [Tables I and J; § COM and Placement Selection].

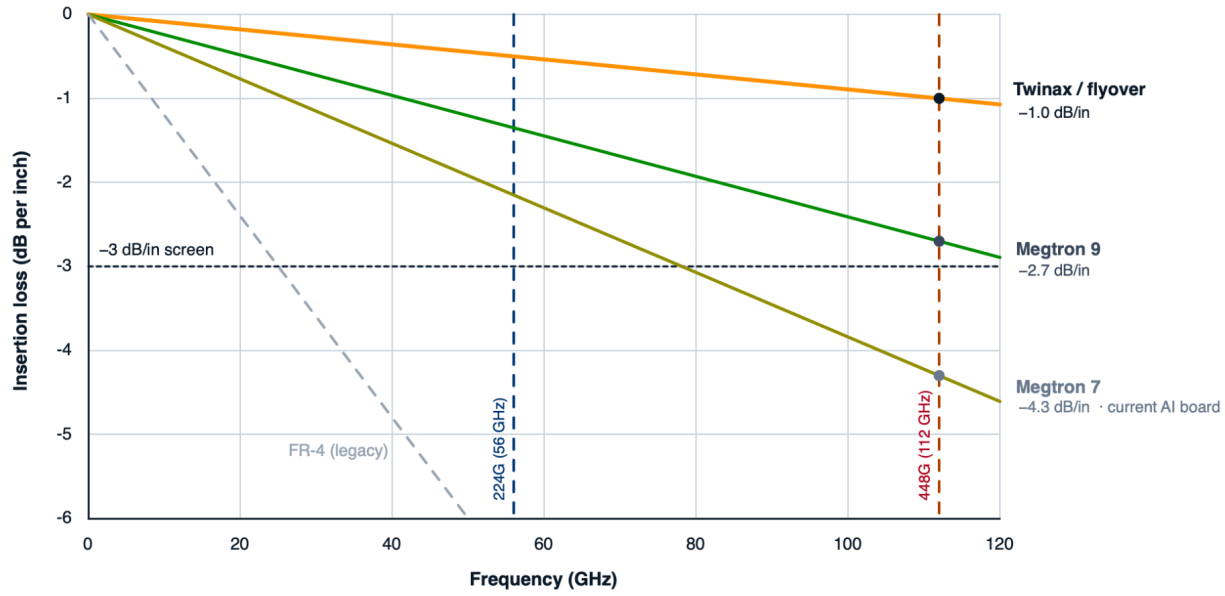


***Figure 3.3**. Per-inch insertion loss at the 448G Nyquist (112 GHz).*

FR-4 and Megtron 7 breach the −3 dB/in screen before the Nyquist; Megtron 9 holds just under it; flyover twinax holds ~1 dB/in, ~63% below Megtron 9. This is loss-only reach; integrated-channel reach (Table J) adds connector and COM-closure penalties.

Figure 3.3 turns material choice into a design constraint. Even the best near-term dielectric clears the 448G screen only thinly, so a better board alone cannot open the link budget. At 224G a board that passes per-inch still buys short reach once the full channel is counted (Table J). Per-inch loss therefore gates how far the OE can sit from the die. Keep it proximate, near-package, and the short electrical path usually holds margin without a retimer. Place it farther out, and the designer reaches it over twinax flyover, where the longest paths still need a retimer to recover that margin. The placement axis below sizes those transitions.

Placement class defines the electrical path from the XPU die to the OE [Fig. 3.4, left segment]. Moving the OE outward from the die adds transitions. Co-packaged

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

placement avoids host-board connector transitions. Near-package placement adds the package-to-host transition. A socketed form adds the socket. A mezzanine form adds the carrier edge. Faceplate placement adds the full host-PCB run and cage.

At 224 Gb/s and above, those transitions become increasingly critical. They have insertion loss and reflections to contend with. Reflection can be as limiting as loss, so the channel will ultimately be judged by a metric that accounts for both. Channel Operating Margin (COM) provides that frame. This architecture uses COM, and standards-maintained equivalents, for electrical-channel evaluation while leaving metric construction to OIF and IEEE 802.3.

Detailed placement selection belongs in the later COM section. The rule here is simpler: the farther the OE sits from the die, the longer the run, and length is paid in both loss and reflection.

3.3. Optical Link Composition

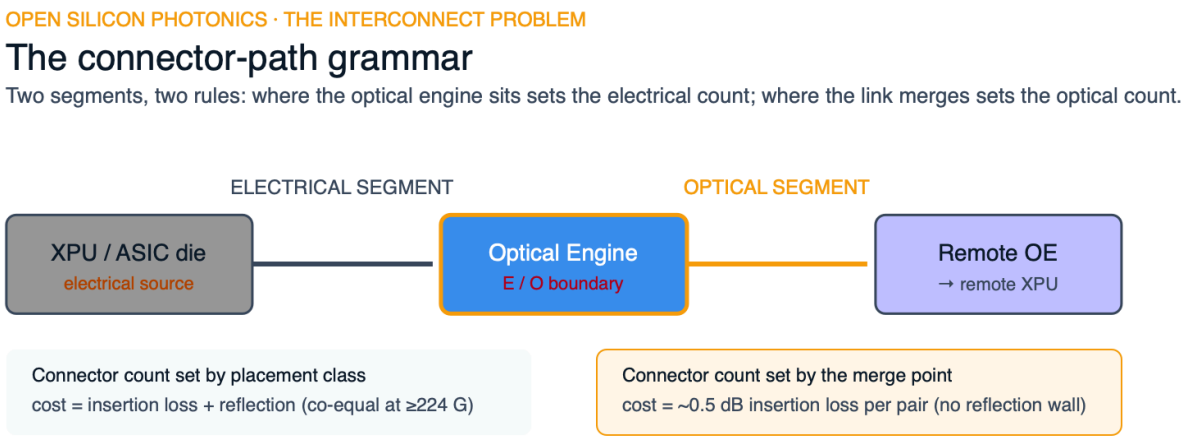


Figure 3.4. *One link, one grammar: loss scales with connector count (N)*

The shared connector-path grammar: XPU / ASIC die → optical engine (OE) → remote OE. Electrical-segment connector count set by placement class (loss + reflection); optical-segment count set by the merge point (~0.5 dB IL each); the OE on the E / O boundary.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

High-speed interfaces are converging in the discipline they demand. Low-speed GPIO once behaved like a simple continuity problem: preserve the path, control capacitance, and the signal arrived. A 224G electrical channel is different. It is a transmission-line network, where insertion loss, reflections, crosstalk, jitter, noise, and equalization limits determine whether the link closes.

Photonics is not simpler. It runs on different physics, but it demands comparable budget discipline. That difference in physics becomes useful because, at rack scale, fiber attenuation is small. The dominant architectural variable is usually connector count.

In practice, a clean production optical connector pair costs about 0.5 dB at these rates. The fiber counts that loss once, and counts it small [Figure 3.4, right segment]. This lets the designer treat a connector as a lever to spend deliberately, not merely as a penalty to absorb.

The count is set by the merge point: the place along the OE-to-OE path where the producer and consumer sides are joined. A bonded engine-to-engine fiber path carries no service connector in the span. A serviceable point on the same tray adds one. Tray-to-tray cabling adds two. Routing through a distribution frame adds three or four. Shuffle-routed paths across racks can reach six connector pairs in line.

Because fiber attenuation over these distances is small, connector count governs the optical budget more than distance. The budget is therefore not mainly a fact of geography. It is a consequence of where the architecture places its service seams.

Table A: Optical Link Topology Composition (Architectural Framework)

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Topology class	Description	Connector pairs in path	Typical span	Dominant loss variable
Direct intra-tray (bonded)	OE-to-OE via continuous bonded fiber path	0	< 1 m	OE-internal coupling
Direct intra-tray (connectorized)	OE-to-OE with single serviceable connector pair	1	< 1 m	OE coupling + single connector
Point-to-point intra-rack	Direct connectorized cable between trays	2	1–5 m	Connector count
ODF-routed intra-rack	Routed through ODF for topology flexibility	3–4	1–10 m	Connector count
Shuffle-routed inter-rack	ODF + inter-rack harness	4–6	5–50 m	Connector count + inter-rack attenuation
Inter-pod with format conversion	XPO/OFC at pod boundary	4–6 + O→E→O stage	5–100 m	Format-conversion stage

This is the tension the section needs to expose. Every connector removed reduces manufacturing complexity and returns about 0.5 dB to the budget. That margin can be spent on reach, denser modulation, lighter FEC, or aging and contamination tolerance.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Every connector also buys something real: tray serviceability, topology flexibility, and fault isolation.

The architecture's position is not that fewer connectors always win. It is that serviceability should be considered deliberately. It is paid from the same optical budget as reach and modulation, in the same units. The merge point is where that bill is set.

Put in numbers, the lever is small but material. At about 0.5 dB per production connector pair, using IEC 61753-1 Grade B loss expectations and manufacturer-independent random-mate qualification under IEC 61755-3-3, the topologies in Table A span roughly 0 to 3 dB of connector loss. Bonded intra-tray links sit at the floor. Shuffle-routed inter-rack links sit near the ceiling.

One property qualifies a link for this table before any budget is counted: tolerance of mid-span mating. The accounting above prices a connector pair as insertion loss and nothing else. That pricing presupposes the mated pair leaves the link's propagating mode set undisturbed. A link class that depends on a launch condition a connector cannot preserve occupies the bonded and single-connector classes only, whatever its loss budget. It crosses higher topology classes through a format-conversion stage, and its energy accounting becomes per-segment rather than per-path. Candidate link profiles should declare this tolerance alongside reach and loss.

At 224G, that 3 dB span is close to the gap between a perimeter-placed engine and one integrated in the area domain. Public results bracket the same design limit from both sides. Production scale-out fabrics avoid routed optical paths beyond about seven connectors to protect signal quality. Co-packaged links have also reported BER below 1e-9 without FEC through about six connectors over more than a kilometer of fiber.

Those two cases are different, but they point to the same rule: at rack scale, connector count usually matters more than fiber attenuation. The architecture should therefore place service seams where they earn their optical cost.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

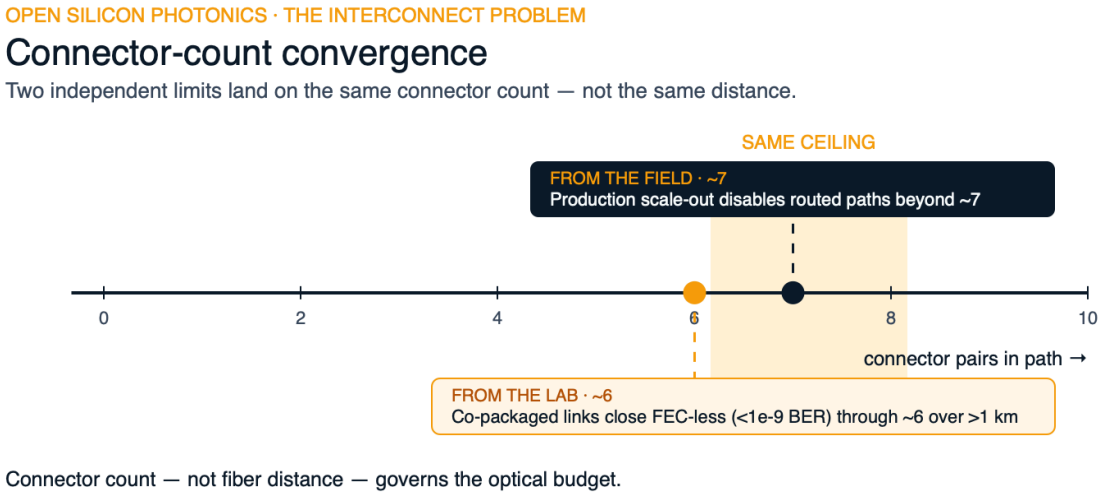


Figure 3.5. *Connector-count convergence*

What remains is density (how much a medium can carry, and how much a chip can release), which is where the gap between copper and light opens widest.

3.4. The Density-Bandwidth Dilemma

For decades, Moore's Law set the pace of density. The pace kept slipping: roughly a doubling every year in Moore's 1965 estimate, every two years by his 1975 revision, and every three years or more in the last decade as each step became harder and more expensive.

Whole industries were built on that cadence while it held. As it has slowed, AI's answer has been to spread the work outward across more silicon, joined by more interconnect. That pushes the problem from transistor density into package density, rack density, and I/O density.

The packages themselves are nearing practical density limits. The I/O that leaves them is bounded by geometry: high-speed pins sit at the die perimeter, and perimeter grows only with the edge. Raising the lane rate from 224G to 448G doubles what each pin can

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

carry, but it also shortens electrical reach. Both apparent escapes, a wider edge or a faster lane, run into physical limits.

AI data-center economics still reward density. The more accelerators a rack can hold, the more of them fall within a signal's time-reach, and the lower the incremental cost to power, cool, and house each accelerator. The rack therefore continues to densify every generation. Public roadmaps make the trajectory plain. NVIDIA's Kyber architecture vision is the clearest marker: compute trays rotated 90 degrees roughly double the in-rack scale-up domain, in a frame drawing 600 kW or more against today's 120 kW racks [Figure 3.6; NVIDIA, GTC].

Density, however, is not free. The bill is paid in interconnect. Every accelerator added must still reach the others, and the bandwidth moving inside the rack rises faster than accelerator count. Tens of terabits per second will leave each part and cross the short distance between tray and backplane.

That compression improves system economics, but it shifts the limiting pressure onto interconnect. More information needs to move through each medium, more channels must escape each die, and more paths will traverse the board and backplane. These constraints rise together, but they do not become limiting at the same rate. Escape is exhausted first, and therefore governs how large the scale-up domain can grow [Figure 3.1].

Tellingly, even the densest publicly shown rack concepts still route this traffic through copper backplane or midplane fabrics. One vendor leans on bidirectionality as the key multiplier, but that multiplier becomes harder to preserve with each millimeter of electrical path as lane rates climb. The next two subsections examine that wall at its sharpest points: escape from the package and reach across the rack.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.



Figure 3.6. *NVIDIA Kyber (public, GTC): vertical compute trays; the in-rack scale-up domain at the leading edge of density.*

3.5. Case Study: The Inter-Canister Tapering Problem

A conventional rack holds on the order of eighteen compute trays; to push past that, the trays are rotated vertically and gathered into canisters, a boundary that is at once physical and electrical. Each canister holds about as many trays as a whole rack once did; two such canisters sit in a typical design, and as many as four at the limit, so a

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

rack's tray count is multiplied two- to fourfold at a stroke.

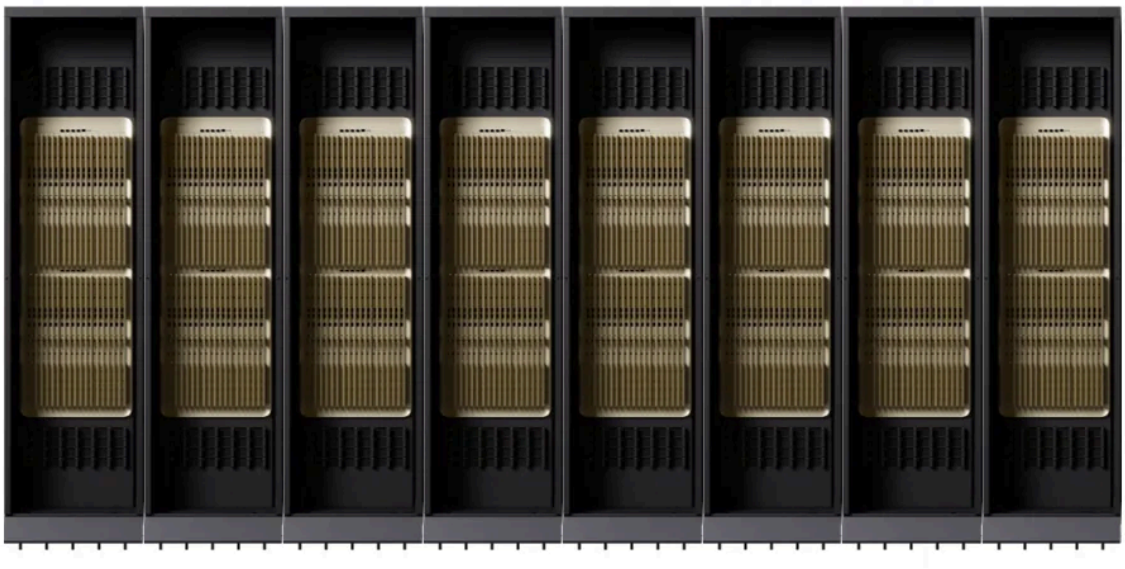


Figure 3.7. *NVIDIA Kyber NVL1152*[4]

The canister does double duty, carrying the rotated trays and holding the midplane that wires their all-to-all connectivity. Every tray-to-tray conversation then runs the same course: out of the tray, through the midplane to a switch, and back through the midplane to its destination. Inside a single canister this is one hop (tray to switch to tray), and the fabric behaves, locally, as one uniform thing.

The difficulty is radix, or, how many connections a single device, here a switch package, can present. The tray count has been multiplied; the ports one switch can offer have not. No single switch can serve every tray in every canister, so the switching is divided, a share to each canister. A conversation that stays within a canister keeps its single hop; one that crosses to another can no longer reach its partner through a single switch. It runs from its tray to its own switch, across to a second switch in the far canister, and only then to the destination tray resulting in two switch hops where the local case had one, and more of the midplane traversed in the bargain. The fabric meant to behave as

[4]https://developer.nvidia.com/blog/nvidia-vera-rubin-pod-seven-chips-five-rack-scale-systems-one-ai-supercomputer/

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

one uniform domain acquires a grain instead: neighbors are cheap, strangers dear, and the line between them is drawn not by the workload but by how many ports a switch can present, the escape-radix limit of the opening figure, met.

The cost lands on the workload, which does not know the grain is there. A model that assumes uniform bandwidth across its domain pays the difference in stalls and imbalance wherever its traffic crosses a canister line, and the difference is not small: in publicly discussed designs the bandwidth across the boundary can fall to about a third of within-canister bandwidth. The obvious remedy (flattening the fabric so every canister talks to every other at full rate) has been estimated to consume on the order of forty percent of a large rack's entire power budget, which is hard to justify thermally and economically. The tapering is invisible on the datasheet and felt in the run: a seam imposed by the radix of the switch and the midplane it crosses, not by anything the computation asked for.

In the public material available to date, the densest demonstrated racks finesse the problem rather than solve it: their canisters run as separate domains with tiers of interconnection; the unified rack-scale domain they are meant to deliver remains under development.[5] The seam is real, and present, and exactly the kind the later sections set out to avoid.

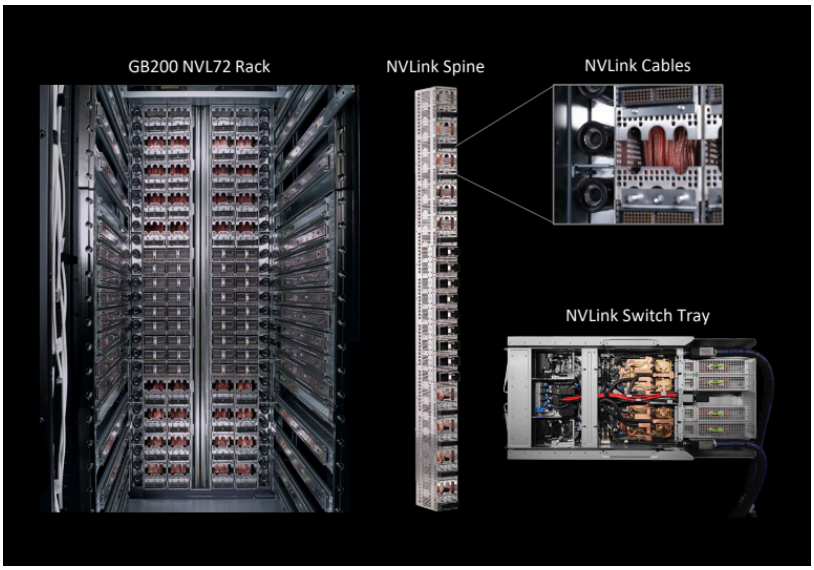
3.6. The Midplane PCB as Single Point of Complexity

In an AI system the network exists to gather compute and memory into one resource, which means the I/O from every switch and every compute die reaches every member inside the desired radius. In today's racks, NVLink™/Ultra Accelerator Link™ (UALink™) traffic converges on a cartridge with kilometers of copper wires that attach as a backplane in an MGX-class system. The proposed next, denser generation retires the cartridge in favor of a midplane: as figure 3.8 shows, a board with routing and connectors on each side, compute on one face and switching on the other, onto which

[5] This analysis is not a judgement on any approach or organization, merely observations to help guide by learning from earlier work.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

the entire scale-up fabric is pressed. It carries every differential pair, every connector, and every routing layer the fabric demands.



A board asked to do this much becomes a single point of complexity in the most literal sense. It may demand up to forty or more layers of controlled-impedance routing, can be built by only a handful of qualified fabricators, and gets redrawn from scratch for every SerDes generation; its signal space is finite, and its routing channels fill before the accelerators they serve are satisfied. And because the whole canister hangs on it, a single bent pin or cracked trace is not a tray's failure but possibly the canister's. The midplane is where the routing-density and reach walls of this section meet and concentrate, it's a design choice that is an improvement over cartridges for the task of increasing density. That a rack at the leading edge of density still rests on potentially a single point of failure that is not optimal for field maintenance is the cautionary note this section ends on. Section 3 has named the problem on each of its axes; the sections that follow are built to avoid these seams rather than to engineer around them.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

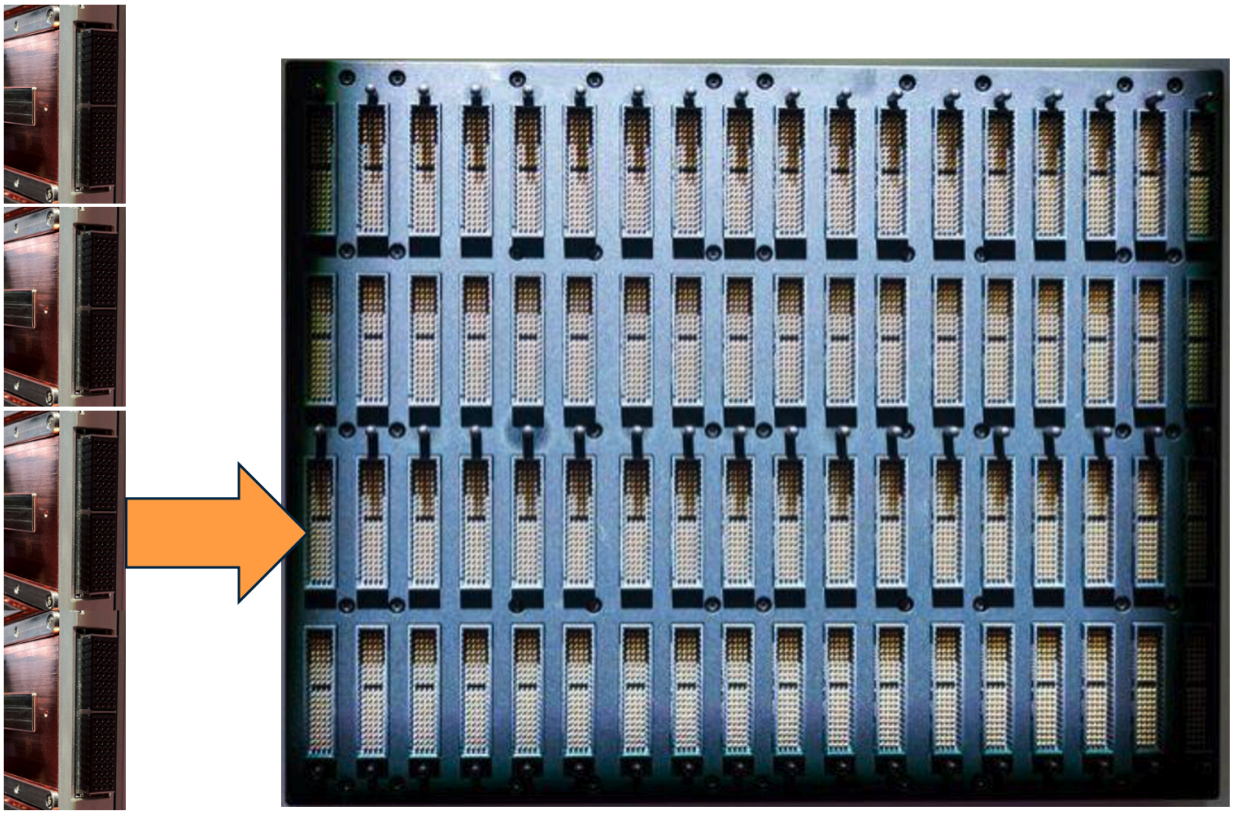


Figure 3.8. *Copper connectors and midplane (GTC, March 2025) — high-density copper connectors (left) and the assembled midplane they populate (right).*

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

4. The Photonic Interconnect System

At its highest level, a photonic interconnect system has three structural elements: a light source, an optical transport medium, and the I/O access points where light enters and leaves the platform. This composition is structurally identical to a copper interconnect system, which has an electrical power source, a copper transport medium (cable or PCB trace), and connectorized I/O at chip packages and faceplates.

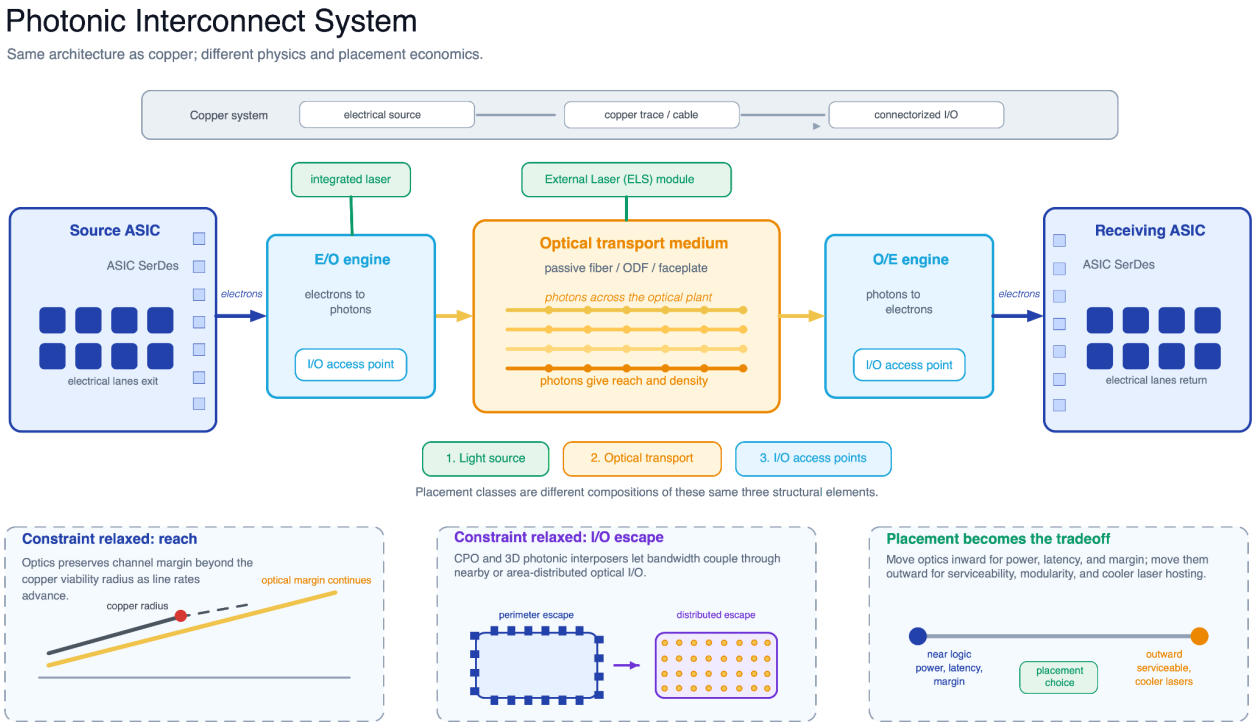


Figure 4.1. *Photonic Interconnect: Same architecture as copper; different economics*

The difference is physics, not architecture: a photonic system converts electrons originating at the ASIC SerDes into photons, moves them across distances and densities that copper cannot serve, and converts back at the receiving ASIC. Everything else in this document — placement classes, link profiles, faceplate geometry, thermal contracts, exchange formats — is the working-out of how those three elements compose at AI-infrastructure scale.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

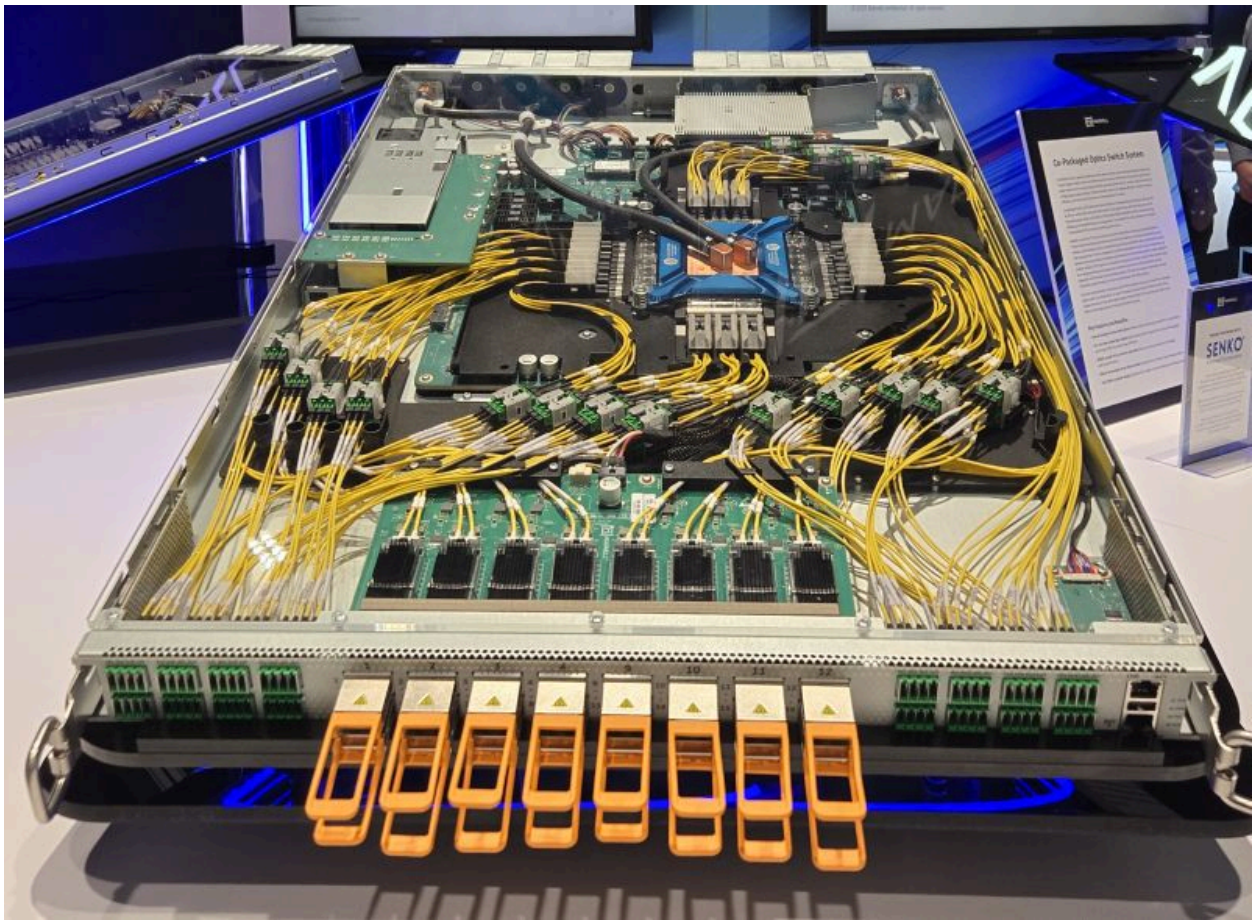


Figure 4.2. *Example Concept CPO Switch with ELSFP Shown at ECOC 2025*[6]

The architectural value of photonics is not that photons replace electrons one link at a time. It is that photonics relaxes two constraints that now shape AI platforms: **reach** and **I/O escape**.

Reach is the simpler axis. At each SerDes generation, optics can preserve channel margin beyond the copper viability radius, and that advantage widens as line rates rise. This is the familiar optical benefit: a signal can travel farther, or close with more margin, than an equivalent electrical path.

[6] https://www.marvell.com/blogs/co-packaged-optics-for-next-wave-ai-data-centers.html

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

I/O escape is a different axis. It asks how much bandwidth can leave the silicon in the first place. When high-speed I/O exits through the package edge, usable escape scales with perimeter, not area. That creates a geometric ceiling: larger compute die may contain more logic, but their edge does not grow fast enough to support proportional bandwidth growth.

Photonics can change that equation in two ways. First, each fiber can carry many electrical-lane equivalents of bandwidth. With an eight-wavelength WDM path carrying 200G PAM4 per wavelength, one fiber carries 1.6 Tb/s in one direction. If the path is bidirectional, the same fiber may carry 3.2 Tb/s of aggregate full-duplex bandwidth. In that framing, one 200G electrical lane is one eighth of the unidirectional optical path, or one sixteenth of the aggregate bidirectional path.

Second, in CPO and 3D photonic-interposer architectures, bandwidth no longer has to escape only through package-edge lanes. Short electrical paths can couple logic to nearby photonic I/O. With deeper integration, the electrical-to-optical boundary can move inward. Photonic I/O can land on bump or ball arrays, or other area-distributed package interfaces, rather than only on die-edge lanes.

That shift makes placement an architectural tradeoff instead of a fixed physics constraint. Optics move closer to logic when power, latency, and channel margin dominate. They move farther outward when serviceability, modularity, and tray economics dominate.

Keeping reach and escape separate avoids a common conflation. Optical reach can support faster or longer links, but it does not by itself increase the number of routable endpoints in a flat scale-up domain. Co-packaging can buy bandwidth inside a perimeter-bound domain. Area-integrated optical I/O is what moves the domain ceiling.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

The governing distinction is therefore **perimeter versus area**, not simply **electrical versus optical** (Figure 4.3). Reticle size marks the radix inflection point for perimeter-bound escape. Appendix D.1 derives the scaling, and Appendix D.2 shows the single-reticle ceiling.

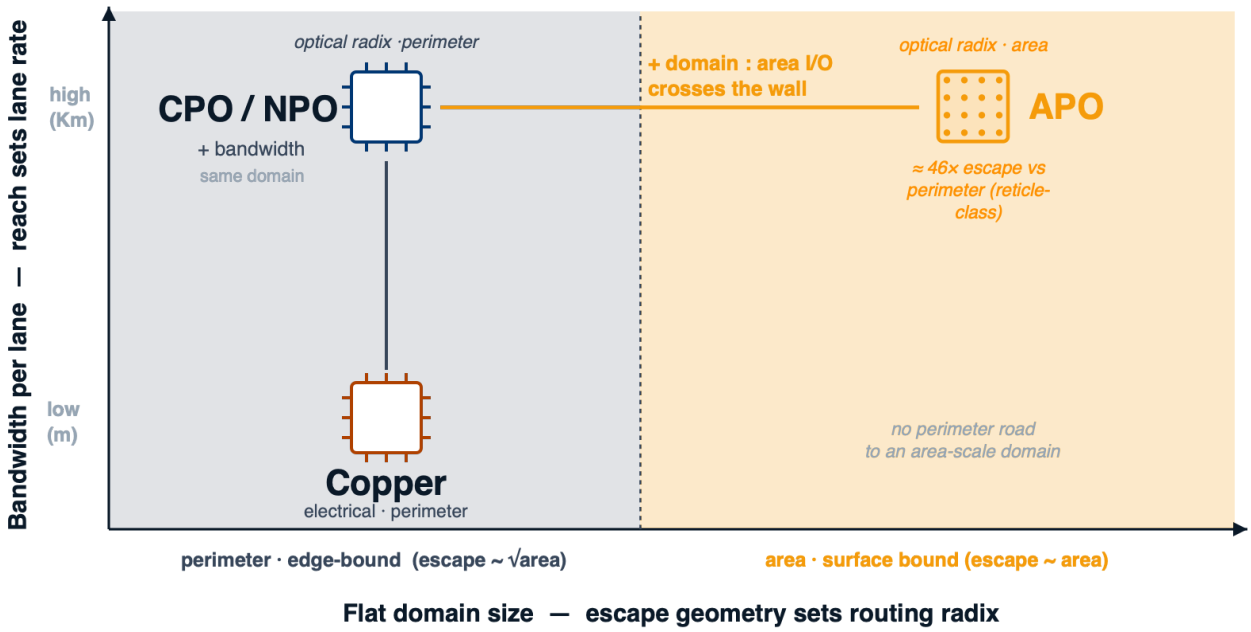


Figure 4.3. *The two axes of scale-up interconnect.*

Reach moves a design upward by supporting more usable bandwidth per path: higher lane rate, longer path, or more link margin. Escape geometry moves a design horizontally by setting how many endpoints can be reached in one flat domain. Copper and CPO/NPO remain perimeter-bound when their routable I/O escapes through the package edge. Area-distributed photonic I/O changes the scaling law by moving escape from the perimeter to the die face.

Three counts stay distinct for the diagram to read correctly.

The first is **physical escape count**: the number of routable interfaces leaving the die or package. For a given perimeter-bound geometry, raising lane rate changes the information carried by each interface. It does not create more physical escape points.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

The second is **optical transport capacity**: the bandwidth carried after the optical engine. This can be increased through lane rate, wavelength count, modulation, bidirectionality, or fan-out. Those changes can make each fiber or optical path carry more traffic. They do not automatically increase the number of destinations the silicon can directly address.

The third is **routing radix**: the number of endpoints reachable in a single flat scale-up domain. This count sets the domain ceiling. A higher-rate fiber plant can make the transport layer look larger, while the silicon-addressable domain remains fixed by the same perimeter-bound radix.

This is the distinction Figure 4.3 preserves. Copper can be rate-lifted until reach collapses. CPO/NPO can recover reach and increase bandwidth while remaining inside the same perimeter ceiling. Only area-distributed optical I/O moves the design across the escape wall, because it changes the geometry from edge-bound escape to surface-bound escape.

Larger systems can still be built beyond one flat element by composing multiple elements with additional hops. Those hops add latency, power, and bandwidth cost. Reach determines how well that approach fits. Escape geometry determines where the flat-domain ceiling begins. The design decisions that follow from this lens are developed in § The Scale Boundary.

The placement of the light source is a separate integration-cost decision. The laser may be integrated with the OE (co-packaged), mounted nearby, or placed in a front-panel module connected by passive fiber using the External Laser Source pattern. Each choice trades optical and thermal performance and serviceability cost.

Laser dies are temperature-sensitive, especially at high performance. Hosting them beside a kilowatt-scale XPU package requires a much harder thermal solution than hosting them in a lower-dissipation module.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

The area regime moves the consumption sites for continuous-wave power from the die edge to the die face, and the supply must follow. Modulator sites scale as L^2, each drawing a roughly fixed optical power, while the fiber feed enters through the edge, which scales as L. The CW supply therefore inherits the same perimeter-versus-area geometry as data escape, with one decisive difference in constants: a data lane consumed a fiber per lane, while one feed fiber at ELSFP-class power serves many sites through passive splitting. The perimeter that was a wall for data is a tax for power.

Two supply topologies serve the area regime, and both close under the confinement and striping contract of § Optical Distribution Frame. Edge-injected external supply keeps the source in a serviceable, thermally independent module and carries CW power across the die face in on-die passive distribution, a pattern demonstrated in published 3D photonic interposer work. It pays waveguide distribution loss and splitter excess against the source budget of § Optical Distribution Frame, bounded by the per-fiber power ceiling that eye-safe exposure, connector power handling, and facet integrity impose. Distributed generation places an integrated source (SiPh-IL, Table C) at the consumption site, removing distribution loss and the polarization-maintaining transport with its connector extinction budget, in exchange for the package thermal environment and package-level FRU. The source-budget inequality decides between them: where compounded split and distribution loss meet the per-fiber ceiling, edge injection ends and integrated generation begins.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

5. Architectural Vision

The previous section separated the two constraints that govern scale-up photonic interconnect: reach and I/O escape. Reach determines how far bandwidth can move with usable margin. I/O escape determines how many endpoints can participate in a flat domain. Together, they define the placement problem this architecture is solving.

This section translates those constraints into a concrete architecture for photonic interconnect in open AI compute platforms. It treats placement as an integration-cost decision across power, latency, serviceability, thermal design, and domain size. The result is not a final product design, but a starting point for community development.

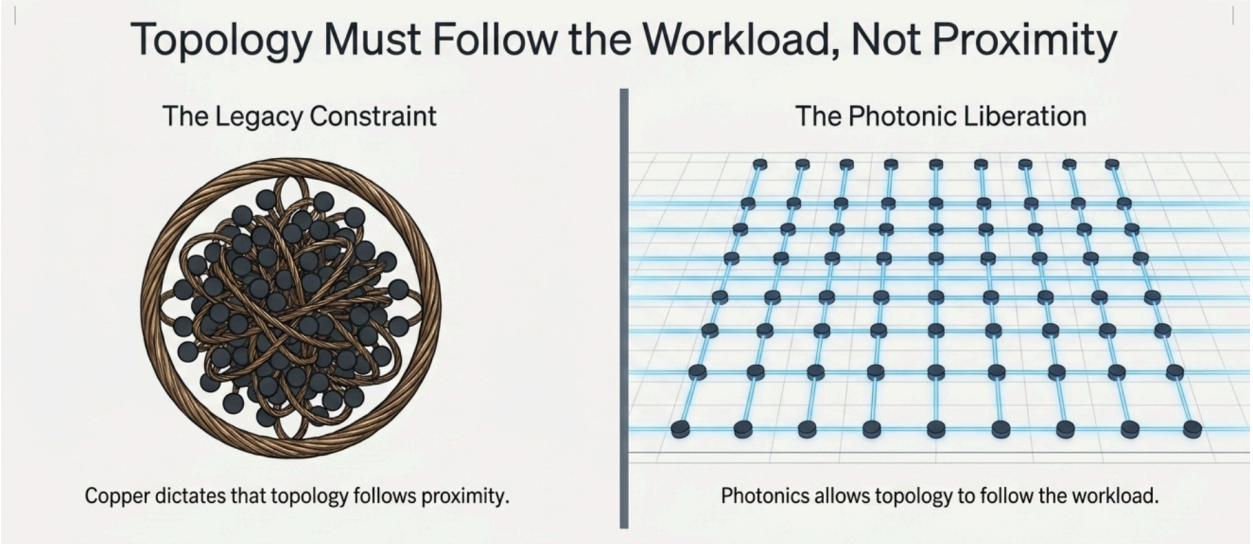


***Figure 5.1**. Reaching for Ideal Topology and Eliminating Constraints*

Topology follows the workload, not proximity. Copper forces proximity-bound clustering on the left. Photonics allows the fabric topology to follow the workload on the right.

The remainder of this section organizes the architecture into design goals, scaling axes, optical-engine and placement classes, validation contracts, and passive infrastructure components.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

5.1. Design Goals

The architecture is organized around ten design goals:

1. **Goal 1 — Photonic integration into AI systems:** Enable large domain scaling in today's AI system form factors with high BW interconnect and networking. This requires coordinated integration of modules, connectors, waveguides (copper or photonic), laser sources and OEs.
2. **Goal 2 — Additive integration:** Modify existing OCP ORv3 and MHS platforms minimally to enable large domain scaling in today's AI system form factors. Existing bus bars, liquid cooling manifolds, rack management interfaces (e.g., OpenBMC), and tray form factors should remain unchanged. Photonic interconnect is added through passive optical infrastructure components, and clearly defined extension classes, including a higher-density Class V path, rather than through redesign of existing subsystems or inclusion of vendor specific components.
3. **Goal 3 — Technology agnosticism at the platform level:** The passive optical infrastructure such as fiber routing, connectors, and distribution frames should accommodate silicon photonics with external lasers, VCSELs, micro-LEDs, coherent optics, and future approaches. Technology selection is a tray-level design decision, not a platform decision.
4. **Goal 4 — Protocol and traffic-class agnosticism at the photonic layer:** The optical physical layer should transparently carry scale-up protocols (NVLink™, UALink™, Ultra Ethernet Consortium™ (UEC), and future protocols), scale-out protocols, and Scale-DSA cross-architecture protocols (i.e. NVLink™ Fusion and open-substrate alternatives in development). AI interconnect ecosystems are converging toward a minimum of 200G/lane-class (such as OIFs Serializers/Deserializers (SerDes) through its Common Electrical I/O (CEI) Interoperability Agreements) physical lanes; the platform should not encode protocol or traffic-class semantics into the optical layer, per IEEE P802.3dj; the photonic layer converts this to optical without protocol or traffic-class awareness.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Multiple independent fabric planes may coexist on the same passive infrastructure, allowing the deployment to separate scale-up, scale-out, Scale-DSA, redundancy, or failure domains without changing the physical contract.

5. **Goal 5 — Generation-invariant infrastructure:** The fiber plant, rack fiber routing, and passive optical components should be permanent infrastructure that does not change when SerDes rates evolve from 224 Gb/s/lane to 448 Gb/s/lane and beyond. Only the OEs at the tray level change per generation.
6. **Goal 6 — Passive optical backplane and optional rail-switch migration:** The architecture should replace multi-layer impedance-controlled copper midplanes and heavy cabled backplane assemblies between trays with passive fiber distribution wherever inter-tray BW exceeds copper viability at the deployment generation. The photonic flyover layer (§ The Flyover Cable Layer) substitutes lightweight fiber bundles for 40+ layer PCBs and rear-of-rack copper cartridges, decoupling per-generation midplane redesign from rack-level mechanical, thermal, and serviceability evolution. In Class V systems, the same goal also permits switching functions to migrate from rear switch cards into deeper front-access compute trays when the implementation benefits from distributed rail switching. The result is a front-access passive infrastructure whose generation cycle is set by the connector and fiber, not the laminate stack or copper cable assembly.
7. **Goal 7 — Front-accessible serviceability with optional rear blind-mate egress:** Every maintenance action, for example, from tray replacement and link troubleshooting to topology reconfiguration and light source service, should be achievable from the front of the rack following OCP's toolless maintenance goals or using standard tooling and deterministic procedures. Front access is the baseline serviceability model. Rear access via blind-mate optical connectors, for example Expanded Beam Optical (EB), is supported as an optional passive egress path, and the architecture treats front-only, rear-only, and split egress

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

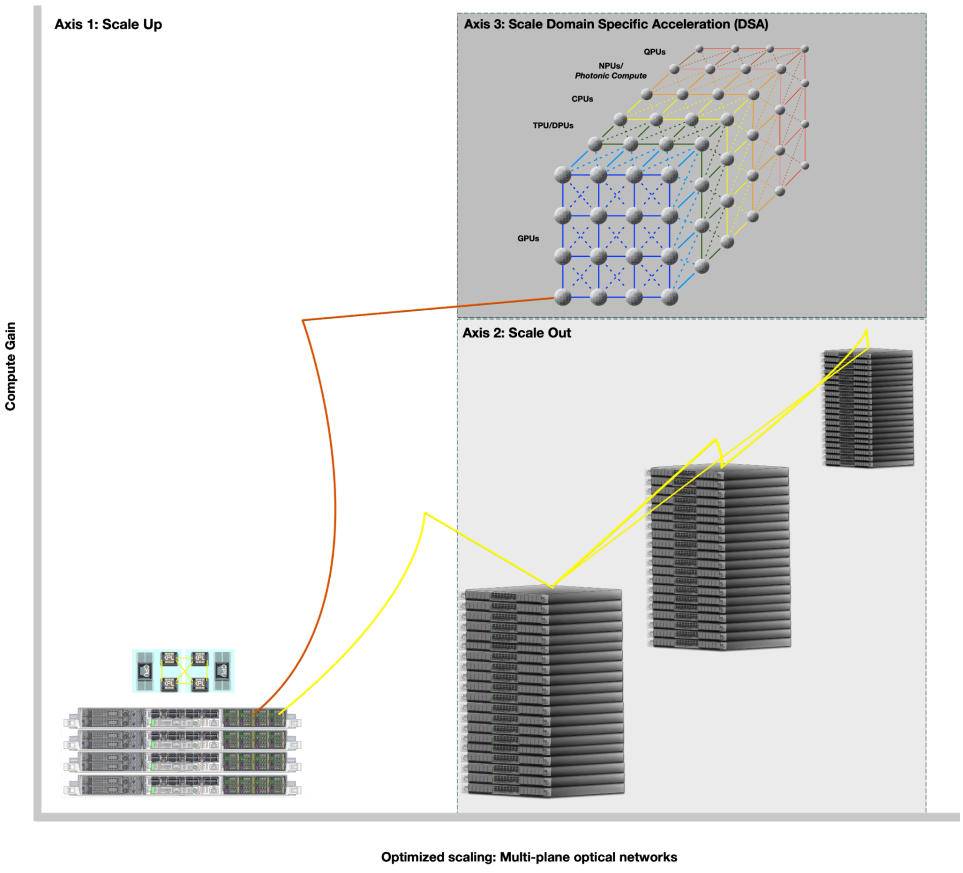
(e.g., scale-up rear / scale-out front) as first-class integration configurations (see § Rear Interface Zone). Rear connector areas may also be partitioned into multiple independent fabric planes, such as Fabric A/Fabric B redundancy or scale-up/Scale-DSA separation, without making the rear zone an electrical midplane.

8. **Goal 8 — Low-latency physical layer:** The architecture should specify non-retimed (linear-drive or direct-drive) link configurations as a primary OE class option, targeting propagation-dominated latency (~5 ns/m) that eliminates the 40-120 ns retimer-stage accumulation characteristic of multi-stage retimed copper scale-up paths, consistent with the scale-up latency requirements defined in OCP ESUN [12] and OIF COI Whitepaper [3]. Retimed configurations remain supported for longer-reach or reliability-constrained deployments, with latency penalty explicitly characterized in the Link Profile definition for each OE class. This latency-stability property is also the precondition for sub-nanosecond time synchronization across the fabric.
9. **Goal 9 — Non-foreclosure of future density regimes:** The architecture should not preclude future step-function density transitions: volumetric compute arrays, area-domain photonic integration, and sub-millimeter optical I/O density. Near-term specifications should document which choices are forward-compatible and which would require successor revisions; choices that close off these regimes should be flagged as foreclosing and require explicit justification.
10. **Goal 10 — Manufacturability through cableless integration:** The architecture should reduce internal cabling in favor of board-level optical routing between OEs and faceplate, blind-mate connectors at module and tray boundaries, and factory-qualified fiber subassemblies where routing flexibility is required. Recent hyperscaler-class platform demonstrations have moved toward cableless tray architectures for the same reasons: fewer manual assembly steps, lower defect rates at production volume, and reduced field-service complexity. Those demonstrations realize cableless integration in copper, which couples the

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

assembly savings to a per-generation laminate redesign. This architecture targets the same savings through a passive, generation-invariant optical plane, so the manufacturability gain does not re-enter the per-generation redesign cycle (§ The Flyover Cable Layer). Manufacturability is co-designed with serviceability (Goal 7): blind-mate optical interfaces, keyed fiber subassemblies, and deterministic inspection/cleaning procedures should reduce assembly variation and field-service risk.

5.2. Scale-DSA: The Third Scaling Axis



AI compute is no longer architecturally monolithic. Workloads now span heterogeneous compute domains within single system boundaries: GPU clusters paired with custom AI accelerators (AWS Trainium, Google TPU, Microsoft Maia, Cerebras WSE, Groq LPU); classical processors bridged to photonic compute substrates; CPU complexes paired with DPUs and domain accelerators; classical control planes coupled to quantum systems. The interconnect requirement that joins these heterogeneous domains is a third scaling problem, distinct from scale-up

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

(within-domain coupling at memory-semantic BW) and scale-out (cross-fabric coupling at Ethernet-class envelopes).

Figure 5.2. *Conceptually, three scaling axes of the AI fabric, scale-up, scale-out, and scale-DSA — over a unified passive fiber plant.*

Scale-DSA, or Domain-Specific Architecture scaling, is this third axis. One property defines it: a memory-semantic envelope at fabric-class reach. It carries memory-semantic bandwidth and latency to endpoints sitting at fabric distance, where neither scale-up's copper distribution nor scale-out's network timing natively reaches. Heterogeneity is the distinguisher: the axis spans different architectural classes, not copies of one. The binding is coherent where the devices support it (Compute Express Link® (CXL®), PCIe®) and composable, re-bound to the workload rather than fixed at build. Whether a link presents as coherent load/store or as general fabric-class bridging follows the capability of the connected devices, not the class boundary. The memory-semantic case is the high-capability subset; the class spans both. Traffic patterns are application-specific bridges: tensor exchange between an LLM training pod and a recommendation-serving pod built on different accelerator architectures; coherent memory access between a general-purpose processor and a domain-specific compute substrate; control-plane state transfer between architectural domains running concurrent phases of the same workload.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

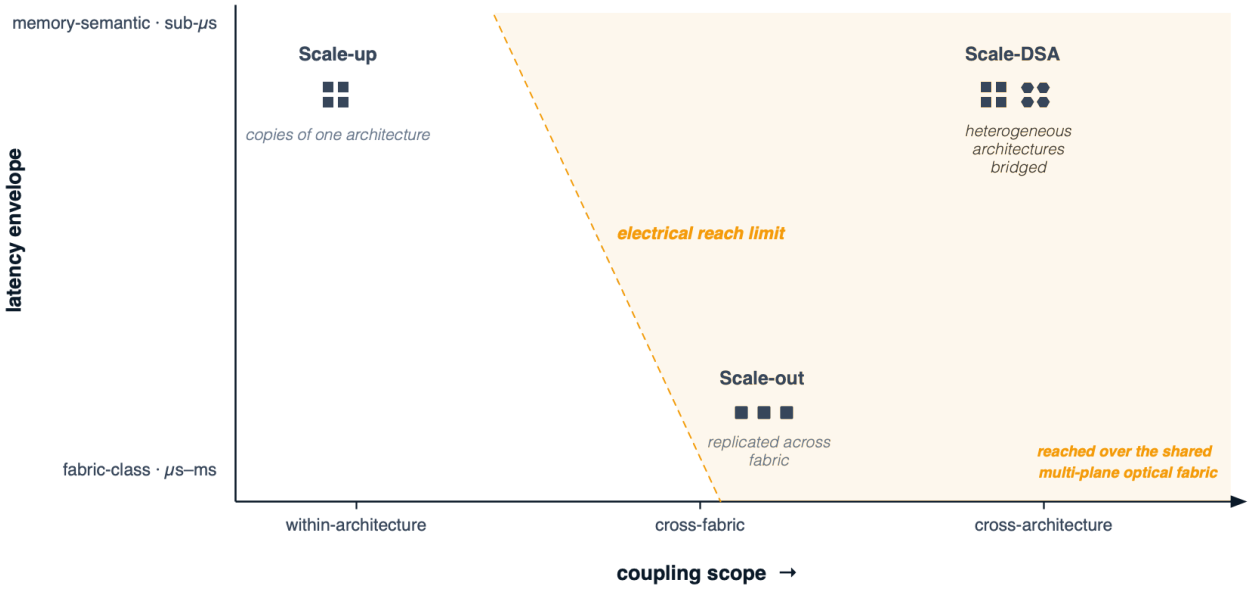


Figure 5.3. *The three traffic classes on coupling scope versus latency envelope, split by the electrical reach limit: scale-up within it, scale-out and Scale-DSA beyond it over the shared multi-plane optical fabric (cf. Table B, Three Scaling Axes Compared).*

Note: the limit leans because two mechanisms bind at opposite ends of the envelope with the retimer-latency budget at sub-μs (only a few 40–120 ns hops), and lane-rate channel loss at fabric distances (Fig. 3.3). The fabric carries past the limit without enlarging the routing domain.

The proliferation is underway now. Custom AI accelerators reached commercial scale in 2023–2024 across multiple hyperscaler ecosystems. Photonic and analog AI substrates entered commercial pilot deployment in 2024–2025. DPU adoption has spread across general-purpose hyperscale infrastructure, and quantum-classical bridging is in active development at all major quantum hardware vendors. Some of these architectures require BW, latency, ordering, or memory-semantics behavior that exceeds conventional Ethernet fabric assumptions without fitting into a same-architecture scale-up domain.

NVIDIA's NVLink™ Fusion, announced at Computex 2025 with commercial deployments spanning AWS Trainium4, Arm Neoverse cores, Intel x86, Fujitsu Monaka, Qualcomm CPUs, and the Marvell custom-silicon partnership, is one industry response

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

to Scale-DSA: open the link layer to heterogeneous silicon while preserving single-vendor protocol semantics. It is a viable solution within the configurations its interface supports. The open architectural question is whether Scale-DSA bridging standardizes at the protocol layer, enabling a single ecosystem player or walled garden to define the contract, or at the physical and link layers, where vendor-neutral substrates allow multiple protocol stacks to coexist on the same infrastructure. Optical circuit switching is one concrete realization of that physical-layer path: it defines no protocol, reconfiguring the optical graph beneath whatever rides it at the workload-phase timescale on which heterogeneous resources are composed. The IOWN Global Forum's all-photonic network and the OCP Optical Circuit Switching subproject are two open, vendor-neutral instances; the OCS module class this architecture already defines for software-controlled topology reconfiguration carries Scale-DSA without a Scale-DSA-specific contract.

The photonic substrate is protocol-agnostic by construction. The Link Profile framework specifies wavelength allocation, connectorization, and telemetry; none of these constrains the upper-PHY protocol. Scale-DSA traffic occupies the same fiber plant as scale-up and scale-out traffic, qualified through the same OE interface contracts, and is served by the same OE class taxonomy. The architectural posture is that the substrate is the right standardization layer for cross-domain BW: protocol stacks come and go on engineering timescales; the photonic substrate carries them without re-specification.

The three axes are concurrent, not sequential. Real AI systems carry scale-up traffic for within-pod collective communication, scale-out traffic for inter-pod and storage connectivity, and Scale-DSA traffic where heterogeneous compute is integrated into the same workload. The same fiber plant carries all three; the Link Profile framework treats each as an independent traffic class with its own contract on wavelength allocation, connectorization, and telemetry. An open photonic substrate therefore solves not one of

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

the three scaling problems but the full integration space that increasingly defines AI infrastructure.

Multiple fabric planes are the mechanical expression of this traffic-class independence. In Class H systems, where dedicated networking trays already exist, the main photonic opportunity is to expose independent fabric planes for scale-up, scale-out, Scale-DSA, and redundancy without major tray-level architectural change. The grouping can happen dynamically in one of two places. A subgroup of connections ingresses and egresses at the tray boundary, or a fabric switch regroups the bandwidth before it reaches the tray. The best fit between these follows from the endpoint XPU or ASIC escape budget, constrained by the electrical lanes it escapes and the fibers they roll up into. That budget stays perimeter-bound through copper and CPO, so until APO opens area escape, the switching function tends to sit in the fabric. Class V inherits these modes and can go further: onboard wavelength or rail switches on the compute trays allocate local and global optical links across fabrics, preserving a single front-serviceable platform. On-tray switching also lets the platform oversubscribe and regroup lanes, stretching the perimeter escape budget and deferring that threshold. Once topology is scheduled rather than wired, connectivity becomes a managed resource: provisioned for aggregate demand, fanned out where one signal should reach many, and returned to the pool when idle.

Table B: Three Scaling Axes Compared

	Scale-up	Scale-out	Scale-DSA
Coupling	Within architecture	Cross fabric	Cross architecture
BW class	Memory semantic	Ethernet class	Application dependent (near-memory to fabric-class)

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

	Scale-up	Scale-out	Scale-DSA
Latency envelope	Sub-µs	µs to ms	Application-dependent
Protocol anchors	NVLink™, UALink™, ESUN	Ethernet, UEC, InfiniBand	NVLink™ Fusion; open-substrate alternatives in development
Example traffic	All-reduce, all-to-all	Storage, gradient sync, fabric routing	Cross architecture tensor exchange, coherent bridges, control-plane transfer
Shared Substrate	One passive fiber plant; per-class Link Profile contract: wavelength allocation, connectorization, telemetry.		

On-tray rail switches multiply the distributed elements. One or two rail-switch ASICs per compute tray reach on the order of 144 switch ASICs inside a dense HBD. The default topology is Clos-equivalent, and independent rails may be bonded or striped as one logical HBD. The rails need not share one personality. A Clos rail carries general scale-up traffic; a collective-optimized rail carries all-reduce and all-to-all; isolated partitions hold tenants or jobs; a Scale-DSA rail reaches disaggregated memory; a spare rail absorbs evacuation and failure response. The scheduler then allocates compute, memory, and fabric together, fixing endpoint set, rail set, bandwidth class, routing mode, isolation boundary, and reconfiguration window.

The reconfiguration layers stack. The passive ODF already spreads each endpoint's lanes across peer ports, giving multipath and graceful degradation with no switching power; optional OCS repartitions the same graph between jobs or after failures. A flat distributed fabric is valuable only if route churn, congestion, and fault isolation stay local. The Base Specification should therefore treat multi-fabric rail operation as a managed resource: telemetry, admission control, partition identifiers, per-rail RAS reporting, and failure containment that keeps a failed tray to its own fabric slice.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

The long-term consequence of Scale-DSA is not only heterogeneous attachment, but workload-defined recomposition. Once high-bandwidth links are no longer bounded by copper proximity, racks can be organized as resource pools rather than fixed homogeneous systems: accelerator-rich racks, CPU/control-plane racks, memory-rich racks, network-service racks, or domain-specific accelerator racks. The passive photonic substrate and Link Profile contracts allow these resources to be composed into logical accelerators whose topology follows workload phase, locality, and latency requirements rather than the mechanical assumptions of a single chassis generation.

This framing aligns naturally with OCP AI HW/SW Co-Design work on polymorphic architecture[7]: hardware/software co-design determines which resources should be discovered, composed, and reconfigured for a workload; the open photonic substrate provides the physical-layer reach, bandwidth, and fabric-plane independence needed to make that composition practical at rack, row, and pod scale.

Recomposition has a measurable payoff, and memory disaggregation is where it lands. A pooled memory-semantic tier earns its place only because Scale-DSA reaches it: load/store-addressable, bounded in latency, and composable at pool scale, the three properties above in order. Pooling the newly reachable endpoints folds many small allocation domains into one, and the tail is the prize. An analytical queueing model puts the straggler that gates a wide synchronized collective at roughly a tenth of its siloed value once the reachable set is pooled. Inserting that tier between local memory and bulk storage cuts effective access latency six- to tenfold, set by the storage path it displaces. Two caveats bound the claim. The pooled-tier latency is modeled, not measured: CXL®-pooled-class latency carried at rack-and-row reach, not lower than electrical coherence. And the gain holds only where the working set is reused at that

[7] https://www.opencompute.org/documents/ocp-document-submission-ai-co-design-docx-pdf

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

capacity, which AI traffic supplies through KV-cache, embedding tables, expert weights, and optimizer state. Appendix E carries the derivation.

5.3. Optical Engine Classes

Table C: Optical Engine (OE) Classes

OE Class	Light Source	Fiber Type	Reach	External Laser	Natural Fit
SiPh-ELS	External CW laser (ex: ELSFP)	SM G.652D	>500m	Yes	Scale-up, high-density BiDi DWDM
VCSEL-DM	Integrated VCSEL array	MM OM3-5 or SM	<100m	No	Ultra-short on-tray, low cost
Micro-LED	Integrated LED array	MM or waveguide	<10m	No	Die-to-die, package-level
Coherent-IL	Integrated tunable laser + DSP	SM G.652D	>10km	No	Inter-pod, DCI, long reach
SiPh-IL	Heterogeneous III-V on Si	SM G.652D	>500m	No	Future: SiPh density, no ELS

Note: OE *technology class and* OE *placement class are orthogonal architectural axes. The OE Class table above describes technology variants; § Placement Class Taxonomy enumerates the five placement classes (CPO, NPO, S-NPO, M-NPO, and FPO), with XPO treated as an MSA-defined faceplate-pluggable/OFC subset of FPO, and their integration boundaries. Any OE class above may be realized at any placement class for*

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

which the channel budget closes, and the platform mechanical integration (Faceplate, Raceway, ODF) is identical across placement classes.

Wavelength-density scaling of OE-internal IL. The structural difference between modulator architectures compounds with wavelength count. Ring-resonator modulator (RRM) cascades co-locate modulation and wavelength multiplexing on a single bus waveguide as each ring modulates on-resonance for its assigned wavelength and passes other wavelengths off-resonance at <0.2 dB per pass whereas MZM-based designs require a separate WDM mux stage (typically 2–4 dB per endpoint at 8–16λ AWG). At single-λ the difference is modest; at 4λ OCI-class density the composition yields approximately 5 dB Tx+Rx OE-internal advantage to ring-based designs, and the gap widens with channel count. This is the physical basis for the architecture's expectation that ring-based SiPh-ELS designs progressively displace MZM-based designs as wavelength count increases. Per-channel insertion loss (IL) values are vendor-determined within published ranges; the structural relationship of ring cascades, that co-locate two functions for which MZM requires separate stages, is not. Appendix C.3.2 develops the comparison quantitatively.

5.4. Fiber Attach Envelope

The fiber attach footprint at the optical-engine boundary is determined by the number of fibers per port, as derived in Appendix A, multiplied by the fiber-array pitch. Standard single-fiber pitch in fiber array units (FAUs) is 250 µm; 127 µm pitch is achievable for higher-density arrays. For a CPO engine serving one 1.6 T port at Profile B 16λ BiDi, the port-side FAU contains a single fiber. At Profile A slow-wide, such as 8×224G DR, the same port presents a 16-fiber FAU, approximately 4 mm wide at 250 µm pitch.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

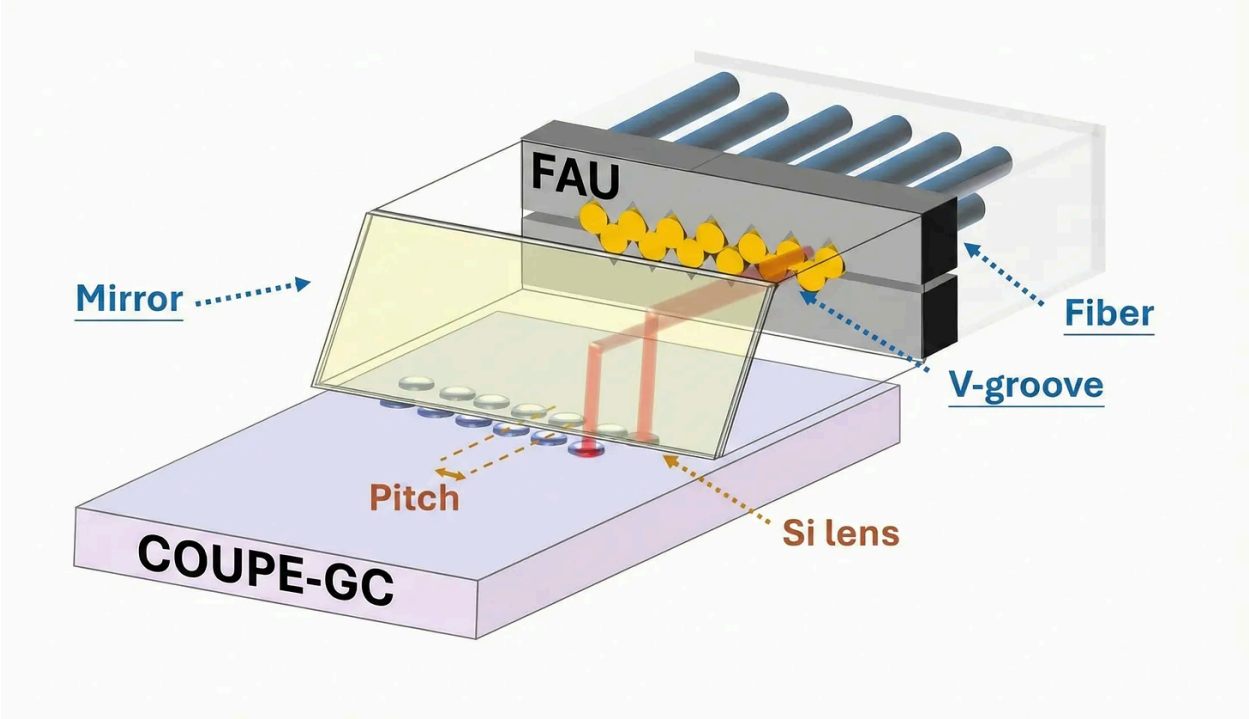


Figure 5.4. *Example of Fiber Attachment Unit Connectivity*[8]

The OE attach envelope is required to accommodate the upper bound required by the supported profiles. The architectural contract is therefore the maximum FAU width per port, together with the applicable alignment-tolerance class. It is not a mandate for a specific pitch. This preserves implementation flexibility while bounding the integration envelope. Specific pitch and tolerance values are defined in the Base Specification.

The FAU also carries an optical and reliability performance contract. The Base Specification should define the FAU loss-parameter requirements, including per-facet insertion loss, both typical and maximum; channel-to-channel uniformity; return loss at the PIC-to-fiber transition; polarization-dependent loss where applicable; insertion-loss drift under thermal cycling and aging; and the operating wavelength band over which these parameters apply. It should also define the relevant FAU physical parameters, including supported fiber types, end-face geometry, optical power handling for the FAU

[8] From The Bull Case of Browave by Bubu Research: https://buburesearch.substack.com/p/the-bull-case-of-browave-3163-tw

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

connector class, attachment-method class, mating-cycle count for re-mateable terminations, and the environmental, mechanical, and reliability qualification regime satisfied by the assembly. The attachment-method class should indicate active or passive alignment, with tolerance class declared under the dimensional contract above. The qualification regime includes Telcordia GR-468 class for the optical subassembly and IEC 61753 grade for any re-mateable connector pair within the FAU.

Physics floors for the loss parameters are derived in Appendix C.2. Deployed values are populated by the Base Specification from ecosystem capability data. As with the dimensional contract, the architecture treats these values as declared parameters rather than prescribed numbers. Different attach methods, including edge-coupled V-groove FAU, grating-coupled vertical attach, glass-bridge or optical-interposer attach, and expanded-beam or photonic-wire-bond approaches, compete on declared performance against a common parameter set.

The FAU has two ends and therefore two roles. The PIC side is intrinsic to the OE and is typically the OE vendor's responsibility. The cable side is the ecosystem boundary. Parameters declared at that boundary, including termination format, position relative to the package, and insertion-loss class of the termination, determine how many distinct tray-internal cable assemblies the ecosystem needs to manufacture, qualify, and inventory. Convergence on a small number of format, position, and IL-class combinations reduces downstream cable SKU count, enables multi-source procurement, amortizes qualification cost, and consolidates sparing inventory. The architecture treats these three parameters as declared values and identifies ecosystem convergence as the operational goal. Specific format selections are deferred to the Base Specification, while preserving a path for enhanced-performance termination classes, such as expanded-beam or photonic-wire-bond-fed terminations, to enter the ecosystem as their qualification matures.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Per-port FAU dimensions are uniform across platform classes at a given generation because the per-port aggregate is the same, as shown in Appendix A. What differs is the number of ports per package. Switch packages generally present a higher port count than XPU packages at the same generation and therefore require a larger total fiber-attach perimeter, even when both use the same per-port FAU envelope. The OE attach contract therefore applies uniformly at the per-port level across reference platforms. Total package-perimeter accommodation is dimensioned per platform class within the bounds of that contract, rather than through separate attach contracts. This preserves a single OE attach standard while allowing platform-class differentiation in implementation.

FAU array geometry is also part of the dimensional envelope. Single-row, 1×N fiber arrays are the standard fabrication approach and support the lowest-count attach configurations. Multi-row arrays, such as 2×N and 4×N, increase fiber count without consuming additional perimeter width, but each added row consumes vertical clearance at the OE boundary and increases optical-coupling and alignment complexity. The architecture therefore treats array-row count as a contract parameter alongside FAU width and pitch. Implementations declare the array geometries they support, and system designs accommodate the resulting envelope at the OE boundary.

OE perimeter available for fiber attach is a finite system resource. Total fibers per package, multiplied by FAU pitch and divided by row count, determines the perimeter consumed at the OE boundary. At the upper bounds permitted by Appendix A, including Profile A slow-wide configurations and extreme high-radix splits, single-row arrays at standard pitch may exceed the perimeter available on practical ASIC packages. Multi-row arrays then become necessary, and at the highest limits even 4×N arrays approach geometric ceilings. System designers should therefore treat OE perimeter as constrained real estate and reserve keep-out zones around the optical-engine boundary that allow migration from single-row to multi-row geometry across generations. Designs

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

that consume the full perimeter in one generation reduce growth options in the next; designs that preserve headroom maintain the architecture's evolutionary path.

This perimeter constraint binds differently across placement classes, as described in the Placement Class Taxonomy. Pure CPO at the package substrate generally provides the largest available perimeter because the substrate envelope is sized for the ASIC and its co-packaged optics. Socketed and substrate-shared placement classes, including S-NPO and OpenCPX-class realizations, work within a smaller envelope since the substrate should also accommodate socket geometry and any copper cable transition alongside the OE. As a result, S-NPO and OpenCPX-class placements are well suited to low- and medium-throughput applications where the FAU envelope remains within the substrate envelope. Higher-throughput applications targeting the upper rows of Appendix A may exceed the FAU envelope and may benefit from migration to CPO-class integration. This is a placement-selection input, not a class limitation: each placement class serves its intended envelope, and the architecture exposes the binding constraint so designers can select accordingly.

5.5. Placement Class Taxonomy

An ideal photonic AI infrastructure makes the integration boundary between ASIC and OE an explicit architectural choice, not an inheritance from prior generations or a constraint imposed by available connectors. Placement class is that explicit choice: where the OE sits relative to the ASIC, which channel connects them, where vendor responsibilities transfer, what field-replaceability the system supports, and the ecosystem boundary for silicon and OE vendors. It is a first-class architectural decision rather than an implementation detail. Placement class is orthogonal to the OE technology classes in the previous subsection (SiPh-ELS, VCSEL-DM, Micro-LED, Coherent-IL, SiPh-IL): a given OE technology may be deployed at multiple placement positions, and a given placement position may host multiple OE technologies. The classes below span the spectrum from in-package integration to faceplate-pluggable

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

integration; rack-boundary aggregation is handled as XPO, a specialized faceplate-pluggable/OFC realization inside FPO. The efficiency and FRU tradeoffs across this spectrum, together with the system-level economic considerations that determine class selection, are developed in § COM and Placement Selection. This taxonomy contracts what the engine is; socket-level vendor technologies or MSAs such as Co Packaged Copper (CPC) and OpenCPX contract how it attaches, and the two layers compose orthogonally.

Co-Packaged Optics (CPO). OEs integrated within the same package substrate as the ASIC, sharing the substrate-level interconnect fabric and the package-level thermal solution. The channel between ASIC SerDes and OE input is internal to the package and traverses no connector — at 224 Gb/s the channel is typically unretimed XSR-class, and at 448 Gb/s it remains unretimed XSR-class with advanced dielectric. CPO is FRU at the package level: a defective OE implies replacement of the package. Both 2D CPO (PIC adjacent to ASIC on substrate, electrical bumps between EIC and PIC) and 3D CPO (PIC under-stacked or otherwise vertically integrated with EIC) are accommodated; the architectural treatment of channel, FRU, and contract is consistent across the two variants, with the structural distinction documented in vendor-specific implementation details.

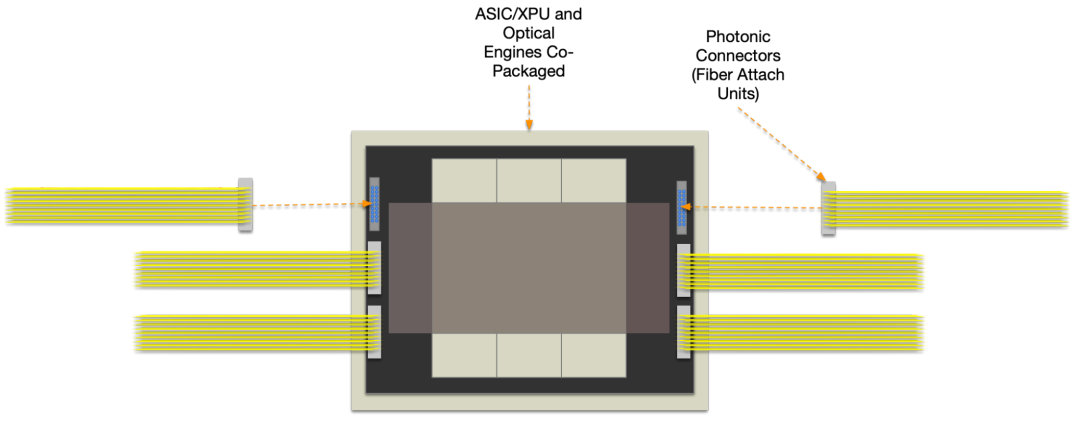


Figure 5.5. *Co-Packaged Optics and Fiber Attach Units.*

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Near-Package Optics (NPO). OE mounted on the host PCB adjacent to the ASIC package, connected via a short-reach electrical channel that crosses the package-to-host transition and a brief PCB trace. Channel parameters per OIF CEI-224G-XSR or VSR depending on host PCB dielectric and trace length; retiming model ranges from unretimed (shortest NPO realizations) to half-retimed (longer NPO with advanced dielectric). NPO is field-replaceable at the OE level provided the OE is socketed or mounted in a serviceable carrier; otherwise FRU level is the host PCB assembly.

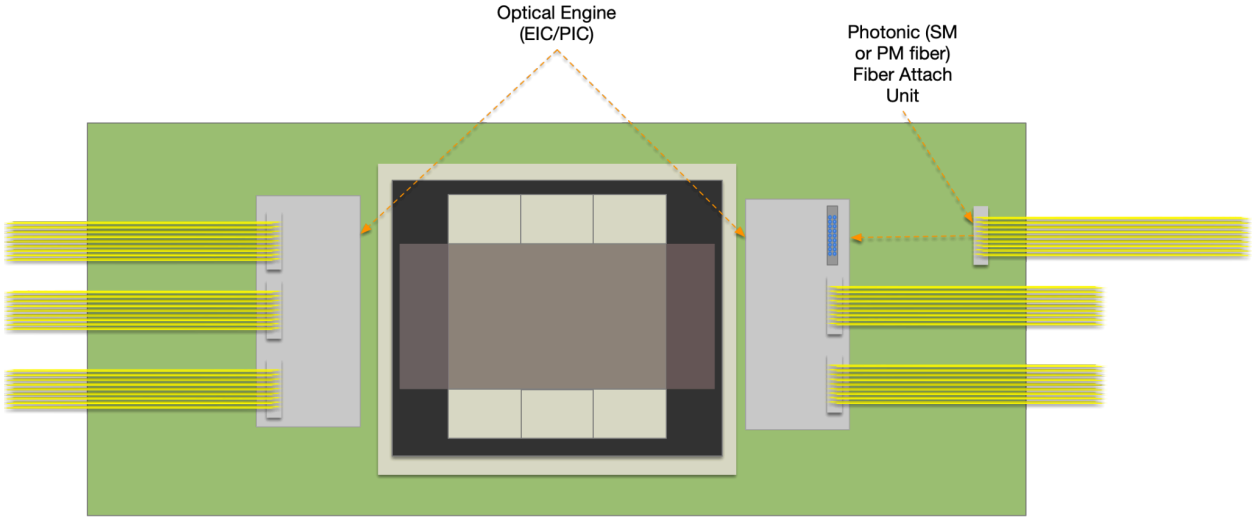


Figure 5.6. *Near Package Optics and Fiber Attach Units.*

Socketed Near-Packaged Optics (S-NPO). OE attaches to the ASIC package substrate through a detachable high-speed socket designed to carry the OE's SerDes channel directly between the ASIC and a substrate-mounted OE or copper cable assembly. The defining property is mode-flexible swap point: the same socket can host an OE in optical mode or a copper cable assembly in electrical mode, allowing the integrator to select the photonic-vs-copper boundary on a per-system basis. Channel parameters per OIF CEI-224G-VSR with the socket as a defined channel discontinuity; retiming model is typically half-retimed or retimed depending on socket implementation. S-NPO is FRU at the socketed-device level. Current realizations include Samtec CPX

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

(SFCM socket plus SFCC cable assembly) and Molex Impress Co-Packaged Copper Solutions; both are mounted on a substrate envelope (e.g., 95×95 mm class) designed to host the ASIC plus socketed OE or cable assembly.

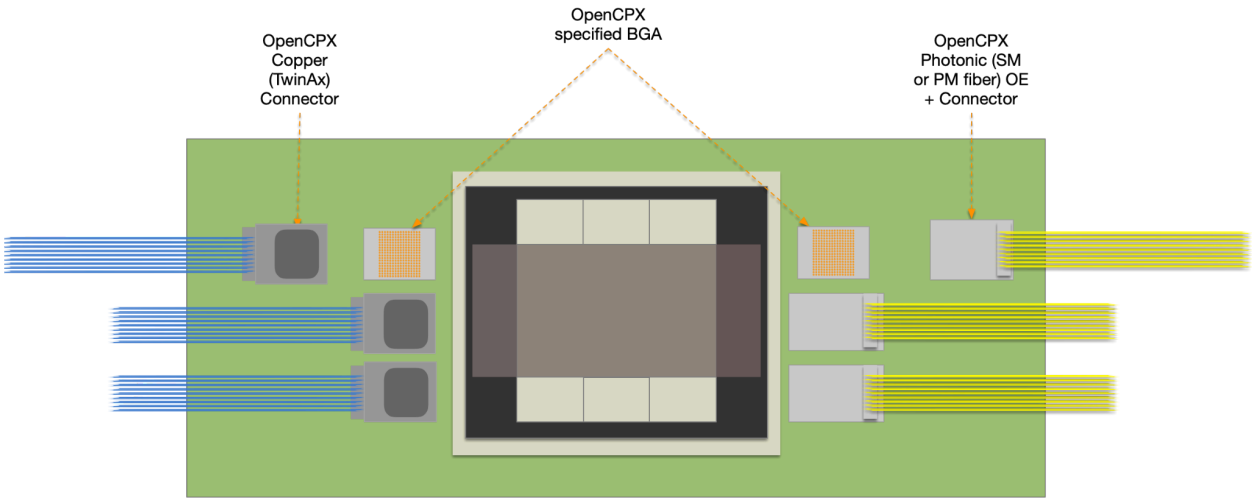


Figure 5.7. *Socketed Optics and Copper High-Speed Interfaces.*

Mezzanine Near-Packaged Optics (M-NPO). The ASIC and its OE(s) are hosted on a standardized mezzanine carrier that mates to the host platform through a defined mezzanine interface. The defining property is FRU-without-mainboard-reflow: the entire ASIC + OE subsystem is field-replaceable as a unit, decoupling the OE service cycle from the host platform service cycle. The mezzanine envelope and interface contract enable an open OE ecosystem in which OE vendors can ship characterized subsystems that integrate with merchant switch silicon. Channel parameters are characterized to the mezzanine edge; the switch silicon vendor publishes the channel budget to the edge, and the OE integrator takes over from there. Retiming model is typically retimed at one or both edges. M-NPO realizations include merchant silicon vendor paddle and mezzanine carriers under development.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

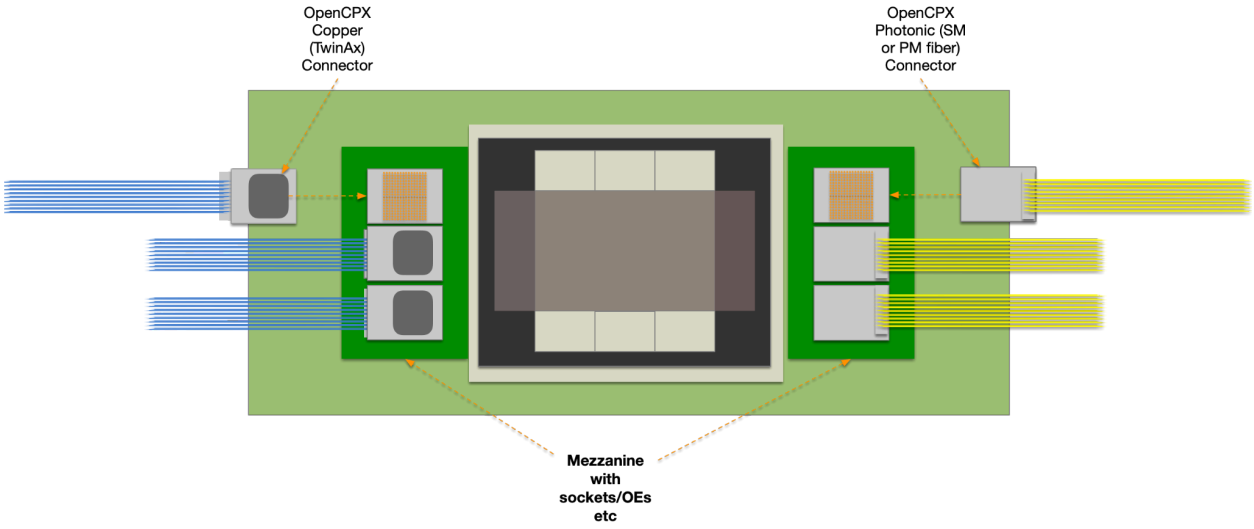


Figure 5.8. *Mezzanine-Socketed Optics and Copper High-Speed Interfaces.*

Faceplate Pluggable Optics (FPO). OEs housed in pluggable modules at the faceplate, connected to the ASIC via a long host PCB channel traversing the full tray depth. This is the reference baseline for current AI compute and switch trays, and the architecture references it as the most-decoupled placement class. Channel parameters per OIF (VSR) Flyover with the host PCB and faceplate cage as defined discontinuities; retiming model is retimed at the host with the option of additional retiming at the cage. FPO is FRU at the module level. Current realizations include OSFP, QSFP-DD, OSFP-XD, and the OCI MSA module envelope; the architecture accommodates all common pluggable module form factors at the faceplate. XPO is treated as a subset of FPO when the faceplate-pluggable module performs rack/pod-boundary optical format conversion.

eXtra-dense Pluggable Optics (XPO). XPO is not a separate placement class. It is an MSA-defined faceplate-pluggable realization within FPO whose distinguishing role is optical format conversion (OFC), typically at rack- or pod-boundary aggregation points. XPO modules deployed at that boundary perform O→E→O conversion between intra-pod scale-up infrastructure and inter-pod scale-out connectivity. The module inherits the FPO mechanical, thermal, power, FRU, and faceplate-serviceability model;

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

its distinction is functional, not a new placement boundary. That boundary placement has channel consequences. Traffic that hairpins through the rack-boundary module will result in the conversion path adding a double-retimer penalty (host SerDes → XPO inbound SerDes → XPO OE). XPO therefore aligns most clearly with switching and aggregation contexts, where rack-boundary format conversion is the intended system function.



Figure 5.9. *XPO Optics with Copper High-Speed Interfaces.*[9]

The placement taxonomy therefore remains five classes: CPO, NPO, S-NPO, M-NPO, and FPO. XPO is a named subset of FPO rather than a sixth class: it uses the faceplate-pluggable module boundary and adds OFC as its defining functional role. ELSFP (OIF-ELSFP-02.0) is also an MSA-defined faceplate-pluggable specification, but it is not enumerated in the

[9] https://www.amphenol-cs.com/product-series/xpo-lpo-optical-transceiver.html

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

placement taxonomy for a different reason :

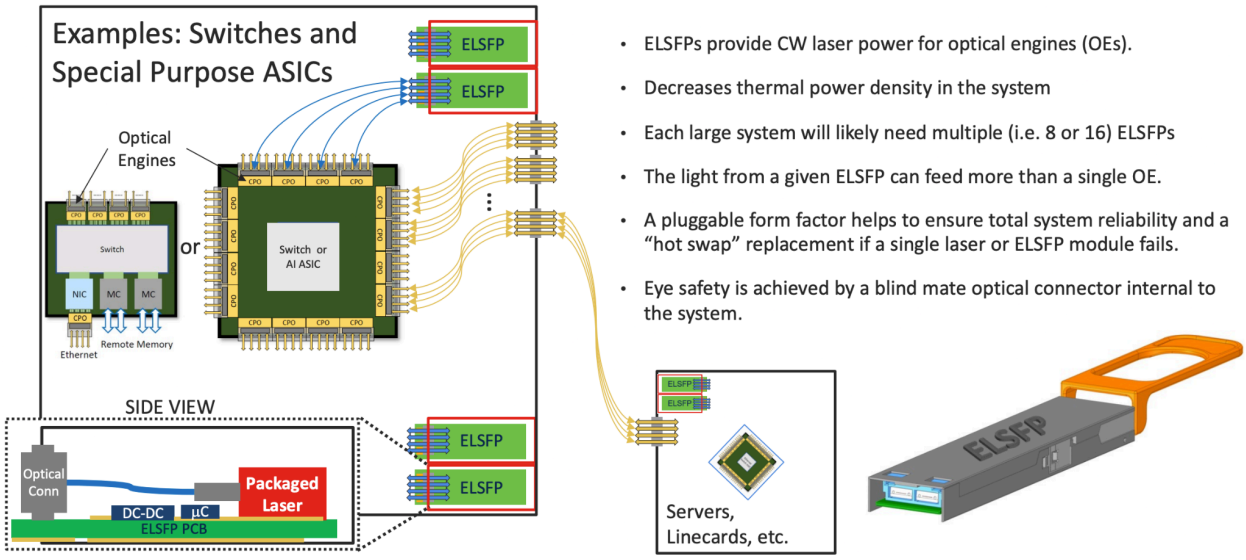


Figure 5.10. *External Laser Small Form Factor Pluggable (ELSFP).*[10]

ELSFP modules sit in the power-delivery path, supplying CW laser power to SiPh-ELS OEs. They do not sit in a data signal channel and therefore are treated in § Optical Distribution Frame as a populated module class of the Optical Services Zone, not as an optical-engine placement class.

Egress direction is likewise a separate architectural axis. For CPO, NPO, S-NPO, and M-NPO, front-faceplate or rear-interface-zone egress (§ Rear Interface Zone) is an integration choice routed through the Raceway. For FPO and XPO, the module occupies the faceplate that also serves as its egress, so placement and egress often coincide in a specific implementation. Even there, egress remains a system-level integration choice, not a placement-class subdivision.

OBO-WG (on-board optics with waveguide coupling) and OFC (optical format converter) are also not separate placement classes. OBO-WG can be treated as a refinement within NPO if a finer-grained distinction is needed. OFC is a functional role independent

[10] Openly presented by OIF Forum: https://www.oiforum.com/wp-content/uploads/OIF_EEI_Interop_Slide_Set_OFC24v2.pdf

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

of placement; this taxonomy captures it only when realized as the XPO faceplate-pluggable rack-boundary subset of FPO.

Areal Packaged Optics (APO) stays outside this list by construction. The five classes place the engine against the die edge, and all five still escape the perimeter. APO is defined by the complementary geometry: I/O leaving the die face. That makes it the perimeter-to-area transition, developed in § The Photonic Interconnect System and Appendix D.1, not a sixth edge-placement. Where this document names APO, it names that geometry.

By naming integration boundaries explicitly and characterizing each placement class consistently - physical structure, FRU level, channel interface, ecosystem context, and current realizations - the taxonomy supports the open-ecosystem goal: any OE technology that satisfies a placement class's contract qualifies for that class regardless of vendor origin, any integrator can select the placement that fits its system economics, and new realizations can enter the architecture without restructuring it.

5.6. Envelope and Contract

An ideal photonic AI infrastructure specifies open and interoperable possibilities that are available at every architectural boundary without dictating how it gets built. This separation between specification and implementation is what lets vendors innovate, standards bodies formalize, and integrators compose without the architecture itself becoming the bottleneck. The architecture distinguishes between the *envelope* as a defined physical and operational space, and the *contract*, the specifications that any implementation occupying that envelope will satisfy. Vendor implementations fill in the envelope according to the contract; the architecture itself does not prescribe how. This is the black-box methodology underlying the rest of this whitepaper: the document defines what each placement class provides and consumes, not how to build the engine that meets those requirements.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

This methodology serves three purposes. It keeps the architecture informative and leading rather than prescriptive, leading the ecosystem instead of dictating to it. It makes the architecture open by construction using the contract as the qualification criterion, so any implementation meeting it qualifies regardless of vendor or MSA origin. This layering also makes the architecture composable. The same envelope-and-contract pattern applies at every architectural boundary, allowing each tier to remain stable as the implementations inside evolve.

The five contract dimensions. Each placement class (CPO, NPO, S-NPO, M-NPO, FPO; see § Placement Class Taxonomy) specifies a contract along five dimensions. XPO is treated as an MSA-defined subset of FPO for rack/pod-boundary optical format conversion, not as a separate placement class:

Table D: Contract Dimensions

Contract dimension	What is specified	What is left to vendor implementation
Mechanical	Outer dimensions, mounting footprint, keep-out zones, height limits, fiber egress geometry	Internal layout, materials, attach method
Power	Voltage rails, current capacity, isolation requirements, sequencing	Regulator topology, decoupling, PMIC choice
High-speed signaling	Lane count, impedance target, COM and return-loss budget (per § Channel Budget Reference), retiming model	Connector geometry, host trace topology, equalization strategy
Thermal	Maximum dissipation per OE, junction temperature limits, airflow direction, fluid coupling where applicable	Heatsink design, thermal interface materials, internal flow paths

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Contract dimension	What is specified	What is left to vendor implementation
Management and telemetry	OCP RAS API endpoint conformance (per OCP RAS API v0.9), CPER-format event reporting, alarm taxonomy, optical power and link-quality telemetry	Internal sensor placement, firmware implementation, vendor-specific diagnostics

The left column is what the architecture publishes; the right column is what each vendor implementation chooses. The architecture's normative contribution is the contract layer; the right column is implementation freedom.

The COM–RAS coupling. The high-speed signaling and management & telemetry dimensions are not independent. They are the design-time and run-time pair for the link layer. COM specifies the channel margin a placement class must provide at design time; the OCP RAS API specifies how the operational performance of that channel is reported at run time. Together they form a closed loop: COM is the design-time guarantee, RAS is the run-time observability, and the BER-margin and effective pJ/bit framing from § COM and Placement Selection is the economic feedback that allows future design to be informed by operational data. The architecture treats these two contract dimensions as coupled axes of the link-layer specification rather than as separate concerns; this coupling is what makes the link layer auditable in operation against its design-time specification, and what extends the OCP RAS discipline from compute and memory to the interconnect.

The interoperability framework. Within the signaling and management & telemetry dimensions specifically, the architecture spans four interoperability layers consistent with how IEEE and OIF describe high-speed serial channels:

- **Electrical attach (ASIC ↔ OE):** retimed, half-retimed, linear-drive, or direct-drive attach models per OIF CEI co-packaging framework; defines where CDR and DSP live and therefore power, latency, and serviceability tradeoffs.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

- **Optical waveform and lane model:** the mapping of SerDes lanes into optical wavelengths and fibers, including directionality (simplex, duplex, bi-directional WDM) and any required FEC or gearbox assumptions.

- **Passive fiber plant inside the rack:** connector families, fiber routing corridors, inspection and cleaning access, fault-isolation breakpoints (ODF).

- **Management, telemetry, and timing:** OCP RAS API endpoint conformance, CPER-format error reporting, minimum required monitoring, and alarm taxonomy across OEs and any external laser subsystem. The architecture preserves a defined path from datacenter timing and management infrastructure to the photonic link, including reference-clock distribution and timestamp insertion; CPO integration relocates the natural management surface inboard of the conventional pluggable boundary, and the architecture specifies how this path traverses the package boundary so the link is not timing- or telemetry-opaque to the datacenter. The current OCP TAP Time Card is a PCIe-attached appliance; realizing photonic-native timing reference distribution is a natural coordination point between this architecture and OCP TAP. Because that surface has historically lived in the pluggable optical module, the module-level management interface does not disappear when optics move inboard; it relocates and nests beneath the system RAS plane. The OIF Common Management Interface Specification (CMIS) characterizes the engine and its telemetry at the module boundary, while PLDM (the DMTF Platform Level Data Model) and the OCP RAS API carry the resulting link-quality and fault state into that plane in CPER format — composable layers of one management path, referenced rather than re-specified.

These four layers refine the signaling and management & telemetry contract dimensions into the specific interoperability points across which IEEE Ethernet physical-layer trajectories (e.g., IEEE P802.3dj for 200G/lane-class PHY evolution), OIF electrical interface agreements (CEI-224G, the CEI-448G framework, EEI), and OCP system specifications align.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Laser Thermal Coupling and Reliability. Placement classes that bring OEs or laser sources closer to the XPU also introduce a thermal-reliability obligation at the placement boundary. For laser-bearing implementations, the relevant contract should account for maximum laser case-temperature range, allowable thermal gradients, transient thermal behavior, wavelength-stability margin, optical-power degradation over life, applicable telemetry, and the photonic thermal-supply assumptions consumed by the implementation.

In high-density systems, liquid cooling may be one realization of that supply. Possible implementations include monolithic cold plates, chambered or zoned cold plates, or directed flow paths arranged to manage mixed thermal limits across the XPU and optical regions. The architecture does not require any specific cooling topology. It defines the thermal conditions, observability, and service boundaries that any conforming implementation is expected to satisfy.

OIF EEI / COI Alignment Considerations. This architecture is intended to remain compatible with the direction of OIF Energy Efficient Interfaces (EEI) and Compute Optics Interface (COI) work without depending on a single OIF implementation choice. OIF EEI/COI frames AI scale-up interconnects around low latency, low energy per bit, high bandwidth density, serviceability, reliability, and multi-vendor interoperability. These priorities align with this architecture's use of a stable optical faceplate envelope, structured internal fiber management, removable ODF or optical-services modules, and modulation-agnostic channel budgeting.

The architecture should therefore preserve clear interface contracts at the optical-services zone, including fiber-count limits, connector positions, labeling, cleaning access, bend-radius egress, diagnostics/test access, and compatibility with blind-mate or high-density connector technologies. It should also reserve design flexibility for external laser source strategies, including front-panel ELSFP-style modules or future

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

optical-services modules, without making laser-source placement normative in this document.

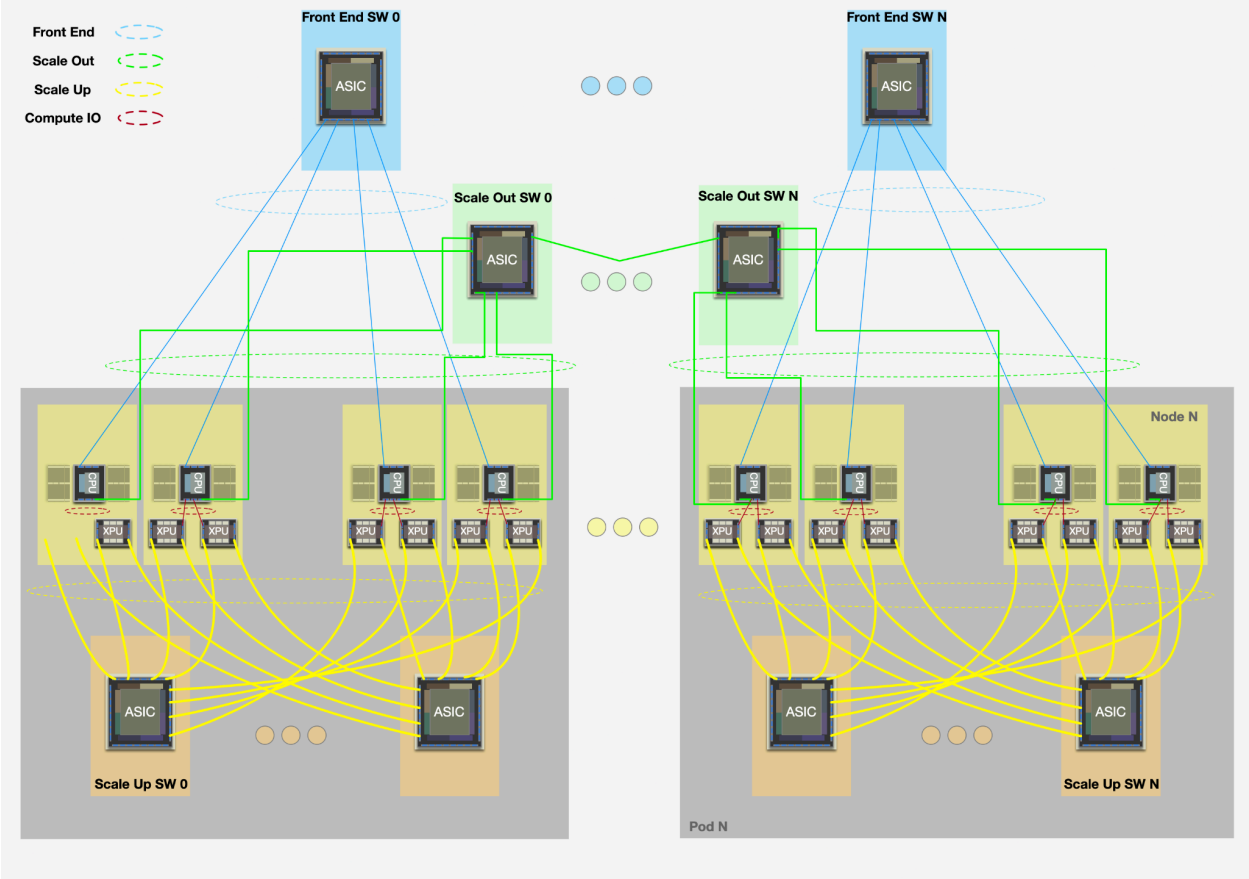


Figure 5.11. *OIF EEI AI Compute aggregates Pods with Scale-out links.*[11]

COI/EEI alignment also reinforces the need to treat the optical path as part of the system architecture rather than as a simple replacement for copper. Retiming strategy, linear versus partially retimed optics, link training, FEC partitioning, diagnostics, management interface, and modulation compatibility should be evaluated together, especially where copper and optical segments may use different signaling assumptions.

Per-class contract realization. Each placement class realizes the five-dimension contract differently. CPO operates with a mechanical envelope at the package

[11] Modeled from OIF EEI COI White paper. See: https://www.oiforum.com/wp-content/uploads/OIF-EEI-COI-WP-01.0.pdf

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

substrate, power delivered through the substrate, signaling at D2D/D2OE distances with no connector in the high-speed path, thermal coupled to ASIC cooling, and management via the host's RAS API endpoint. NPO operates with a mechanical envelope adjacent to the package on the host PCB, power from host rails, signaling across the short host trace, thermal independent of ASIC, and management via a host-coordinated RAS endpoint. S-NPO and M-NPO add socket and mezzanine interfaces respectively at the mechanical and signaling dimensions, with management routing through the respective interface and RAS event reporting traversing the additional boundary. FPO operates with the most decoupled contract, a mechanical envelope at the faceplate cage, independent power and thermal, signaling across the full host PCB depth, and management via the cage-level or front-panel module interface. XPO inherits this FPO placement contract and adds a rack/pod-boundary optical-format-conversion function, typically with an internal retiming or format-conversion stage and management via a rack-aggregation RAS endpoint. The Channel Budget Reference Table characterizes the signaling dimension across all placement classes and calls out XPO as a higher-penalty FPO subset where relevant; § Reference Platform Scope describes how the mechanical, power, thermal, and management dimensions integrate at the tray and rack levels.

The boundary transition pattern. S-NPO, M-NPO, and the XPO subset of FPO all share a structural pattern: each defines a characterized interface between two architectural domains with explicit vendor handoff at the boundary. S-NPO defines the boundary at the substrate socket, with single-device handoff (OE or copper cable plugs into the same socket). M-NPO defines the boundary at the mezzanine edge, with subsystem-level handoff (ASIC and OE as a separable carrier; the switch vendor characterizes the channel to the mezzanine edge, the OE integrator takes over from there). XPO defines the handoff at the rack aggregation point, where the faceplate-pluggable module performs cross-tier optical format conversion between intra-pod scale-up and inter-pod scale-out domains. The same envelope-and-contract pattern therefore operates at multiple system boundaries, while the placement

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

taxonomy keeps XPO inside the FPO family rather than creating a parallel placement class.

Vendor implementations as envelope occupants. Current vendor product families occupy specific points within these contracts. Samtec CPX (SFCM + SFCC) and Molex Impress Co-Packaged Copper Solutions are realizations of the S-NPO contract, distinguished by their internal mechanical and signaling implementations while sharing the contract's external interface. Merchant silicon vendor paddle and mezzanine carriers under development are realizations of the M-NPO contract. The OCI MSA, OSFP-XD module specifications, and others are realizations of the FPO contract. XPO MSA and rack-boundary aggregation products are FPO realizations specialized for optical format conversion. Each MSA defines a specific way to fill in the envelope; the architecture's contract layer ensures interoperability across MSAs occupying the same placement class. New implementations enter the architecture by meeting the contract, not by being predicted by the document.

The architecture's normative contribution is the contract layer: the five contract dimensions per placement class and the COM–RAS coupling as the design-time and run-time pair for the link layer. The implementations behind the contract are the ecosystem's work: vendors, MSAs, standards bodies, and integrators fill in the envelope according to their respective competencies. This separation makes the architecture vendor-neutral by construction and durable across implementation generations. It is what makes an ideal photonic AI infrastructure buildable, so ODMs, OE vendors, switch silicon vendors, and system integrators can each commit to their portion of the system with confidence that the others will interoperate at the agreed interfaces.

5.7. System Integrator Validation Matrix

Each placement class should include a validation matrix covering electrical channel margin, optical link budget, thermal operating envelope, power delivery, management telemetry, service motion, and fault-containment scope. A placement class is considered

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

reference-platform-ready only when all seven dimensions close simultaneously under the stated workload and environmental assumptions.

A placement class can satisfy all seven dimensions and still fail on a line. Manufacturing readiness is the eighth closure dimension, and unlike the others it is not settled on paper. It is established in hardware: through test vehicles and production runs that show the class can be built and serviced the same way at volume, not once.

A major risk is optical coupling. Single-mode attach is sensitive to very small lateral offset, and the loss it adds is governed less by the nominal design than by what the process holds across many parts: fiber and FAU placement, adhesive shrinkage during cure, and the drift left after thermal cycling and handling. Closing the dimension means characterizing that distribution, not a first-pass best case, and showing where yield actually falls.

The evidence comes in stages. A production-host vehicle fixes the chassis or tray shell, the engine keep-out, fiber routing, cooling access, connector access, factory handling, and datums; it does not prove coupling yield. Coupling and attach are developed on separate vehicles, narrow at first, from single-fiber edge or grating coupling to FAU attach, detachable interfaces, and reattach repeatability, so variables are isolated before density is added. The shared passive plane — faceplate, ODF or shuffle, raceway, rear interface, and ELS or ELSFP service zone — is validated on its own with loopbacks or surrogates, ahead of any final engine.

End-of-line measurements should reach platform management rather than stop at the test station: Tx/Rx power, laser bias current, module temperature, received signal quality, and pre- and post-FEC behavior, mapped so a later fault resolves to a physical boundary — engine, FAU bond, connector, ODF cassette, fiber plane, blind-mate interface, ELS module, or thermal-control issue. At hyperscale fiber counts, the

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

serviceable unit is a qualified plane, cassette, cartridge, or keyed harness, not a loose fiber, and should be validated for replacement, cleaning, inspection, and post-service verification.

5.8. Reference Platform Scope

This section translates the preceding taxonomy into a reference deployment envelope. It defines the mechanical limits, optical-placement assumptions, and rack-scale targets that the reference platform supports.

The baseline envelope is anchored to Open Rack v3 and the MHS 1OU common tray and chassis geometry. Table E defines the mechanical footprint used throughout this section.

Table E: MHS 1OU Common Tray and Chassis Envelope

Metric	Dimension (mm)	Description
Height	41.50	Internal vertical clearance
Full Depth	808	Total depth of the chassis
Chassis Depth	766	Depth excluding specific rear/front protrusions
Total Width	438.00	Maximum external width
Inner Width	431.40	Width between main chassis walls

Current rack-scale systems often depend on large copper cable and cartridge assemblies. These structures consume volume, complicate airflow, constrain routing, and create long-term serviceability challenges. The reference architecture reduces that overhead by moving toward a modular, layer-routed photonic stack: optical engines stay on the board plane, and fiber routes through thin shuffle and raceway layers. The

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

architecture stays within the Table E envelope while preserving open, multi-vendor silicon integration.

Figure 5.12 contrasts these regimes. The copper path carries layer, retimer, regulator, and pluggable-cage overhead; the photonic path collapses much of that structure into the board stack and recovers tray volume.

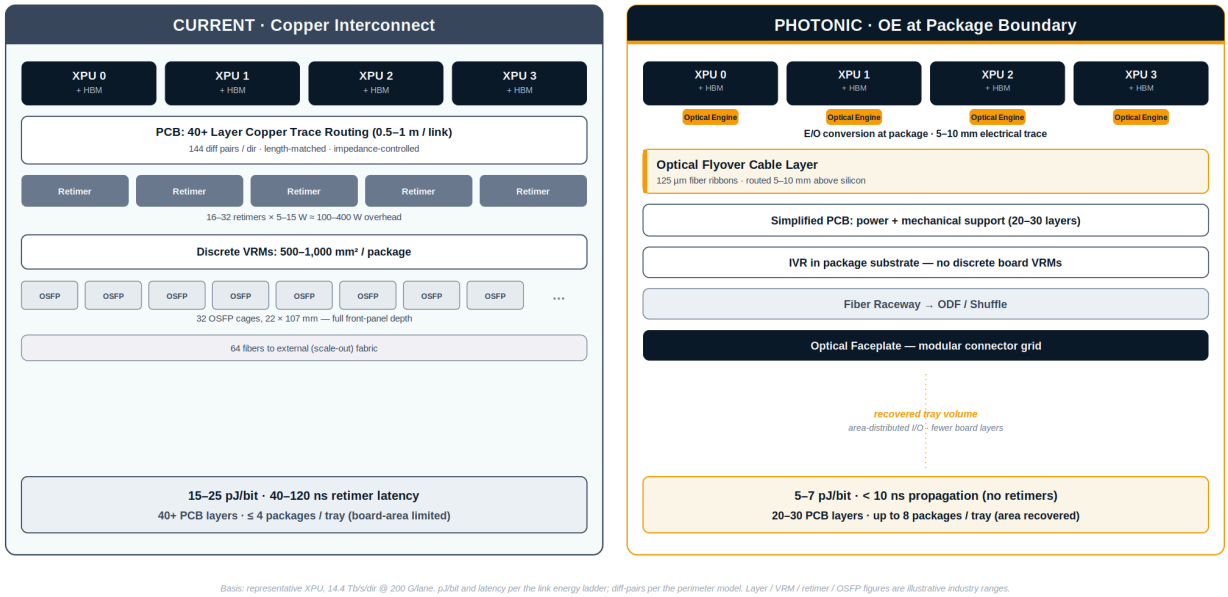


***Figure 5.12**. The interconnect stack, from Copper to Photonics.*

Classes are defined by tray orientation relative to this envelope. Class H is the horizontal MHS/Oberon-style tray that operates inside the baseline footprint. Class V mounts compute trays vertically in a canister (the vertical-tray style NVIDIA calls Kyber) to use rack depth for higher density. The detailed form-factor treatment appears in Platform Form Factor Evolution; here, Class H and Class V are the deployment targets the reference platform should support.

Class H follows the horizontal compute and switch-tray pattern. Within the Table E footprint, the layer-routed approach places optical engines on the board plane and routes fiber through shuffle and raceway layers. Integrated voltage regulation reduces

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

board-level VRM area, providing a path for the tray move from four packages toward eight. It supports the Dense-class tier at a roughly 72-XPU-per-rack baseline, with ELS provisioning and near-package or faceplate-pluggable placements (NPO, XPO) over DR- or OCI-class links. Class H is therefore the near-term open-rack foundation, bounded by board area: it reaches the practical horizontal envelope before Ultra-dense rack counts. Rear copper cartridge assemblies are removed; scale-out exits the tray as fiber, while short intra-tray scale-up may remain copper.

Class V follows the vertical-canister pattern, claiming rack depth the horizontal envelope leaves unused. Two HPM depth designators define the main variants. V350 places two ~350 mm trays back-to-back in each slot, with front and rear service access, and stays within the existing Class L envelope. V700 places one ~700–800 mm front-access tray per slot, which exceeds the Class L 365 mm depth limit and requires a new MHS HPM class near 800 mm.

Together, these variants raise the rack target toward roughly 288 XPUs, about three to four times the Class H baseline, and the optical I/O plant scales with them. At eight packages per tray, the scale-up plant alone runs on the order of 4,000 fibers, ~100 laser modules, and 7–9 kW of link power. At the co-packaged-optics ceiling, that already exceeds the horizontal 1OU faceplate and rear-termination budget by the same three-to-four-fold margin. Area-distributed escape (APO), by lifting per-package bandwidth past that ceiling, scales it further. The canister's reclaimed depth is what terminates, routes, powers, and cools that plant, the volume the horizontal envelope cannot supply. Class V is therefore the forward planning path for systems that exceed the practical Class H limit, with the copper midplane replaced by a passive photonic coupling zone.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

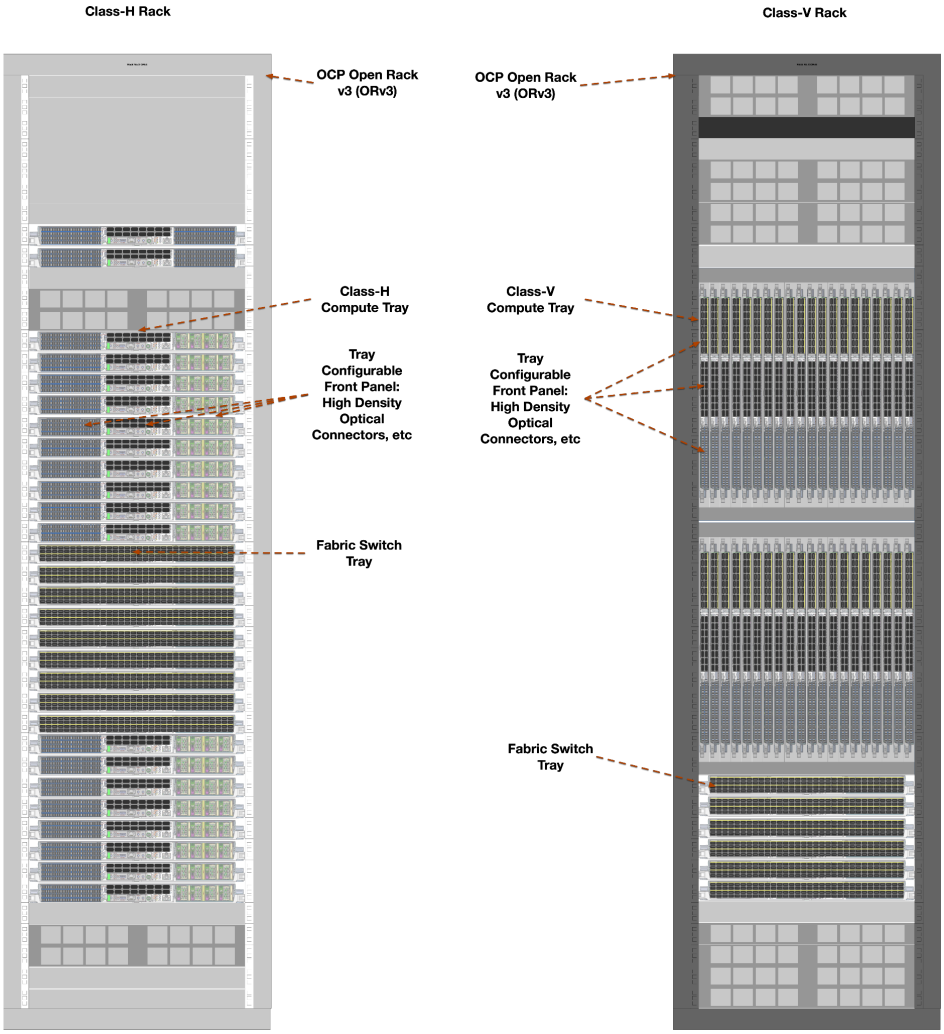


Figure 5.13. *Class-H and Class-V trays composing rack architecture.*

With this system objective in view, the reference platform scope is organized around two architectural paths, each defined by its system role, density envelope, and photonic integration profile:

Compute (XPU) Reference Platform: Defines the photonic interconnect integration for accelerator compute trays and trays. This platform carries scale-up traffic (XPU-to-XPU and XPU-to-switch via NVLink™, UALink™, or equivalent), scale-out traffic (XPU-to-fabric via Ethernet/UEC), and memory disaggregation traffic (XPU-to-remote-memory via CXL® or equivalent). The compute platform is the

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

higher-density, higher-complexity case: it hosts heterogeneous OE classes, integrates with the densest packaging (multi-die XPU modules with HBM), in order to accommodate the widest range of thermal and mechanical constraints.

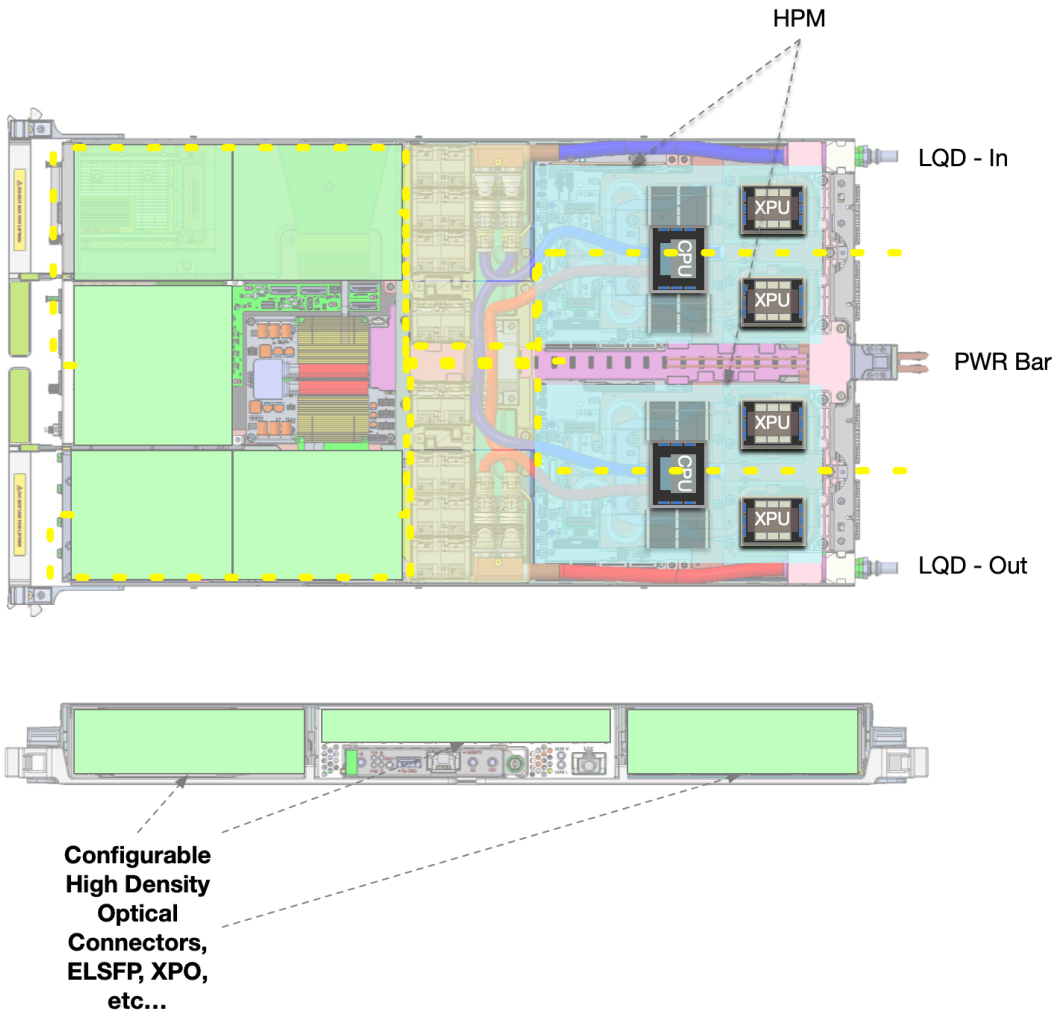


***Figure 5.14**. Example Compute Tray (Class-H) with 4 XPU.*

Switch Reference Platform: Defines the photonic interconnect integration for switch trays (both scale-up NVLink™/UALink™ switches and scale-out Ethernet/UEC switches). Switch platforms have different characteristics: higher port count but more uniform OE requirements, different thermal profiles (switch ASICs vs XPU die), and a more advanced deployment timeline as hyperscalers have already validated CPO in

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

production switch deployments. The switch reference platform can leverage this operational experience.

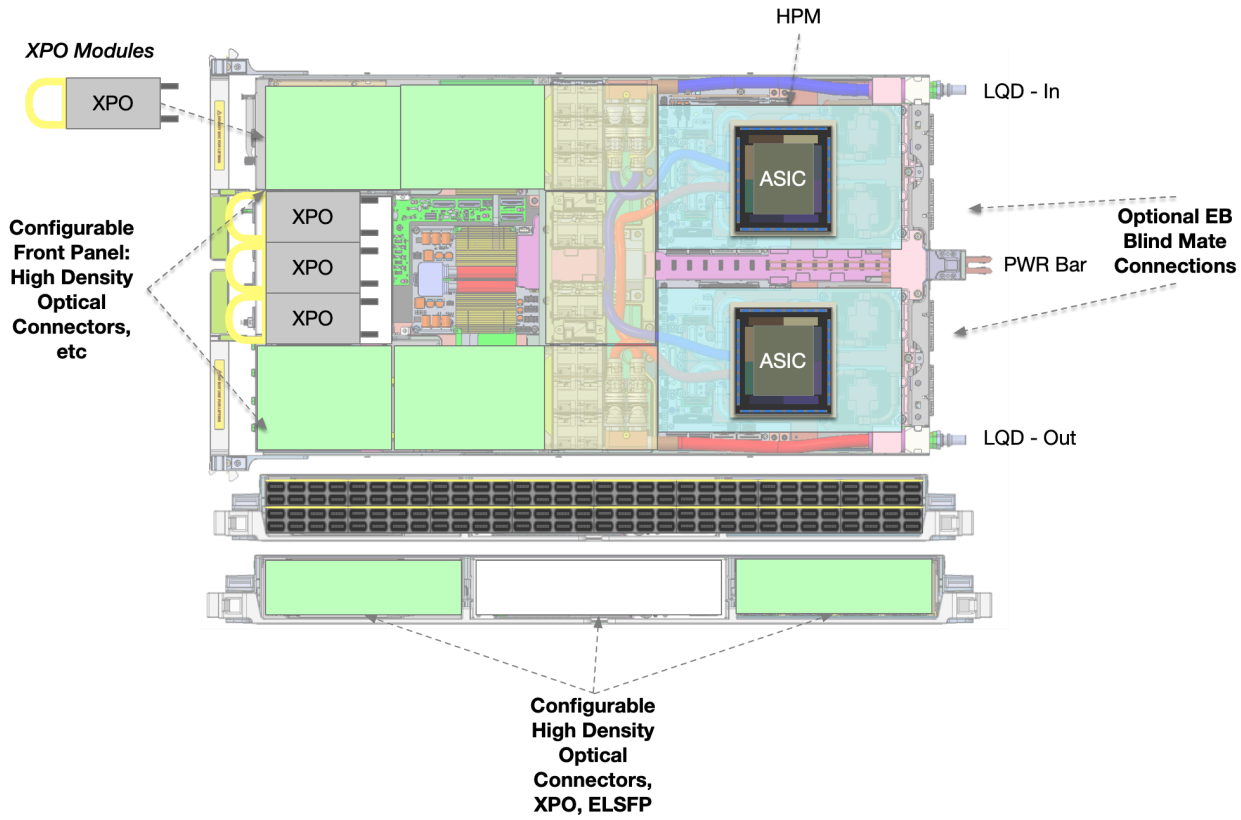


Figure 5.15. *Example 2-ASIC (Rail Type Switch) Configuration with XPO.*

HPM Configuration Progression: The reference platforms above describe what each role carries; the configuration progression below describes how many packages a single Host Processor Module (HPM) lands and how they are arranged. The architecture defines a four-tier configuration ladder that starts with the base MHS form factor and progresses into density configurations that may invoke MGX-class provisions per the OCP MGX Accelerated Computing Rack and Trays Specification (which aligns with the NVIDIA MGX product family) where the package combination exceeds the base envelope. Each tier consumes the same passive optical infrastructure from § Passive Optical Infrastructure Components including Faceplate, ODF, raceway, and optional Rear Interface Zone without modification. The configuration choice is an ODM and

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

integrator decision within the published envelope; the architecture neither prescribes a single tier nor requires migration between them. To support this variation over successive deployment waves, the architecture assumes a **Dynamic Grid framework**. Rather than forcing a static, coordinates-locked board layout, a Dynamic Grid defines a flexible spatial and routing envelope. This framework allows system layout designers to adjust component boundaries, pitch, and package allocations dynamically as hardware configurations evolve.

MVP-class: 1–2 XPU or Switch packages on the base MHS form factor. No mechanical or power additions beyond the photonic faceplate, raceway, and ODF specifications from § Optical Distribution Frame and § Passive Optical Infrastructure Components are required. This is the lowest-barrier entry point and the recommended starting configuration for integrators developing their first photonic-enabled platform; it allows early adoption of the passive optical infrastructure without invoking density-progression mechanical changes.

Compound-class: 2 XPU + 1 CPU on the base MHS form factor where signal integrity, power, and mechanical envelopes accommodate the combined footprint. Configurations exceeding two high-pin-count packages on a single HPM may invoke MGX-class mechanical and power provisions where the base MHS envelope cannot accommodate the combined package set. The architecture is neutral on which modular-server-platform specification is referenced and supports any that satisfies the contract dimensions of § Envelope and Contract; OCP MGX is the canonical reference and aligns with the NVIDIA MGX product family.

Dense-class: 4 XPU + 1 CPU + memory in a centered landing-zone topology. The CPU occupies the center landing zone with memory flanking it on two sides, preserving the shortest CPU–memory propagation path (the CPU is the latency-sensitive consumer of memory access), and the four XPUs are placed at equal radius around the CPU. XPU-to-XPU links are carried by photonic flyover above the board plane per § The Flyover Cable Layer ; the four-XPU equidistant topology is the architectural

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

consequence of removing the perimeter routing constraint that copper imposes, since photonic flyover does not consume PCB layer budget or perimeter escape on the host board. Signal propagation from CPU to each XPU is symmetric by construction.

The reference platform scope is intended to cover current and N+1 rack-scale architectures, which are best fit for current and imminent to market technologies and systems based on NPO, OCI, XPO, or similar integration approaches. In this envelope, Class H provides an open, silicon-agnostic rack architecture that supports heterogeneous compute, memory, I/O, and switching elements while preserving implementation choice across suppliers. Its physical and logical density targets are intended to be comparable to current state-of-the-art rack systems, with implementation-specific headroom where power, thermal, and interconnect budgets permit.

As rack-scale designs move toward substantially higher-density optical, thermal, or power-delivery regimes, including CPO-oriented architectures, Class V should be considered the forward reference architecture. In this sense, Class H provides the near-term open rack foundation, while Class V establishes the planning path for architectures that exceed the practical Class H design envelope.

Ultra-dense class: Higher XPU counts (illustratively, 8–12 XPU per HPM) enabled by smaller-footprint CPU classes (e.g., ARM-class or equivalent) that free landing-zone area for additional XPU placement. The architecture's contract dimensions remain unchanged at this tier; the binding constraints become thermal (per-package and aggregate per-HPM TDP) and power delivery (high-current distribution to densely packed XPU positions), addressed by the cooling and power infrastructure documented in § Passive Optical Infrastructure Components and OCP Cooling Environment. Higher-density tiers (Dense-class and Ultra-dense) naturally favor Class V (vertical tray) form factor, where the canister chassis provides the volume, cooling, and power infrastructure these XPU counts require; Class H accommodates Dense-class within board-area constraints but typically reaches its envelope before Ultra-dense densities

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

(see § Binding Constraints at 8-Package Density). The Class V HPM class supporting these tiers is formalized in § Proposed Modifications to ORv3/MHS with two depth designators: V350 at the Dense-class tier (4 XPU + 1 CPU per compute tray; MHS Class L envelope at L350), and V700-class at the Ultra-dense tier (8 XPU + 2 CPU + on-tray rail-switch per compute tray; new MHS HPM class proposed). CPU↔GPU links use short-reach copper (NVLink™-C2C class) by default; at V350 the per-tray area headroom preserves the option for an optical CPU where the integrator requires CPU participation in an optical fabric, while V700-class tray content density binds the CPU to copper-only. Per-rack XPU count at the 1:4 CPU:GPU ratio: V350 = 288 XPU/rack, V700-class = 576 XPU/rack. Two Dense (V350) racks compose the same true 576-XPU NVL576-class flat domain (§ Path to High-Radix Switching) within the existing MHS Class L envelope — the Ultra-dense tier seats it in one rack but needs the new HPM class. The V700 × 3-rack photonic scale-up domain reaches ~1,728 XPU.

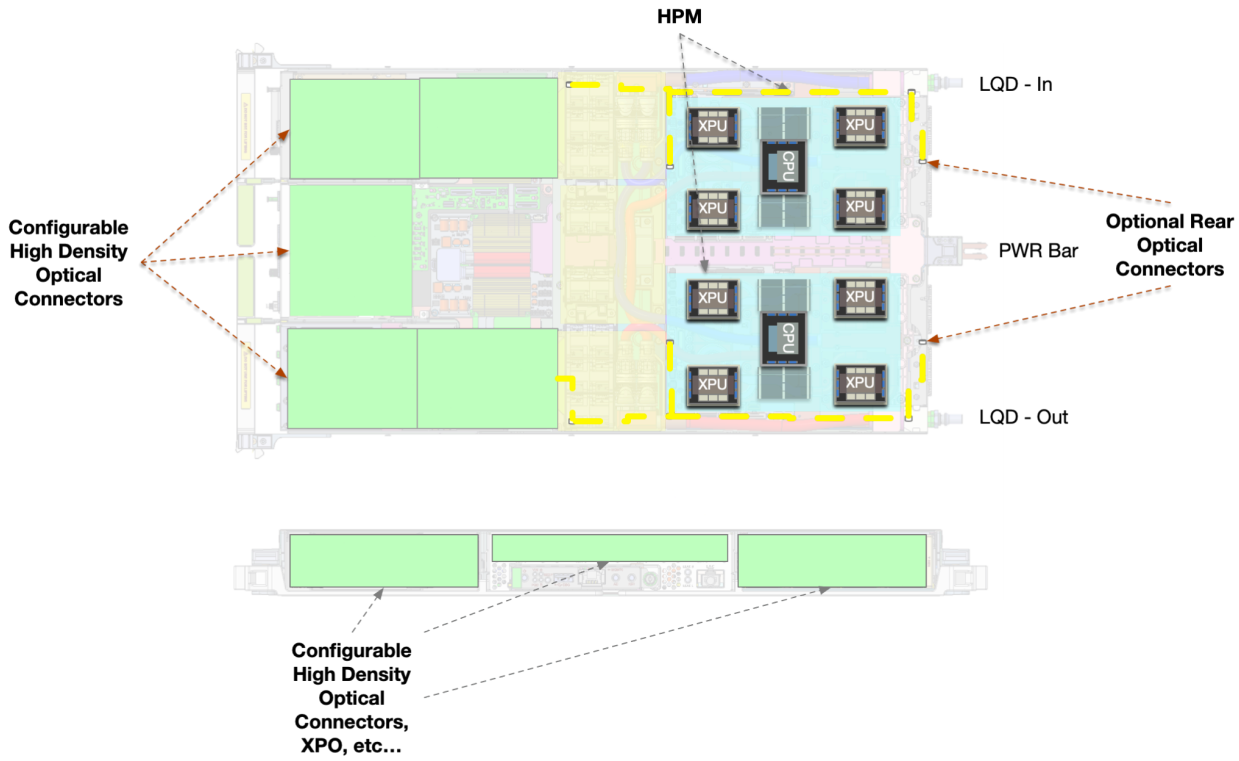


Figure 5.16. *Example 8-XPU MHS Configuration.*

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

The configuration progression is contract-compatible end to end: a Dense-class implementation and an Ultra-dense-class implementation occupying the same rack present the same fiber-plant interface to the ODF, the same thermal interface to the cooling manifold, and the same RAS API endpoints to the management plane. The progression is a buildable ladder, not a forced migration path; integrators select the tier appropriate to their workload, density target, and supply-chain readiness, and the passive optical infrastructure investment is preserved across tier transitions.

Both reference platforms share the same passive optical infrastructure (Optical Faceplate, ODF, Raceway, Laser Shelf) and the same Link Profile framework. They differ in tray-level OE integration, thermal management, and the specific extensions to existing OCP specifications they require.

5.8.1. Photonic Benefits for Compute (XPU) Platforms

The benefits of photonic interconnect for compute platforms are detailed throughout this white paper. In summary: photonics extends the scale-up domain beyond copper reach (72 to 1,024+ XPUs), eliminates the copper midplane in dense canister architectures, flattens inter-canister BW within the rack power envelope, reduces interconnect energy from 15–25 pJ/bit toward <5 pJ/bit, and enables heterogeneous OE integration for different link classes on a single tray. The primary measurable outcome is MFU improvement from 38–43% toward 55–65% through larger, flatter scale-up domains.

For Class V compute platforms specifically, photonics also enables a deeper compute-tray option in which rail-switch ASICs are integrated onto the compute tray. The compute tray then becomes both endpoint and fabric participant: XPU injection BW, local tray-to-tray BW, and global rack-to-rack BW are co-designed against the same optical Link Profile envelope.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

5.8.2. Photonic Benefits for Switch Platforms

The switch case is equally compelling but for different reasons. Switch ASICs are the most I/O-intensive components in the system—a current-generation 51.2T switch has 512 SerDes lanes (at 100 Gb/s) or 256 SerDes lanes (at 200 Gb/s), and next-generation 102.4T switches will double aggregate throughput at constant lane count by moving to 400 Gb/s/lane. At 200 Gb/s and 400 Gb/s per lane, the switch ASIC's I/O challenge is dominated by three constraints that photonics directly addresses:

High radix through SerDes liberation: A switch ASIC's radix (number of ports) is fundamentally limited by how many SerDes lanes can be routed off-package through copper traces to front-panel connectors. Copper at 224G requires controlled-impedance differential pairs with strict spacing, length matching, and via management. Each lane consumes precious PCB real estate and package escape routing. At 448 Gb/s, this problem intensifies: the usable trace length shrinks, the routing density decreases, and the practical radix of a copper-connected switch drops. Photonic interconnect breaks this constraint. OEs co-packaged with the switch ASIC convert electrical SerDes to fiber at the package boundary, eliminating the long copper traces entirely. This enables the switch designer to use the full SerDes lane count of the ASIC as usable ports rather than derate for SI margin, restoring the routable electrical radix — the SerDes lanes that actually escape the package — to the silicon's nominal lane count. Performing the optoelectronic conversion at the package boundary also shortens the electrical channel, relaxing the host-SerDes equalization burden and enabling simpler, lower-power SerDes IP at the same closure target.

Fabric flattening and tier reduction: Today's large-scale AI fabrics require multi-tier Clos topologies (leaf-spine-superspine) because no single switch has sufficient radix to connect all endpoints directly. Each tier adds hop latency, cabling complexity, and power consumption. Every doubling of effective switch radix through photonic I/O can potentially eliminate an entire tier from the fabric. A 102.4T switch with photonic ports

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

operating at Profile B (BiDi DWDM) could support significantly more endpoints per switch than the same ASIC constrained by copper fan-out. At scale, tier elimination is transformative: removing one switch tier from a 100,000-GPU fabric eliminates thousands of switch ASICs, tens of thousands of cables, and megawatts of power.

The latency and performance benefits of flattening are measurable and significant. Each switch hop in a multi-tier fabric adds both serialization latency (nanoseconds per hop through the switch ASIC pipeline) and queuing latency (microseconds under congestion). In scale-up networks where all-reduce and all-to-all collective operations dominate, every hop multiplies tail latency across thousands of participating XPUs. A two-tier fabric that could be flattened to a single tier through higher switch radix halves the worst-case hop count, which directly reduces collective completion time. For large-model training, collective operations can consume 30–50% of total step time; reducing collective latency by even 20–30% through fabric flattening translates to measurable MFU improvement. In scale-up domains specifically, where the traffic pattern is dense all-to-all between tightly coupled XPUs, a flat single-hop fabric is the ideal—every additional hop is pure overhead with no functional benefit.

5.8.3. Merchant Silicon Evolution Reinforces the Photonic Case

Merchant switch silicon is evolving rapidly. Switch ASIC throughput has been doubling roughly every generation: 12.8T to 25.6T to 51.2T to 102.4T, with 204.8T on the horizon. SerDes rates track a similar doubling cadence moving from 100G to 200G to 400G per lane. Switching latency continues to decrease with each generation. These improvements are welcome and this architecture does not position photonics as an alternative to merchant silicon advancement—rather, photonics is the enabling technology that allows each generation of merchant silicon to deliver its full capability to the system.

The paradox of this doubling cadence is that it keeps the lane count roughly constant (a 51.2T switch at 200G/lane and a 102.4T switch at 400G/lane both have approximately

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

256 lanes) while making each lane dramatically harder to route in copper. At 200G SerDes, copper traces can reach roughly a meter with careful SI engineering. At 400G, that shrinks below half a meter. At 800G (the generation after next), copper may not close for any meaningful board-level distance without retimers. The switch ASIC's throughput doubles every generation, but the copper infrastructure's ability to deliver that throughput to the fabric degrades at every step. Photonic interconnect resolves this divergence: each merchant silicon generation gets faster SerDes and more throughput, and the photonic layer transparently converts that to optical without the copper derating. The same fiber infrastructure serves the 51.2T generation, the 102.4T generation, the 204.8T generation, and whatever follows, unchanged.

This dynamic also applies to emerging switch architectures that integrate multiple ASIC dies in a single package or tray to achieve even higher aggregate throughput. Multi-die switch configurations multiply the I/O challenge: more dies mean more total SerDes lanes requiring off-package routing. Copper routing for multi-die switch packages at 224G+ is among the most challenging PCB design problems in the industry. Photonic I/O at the package boundary makes multi-die switch scaling practical by eliminating the inter-die and off-package copper bottleneck.

5.8.4. Path to High-Radix Switching (512, 1,024+ Ports)

The architectural prize in scale-up networking is the flat, single-hop domain: every XPU one hop from every other, uniform latency, full bisection bandwidth, no oversubscription. The size of that domain is set by a single number — switch radix. This section addresses how high radix can be driven, what physically bounds it, and what to do once one switch can no longer hold the whole domain. Two stories run in sequence: climbing the radix ladder to the physical limit of a switching element, then composing finite-radix elements into larger fabrics at the minimum hop cost.

Climbing the radix ladder

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

One distinction the rest of the section turns on: there are two radices, and they need not match. The host silicon has an electrical radix — the number of SerDes lanes it can present, set by its I/O limits. That electrical I/O drives an output stage with its own photonic radix — how many independent fibers can leave the package or Fiber Array Unit (FAU). NPO with single-λ DR and one fiber per electrical trace is the 1:1 case: photonic radix equals electrical radix, and the output carries the silicon's full I/O, no more and no less. The output stage need not be photonic — copper, with OpenCPX as the next horizon, is the alternative — but copper output is perimeter- and reach-bound, so it shares the ceiling photonics is trying to escape. Everything below concerns what happens to the electrical-to-photonic radix match as electrical radix climbs, and there are exactly two levers on the photonic side: wavelength-scaled radix (WSR — more wavelengths per fiber, so a fixed fiber budget keeps pace) and escape dimension (CPO perimeter → Areal Packaged Optics (APO) area, which lifts the fiber-escape ceiling itself, $L \rightarrow L^2$). Photonic radix is their product; the wall is where it can no longer keep pace with electrical radix.

Electrical radix grows along two orthogonal knobs — more lanes (wider) and faster lanes (hotter). Hold the lane rate and the ladder is a sequence of lane counts; the rungs that matter are the ones that exactly seat an industry-standard scale-up domain in a single hop. Two domain-sizing conventions coexist in the ecosystem: a power-of-two convention with UALink™/Ethernet lineage (512, 1,024, 2,048) and a 72×2^k convention with NVLink™ lineage anchored on the 72-XPU base (576, 1,152, 2,304). A switching element earns a generation's flat domain only when its radix reaches the larger member of the relevant pair — which is precisely why a 512-radix part falls 64 ports short of an NVL576 domain while a 576-radix part seats it exactly. The ladder below is the union of the two conventions, generation by generation.

Table F: The Radix Ladder at 200G and 400G per Lane

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Radix	Domain seated in a single hop	Capacity @ 200G/lane	Capacity @ 400G/lane	I/O-escape regime
512	≤ 512 XPU (2^9, UALink™-class)	102.4 T	204.8 T	Monolithic die; full beachfront via CPO (edge SerDes, ≈0 optical hop)
576	NVL576 (72×8)	115.2 T	230.4 T	Monolithic die at the practical beachfront edge for 400G (CPO)
1,024	1,024 XPU (2^{10}, UALink™ 1.0 target)	204.8 T	409.6 T	Reticle / beachfront ceiling — multi-die buys edge; APO (area) extraction begins
1,152	NVL1152 (72×16, Feynman-class)	230.4 T	460.8 T	Beachfront saturated — APO (any-pin area extraction) preferred
2,048	2,048 XPU (2^{11})	409.6 T	819.2 T	APO mandatory — I/O from the die area, not the edge
2,304	72×32 (next NVLink™ horizon)	460.8 T	921.6 T	APO mandatory — area extraction / stacked multi-die

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Climbing by adding lanes runs into geometry, not transistor count. SerDes lanes escape through the die edge — the beachfront — and beachfront is a perimeter resource that grows as L. Switching throughput, and the logic and buffering that support it, are area resources that grow as L^2. The two cannot track: doubling die area buys only $\sqrt{2}$ more perimeter. A single reticle (≈858 mm^2) seats on the order of 512 lanes at practical escape pitch; pushing toward 1,024 lanes and beyond exhausts the edge of even a stitched, reticle-limited die. Co-packaging (CPO) relaxes the copper ceiling — by sitting the OE ≈0 mm from the perimeter SerDes it removes the trace-reach and equalization penalties, drops the faceplate-cage area limit, and lets the designer run the full beachfront at rate rather than derate for SI margin. But it does not change the topology of the problem. The SerDes still line the edge and the I/O still leaves through the perimeter; CPO maximizes radix within the L regime, it does not exit it.

The optical profile decides how punishing that perimeter is — it sets photonic radix against electrical radix. A switch ASIC's electrical radix is N SerDes lanes; the photonic radix that actually reaches the fiber plant — and so the single-hop domain that can be stood up — is the package's fiber-escape budget divided by fibers per port, where fibers per port is a property of the profile, not the silicon (Table F; Table P). The span is wide. A single-λ DR port carries one wavelength per fiber with separate Tx/Rx, and IEEE single-λ DR tops out at 200 G per wavelength — so a 1.6 T port on the familiar standardized profile spends 16 fibers, while an 8+8 ×200 BiDi port (eight wavelengths per direction, sixteen total) collapses the same 1.6 T onto a single fiber: a 16× span. For a package with a fixed fiber-escape budget, the profile therefore sets the realizable radix: an envelope that escapes ≈1,024 fibers stands up roughly 64 ports on single-λ DR8 and the full 1,024-port single-hop domain on 8+8 ×200 BiDi. The ASIC did not change. On fiber-heavy profiles, fiber escape rather than silicon becomes the wall, and it binds at a fraction of the nominal radix. The familiar PMD is itself the radix hog, which is the cleanest motivation for what follows. OCI sits between these endpoints rather than at the dense extreme: its v1.0 PHY carries 212.5 Gb/s bidirectionally (425 Gb/s

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

aggregate) on a single fiber via four DWDM wavelengths per direction (4+4, Profile E), so a 1.6 T port costs about 8 fibers; its bidirectionality already halves the single-λ DR count, and the per-fiber rate doubles a generation at a time.

Directionality is the quieter scaling axis. Adding wavelength depth increases density, but also adds interoperability burden: every additional wavelength is another operating point that both ends of the link are to support and validate. In an open scale-up fabric spanning multiple switch vendors, that validation burden is significant. Bidirectionality avoids this class of agreement. It carries transmit and receive over the same fiber, reducing fiber count by half without expanding the wavelength contract. That is why single-λ DR, alone among the profiles, still carries separate Tx and Rx. It is the one place where the lever remains unspent, and when folded, it meets OCI's fiber count without adding a wavelength: the same density is bought on the axis the interoperable fabric is free to use. Photonics therefore moves from uni- to bidirectional as the efficiency default; the profiles above already sit there. The point sharpens as integration moves toward the area domain: the bandwidth a reticle generates is still constrained to leave through a fiber-attach footprint of fixed pitch, and on the sparse interoperable profiles it is bidirectionality that keeps that egress inside the connector and FAU area the package can present.

The second knob, lane rate, is explored on its own axis. It raises capacity without adding lanes, but every doubling collapses the copper reach and tightens the optical coupling pitch:

Table G: The Lane-Rate Axis (radix held at 1,024)

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Lane rate	Capacity @ radix 1,024	Copper board reach	Escape regime
200G	204.8 T	~1 m (SI-engineered)	Copper marginal at this radix; photonic edge preferred
400G	409.6 T	< 0.5 m	Copper infeasible at radix; photonic edge required
800G	819.2 T	cm-class → negligible	Photonic only; edge-coupler / FAU pitch pressured
1.6T	1.64 Pb/s	none	Photonic only; perimeter escape exhausted → area-domain extraction

The lane-rate rungs above are electrical SerDes rates; single-λ DR optics realize only ≤ 200 G per wavelength, so per-fiber payload at 400 G and above is reached through wavelength-scaled profiles (Table F), not a single 400 G optical lane.

Power efficiency at extreme I/O density: Switches move more bits per watt than any other component in the system because their entire function is I/O. The power consumed by SerDes and associated retimers, cable drivers, and signal conditioning on a copper-connected switch tray is a large fraction of total switch power. Photonic I/O at ~5 pJ/bit versus 15–25 pJ/bit electrical delivers proportionally greater savings on switches than on compute, because a higher percentage of switch power is I/O-related. For a 102.4T switch tray, the estimated power reduction from moving to CPO is on the order of hundreds of watts per tray, compounding across the hundreds or thousands of switch trays in a large fabric.

Cabling complexity collapse: A single 51.2T switch with pluggable optics today requires 64 OSFP cages and 64 fiber cables at the front panel. With DWDM CPO (Profile B, 16λ BiDi), the same BW can be carried on significantly fewer fibers,

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

potentially a 4–16x reduction in physical cable count depending on wavelength density. At fabric scale, this transforms cable management from a dominant operational burden into a manageable infrastructure task. The ODF/Shuffle architecture described in § Optical Distribution Frame is particularly valuable for switches, where the cable count reduction and topology reconfigurability have immediate operational impact.

Proven deployment path: Unlike compute CPO, which is pre-production, CPO switch deployment has been validated in production by hyperscalers. This operational experience provides concrete reliability data (2.6M hour MTBF demonstrated), deployment methodology, and failure mode characterization that can be leveraged directly by the switch reference platform. The switch platform is therefore a lower-risk entry point for organizations adopting photonic interconnect, and lessons learned from switch deployment can inform the compute platform specification.

Class V Canister Geometry

The Class V frontside to backside arrangement has a physical depth budget that merits explicit characterization, because it determines what switch tray form factor is achievable and what optical coupling geometry is practical at the rack interior.
An extended ORv3 derivative rack envelope provides approximately 960mm of usable internal depth rail-to-rail. After allocating for front door and cable management clearance, the compute tray depth, and rear door and ISL cable clearance, the residual available to the rear-loading switch assembly is approximately 350mm of board and chassis depth. This residual is not incidental, it corresponds naturally to established MHS form factor targets for rear loading board assemblies (Class N or L at the 350mm depth point), and it accommodates a full switching ASIC board with photonic interfaces, power delivery, and liquid cooling termination within existing mechanical conventions.

The zone at the rack interior where compute tray rear faces and switch tray interior faces oppose each other is the optical coupling zone. In the copper midplane model, this zone is occupied by a PCB with high-density connectors. In the photonic model, it is

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

a passive fiber distribution surface: MPO or equivalent connectors on the switch assembly's interior-facing edge mate with fiber ribbons from the compute tray raceways, establishing the compute-to-switch paths without signal regeneration. No active components occupy this zone. The fiber paths are short, loss budgets are trivially satisfied, so no retimer latency is introduced.

The switch assembly's second fiber boundary faces the rack exterior at the rear. All inter-switch links, both inter-canister links within the same rack row and scale-out links to remote racks, exit through this face. This satisfies the condition above in a specific direction: scale-up traffic between compute and switch stays entirely within the canister and is managed at the interior coupling zone; ISL traffic exits at the rear and is managed at the rear patch point. These are structurally distinct fiber boundaries serving distinct traffic classes, and treating them as a single unified surface would conflate two different serviceability and topology management problems. The rear ISL exit also has the operational advantage that inter-switch cabling is accessible and reconfigurable from the rack rear without any interaction with the compute side of the rack.
Thermal management for the switch assembly in this position is handled by the same direct liquid cooling manifold infrastructure that serves the compute side. A 102.4T (and greater as roadmap defines) class switching ASIC with co-packaged optics operates within the per-slot thermal budget that DLC-capable ORv3 rack configurations are designed to sustain. The rear-loading orientation eliminates the reverse-airflow conflicts that have historically complicated placement of network equipment in front-to-rear airflow racks, because DLC removes airflow from the primary thermal path entirely.

The same depth budget also defines the distributed Class V option. Instead of assigning the residual rack depth to a separate rear-loading switch assembly, the implementation may allocate additional depth to the front-inserted compute tray and place rail-switch ASICs on that tray. This removes active rear networking cards while retaining the rear optical zone as a passive egress, shuffle, or cartridge surface. Failure modes are

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

functionally comparable to rear switch-card designs because the active switch ASICs still reside on replaceable cards; the difference is that the service boundary moves to the compute tray and the midplane blast radius is removed. These canister populations, the V350 and V700-class depth designators with their per-tray content and per-XPU footprint, are specified once in § Proposed Modifications to ORv3/MHS.
PCB area available per tray is approximately 158,000 mm² for V350 (350 mm × ~450 mm) and 315,000 mm² for V700 (700 mm × ~450 mm).

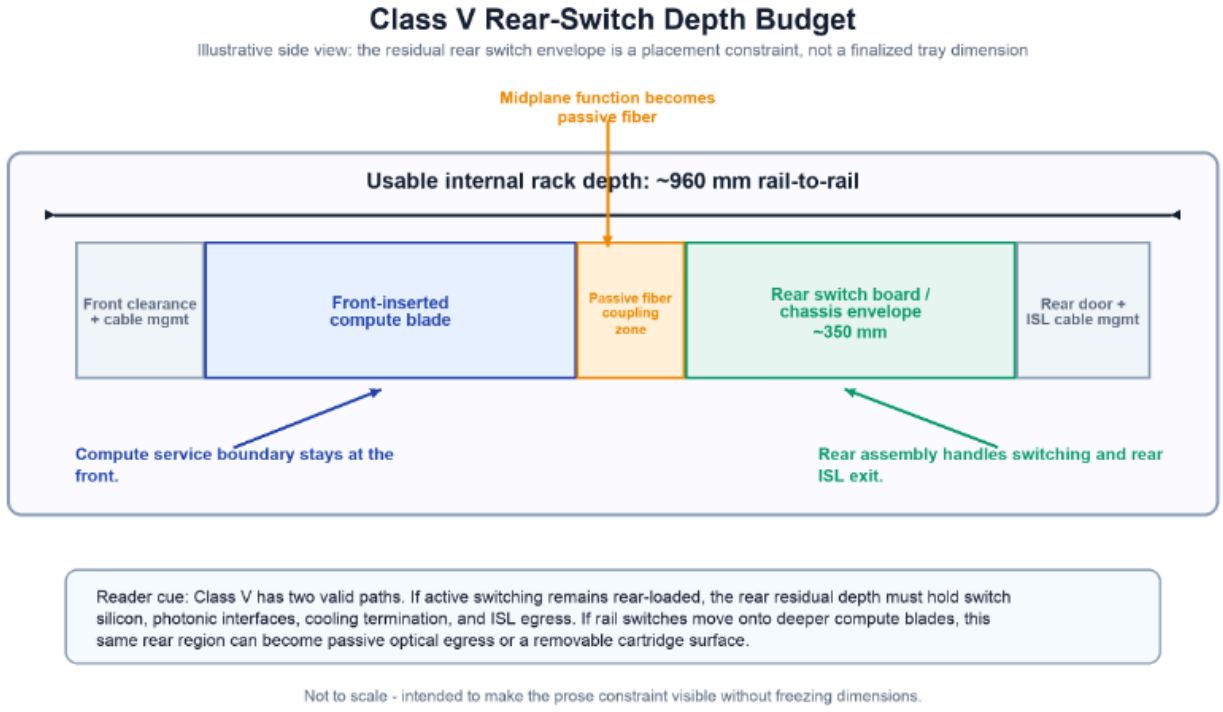


***Figure 5.17**. Class V canister depth-budget view*

V350 back-to-back (~350 mm front compute tray + ~100 mm photonic coupling zone + ~350 mm rear switch tray) versus V700-class single front-loaded (~700–800 mm compute tray with on-tray rail-switch; rear reserved for passive fiber egress). Two stacked rows of 18 tray-slot positions per canister, 26–28 mm slot pitch. Dimensions are illustrative pending Design Specification iteration.

The Photonic Density Thesis

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Photonic interconnect makes physical density and logical disaggregation independent design knobs, not competing ones. This section develops the density side; § System-Level Impact on AI Workloads takes up the logical-disaggregation consequence.

Rack-scale processor density binds on the min-of-five constraints of the §4 lens: interconnect reach, PCB signal-layer count, package-perimeter egress, PDN headroom, and thermal-manifold access. In dense racks these couple through the midplane. A 40+ layer impedance-controlled PCB carries the differential pairs, consumes the layer budget PDN would otherwise use, fixes the perimeter escape geometry, and defines the manifold footprint. Tightening one tightens the others.

Flyover interconnect, independent of medium, relaxes three of the five simultaneously by moving the signal off the PCB. Layer count collapses toward the minimum required for power and low-speed control. Layers vacated by signal duty become power planes, raising the local current density the package can sustain and therefore the per-package TDP that survives at tight pitch. Perimeter egress becomes a connector stack-up problem rather than an edge-routing problem. The midplane ceases to be a 40+ layer artifact that must be redesigned per SerDes generation; it is replaced by passive cabling whose generation cycle is set by the connector and the fiber, not the laminate stack.

What photonic flyover adds is the fifth: reach. Copper flyover at 224 Gb/s/lane using PAM4 reaches approximately 1 m on 32 AWG and falls below 25 cm at 448 Gb/s/lane; photonic flyover extends reach by two to three orders of magnitude with no SerDes-generation coupling. Copper flyover alone yields a denser canister; photonic flyover yields a flat rack-scale domain. Both lifts are required for the volumetric arrangement described in Phase 3, where all five constraints are simultaneously slack.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

The architectural implication: optimize compute density against cooling and power delivery, and let the photonic fabric carry the BW uniformly across whatever physical arrangement those constraints permit. Architectures that spread compute to recover signal integrity margin pay infrastructure cost for a constraint the photonic link has already removed. The next subsections develop this consequence at the tray level (§ Density Doubling Through Board Area Recovery), the mechanical enabler of cabled flyover routing (§ The Flyover Cable Layer), and the new binding constraints at increased density (§ Binding Constraints at 8-Package Density).

5.8.5. Density Doubling Through Board Area Recovery

The photonic density thesis has a concrete near-term implication: doubling the number of XPU packages per tray from 4 to 8. It follows directly from the board area freed when copper interconnect, pluggable transceiver infrastructure, and discrete power delivery give way to photonic and advanced-packaging alternatives.

Today's 4-package compute tray dedicates the majority of its board area and volume to four functions that photonics and advanced packaging can eliminate or dramatically reduce:

Copper trace fan-out: Each XPU package potentially requires hundreds of controlled-impedance differential pairs routed through the PCB to reach FPO connectors or midplane interfaces. These traces demand strict spacing, length matching, and via management, consuming substantial board area and requiring 40+ layer PCB stackups. A typical package with 256 SerDes lanes (512 differential pairs for Tx+Rx) monopolizes a routing corridor far larger than the package itself. With co-packaged optics, SerDes conversion occurs at the package boundary—electrical traces shrink from 0.5–1 meter board traversal to 5–10 mm to the adjacent OE. Board layer count can drop from 40+ to 20–30 layers, and the routing corridor between packages nearly disappears.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Pluggable transceiver cages: A compute tray with 32 OSFP ports dedicates its entire front-panel depth to transceiver cages, each approximately 22 mm × 107 mm × 13 mm. This volume is recovered in a CPO system, the pluggable function moves to the package, and the front panel becomes the thin Optical Faceplate and ODF described in § Passive Optical Infrastructure Components . The recovered depth is available for additional package positions or for the fiber flyover routing layer.

Retimers and signal conditioning: Copper links at 224G and above frequently require retimer ICs to extend reach. A 4-package tray may contain 16–32 retimers, each consuming 5–15 W and approximately 200 mm² of board area with associated passives. Photonic interconnect at CPO, NPO, or S-NPO eliminates retimers entirely—fiber does not require signal conditioning regardless of distance within the rack. This frees both power budget (potentially 100–400 W per tray) and board area.

Discrete voltage regulators (VRMs): Each XPU package requires multiphase voltage regulators with large inductors and capacitor banks, typically occupying 500–1,000 mm² of board area adjacent to the package. Integrated Voltage Regulators (IVR), which move final-stage voltage conversion into the package substrate or interposer, can eliminate this footprint. IVR is an independent technology trend driven by power delivery efficiency and package density, but it is synergistic with photonic interconnect: both reduce the board's role from complex multi-function substrate to simplified power distribution and mechanical support.

When these four area consumers are reduced or eliminated, the recovered board area and volume are sufficient to accommodate a second row of XPU packages. An 8-package tray retains the same mechanical envelope as today's 4-package tray and fills the freed area with additional compute, given the engineering to manage low-speed connectivity and power at that density. The packages connect to the photonic fabric through the same optical faceplate and ODF while the fiber infrastructure is indifferent to whether 4 or 8 packages share it.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

5.8.6. The Flyover Cable Layer

A key mechanical enabler for 8-package density is a structured Flyover Cable Layer that removes high-speed signaling lines from the structural limits of the host PCB stackup. This volume is not a new mechanical layer: the DC-MHS form factors already reserve it as Layer B which is the cabling-and-system-mechanical space between the host board's height-restriction zone (Layer A) and the major-component layer of cards, heat sinks, and cold plates (Layer C). The architecture populates that reserved layer rather than defining one of its own, and references every Z-height in this section to the host board plane, the M-DNO datum, not to the cold-plate surface. Rather than assuming a purely optical medium, the architecture treats this spatial envelope as a generalized multi-tiered routing volume capable of accommodating heterogeneous high-speed cabling options. To manage mechanical interference and tolerance stack-ups across different media, the architecture defines a multi-layer Z-height allocation framework:

- **Primary Routing Zone (e.g., HSIO-Zone 1):** The board-near tier, nominally 5–10 mm above the host board, routed among the cold-plate borders and coolant fittings. Flyover copper claims it because copper is the reach-limited medium and needs to take the shortest, most direct path, package edge to package edge. What copper carries here are the host-class links the DC-MHS form factors already treat as ultra-short-reach cabled host connections — PCIe®, the CPU↔XPU coherent link (NVLink™-C2C class), and CPU-to-CPU coherent links; routing this mesh through the copper tier rather than the host PCB is what frees the board's residual signal layers once the scale-up SerDes has gone to fiber. These links stay copper until their lane rate outruns flyover-copper reach — the same threshold that governs the scale-up lanes, crossed later here because host-class links signal slower. Where no flyover copper is present, photonic waveguides default into this tier instead.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

- **Secondary Routing Zone (e.g., HSIO-Zone 2):** A vertically stacked parallel volume positioned immediately above the primary zone. In hybrid architectures utilizing both media, this layer is explicitly reserved for flexible, lightweight photonic waveguide ribbons (which, at ~125 µm individual fiber diameter and ~3 mm wide for a 16-fiber ribbon, map smoothly into tight vertical constraints). Fiber takes the upper tier for the converse reason copper takes the lower one: it has reach to spare and can afford the longer path, up and over the cold-plate tops. Two consequences follow for free — the passive optical sits clear of the coolant plumbing that is the vibration source in this volume, and a ribbon bundle a few millimeters thick adds almost nothing to the tier it occupies, where a copper bundle would not.

The transition between the two tiers is the one place this ordering costs anything. Each package's fiber leaves its OE at the board-near attach, so reaching the upper tier requires a short vertical riser that clears the cold-plate border — a bend at the most delicate point in the path. This riser is the case the DC-MHS height exception already covers: the SFF-TA-1016 and SFF-TA-1026 connectors are granted purposeful intrusion from the board interface into Layer B, and the optical attach inherits that allowance. One difference is worth stating for the OCP MHS workgroup: those connectors intrude into Layer B occasionally, whereas flyover promotes Layer B to a primary routing plane, so its height budget is plane-wide rather than per-connector. The plane also depends on a defined cold-plate top-surface envelope to give the upper tier a floor; specifying that envelope is companion work for the OCP Thermal and Packaging working group.

The optical flyover layer is a thin mechanical channel—typically 5–10 mm in height—above the package and cold plate surface, through which fiber cables route from each package's OE to the raceway and ultimately to the ODF. This layer is structurally simple (guide clips or a thin sheet-metal channel) and carries no electrical signals so it is purely passive fiber routing.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

The architectural impact is that I/O routing is decoupled from the PCB layout. The board becomes primarily a power distribution structure (high-voltage DC bus, IVR power planes) and mechanical support platform (package sockets, cold plate mounting). Signal integrity—the dominant constraint in copper board design—is no longer a board-level concern. It is handled entirely within the OE at the package boundary. This decoupling is what makes 8-package or more density physically feasible within existing tray envelopes. The packages can be positioned at minimum pitch (limited only by HBM stack footprint and cold plate dimensions), with fiber flyovers routing above them to the faceplate. No trace routing channels are needed between packages. No retimer positions are needed. No OSFP cage depth is needed.

5.8.7. Binding Constraints at 8-Package Density

The transition from 4 to 8 packages per tray shifts the binding constraints from interconnect to cooling and power delivery:

Cooling: 8 packages at 700–1,200 W each produce 5.6–9.6 kW per tray, compared to 2.8–4.8 kW for 4 packages. Liquid cooling scales linearly—double the cold plate coverage, double the flow rate through the tray's liquid loop. Some current and emerging CDU and rack manifold architectures can support this thermal load. The cold plate design should accommodate the denser package pitch, but this is a mechanical engineering problem with known solutions, not a fundamental physics constraint.

Power delivery: With IVR, board-level power delivery may simplify toward a higher-voltage DC distribution bus (48 V or potentially 400–800 V DC) with final-stage conversion at each package. The total per-tray power draw of 5.6–9.6 kW requires adequate bus bar or power connector capacity, but this is within the ORv3 rack power architecture's capability, particularly with the transition to high-voltage DC distribution that dense rack architectures are already driving. What paces this tier is the conversion stage rather than the bus: the high-density in-package and point-of-load regulation that

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

delivers this current within the board area photonics has freed is built on wide-bandgap power devices — GaN at the point of load, SiC at the higher-voltage rack tier — whose volume maturity is still advancing: 2026 disclosures already show point-of-load conversion at 2+ A/mm² and 85–90% efficiency in shipping rack-scale designs, though not yet uniformly across the ecosystem. The 8-package tray is therefore achievable rather than aspirational, but its near-to-mid-term realization is paced by that maturity reaching the per-tray current envelope, not by the interconnect, which photonics has already removed as the binding constraint.

Managing Thermal Gradients and Cooling Topologies: As compute trays scale to higher densities, driving power configurations producing 5.6–9.6 kW per tray, the thermal gradient between the high-TDP ASIC/XPU and adjacent co-packaged OEs becomes a primary design consideration. Because OE thermal tolerances vary significantly across vendor implementations (with some engines highly resilient to elevated case temperatures and others requiring strict isolation), this architecture remains explicitly non-prescriptive regarding the exact mechanical cooling topology.

System integrators have the flexibility to align their mechanical cooling choices with their specific thermal masses and component selections, navigating a continuum of ecosystem innovation:

- **Monolithic Cooling:** Traditional baseline designs utilizing a single-piece cold plate to create a unified thermal mass. *Note: Monolithic designs need to carefully account for the significant Z-height profile of emerging Open CPX MSA and Co-Packaged Copper (CPC) flyover solutions (based on specifications of commercially available realizations), where the* OE *or copper connector sits directly on top of a BGA substrate.*

- **Decoupled or Chambered Designs:** Evolved mechanical topologies featuring discrete or multi-chambered cold plates that physically fit together. This approach helps isolate distinct thermal zones and accommodates the varying

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

vertical clearances (Z-heights) of the XPU relative to adjacent interconnect assemblies.

- **Advanced Flow Routing:** Strategies such as spiraling the inlet coolant from the cooler outer periphery (where temperature-sensitive optics are typically placed) inward toward the high-flux XPU die.

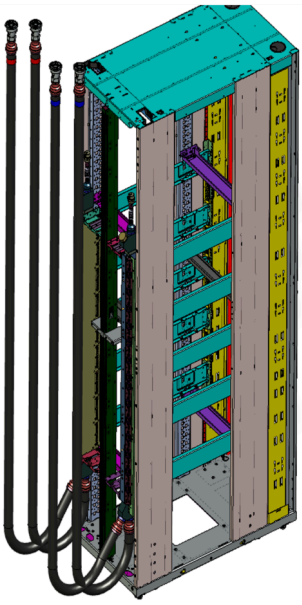
By keeping the thermal envelope open and framed as guidance, the architecture seamlessly accommodates both highly robust OEs capable of withstanding elevated temperatures and components requiring strict thermal isolation. This space remains identified as companion work for the OCP Thermal and Packaging working group.

The practical maximum package count per tray remains a validation outcome, not a fixed assumption. Where cooling and the power bus are not the binder, four items still require deployment-specific closure. Two are thermal: crosstalk between adjacent packages, and OE isolation, both sharpened by the denser pitch. Two are mechanical: fiber-routing clearance including Z-height, flyover, bend-radius at that pitch, and connector current rating. Serviceability closes the list.

IVR adoption is highly synergistic with photonics (such as CPO) and, to some extent, flyover copper, because it can free board area and simplify high-current distribution. However, their respective maturity timelines should not be assumed to align. Therefore, the architecture should support a dynamic grid, scaling from the MHS CPU allocation up to an envelope for 8-package-class and higher-density trays. The resulting Base and Design Specifications from this work will intercept the right capacity-to-capability points. As the optimum will likely shift depending on actual density, this flexibility allows board designers to validate and reduce risk across various package counts.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

5.9. Integration by Reference



This architecture does not re-specify capabilities that are already defined in existing OCP project contributions and external standards. Instead, it extends applicable specifications by reference, adding only the photonic-specific elements that existing specifications do not cover. The following OCP contributions are incorporated by reference and extended:

Modular Hardware System (MHS): The tray mechanical form factors (1U, 2U, 4U), blind-mate power and cooling interfaces, and tray-to-rack mechanical interfaces defined in MHS are retained without modification. This architecture adds the Optical Faceplate specification to the MHS tray front-panel definition and the Raceway specification to the MHS tray chassis definition.[12]

Open Rack v3 (ORv3): The rack mechanical dimensions, power bus bar specifications, cooling manifold interfaces, and rack management are retained without modification. This architecture adds an optional rack-level ELS Shelf position reservation and rack-level fiber routing channel specifications.

Open Systems for AI (OSAI): High Bandwidth Domain (HBD) concepts, compute tray and switch tray functional definitions, and pod architecture are retained and extended. This architecture extends the HBD definition to include photonic interconnect as the physical layer enabling multi-rack HBD, and adds the ODF and Optical Faceplate as new functional components within the OSAI tray and rack architecture.

OCP Cooling Environment: Cold plate specifications, CDU interfaces, and liquid cooling loop definitions are retained. This architecture identifies additional thermal management requirements for co-packaged OEs, the ELS shelf, and the front-panel

[12] https://chevalstore.com/wp-content/uploads/2024/10/Nvidia-MGX-with-rack-mounting-kit22-478x1024.png

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

module thermal envelopes as companion work items to be developed in coordination with the Cooling Environment project.

External standards and consortia, including OIF, IEEE, UCIe®, CXL®, UALink™, and UEC, are similarly referenced, not re-specified. IEEE 802.3 governs Ethernet PHY and management evolution; OIF governs CEI/EEI electrical attach, co-packaging, ELSFP, and related interoperability work; UALink™ and UEC define scale-up and scale-out fabric semantics; and OCP defines the rack, tray, power, thermal/liquid cooling, network interface, firmware management, RAS and manageability, serviceability, Open Chiplet Economy (chiplets), and platform-integration contracts that compose these technologies into open AI systems. The photonic architecture defines how these standards' interfaces connect to the platform infrastructure; it does not redefine the standards themselves. This integration-by-reference approach minimizes specification overlap, reduces the risk of conflicting requirements, and ensures that improvements to referenced specifications automatically benefit the photonic architecture.

5.10. Passive Optical Infrastructure Components

The architecture envisions five passive optical infrastructure components that together replace the function of a copper midplane while adding capabilities copper cannot provide. These components are described conceptually below; they replace the generation-specific reach-extension role of dense copper midplanes and copper cable assemblies with a passive, serviceable, and generation-invariant optical infrastructure. They do not prescribe a single OE, connector family, wavelength plan, or topology. Instead, they define the mechanical, optical, serviceability, thermal, telemetry, and recovery-domain envelopes into which different Link Profiles and optical-engine implementations can be qualified.

The passive optical infrastructure is intentionally separated from the active OEs. OEs may be CPO, NPO, S-NPO, M-NPO, FPO, or XPO realizations; the infrastructure defined here provides the common tray, canister, and rack-level substrate that lets those

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

implementations coexist. This distinction is central to the architecture: the active OE evolves with SerDes generation and photonic technology, while the passive fiber plant, faceplate, raceway, ODF, and optional rear egress are intended to remain stable across multiple active-component generations.

Detailed specifications would be developed collaboratively in a subsequent Base Specification. They should convert this section into *normative* requirements for mechanical envelopes, connector qualification, fiber population, bend-radius control, insertion-loss allocation, inspection and cleaning access, service motion, labeling, topology mapping, optical telemetry, and RAS reporting.

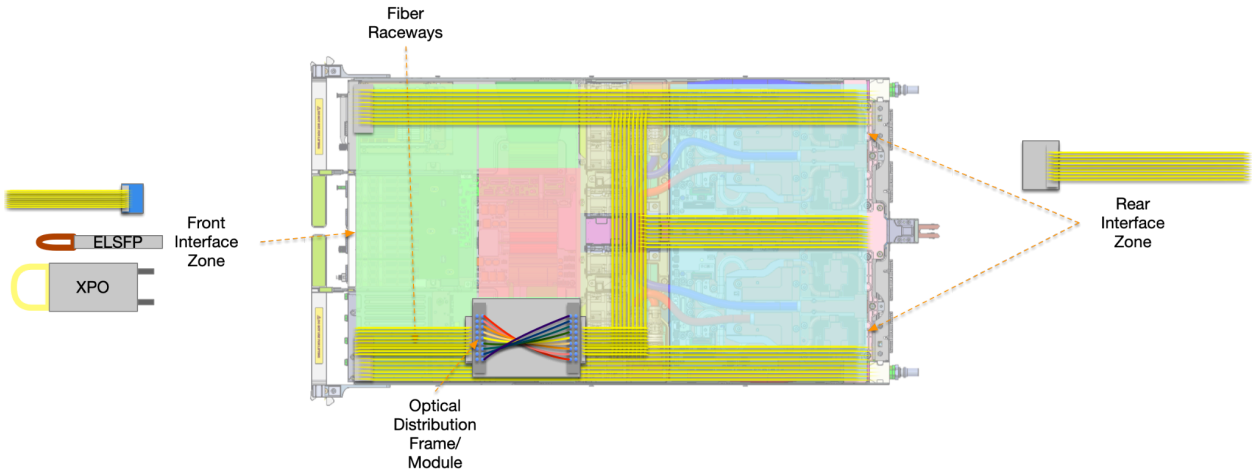


***Figure 5.18**. Passive optical infrastructure components — Optical Faceplate, ODF, Fiber Raceway, Rear Interface Zone, and Front-Panel Module Thermal Envelope.*

Figure 5.17 illustrates the passive optical infrastructure family and the optional optical-services population. The Optical Faceplate provides the front connector boundary; the ODF / Optical Services Zone provides shuffle, patch, module, and optional ELSFP / OCS population; the Fiber Raceway provides protected internal routing; and the Rear Interface Zone provides optional blind-mate rear egress. The Front-Panel Module Thermal Envelope reserves the cooling and service space required for active front-panel optical-service modules such as ELSFP, XPO, or future OCS implementations.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

5.10.1. Optical Faceplate

The Optical Faceplate is a modular front-panel connector interface that occupies the mechanical envelope currently used by pluggable transceiver cages (e.g., OSFP or QSFP-DD bays) in MHS-class trays. In CPO or NPO systems, the pluggable transceiver function moves closer to the ASIC package or board-level OE, freeing this front-panel pluggable volume for passive fiber connectivity, optical distributions, and service access.

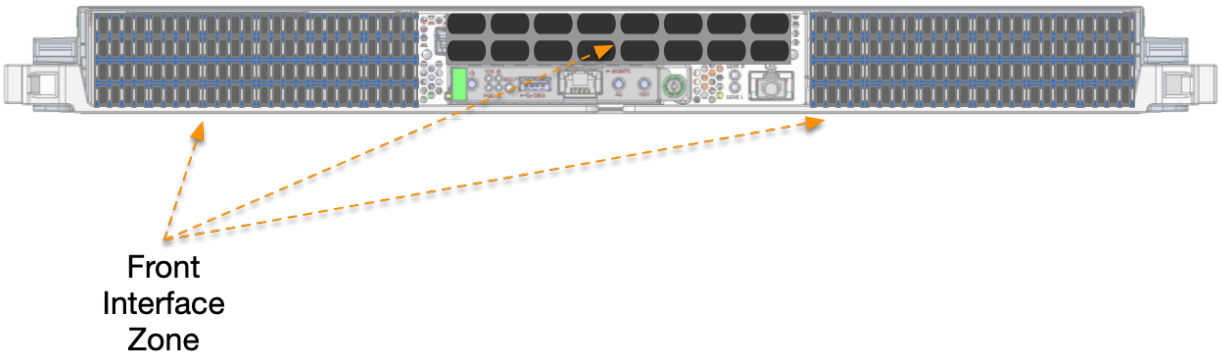


Figure 5.19. *Example of Modular and Configurable Optical Faceplate for Compute Tray showing ELSFP and VSFF (Very Small Form Factor) Connectors.*

The optical faceplate is a passive mechanical panel organized as a stable grid of zone envelopes rather than as a deployment-specific connector layout. Each zone has clearly bounded mechanical datums, adapter-envelope limits, retention-force limits, extraction clearance, labeling space, cleaning and inspection access, and bend-radius egress. Within that published envelope, different connector module populations can be qualified without redesigning the tray faceplate.

A zone may be populated at integration time with high-density multi-fiber adapters for scale-up links, standard duplex adapters for scale-out fabric, or management connectors for service access. Candidate connector populations may include MPO/MTP, MMC, SN-MT, Expanded Beam (EB) Optical, or future high-density adapter technologies, subject to the adapter contract for that zone.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

This is the envelope-and-contract methodology (§ Envelope and Contract) applied to the front-panel layout: the architecture fixes the zone envelope and interface constraints, while the integrator selects the connector technology appropriate to the deployment. Connector counts per zone are typically dimensioned against the worst-case fibers-per-tray-face derived in Appendix A, with normative connector counts defined in the forthcoming derivative Base Specification(s). The faceplate should not assume upper-layer protocol semantics; those are defined by the Link Profile and active optical-engine implementation. The passive faceplate contract instead defines the physical, optical, service, and telemetry boundary that makes those profiles deployable in an open platform.

The faceplate and the rear interface bound the tray depth, and between them sit five functional zones: Front IO, Mid IO, LC Services, Compute, and Rear Interface (Figure 5.19). The grid names what each depth zone is for; the placement axis names where optical modules sit. A position therefore carries a zone-and-slot coordinate: optical modules occupy the I/O zones, while the Compute and LC Services zones host none.

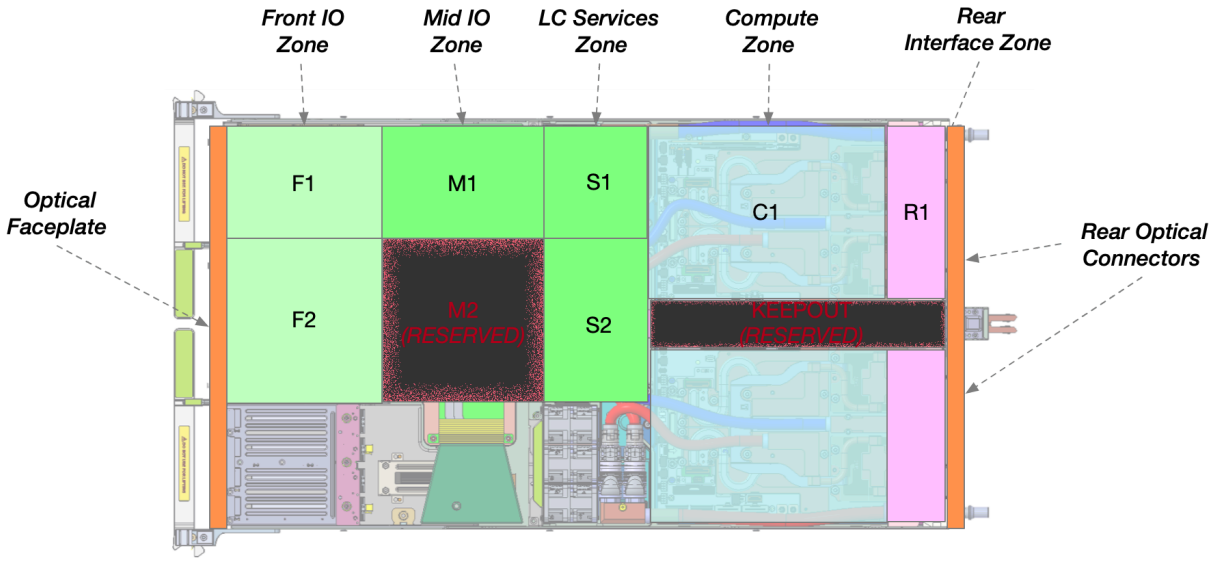


Figure 5.20. *Tray Depth Zones and Placement Coordinates.*

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

The optical faceplate and rear interface bound five functional zones along the tray depth, each position named by zone and lateral slot, with optical modules placed only in the I/O zones.

Module placement is an axis distinct from egress direction, and the faceplate sets its origin. The faceplate is the scarce front shoreline, reserved for the scale-up traffic that must reach the front to egress; the tray depth behind it carries the rest. Three serviceable positions follow: a front position at the faceplate, a mid position behind it within a long bay and withdrawn from the front for service, and a rear position served from the rear through the Rear Interface Zone. Light sources and lower-rate scale-out traffic, including optional CPU fabric participation, take the mid and rear positions and leave the faceplate to scale-up. Table H summarizes the three positions and their service, handling, cooling, and occupant contracts.

The front position is the front-panel module class. The ideal vision of this is a self contained module that toollessly mounts at the faceplate, latches securely with a handle, is connectorized and is serviced from the front without tray or rack removal, handling the scale-up data that egresses forward. An air-cooled occupant stays in the faceplate cooler zone, while a liquid-cooled occupant draws on the front-panel module thermal envelope (§ Front-Panel Module Thermal Envelope). Its depth-side counterpart, the internal-mount module class, is defined with the mid and rear positions at the Rear Interface Zone (§ Rear Interface Zone).

Table H: Tray Placement Positions

Position	Location	Serviced from	Front handle	Cooling	Typical occupant
Front	At the faceplate	Front	Yes	Air, or faceplate thermal envelope	Scale-up data

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Position	Location	Serviced from	Front handle	Cooling	Typical occupant
Mid	Long-bay interior, behind the front	Front, by withdrawing the bay	No	Liquid, internal-mount	Light sources, internal data
Rear	Bay rear	Rear	No	Liquid, internal-mount	Light sources, CPU scale-out

5.10.2. Optical Distribution Frame (ODF / Fiber Shuffle)

The internal fiber management challenge is substantial: a 102.4T switch tray using conventional per-lane fiber routing would require over 1,000 single-mode signal fibers plus approximately 128 polarization-maintaining fibers for laser sources, all navigating complex paths with tight bend radii within the tray envelope. This fiber density is a first-order driver for the architecture's emphasis on WDM-based fiber-count reduction and structured internal fiber management via the Raceway specification.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

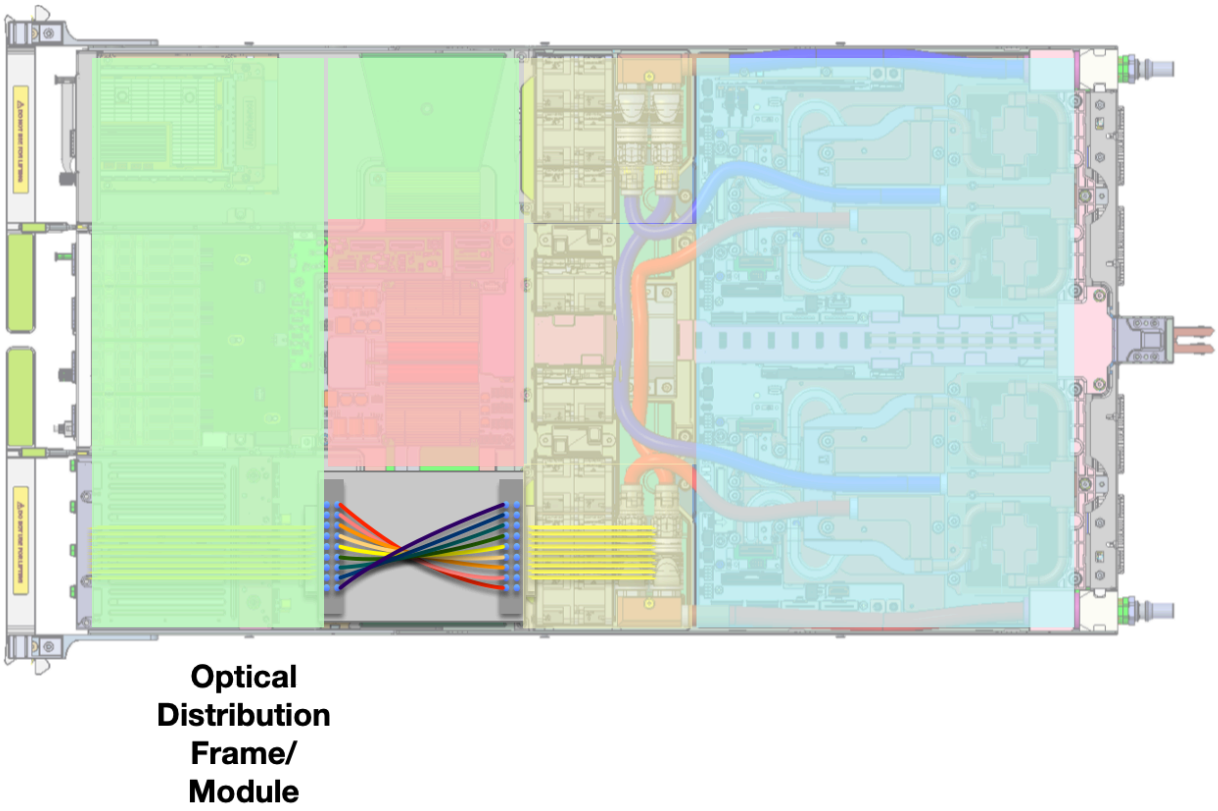


Figure 5.21. *Example of Optical Distribution Frame*

The ODF is a defined optical-services zone within the MHS bay, positioned behind the Optical Faceplate in the recovered OSFP and QSFP-DD transceiver bay area. Its baseline function is passive fiber routing: it maps front-panel port positions to internal tray destinations according to the network topology, performing the optical routing role that a copper midplane performs electrically.

This optical-services-zone treatment is consistent with OIF EEI/COI concerns around fiber shuffles, connector density, diagnostics, serviceability, and future scale-up optical topology options.

The ODF specification should define a normative interface contract and a recommended mechanical realization. The normative contract includes the zone envelope, mechanical datums, connector positions, fiber-count limits, ingress and

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

egress paths, minimum bend-radius constraints, labeling, cleaning and inspection access, test-access break points, retention and extraction limits, and any allocated loss-budget assumptions. Within that contract, different ODF implementations can be qualified without redesigning the tray or faceplate.

The preferred mechanical realization is a removable cassette or module that plugs into the optical-services zone. In the baseline configuration, this cassette contains passive fiber shuffles and connectorized patch points. This approach improves serviceability by allowing fiber routing assemblies to be installed, inspected, cleaned, replaced, or reconfigured as field-replaceable units rather than as fixed tray infrastructure. This is not the only realization possible.

Structured Optical Cabling Plane. At hyperscale fiber counts, the ODF should not be treated as a field-managed collection of individual jumpers. The Base Specification should define a Structured Optical Cabling Plane: a factory-terminated, keyed, labeled, and loss-qualified passive waveguide assembly that carries a defined fabric plane between tray OEs, the Optical Faceplate, the ODF/shuffle position, rear interface zones, and optional laser-source positions. The assembly may be implemented as flat ribbon fiber, bonded fiber sheet, molded fiber harness, planar polymer/glass waveguide, or another manufacturer-selected passive waveguide format, provided it satisfies the published envelope, bend-radius, insertion-loss, cleaning, inspection, and replacement contracts.

Laser-delivery and data-transport waveguides should be physically distinguishable and operationally segregated. The Base Specification should define color, mechanical keying, labeling, connector, and routing conventions that distinguish continuous-wave laser distribution from bidirectional data transport; color coding is useful, but service correctness and eye safety should not depend on color alone.

For production serviceability, the replaceable unit is the plane, cassette, cartridge, or bound harness, not the individual fiber. Individual fiber access may be retained for manufacturing, qualification, and lab debugging, but fleet operation should support

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

replacement of a qualified plane as a unit. This supports faster XPU or tray replacement, reduces field-touch labor and contamination risk, and lets hyperscale operators swap the failed optical plane without tracing thousands of individual links.

Because the ODF interfaces connectorized fiber, it should provide three capabilities that a copper midplane cannot:

- Fault isolation, since every link has a test-accessible break point at the ODF;
- Reconfigurability, since topology changes can be made by re-patching or replacing a shuffle cassette rather than replacing hardware;
- Front serviceability, since fiber connections are accessible without rear rack access.

The ODF also manages the transition between external cabling conventions and internal tray-level fiber routing. As wavelength density increases from single-wavelength DR to 4-wavelength FR4 to 16+ wavelength DWDM BiDi, the physical fiber count through the ODF decreases for the same aggregate bandwidth, making the ODF progressively simpler over time. The ODF per-tray fiber capacity is dimensioned against the fibers-per-tray-face envelope derived in Appendix A across the supported profile and path combinations.

The optical-services zone is the architectural primitive: a cage envelope in a module bay behind the Optical Faceplate. The ODF passive shuffle is its baseline population. The same envelope can host additional module classes to the same cage contract:

- **External Laser Source (ELS) modules** for centralized continuous-wave laser delivery to silicon photonic OEs that require an external CW source (SiPh-ELS class). ELSFP [18] is the reference form factor; the cage contract is form-factor-agnostic. When aggregate laser-source density exceeds tray-level zone capacity, ELS modules may populate a dedicated rack-level shelf to the same cage contract.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

- **eXtra-dense Pluggable Optics (XPO) modules** providing faceplate-pluggable optical format conversion (O→E→O) at rack/pod boundary aggregation positions, per the XPO MSA [19] (see § Placement Class Taxonomy).
- **Inline passive shuffle modules** for fan-in/fan-out, polarization-maintaining fiber routing, or topology-specific patching alongside or in place of the baseline shuffle cassette.
- **Optical Circuit Switch (OCS) modules** for software-controlled topology reconfiguration. The OCS class spectrum (robotic, MEMS, liquid crystal, piezoelectric, silicon photonic, metasurface) varies in switching time and radix; the zone envelope preserves compatibility without committing to any specific class.

The External Laser Source is a first-class, field-replaceable resource: any vendor may supply it, and any compatible engine may draw from it. Its contract follows the module model, with one defining change. Because the ELS carries no data signaling, the high-speed lane axis is replaced by optical-power class per wavelength. That class is the boundary that lets link, splitter, and aging budgets close across suppliers.

Mechanical, electrical power, thermal, and telemetry requirements remain aligned with the module contract. The ELS adds two source-specific boundaries: eye-safe exposure in every mated and unmated state, and distribution mapping that limits a source loss to one recovery domain. The architecture defines these boundaries; the Base Specification completes them.

OIF's ELSFP work standardizes front-panel placement (historically the cooler, air-cooled zone; § Front-Panel Module Thermal Envelope contracts active cooling here as ELSFP power and NIC-class density rise), hot-swap service model, and a blind-mate optical connector strategy that mitigates eye-safety risks at elevated optical powers. Power classing should be compatible with ELSFP optical power classes (spanning low-power

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

through very-high-power regimes) so that system designers can budget link loss, splitting loss, and aging margin coherently [17].

Figure 5.21 shows one possible placement of an ELSFP-compatible cage within the Optical Services Zone. The figure is *illustrative*, not *normative*. The Base Specification should define the cage envelope, service clearance, optical blind-mate path, airflow or liquid-cooling boundary, power connector, management connector, and interlock behavior.

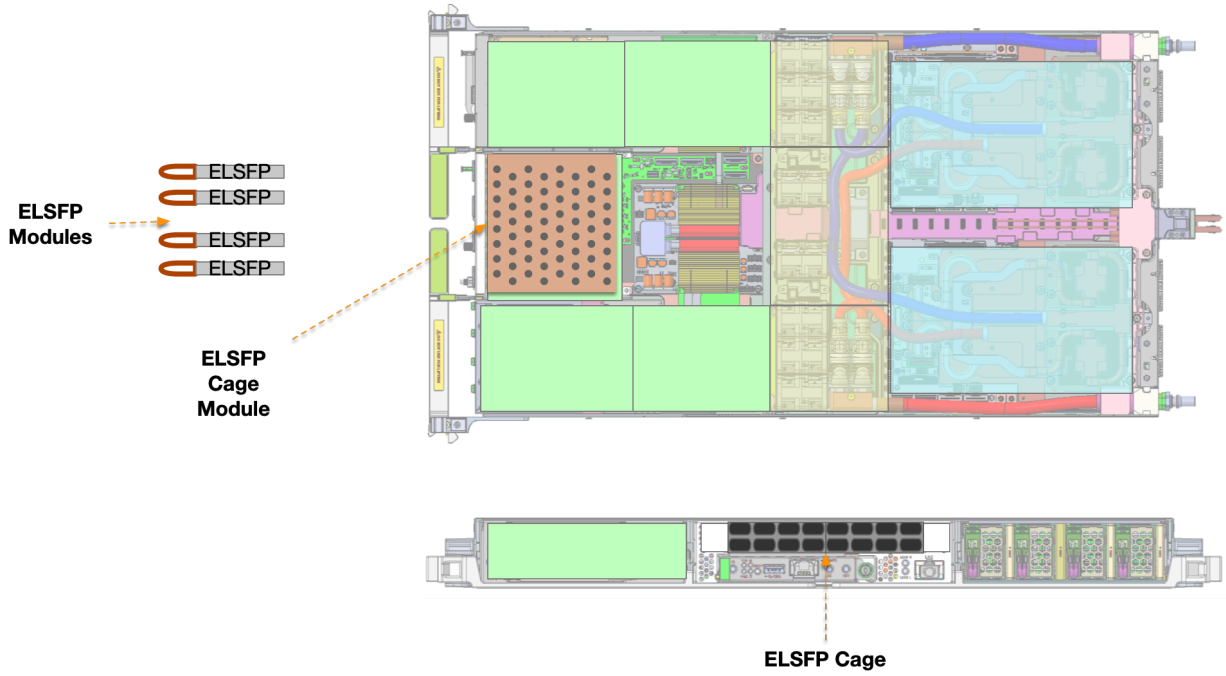


Figure 5.22. *Example of ELS Cage positioning in Compute Tray.*

Figure 5.22 illustrates a higher-power ELSFP-compatible cage where liquid cooling is required, which is increasingly likely given the trend towards Ultra High Power (UHP) laser requirements. The thermal design is implementation-specific, but the platform contract should reserve the mechanical envelope, service motion, thermal interface, and RAS-visible thermal telemetry required for high-power laser-source operation.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

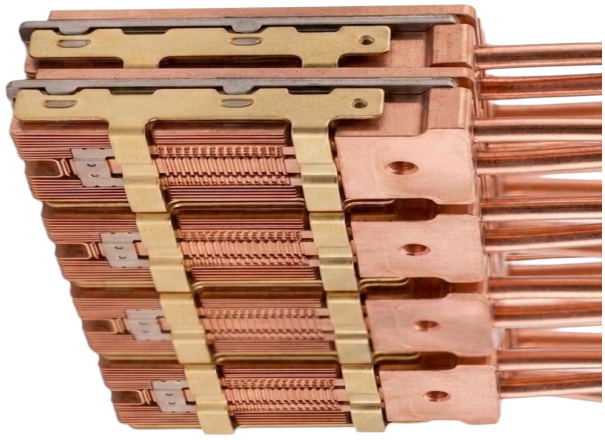


Figure 5.23. *Example of ELS Cage (Liquid Cooled)*

Laser distribution should be dimensioned by recovery domain rather than by mechanical convenience. The recovery domain is intended to bound the ‘blast radius’ of outage and is the smallest scheduler-visible or service-visible unit that the platform can isolate, degrade, or recover without treating the event as a larger resource outage. For a switch, the recovery domain may be the switch ASIC, rail-switch group, or fabric slice. For an accelerator, the recovery domain should be the XPU node or tray-level endpoint.

The governing rule is confinement with striping. Confinement means that a single ELSFP optical power resource should feed only one recovery domain wherever physically and economically feasible. It should not span independent scheduler-visible resources unless the Base Specification explicitly declares the exception and quantifies the residual blast radius. Striping means that, within a recovery domain, optical power from multiple ELSFPs should be distributed across OEs, lanes, or wavelength groups so that loss of one ELSFP causes a bounded bandwidth reduction rather than complete loss of the domain. The failure-impact target should be expressed as:

$$\frac{BW(lost,\ one\ ELSFP)}{BW(recovery\ domain)} \leq \alpha$$

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

where BW(lost, one ELSFP) is the bandwidth removed by one ELSFP failure, BW(recovery domain) is the total bandwidth of the recovery domain, and α is a platform-defined service target. For example, a four-way striped laser plan targets approximately 25% bandwidth loss for one ELSFP failure before higher-level redundancy, failover, or workload rescheduling is applied. This makes the blast radius visible to the scheduler and service tooling. For each wavelength supplied by an ELSFP, the optical source budget should satisfy:

$$P(ELSFP, \lambda \geq P(OE_{required}, \lambda) + L_{distribution} + L_{split} + L_{connector} + M_{aging} + M_{service}$$

where L_split includes the ideal splitter penalty, for a 1 split plus splitter excess loss; L_connector includes connector, blind-mate, and distribution losses; and M_aging and M_service cover laser aging, contamination, repair variation, and operational margin.

For switch-domain mapping, ELSFP outputs should be distributed as uniform stripes across OEs so that a source failure appears as a controlled aggregate bandwidth reduction rather than a partition of the switch fabric. For XPU-domain mapping, ELSFP outputs should be confined to one XPU node or tray-level endpoint so that a source failure degrades or isolates that endpoint only, allowing the scheduler to reschedule work at a boundary it already understands.

Confinement cannot always be perfect. As a tier approaches its throughput ceiling, integer packing of ELSFP sources into recovery domains may leave no whole spare to hold, and limited sharing across resources may become unavoidable. In that regime, the rule relaxes from eliminating cross-domain sharing to minimizing it. Any implementation that allows one ELSFP to span multiple recovery domains should explicitly declare the exception, quantify the affected resources, and expose the residual blast radius to the management plane, scheduler, and service workflow.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

A confinement guarantee is only as strong as its observability. The optical RAS model should make each ELSFP-to-domain mapping and the state of the laser behind it visible, so a failure presents as a bounded, attributable event at its recovery domain rather than a masked degradation. The Base and Design Specifications fix the telemetry field set.

The confinement and striping contract binds the laser source as a class, not the ELSFP form factor. An integrated source (SiPh-IL) fails the same way an external one does: it dies, it takes bandwidth with it, and its blast radius must land inside one recovery domain. The same α target applies, with striping realized across integrated sources on a package rather than across ELSFP outputs, and with the recovery-domain boundary drawn at the OE or package FRU that a source failure condemns. What changes with placement is the repair action, not the failure accounting.

The structural difference is FRU coupling, and the design space is small. A mateable boundary sits between source and engine, around both, or nowhere. The first is the external-source pattern regardless of reach: the source holds its own repair quantum, and pairing it with a socketed OE yields a uniform FRU scheme with source, engine, and shuffle each replaceable at its own boundary and MTTR bounded by a swap. The second is the pluggable pattern regardless of placement: a socketed source-plus-engine unit is a pluggable at the logic level, whether it sits at the faceplate or beside the package, and the socket caps the repair quantum at the module boundary. Only the third, the fused source sharing its host's silicon, lets the quantum ride the placement ladder, reaching package scale at CPO, where design-time redundancy, sources striped across an on-die distribution network so that a source death is a bounded derate, is the only lever left. Three repair patterns exhaust the space: swap the source, swap the unit, or provision redundancy. A design mixing source classes should state, per placement class, which pattern pays the α target and what quantum a failure condemns.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

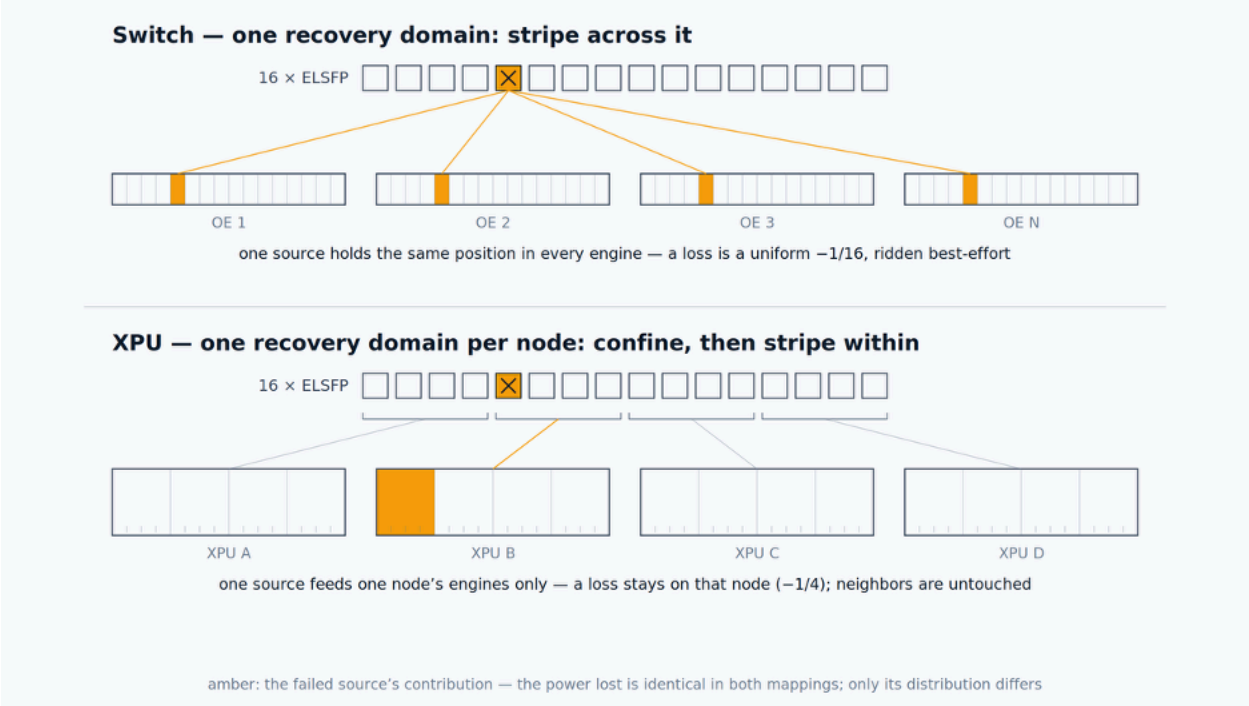


Figure 5.24. *Laser distribution and recovery domains.*

A switch spreads a single source's loss uniformly across the whole device; an XPU confines it to one node.

This is not a midplane. The ODF / Optical Services Zone does not define a fixed electrical topology, does not impose a generation-specific copper channel, and does not create a hidden single point of failure equivalent to a copper midplane. It is a serviceable optical mapping layer with explicit topology, failure boundary, and service behavior, which makes it easier to audit and qualify across vendors than a copper midplane.

5.10.3. Fiber Raceway

The Fiber Raceway is a passive mechanical channel built into the tray chassis or canister structure that routes Structured Optical Cabling Planes, bound fiber harnesses, or fiber cables from tray OE positions to the ODF / Optical Services Zone at the front. In a vertical canister architecture, the raceway provides a guided 90-degree turn from the

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

vertical tray slot into a horizontal run toward the front-mounted ODF. Where the raceway carries structured planes rather than loose fibers, the contract should define plane keep-outs, bend radii, retention features, extraction direction, and mate/de-mate clearance so an XPU tray or optical plane can be serviced without disturbing adjacent planes.

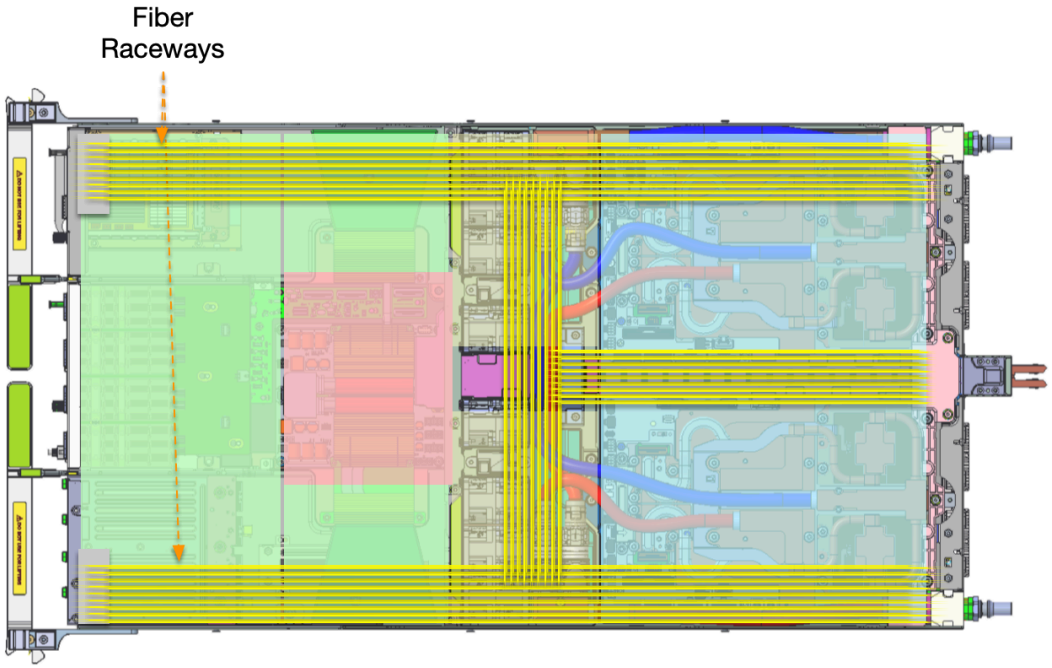


Figure 5.25. *Example of Optical Distribution Frame showing main fiber routes.*

The raceway is the simplest component in the architecture: it is sheet metal or molded channel with defined bend radius minimums, cross-section dimensions, and designed clearances to avoid interference with liquid cooling loops. It routes any fiber type (single-mode, multimode) and any fiber count[13]. It does not change with SerDes generation, OE technology, or network topology. It is structural infrastructure. The capacity ceiling derives from the slow-wide upper bound of Appendix A; the cross-section is dimensioned to accommodate that ceiling with a routing margin defined in the Base Specification.

[13] Raceway should support a defined maximum fiber population per tray class (to be derived from Link Profiles) and reserve fiber corridors that do not interfere with liquid QDs/manifolds.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

The raceway inside the tray also integrates with a rack level raceway. This is a first-order benefit: the highest-cost, lowest-yield, most generation-specific component in a dense rack architecture is replaced by commodity sheet metal and passive fiber cable. It also improves serviceability by giving optical cables a defined front-to-rear path that avoids unnecessary interaction with the interior tray and can reduce reliance on rear cable-cartridge routing.

The raceway should be generation-invariant. It should not require redesign when OEs evolve from 224 Gb/s/lane to 448 Gb/s/lane and beyond, provided the supported Link Profiles remain within the published fiber-count, bend-radius, and service-motion envelope. Its capacity should be dimensioned against the worst-case supported Link Profile and routing topology, with margin for manufacturing tolerance, service motion, and future connector or fiber variants.

5.10.4. Rear Interface Zone (Optional)

The architecture supports both front-access egress (the default service model, per Goal 7) and **rear-access egress via blind-mate optical connectors**. Rear access is intended to support direct service of the interconnect plane and associated field-replaceable elements. By replacing large free-routed high-speed cable and cartridge assemblies with an organized planar structure, the architecture creates clearer maintenance zones and reduces service complexity at the rear of the rack.

The rear interface zone is an architecturally first-class egress path, allowing front-only, rear-only, and split egress (e.g., scale-up rear / scale-out front) configurations as deliberate integration choices that the architecture publishes as supported configurations. The rear zone applies the same zone-modularity principle as the front faceplate: the architecture publishes the rear envelope and the blind-mate contract; integrators populate the zone with the connector technology appropriate to their

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

deployment. Expanded Beam Optical (EB) connectors are the representative blind-mate technology, well-suited to the alignment-tolerance and particulate-tolerance characteristics of blind-mate operation, though the architecture is connector-agnostic at this contract layer.



Figure 5.26. *Example of Expanded Beam Optical Connectors.*[14]

Regarding the discussion on Expanded Beam optical connectors as shown in Figure 5.25, there has been a recent formation of the Expanded Beam Optics Multi-Source Agreement (EBO MSA)[15] which came up long after this whitepaper was started. It reinforces a physical-layer trend: AI data-center operators are seeking open, multi-vendor connector specifications that reduce fiber-handling complexity and improve Layer-1 deployment resilience. Our recommendation for Expanded Beam is based on the perceived operational advantages for end users in data centers. Other types of connectors can and will be integrated in systems. As expanded beam matures, it is poised to become a market dominant solution.

Figure 5.26 illustrates how a compute tray or tray may expose front optical access, rear optical access, or a split-egress configuration. The figure is *illustrative*, not *normative*. A

[14]https://www.molex.com/en-us/products/optical-solutions/fiber-optic-connectors-and-adapters/expanded-beam-connectors
[15] https://ebomsa.org/

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

compliant implementation may choose front-only, rear-only, or split egress if it satisfies the same Link Profile, optical budget, serviceability, and RAS contracts.

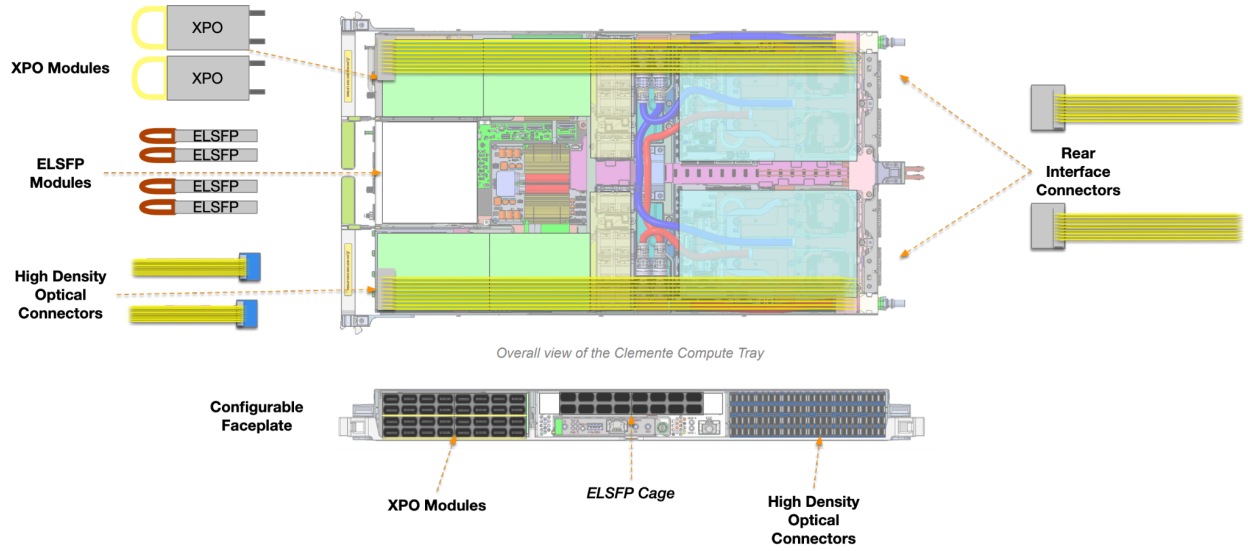


Figure 5.27. *Example of Front and Rear I/O Access in Compute Tray.*

The Rear Interface Zone completes the passive optical infrastructure by providing an optional rear optical boundary for deployments that benefit from blind-mate serviceability, rear cartridge insertion, rear ODF placement, or separation of scale-up and scale-out egress paths. The architecture does not require rear egress, nor does it prescribe a specific connector family or routing topology. Instead, it defines the mechanical, optical, serviceability, and telemetry contracts that allow rear-access optical infrastructure to be integrated consistently across tray classes and future optical-engine generations. The Rear Interface Zone supports several deployment models:

- **Direct cabling** point-to-point to a destination tray's rear connector. Suited to peer-to-peer scale-up topologies and minimum-infrastructure deployments.

- **Harness to ODF or Shuffle**, where rear-egress fiber runs to a separate or shared fiber distribution frame for topology-defined routing. The shuffle position can be top-of-rack, mid-rack, or co-located with the front-egress ODF.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

- **Cartridge assembly**, where a pre-assembled fiber cartridge mates blind to a group of trays simultaneously, presenting a factory-assembled scale-up plane that is swappable as a unit.

The cartridge mode warrants a specific note: it inherits *none* of the SI, yield, or generation-coupling pathologies of today's rear copper cable cartridges and custom midplane PCBs. It is passive single-mode glass with blind-mate connectors, generation-invariant, low-yield-risk, and substantially simpler in mechanical and electrical terms than the copper assemblies it can replace.

The same property collapses assembly labor. A cabled scale-up generation presents thousands of discrete links per rack, each routed, dressed, and tested by hand, at hour-scale per-tray assembly. A bound plane presents one keyed blind-mate, landing per-tray assembly in the minutes regime that recent cableless platforms report. Cableless copper midplanes reach that regime too, but only by folding the interconnect into the multi-layer laminate stack the optical flyover layer retires. The bound optical plane captures the labor collapse and keeps the fiber plant generation-invariant. The minutes-scale figure is vendor-disclosed for a cableless-PCB realization. The bound-plane figure is argued by analogy here, stated at analytical confidence pending a cartridge mate-time measurement and a pilot-volume defect-rate study.

The OIF EEI COI Whitepaper §6 identifies blind-mate optical connectivity as a requirement for dense scale-up rack architectures, citing fiber count (4,000–8,000 fiber link pairs per rack at server-tray density), reliability, and contamination tolerance as the driving considerations. The rear interface zone provides the architectural reservation under which any compliant blind-mate technology can be deployed.

The architecture specifies the **interface contract**: position envelope within the chassis, keep-outs against ORv3 rear power bus bars and liquid-cooling blind-mate features, fiber-count capacity targets, and alignment tolerances appropriate to blind-mate operation. It does not specify the connector. Implementations may select connector

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

technology appropriate to their density, cleanliness, and mate-cycle requirements; expanded-beam optical (EB) connectors are one execution path that matches the alignment-tolerance and particulate-tolerance characteristics of blind-mate operation, but the architecture is connector-agnostic.

Capacity guidance: Reserve 3–4 connector areas within the rear zone, with per-area fiber capacity sufficient to carry full scale-up egress for the highest-density tray class. Exact dimensions, connector counts, and fiber-count-per-area targets will be defined in the Base Specification.

Fabric-plane partitioning. A rear zone with four connector areas, for example, may allocate two connector areas to Fabric A and two to Fabric B, or split capacity among scale-up, scale-out, and Scale-DSA planes. Current EB-class blind-mate connector envelopes reaching roughly 144 fibers per connector make this split meaningful at the tray scale. The Base Specification should therefore define connector-area capacity and fabric-plane labeling independently: connector geometry is the mechanical contract; fabric assignment is a topology contract.

Internal-mount module class. A module in a mid or rear position (§ Optical Faceplate) needs no faceplate, latch, or front handle; removing them returns the carrier to a bare internal module and recovers the front-service overhead. This defines an internal-mount module class as the depth-side counterpart to the front-panel module class: one liquid-cooled envelope, served from the rack interior, carrying a light-source, data, or photonic-engine function. The class depends on liquid cooling. An air-cooled module remains bound to the faceplate cooler zone, while a liquid-cooled module carries its own cold plate and rides the depth.

Module rear contract. Each module presents one invariant rear contract to its carrier: a blind-mate liquid disconnect, a power-and-management card edge per SFF-TA-1002, and a high-speed escape path sized to its medium rather than a fixed connector. The card edge is a 2C interface per module. A long bay holding a front and a mid position

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

aggregates the two into a single 4C host connector at the bay rear, so each module presents one 2C face and each long bay presents one 4C face; short or single-position bays under-fill the 4C. Power aggregation is a published threshold rather than a fixed rule: the 4C carries both positions while their combined draw stays within its power-contact rating, and above that the binding position takes a dedicated supply feed while the 4C reverts to signal and management. The integrator selects the branch from the deployment's power class.

5.10.5. Front-Panel Module Thermal Envelope (Optional)

The front interface zone once belonged to the cold aisle, holding low-power pluggables in the cool intake. That no longer holds. ELSFP modules, faster NICs, and XPO raise both power and connector density at the faceplate. Rising tray density narrows the airflow that once cooled it. The industry is already moving its high-power pluggables and NICs onto liquid. The faceplate has been looped in.

The architecture defines a thermal envelope at the faceplate serving any actively-cooled front-panel module. Motivating classes include faceplate-mounted ELSFP laser sources, XPO format converters at the rack boundary, linear-driven optical NICs, and the front-panel classes still to come. It is the front-panel counterpart to the Rear Interface Zone, the same envelope-and-contract methodology turned to a different surface.

A coolant manifold is reserved in the tray volume between the HPM rear edge and the faceplate, at the position air-cooled trays use for fans. Direct liquid cooling recovers that volume, and a branch off the tray's main loop carries coolant forward to the module positions on short runs.

The five contract dimensions are those of § Envelope and Contract, with power carried as a small set of discrete module-class bands parallel to the OIF ELSFP taxonomy, so rack budgeting stays bounded. The optical thermal supply is for the light source, drawn

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

through a front-zone liquid manifold with a hot-swap quick-disconnect and leak detection reported through the management plane. Three things are specific to this surface. The mechanical envelope clears the faceplate connector grid of § Optical Faceplate. The serviceability invariant keeps the service path and the thermal path from intersecting. Field replacement opens no fluid path and exposes no energized optical path, so the module stays dry and the chassis keeps the wet side of every coupling. For laser-bearing classes, an interlock ties emission to both fluid coupling and insertion state, per OIF ELSFP eye-safety provisions and applicable IEC standards.

5.10.6. Rear Thermal Envelope

Removing copper's mass and consolidating high-speed interconnect into a planar structure clears the critical airflow regions and improves the predictability of thermal design as rack density increases.

Rear tray and rack zones should preserve clear liquid-cooling ingress and egress paths, including access to blind-mate cooling interfaces and service clearances.

5.10.7. Photonics Thermal Supply

Photonic components carry thermal requirements that differ qualitatively from XPU silicon: laser wavelength, modulator bias, and photodetector responsivity all drift with temperature in ways that directly hit link budget and BER. DWDM-class engines in particular need stability beyond what XPU cold-plate loops are designed to deliver, since those loops are sized for thermal mass and dissipation density, not wavelength anchoring.

The architecture therefore calls for a photonic thermal supply path tailored at the CDU rather than sharing setpoint and flow with the compute loop. Realization of a dedicated loop, a conditioned branch off the main CDU, or a heat-exchanger-based pre-conditioner is a rack and CDU integration decision consistent with OCP Cooling

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Environment specifications. The architecture publishes what the supply is to deliver, not how it is built.

Supply contract dimensions:

- **Inlet temperature:** tailored to photonic reliability and wavelength stability; typically lower than the compute-loop setpoint.
- **Stability:** approximately ±1 °C at the supply manifold, tighter where the deployed DWDM grid spacing or laser regulation headroom warrant.
- **Flow rate:** matched to photonic dissipation per tray and tray, typically much lower than the compute loop.
- **Telemetry:** supply temperature, return temperature, and flow rate reported via the OCP RAS API endpoint, per the management-and-telemetry dimension in § Envelope and Contract.

The contract applies wherever photonics is co-located with the rack: OE cold plates across all placement classes, the External Laser Source shelf, and the Front-Panel Module Thermal Envelope. Treating photonic thermal supply as one rack-level concern with multiple consumption points avoids fragmenting the thermal design into per-consumer solutions. Coupling-class consistency follows the same logic: photonic FRUs participating in the liquid loop should use the rack's DLC quick-disconnect class, with vendor choice operating within that class rather than across classes.

6. Optical Engine Selection and Economics

The link types within an AI compute system differ in reach, BW density, power, and cost requirements depending on network graph. For example, a CPU to GPU link vs a GPU to memory link will compare on these metrics very differently. A central principle of this architecture is that no single photonic technology serves all link classes optimally. The platform should accommodate multiple photonic technologies simultaneously, allowing tray designers to select the appropriate OE for each link class.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

This section describes the landscape of photonic technologies and how they map to system link classes. They are not ranked or preferred; they are options within the degrees of freedom the architecture provides.

6.1. Channel Operating Margin and Placement Selection

An ideal photonic AI infrastructure derives its integration choices from a channel evaluation function applied deliberately to the parameter space, rather than from inherited assumptions or the path of least resistance. The efficiency and serviceability tradeoffs are made and operational data informs the suitability. This subsection establishes that selection framework. The placement class taxonomy in the preceding subsection enumerates what the integration positions are; this section establishes the evaluation basis among them. This basis includes a channel evaluation function that maps (line rate, cable, retiming model, and host PCB) onto the set of viable placement classes, and the tradeoff curve that determines which viable class is preferable for a given system.

COM as the selection function. The architecture treats Channel Operating Margin (COM) and equivalent metrics defined by IEEE 802.3 and the OIF CEI implementation agreements as the canonical channel evaluation function. At 224 Gb/s and above, reflections become co-equal with insertion losses as a determinant of channel viability, and COM-class metrics that integrate both into a single figure of merit are the appropriate evaluation basis. The architecture is not normative about the specific COM implementation (IEEE 802.3 COM, OIF CEI ERL/IL/RL, vendor-equivalent metrics); it is normative about the principle that placement-class selection is grounded in a channel evaluation function applied to the published parameter space. This framework adapts the COM construction for placement-class selection by treating the OE-internal channel and any flyover, socket, or substrate transition as defined channel segments with their own IL and noise contributions.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

How COM selects placement classes. Each placement class presents a characteristic channel signature of channel length, connector count, dielectric class, and retiming model and therefore has a characteristic COM signature. APO presents the shortest channel with no connector discontinuity and the lowest reflection signature; CPO can be from less than 1mm to very few mm with no or very few intermediate impairments. NPO adds the package-to-host transition; S-NPO adds the socket discontinuity; M-NPO adds the mezzanine edge interface; FPO adds the full host PCB length plus the faceplate cage; XPO, as an FPO subset, adds the rack-boundary format-conversion stage with its double-retimer penalty on top of the FPO channel. For a given line rate, the COM threshold determines which placement classes close the channel budget and which do not. Closing the budget is the necessary condition for viability, not the sufficient condition for selection.

The efficiency hierarchy. Among viable placement classes, the architecture observes a consistent efficiency hierarchy: the more direct the channel, the lower the effective energy-per-bit and the lower the structural latency. APO and CPO occupy the high-efficiency end with no connector and no retimer (assuming no gearboxing, it is the direct-linear floor for the architecture); each subsequent placement class adds connector loss, retiming power, or both, with FPO occupying the lower-efficiency end and XPO representing the highest-penalty FPO variant when optical format conversion is required. The hierarchy is not a strict ranking, as system-level FRU economics, vendor-ecosystem participation, and serviceability considerations may favor a less efficient class for legitimate reasons. The underlying physics, however, is uniform across the spectrum.

BER margin and goodput economics. A second consideration sits alongside the efficiency hierarchy: bit error rate (BER) margin translates directly into goodput. AI training workloads are exquisitely sensitive to silent data corruption, Timing skew, Not a Number (NaN) propagation, and step-replay/checkpoint-reload events caused by interconnect errors that escape FEC. The OCP RAS effort has explicitly targeted these

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

failure modes. The Server Component Resilience SDC Specification (AMD, ARM, Google, Intel, Meta, Microsoft, NVIDIA), the GPU & Accelerator RAS Requirements v1.7, and the SDC in AI whitepaper (NVIDIA) converge on one position: interconnect-driven goodput loss is a first-order economic concern. A placement class that provides more BER margin enables tighter FEC overhead, reduces the rate of error-driven workload disruption, and improves the effective goodput per nominal link BW.

The integrating metric is effective pJ/bit: the energy consumed per delivered bit divided by the effective goodput. A placement class with somewhat higher nominal pJ/bit but substantially better BER margin can deliver lower effective pJ/bit because it loses less of its nominal BW to error-driven goodput loss. The architecture treats effective pJ/bit as the canonical efficiency metric because it integrates the direct energy cost with the indirect cost of interconnect-induced workload disruption. The OIF CMIS and OCP Hardware Management RAS API v0.9 (referenced in § Envelope and Contract) provides the run-time observability layer that makes effective pJ/bit measurable in deployment, closing the loop between design-time COM analysis and operational data.

The FRU spectrum. Parallel to the efficiency hierarchy is the FRU spectrum: APO is at the same level as an ASIC or XPU (thus, for example, if the GPU FIT rate is higher than the APO OE, the FRU granularity may become more important.) may while CPO and NPO can be FRUs at the package level. S-NPO can be a FRU as a socketed-device such as with openCPX; M-NPO can be a FRU at the mezzanine subsystem level (and more granularly if mezzanine carries a socket); FPO including XPO is a FRU at the module level. The FRU level determines the field-service economics, the failure blast radius, and the operational labor profile of the system. The architecture does not prescribe a single FRU recommendation because different system economics

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

legitimately favor different FRU levels.

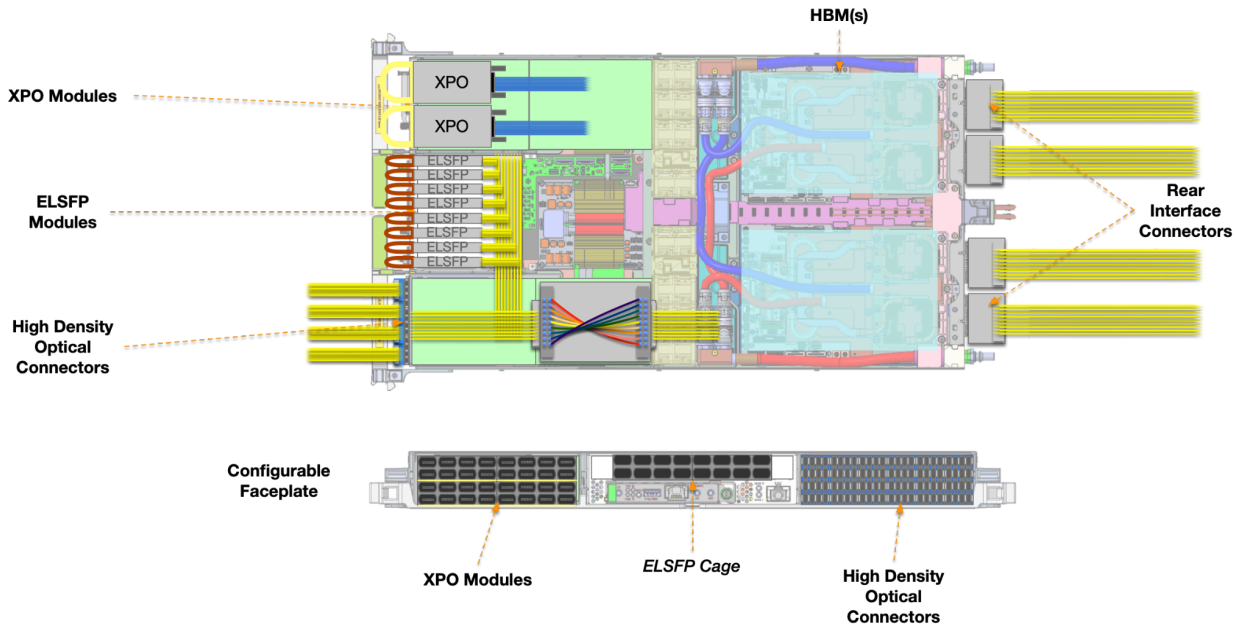


Figure 6.1. *Example ingress and egress I/O FRU Placement.*

The FIT-rate consideration. Conventional wisdom has historically held that CPO-class implementations present higher operational risk than pluggable optics because the failure of an integrated OE has larger system consequences. Interestingly the empirical record argues against this conventional wisdom: hyperscaler-published reliability data for CPO-class deployments shows MTBF on the order of 2.6M+ hours, compared to ~0.55M hours for pluggable optics (a roughly 5× advantage; see the Target Outcomes table). The architecture is informed by this empirical inversion. Additionally, Field Replaceable Unit (FRU) strategy provides a way to adjust the impact of FIT, in multiple ways, for example by reducing the Mean Time to Restoration (MTTR), and increasing availability with controlled degradation without failure. The placement class selection should accordingly be grounded in measured FIT rates rather than inherited assumptions.

Two structural consequences for the architecture. First, far fewer placement classes close the channel budget at 448 Gb/s than at 224 Gb/s. The integration boundary

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

therefore migrates inward as channel margin tightens, and the architecture's center of gravity moves toward the more-integrated end of the placement spectrum as supported line rates rise. Second, the efficiency-and-FRU economic case is not statically optimized against a single placement class: different system configurations legitimately favor different classes while keeping the framework for evaluating the tradeoff uniform across configurations.

The link-budget lever framing. The proposed system architecture exposes four independent levers against the optical link budget: link architecture which may include gearboxing (the OCI MSA [9] targets 53G NRZ lanes rather than 224G PAM4); the OE-internal design (≈5–7 dB span across publicly characterized PAM4 modulator and coupling ranges); topology selection (2–3 dB span across the six topology classes in § Optical Link Composition); and the tolerance envelope (3 dB worst-case planning value, recoverable to ≈1.5 dB under qualified-deployment terms). Link architecture is the first-order lever: it fixes the required budget the other three work against. Those three are margin levers. They compose multiplicatively (in the linear regime and additively in the logarithmic regime such as dB) when an integrator coordinates them, giving 8–11 dB of qualification-coordinated swing, larger than any single margin lever's range. Each lever sits at a distinct layer: link architecture is a specification choice; OE-internal is a device choice (vendor-determined within the OE design); topology is a system choice (integrator-determined per § Optical Link Composition and the placement-class taxonomy above); envelope is a qualification choice (deployment-context-determined). The four are set independently, which is the structural reason § Optical Link Composition commits to a budget framework rather than to specific dB allocations. Appendix C develops each lever quantitatively.

This subsection establishes COM as the canonical channel evaluation function for the architecture and the efficiency-vs-FRU tradeoff curve as the framework for selecting among viable placement classes. Subsequent sections reference COM and equivalent metrics as system-level requirements, and reference the tradeoff curve when discussing

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

why specific placement classes fit specific system contexts. By grounding placement selection in COM-driven analysis with effective pJ/bit as the integrating metric, and by replacing perceived-risk-based serviceability arguments with operational FIT-rate data, the architecture transforms placement-class selection from convention into transparent, data-grounded engineering — the discipline an ideal photonic AI infrastructure requires from its design layer.

Table I: 224 Gb/s PAM4 per lane

Placement class	Reach class	Retiming model	Channel & host parameters	Reach	Source / status
CPO	XSR	Unretimed	Substrate-internal, no connector in channel	≤ 50 mm	Vendor demonstrations
NPO	XSR / VSR	Unretimed or half-retimed	Short host PCB trace on Megtron-class dielectric	≤ 100 mm	OIF CEI-224G-XSR, vendor characterization
S-NPO	VSR	Half-retimed or retimed	Substrate-mounted socket (e.g., Samtec CPX, Molex Impress); 95×95 mm substrate envelope	~ 1 m on 32 AWG	Samtec, Molex public data
M-NPO	VSR / MR	Retimed	ASIC + OE on shared mezzanine carrier; characterized to mezzanine edge	~ 1 m intra-mezzanine	Merchant silicon vendor disclosures

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Placement class	Reach class	Retiming model	Channel & host parameters	Reach	Source / status
FPO	LR Flyover	Retimed	32 AWG twinax flyover from package to faceplate cage	~ 1 m	Meta empirical anchor (Amiralizadeh & Lowalekar, Chiplet Summit 2026)
XPO (FPO Subset)	LR Flyover + OFC	Retimed (double, OFC)	Faceplate-pluggable rack/pod-boundary OFC module; FPO channel plus format-conversion stage	~1 m intra-rack plus format conversion	XPO MSA / switching aggregation realizations

Note: loss-only reach; integrated-channel reach in Table J includes connector and COM-closure penalties.

Table J: 448 Gb/s PAM4 per lane (PAM4 baseline; see Forward-looking note on modulations beyond PAM4)

Placement class	Reach class	Retiming model	Channel & host parameters	Reach	Source / status
CPO	XSR	Unretimed	Substrate-internal, no connector in channel	≤ 25 mm	OIF CEI reach classes (224G published), vendor projections

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Placement class	Reach class	Retiming model	Channel & host parameters	Reach	Source / status
NPO	XSR / VSR	Half-retimed	Short host PCB trace on advanced dielectric (Megtron 9 or better)	≤ 50 mm	OIF CEI reach classes (224G published)
S-NPO	VSR	Retimed	Substrate-mounted socket; channel loss budget under active development	Output-power-limited	OIF CEI reach classes (224G published), vendor characterization in progress
M-NPO	MR	Retimed	ASIC + OE on shared mezzanine; advanced dielectric required	Output-power-limited	Merchant silicon vendor projections
FPO	LR Flyover	Retimed	28 AWG twinax flyover, advanced connectors	~ 0.254 m (28 AWG) to ~ 1 m (24 AWG, Megtron 9, aspirational)	Meta empirical (28 AWG); OIF aspirational target (24 AWG, advanced)

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Placement class	Reach class	Retiming model	Channel & host parameters	Reach	Source / status
XPO (FPO subset)	LR Flyover + OFC	Retimed (double, OFC)	Faceplate-pluggable rack/pod-boundary OFC module; FPO channel plus format-conversion stage	~same as FPO plus format conversion	XPO MSA / switching aggregation realizations

Forward-looking note on modulations beyond PAM4. OIF and the broader industry are actively exploring moving modulation formats beyond PAM4 for copper channels specifically. One potential path is PAM6 modulation at 448 Gb/s/lane, which would reduce the Nyquist frequency requirement from approximately 112 GHz for PAM4 (currently in specifications) to approximately 86.5 GHz by lowering the symbol rate while increasing modulation complexity. PAM8 cuts the Nyquist further, to roughly 75 GHz, but at a price: about 2.5 dB of additional vertical SNR penalty beyond PAM6, plus tighter transmitter linearity, higher DAC/ADC resolution, and more FEC margin. It earns its place only if PAM6 cannot deliver the needed reach or margin, and it carries higher implementation risk. If either format matures within the deployment window, the 448 Gb/s budget tables may need separate rows by modulation; the framework absorbs that without restructuring. One coupling matters most. If copper adopts PAM6 or PAM8 while the optical interface stays PAM4, the electrical-optical boundary needs translation, retiming, or gearbox/DSP — and that erodes the case for the linear, low-complexity OEs that depend on a modulation-transparent path. Modulation is therefore an optical-interface decision as much as a copper one, weighed against FEC partitioning, latency, and power.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

6.2. Effective pJ/bit and Effective pJ/token

For AI systems, interconnect efficiency should be reported both as pJ/bit at the link layer and as effective pJ/token or pJ/training-step at the workload layer. The latter captures retransmission, FEC overhead, congestion, topology-induced stalls, and recovery from link faults. This distinction prevents component-level energy improvements from being overinterpreted when the workload-level bottleneck lies elsewhere.

6.3. Heterogeneous Optical Engine Integration

A single compute tray may host multiple OE classes simultaneously. Consider a tray with four GPU packages and scale-out networking: the NVLink™-facing OEs might use dense BiDi DWDM (SiPh-ELS class) optimized for maximum BW density on the scale-up links, while the scale-out networking ports use simpler single-wavelength optics (VCSEL-DM or SiPh-ELS DR class) for interoperability with the existing switch ecosystem. Memory disaggregation links, if present, might use yet another OE variant optimized for the asymmetric traffic pattern of remote memory access.

All of these coexist on the same tray. All route through the same raceway. All terminate at the same ODF. All exit through the same faceplate. The passive infrastructure is OE-agnostic by design. This means OE vendors can specialize: one vendor might build the best dense DWDM engine, another the most cost-effective VCSEL array, and a third optimized coherent module set. They all target the same electrical SerDes interface on the ASIC side and the same fiber connector interface on the network side.

> *The architecture shifts optical modularity from the transceiver cage (where you plug in different OSFP/QSFP modules) to the tray design (where you select OE classes per link type) and the faceplate (where you populate connector adapters per deployment). The* OE *becomes part of the compute BOM, not the networking BOM.*

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

6.4. Adoption Sequencing

Technology diversity enables incremental adoption. A hyperscaler need not wait for a single photonic technology to cover all link classes before adopting any. A plausible sequence:

- Begin with proven single-wavelength (DR-class) OEs for scale-out fabric links, replacing pluggable transceivers with near-package optics. This is the lowest-risk entry point.
- Starting with DR or OCI for interoperability, increase to moderate-density multi-lambda BiDi or multi-wavelength optics for scale-up links within the rack, enabling larger scale-up domains. This phase delivers the primary MFU improvement.
- Push to high-density DWDM (16λ+ BiDi) for on-tray and inter-canister links as the SiPh ecosystem matures. This phase enables flat-BW dense racks.
- At the platform level, adoption diverges by orientation class. Class H primarily uses this stage to add independent fabric planes and reduce pluggable/copper dependencies around existing network trays. Class V can use the same optical infrastructure to choose between rear optical switch trays and distributed rail switches on deeper compute trays, depending on radix, serviceability, and thermal tradeoffs.
- Explore ultra-short-reach options (micro-LED, advanced VCSEL arrays) for optimization of package-level and die-to-die optical interconnect as these technologies mature.

Each phase adds capability to the same physical infrastructure. The raceway, ODF, and faceplate installed in Phase 1 continue to serve in Phase 3 without modification.

Beyond the component-level foundation. The architecture this whitepaper establishes is, by design, a component-level discipline with open interfaces, contract-based qualification, and vendor-neutral integration. All three are required to get

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

the silicon photonics AI ecosystem started at production scale. This is the right starting point because stable, abstracted interfaces are what enable multiple vendors and standards bodies to converge on interoperable infrastructure. The next generation of optimization will benefit from co-design between system interfaces and components: joint optimization across ASIC SerDes, OE, channel, and FEC that captures cross-component performance and operational opportunities the abstracted-interface approach does not reach. This next generation includes emerging ultra-dense vertical-tray compute architectures and designs deploying after this specification matures. The contract layer established here is the foundation co-design builds upon. Co-design refines the implementations behind the contracts rather than replacing them, and new contract layers emerge as the co-design learnings are codified.

Historically, OE selection has been driven as much by supply-chain logistics, integration, and vendor preference as by engineering. Open integration ends that. Signal integrity and economics now couple too tightly to design a system first and select a component second; they co-determine. That coupling tightens with scale.

7. Link Profile Framework

Rather than specifying a single optical link architecture, this vision proposes a framework of Link Profiles that describe optical link configurations as bandwidth-and-fiber-count contracts. Each profile defines what the system integrator sees at the fiber connector; how the OE achieves that contract is an implementation decision.

Two endpoints of a physical fiber link need to align on the same Link Profile. However, different links within the same system can use different profiles simultaneously.

7.1. Candidate Link Profiles

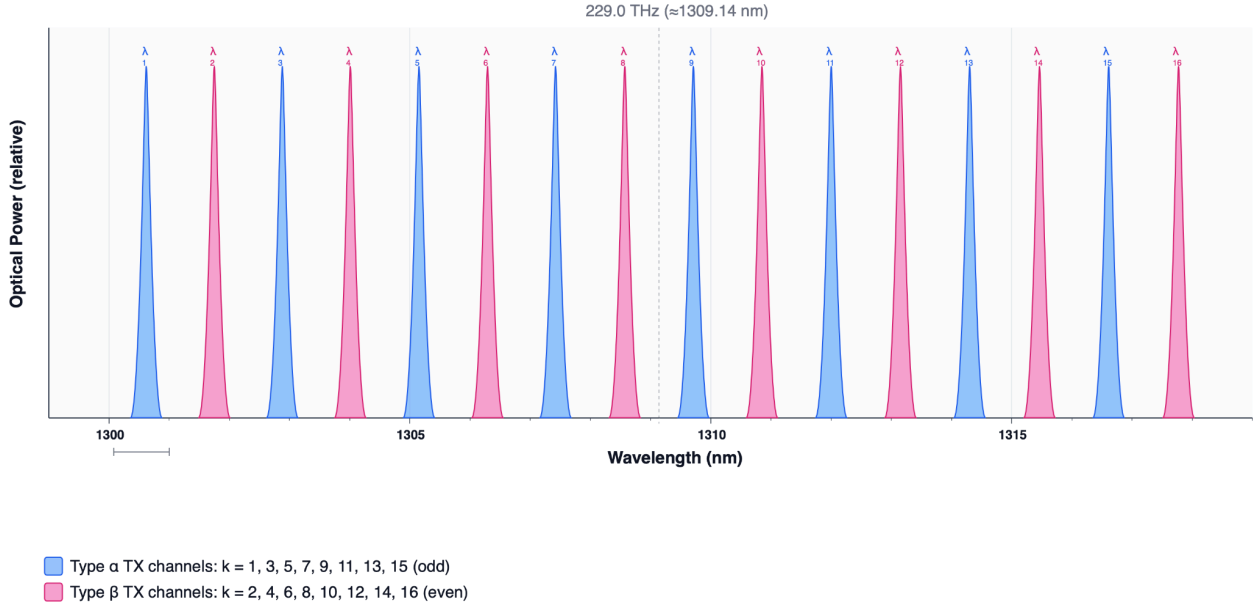
A Link Profile is the minimum interoperability contract between (a) a tray-integrated OE and (b) the platform fiber plant (ODF + faceplate) such that the platform can support

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

multi-vendor OEs without redefining the rack-level passive infrastructure each generation. Link Profiles are protocol-agnostic (they carry scale-up, scale-out, or Scale-DSA traffic without embedding protocol semantics) but lane-explicit: each profile specifies λ count and directionality so fiber counts, connector density, and link-budget margins can be engineered deterministically.

Scale-up HBD links are currently emerging at single-wavelength parallel using DR-class and OCI-class profiles, with single-fiber-per-lambda assignments and parallel multi-fiber connectorization. These profiles are sufficient at near-term HBD sizes but become fiber-count-limited at the 1024+ XPU HBD scales this architecture is sized for: at 200–400G per lambda, the connector and faceplate area required for single-lambda parallel exceeds what current mechanical specifications support before BW saturates.

The architectural response is to maximize *information density* per fiber resource along three orthogonal axes the industry is actively exploring. The **wavelength-count axis (WDM)** consolidates multiple wavelengths per fiber, progressing from the current 1λ DR baseline through OCI (4λ), 8λ, and toward 16λ per direction and beyond, gated by optical-engine integration maturity and per-channel power budget.



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Figure 7.1. *Representative Example: Optical 16-λ DWDM Bi-Di Link Architecture.*

The **directionality axis (BiDi)** carries counter-propagating wavelength sets on the same fiber pair, halving fiber count per link of a given aggregate BW, gated by isolation, crosstalk, and Tx/Rx co-packaging in the OE.

The **waveform axis (modulation order)** extracts more bits per symbol per wavelength. The industry is extending the PAM4 baseline along several directions: PAM6/PAM8 for incremental gains within IM/DD constraints, and coherent-lite variants under active exploration for short-reach intra-datacenter and scale-up applications. Each is gated by DSP power budget. The three axes are multiplicative, and the trade-offs are independently negotiable: an integrator can scale WDM aggressively while keeping waveform conservative, extract spectral efficiency from waveform while keeping λ count modest, or combine both. The architecture treats each axis as an independent design knob the integrator selects against the placement class, channel-operating-margin envelope, and effective pJ/bit target of § Power Efficiency at Scale, recognizing that information density and energy density are jointly optimized: coherent-lite, for instance, increases bits per fiber at a cost in pJ/bit that the placement-class and COM framework needs to accommodate.

These axes define wavelength-scaled radix: the effective topology radix achieved within a fixed connector, FAU, or tray-face envelope when fiber count is multiplied by wavelengths per fiber, payload rate per wavelength, and bidirectional directionality. For a target logical link BW, the number of fabric ports supported by a connector area is approximately (fiber count × wavelengths per fiber × per-wavelength payload rate × directionality factor) divided by target link BW. This is why multi-wavelength BiDi profiles change the architecture: the limiting object becomes the information rate per fiber bundle, not the connector count alone.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Each profile is a contract on three dimensions: **wavelength allocation** (grid family focused on CWDM, LAN-WDM, or O-band custom, and per-direction wavelength assignments); **connectorization** (single-fiber versus fiber-pair, polarity convention, connector class); and **telemetry** (per-lambda OSNR, BER, temperature reporting to the OCP RAS endpoints defined earlier). Waveform class is an OE-internal implementation choice the profile contract does not constrain: any waveform that satisfies the channel-operating-margin envelope for the chosen placement class qualifies. This keeps the contract vendor-neutral and waveform-neutral simultaneously, accommodating PAM-family modulations, coherent-lite, and future waveform innovations as the field develops without requiring profile revisions when a new modulation qualifies.

Latency receives similar treatment: it is OE-class-determined rather than profile-discriminating. Each profile supports linear (unretimed), half-retimed, and fully-retimed configurations per Goal 8, with retiming adding ~8-10 ns per DSP/retimer stage, accumulating to ~40-120 ns across multi-stage retimed copper path, as quantified in the OIF COI [3] work; per-link latency is reported through the telemetry dimension so implementations are verifiable against application latency budgets without proliferating profiles by latency tier.

For scale-out (fabric) links, Link Profiles carrying Ethernet, including the Ultra Ethernet Consortium (UEC) transport extensions for AI fabrics, should align with existing IEEE Ethernet PMDs where feasible (e.g., DR/FR-class optics) to preserve ecosystem reuse. They may still allow co-packaged placement and alternative attach models, such as linear or half-retimed interfaces, when power or latency targets justify them.

Table K: Network Profiles

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Profile	Description	Fibers/Link	BW @ 224G	BW @ 448G	Primary Use
A: Baseline	Unidirectional, 4–8λ per fiber	2 (Tx + Rx)	1.6–3.2T	3.2–6.4T	Scale-out fabric · Scale-DSA fabric-class
E: OCI 4+4 (DWDM BiDi)	Bidirectional, 4+4 DWDM	1	1.6T	3.2T	Scale-up baseline (current OCI 200G)
B: High-Density	Bidirectional, 8+8 DWDM	1	3.2T	6.4T	Scale-up, memory · Scale-DSA memory-semantic
C: Extended	Bidirectional, 16+16 DWDM	1	6.4T+	12.8T+	Future high-radix · Scale-DSA cross-architecture
D: High-Cap	Unidirectional, 16–32λ DWDM	2 (Tx + Rx)	6.4–12.8T	12.8–25.6T	Fabric backbone · Scale-DSA aggregation

BW columns show aggregate bidirectional payload throughput per link, computed as (total wavelengths across all fibers in the link) × (per-λ payload rate: ~200 Gb/s at the 224G PAM4 column; ~400 Gb/s at the 448G PAM4 column, after FEC overhead). Demonstrated versus target: Profile B has been publicly demonstrated at 16-wavelength bidirectional operation on a single fiber, at approximately 800 Gb/s aggregate at current per-lane rates; the 3.2–6.4T column values are per-lane-rate roadmap targets rather than demonstrated points. Bidirectional profiles list wavelengths per direction in N+N form (e.g., 8+8 is 8 wavelengths each way, 16 total on one fiber); unidirectional profiles list wavelengths per fiber.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Scale-DSA mapping is by traffic envelope rather than dedicated profile: memory-semantic Scale-DSA shares the Profile B/C contract; fabric-class Scale-DSA shares the Profile A/D contract. No Scale-DSA-specific profile is required at the physical-contract layer.

7.1.1. Latency Characterization Across Candidate Link Profiles

In addition to BW, fiber count, and reach, candidate Link Profiles should define their latency characteristics, as latency directly impacts the efficiency of tightly coupled AI workloads. Latency in high-speed interconnects is determined by physical layer processing, FEC, propagation delay, and, where applicable, switching and buffering, as defined in IEEE 802.3 and OIF specifications. Latency characteristics vary across Link Profile classes:

- **Scale-up (intra-rack, CPO / BiDi optical):** These profiles are optimized for low-latency operation with minimal or no retiming and optional reduced-latency or bypassed FEC modes. Due to short physical reach (typically <10 m), propagation delay is negligible (≈5 ns/m). Latency is therefore dominated by PHY processing and implementation-dependent components (e.g., SerDes and attach), which are commonly in the *few to tens of ns range* based on high-speed electrical and optical interface implementations described in OIF CEI frameworks. Deterministic and tightly bounded latency is required for synchronization-sensitive workloads.
- **Scale-out (Ethernet / fabric-based interconnects):** These profiles incorporate standardized PHY and FEC pipelines (e.g., RS-FEC / KP-FEC defined in IEEE 802.3). RS-FEC implementations can introduce *tens to hundreds of ns latency*, depending on code parameters (e.g., block size and rate as defined in IEEE Clause 91[16] and related clauses). In addition, switch traversal introduces *sub-µs to µs-scale latency*, depending on architecture and buffering conditions. As a result, E2E latency is higher and less tightly bounded compared to scale-up interconnects.

[16] https://ieeexplore.ieee.org/document/9988984

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

- **Long-reach optical transport (e.g., DWDM, APN-class networks):** In these profiles, propagation delay becomes a dominant component, scaling linearly with distance at approximately *5 µs/km* in optical fiber. Optical modules and transport layers introduce additional latency contributions (e.g., O-E conversion and signal conditioning), typically in the *tens of nanoseconds (ns) range*, as characterized in OIF implementation agreements and optical interface specifications. These links therefore operate in a fundamentally different latency regime compared to intra-rack interconnects.

FEC contributes a measurable portion of total latency. For example, RS-FEC schemes defined in IEEE 802.3 introduce latency on the order of ~100 ns, with variation depending on implementation and data rate (e.g., higher-rate interfaces such as 200G/400G/800G). The table below summarizes representative latency characteristics across selected Link Profiles.

Table L: Latency Characteristics

Link Profile	Retiming Model	FEC Usage	Representative Latency Characteristics
SU-1 BiDi **DD**	None (linear / DD)	Optional / reduced	~10–50 ns total (PHY-dominated, negligible propagation)
SU-2 BiDi Half-Retimed	Partial retiming	Standard or reduced FEC	~50–150 ns including FEC and retiming
SO-1 Fabric Retimed	Full retiming	Standard RS-FEC / KP-FEC	~200 ns to >1 µs including FEC and switching

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Latency directly impacts distributed AI workload efficiency: synchronization, collective communication, and memory access patterns. Accordingly, Link Profiles should define bounded and predictable latency characteristics, particularly for HBD deployments in OCP architectures.

7.2. Profile-Invariant Infrastructure

The following infrastructure elements are proposed as common across all Link Profiles and independent of OE class. Together, they represent the generation-invariant layer that, once installed, does not change:

- Fiber routing: Standard fiber cable in the raceway. The raceway dimensions accommodate both single-mode and multimode fiber of standard outer diameter.
- Connector interface: A standardized adapter position in the Optical Faceplate and ODF. The adapter type may vary by deployment, but the mechanical grid is common.
- Power and thermal envelope: The ORv3 tray specification allocates a power and cooling budget for OEs. Whether a vendor fills that budget with a dense DWDM engine or a simple VCSEL array is a vendor implementation choice.
- Management and monitoring: Link-level optical power monitoring, bit-error-rate reporting, and wavelength identification, accessible through standard management interfaces.

7.3. Protocol Accommodation

The architecture explicitly accommodates the major scale-up and scale-out protocol families: NVLink™ (NVIDIA proprietary),[17] UALink™ (open consortium, 85+ members),[18] and UEC/SUE (Ultra Ethernet Consortium™, 100+ members).[19] All three converge on 200G/lane (and future 400G/lane) SerDes at the electrical level. The photonic layer is

[17]NVIDIA, “NVLink™ and NVLink™ Switch.” Fifth-generation NVLink™ (Blackwell): 1.8 TB/s per GPU; sixth-generation NVLink™ (Rubin): 3.6 TB/s per GPU. www.nvidia.com/en-us/data-center/NVLink™/.

[18]UALink™ Consortium, UALink™ 1.0 Specification, April 2025. Supports up to 1,024 accelerators.

[19]Ultra Ethernet Consortium, UEC 1.0 Specification, June 2025.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

transparent transport at the PHY level. It carries whatever the SerDes presents, without protocol awareness.

This means the same optical infrastructure supports NVLink™-based, UALink™-based, and UEC-based deployments without modification. The protocol choice is orthogonal to the photonic choice. A system integrator can deploy NVLink™ scale-up with UALink™ scale-out, or all-UEC, or any combination, on the same optical platform.

Additionally, Compute Express Link (CXL®) is emerging as a key protocol for memory disaggregation and resource pooling. CXL®.mem over photonic links is a natural fit for the photonic HBD concept described in § Photonic HBD : the same fiber infrastructure that carries scale-up traffic can carry CXL® memory traffic to remote memory trays, enabling cache-coherent access to disaggregated memory pools across optical reach. The architecture's protocol agnosticism ensures that CXL®, PCI EXPRESS® (PCIe®), and future memory-semantic protocols are accommodated without platform modification.

7.4. Time and Synchronization

For a generation the compute tray took its time through the front-panel network interface: a NIC carried the reference in, the host disciplined to it, and everything downstream followed. Co-packaged optics makes that interface optional, and a design that depends on it breaks when the card is absent. The function does not disappear — the accelerators still need a common time, and need it more tightly than copper ever asked, in the sub-nanosecond range that synchronized collectives and scheduled fabrics demand. Time stops being a service the network quietly supplies and becomes a subsystem the tray supports.

The recommended default for next-generation AI trays is IEEE 1588-2019 High Accuracy, the standardized form of the White Rabbit technique, which measures the one-way link delay in hardware rather than assuming the forward and return paths are symmetric. That is why it reaches sub-nanosecond accuracy where ordinary two-way

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

time transfer stalls at tens of nanoseconds: it knows the delay rather than inferring it. The recommendation fixes a capability, not an accuracy — a tray should be built High-Accuracy-capable so the sub-nanosecond option exists, while a workload that tolerates coarser alignment may run that hardware relaxed. What it establishes is the single time domain everything on the tray is traceable to, and the protocol the network-terminating elements speak.

The reference is best carried as one signal: a frequency with a one-second phase marker folded into it, so that phase becomes a feature of the waveform the link already carries rather than a second cable run beside the first. Because frequency and phase ride the same waveform they cannot skew against one another, and skew between separate signals was the dominant error a dense tray used to fight. The principle holds on both sides of the board edge. On copper it is mature, available from more than one timing-silicon supplier, and distributed through ordinary clock-tree fan-out over connectors already present, where a matched-delay buffer adds tens of femtoseconds. On the optical plane the same idea applies — embed the marker in the link's own physical layer so it rides the infrastructure that already reaches every accelerator, crossing the optical domains without a separate timing path. One method, two media, one boundary (Figure 7.2).

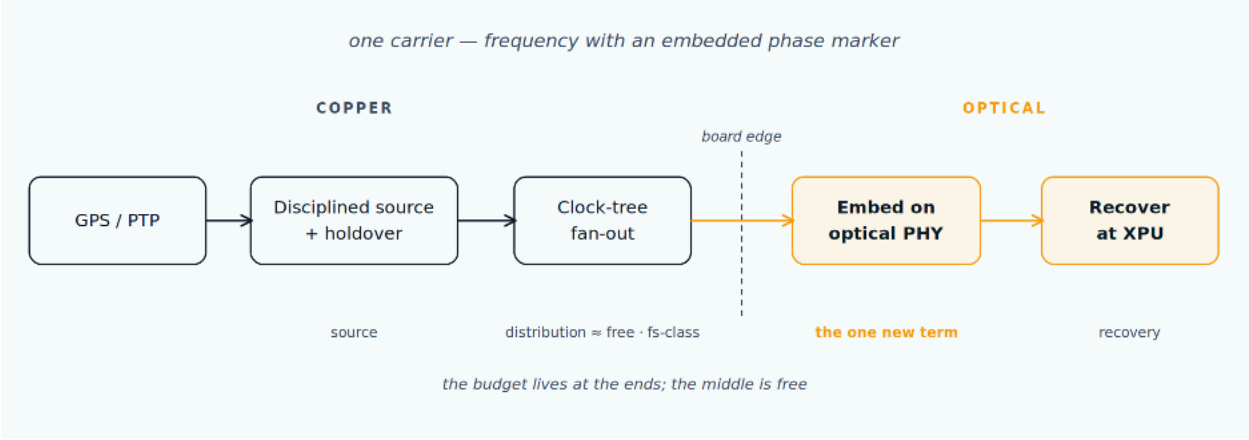


Figure 7.2. *One reference, two media, one boundary: a frequency carrying an embedded phase marker, distributed over copper and embedded onto the optical plane at the board edge.*

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Three carriers compose rather than compete. The standard High-Accuracy protocol is the time domain and the interoperable baseline; it runs natively wherever Ethernet terminates and serves the host, the management module, and any network endpoint directly. The copper carrier delivers the embedded frequency-and-phase signal across the board at low pin count, reaching the host, the control module, and the board-edge elements. The optical carrier extends the same embedded principle onto the photonic plane and is the only one that reaches the memory-semantic accelerators, where neither network timing nor copper distribution natively goes. Read as a baseline with two enhancements, the genuine design choice lives in one place, the optical leg, where it trades maximal interoperability today against the distribution-free economy of carrying time on the links the accelerators already have.

7.4.1. What Needs To Be Synchronized, and What Provides It

The consumers divide by what they draw. The accelerators, the host, and the secure control module (DC-SCM) need absolute time and phase, a time-of-day they can stamp and align against; the high-speed serial interfaces and retimers on the same buses need only frequency and ride the same distribution tree without taking the phase marker. Everything on the system bus is therefore in scope for distribution, while the phase consumers are the named subset the time-of-day must reach. The rail or top-of-rack switch is both at once: a consumer, because it needs to hold time to act as a boundary clock, and a distributor, because it fans that time onward.

Time enters the tray through the front- or rear-panel network, through the fabric uplink as an embedded marker from a parent that is already disciplined, or through a dedicated timing source with holdover and external (GNSS or network) discipline. Because co-packaged optics makes the network interface optional, the ingress should not depend on it; the dependable home for the reference is a dedicated source folded onto the always-present control module, with the network interface contributing where it is present. The control module thus does double duty — a consumer of time for management and reliability reporting, and the natural home for the tray's reference.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

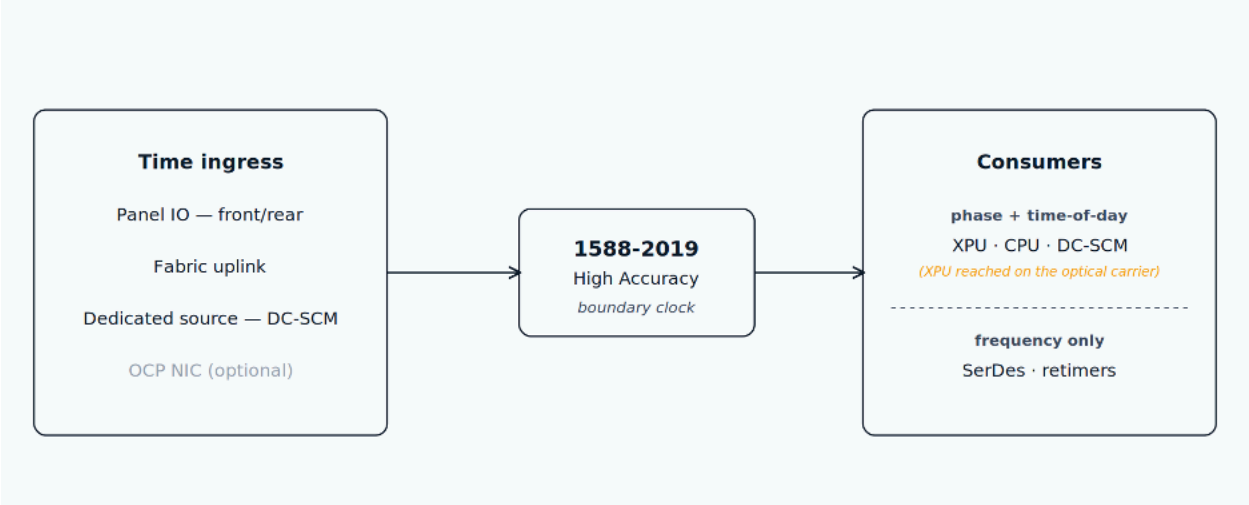


***Figure 7.3**. What needs to be synchronized, and what provides it.*

7.4.2. Eye Safety and the Synchronization Budget

On the optical leg one constraint binds before any other, and it is best read as an account. Eye safety caps the total optical power accessible at an exposed fiber end, summed across every wavelength on that fiber, and dense wavelength multiplexing has already drawn that account down. This is why embedding is the correct choice and not merely an elegant one. A separate timing wavelength is a real withdrawal: an added channel that forces every data channel down a few tenths of a decibel to stay under the cap, landing hardest on the dense rollups that already run eye-limited. An embedded marker, by contrast, modulates light that is already present, and its added power rounds to nothing. Where eye safety is held by reducing power on an open connector, the timing needs to fall back to local holdover during service rather than go dark.

Distribution itself is the cheap part — a matched-delay fan-out contributes a negligible fraction of the budget — so accuracy is set at the two ends, the disciplined source with its holdover and the recovery at each endpoint. By first-principles accounting the forwarded-calibrated carrier sits comfortably inside the sub-nanosecond target, while the inferred two-way carrier runs far tighter and can fail as separate-fiber skew grows; the carrier, not the recovery precision, is the binding choice. The precise budget

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

allocation — and the measurement of the optical embed-and-recover term — is left to the normative specification.

In summary: use 1588-2019 High-Accuracy-capable as its default time domain and run the system to the accuracy the workload requires; carry phase as a physical-layer feature of the link on whichever medium each leg uses; embed the optical reference rather than spend a separate wavelength, keeping it within the eye-safety account; and rest the timing ingress on a dedicated source on the control module, no longer assuming the network interface will provide it.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

8. The Scale Boundary

8.1. Premise: Two Trajectories, One Wall

This chapter turns the architecture developed in the preceding sections into provisioned systems. It applies the defined placement classes and link profiles to a small set of representative compute and switching systems, then asks what each system physically requires: the bandwidth that leaves every accelerator, the fibers and laser sources that carry it, the connectors that terminate it, and the faceplate or rear surface area it consumes. The goal is not to specify a product. It is to make the binding constraints visible: for each system, which constraint is reached first, and what relaxes it?

Two roadmaps drive the demand and are best read together.

On the accelerator side, domain sizes are expanding from today's 72-package generation toward announced 144- and 288-package architectures[1], while open interconnect specifications target 1,024-accelerator domains[2]. Per-package scale-up bandwidth is rising from roughly 7.2 Tbps per direction today toward 14.4 Tbps per direction and beyond, with aggregate domain bandwidth approaching 12 Pbps. These accelerators also expose flexible I/O: the same SerDes lanes can carry PCIe®, native accelerator interconnect, or Ethernet-based fabrics. That makes them a natural fit for a photonic endpoint that can convert any of those electrical protocols to optical transport transparently.

On the switching side, the elements that seat these domains are advancing on a parallel ladder. Merchant switch throughput has roughly doubled each generation: 12.8T, 25.6T, 51.2T, 102.4T, with 204.8T on the horizon. SerDes rates have followed the same cadence, moving from 100G to 200G to 400G per lane. For a flat, single-hop scale-up domain, the governing number is switch radix. The ecosystem carries two domain-sizing conventions: a power-of-two lineage at 512, 1,024, and 2,048 endpoints, and a 72×2^k

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

lineage at 576, 1,152, and 2,304 endpoints. A switching element earns a generation's flat domain only when its radix reaches the larger member of the relevant pair, as shown in Table F.

Those two trajectories meet at the same constraint. The throughput-doubling cadence keeps the lane count roughly constant: a 51.2T switch at 200G per lane and a 102.4T switch at 400G per lane both carry on the order of 256 lanes. But each faster lane becomes much harder to route in copper. Copper reach falls from about a meter at 200G to under half a meter at 400G; at 800G, it may not close for any board-level distance without retimers, as shown in Table G. The market therefore demands higher radix at faster lane rates, while the faster lane rates are precisely the ones copper can no longer carry over the required distance. Demand and implementation run into the same wall.

The present record suggests that this wall is already visible, not merely projected. Hyperscalers have validated co-packaged optics reliability in production switch deployments, including multi-million-hour MTBF demonstrations. Next-generation dense-rack architectures already show optical inter-canister links in roadmap material, and supply-chain indications point to configurations reaching 288 multi-die packages, or more than a thousand accelerator dies, in a single rack. That reinforces the 1,024-plus domain target. At the same time, scale-up links inside current dense racks remain copper and are already on the critical path for the next density doubling, where copper midplane reach becomes untenable. SerDes at 448G and beyond effectively requires optical transport for any link beyond about a meter, and open consortium roadmaps name optical reach as an explicit longer-term scaling vector.

Taken together, these signals point in one direction: the demand for larger scale-up domains is committed, even where the corresponding optical commitment has not yet been made public.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

The argument now moves linearly instead of circling: accelerator demand rises, switch capacity rises, copper reach collapses, and optics becomes the escape path.

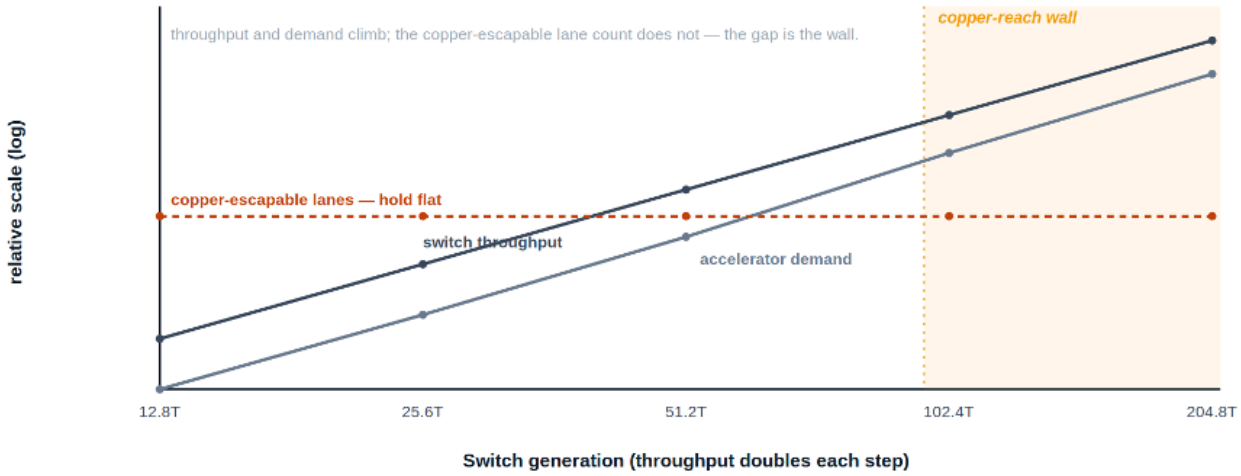


Figure 8.1. *Demand versus the seating element, converging on the copper-reach wall.*

8.2. How to Read These Tables

Each system below is evaluated through the same provisioning chain so the results can be compared directly. For each endpoint, the chain calculates the required scale-up bandwidth, converts it into optical fibers and laser sources under the selected scheme, counts it into front and rear connectors, translates that into faceplate and rear-zone area, and checks the result against the first constraint the design reaches.

The calculations are first-principles and grounded in public sources. Named components are used only as public examples of capability classes, not as product claims. Every figure should be read as a minimum architectural requirement, not as a vendor specification.

Each system is tested against five possible binding constraints: interconnect reach, PCB signal-layer count, package-perimeter egress area, power-delivery headroom, and thermal-manifold access. Rack-scale density is bounded by whichever of these limits is reached first. Copper flyover relaxes three of the five; photonic flyover also relaxes

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

reach. Each provisioned system therefore names its binding constraint, so the chapter can show when a denser organization is forced by physical limits rather than simply asserted.

Two kinds of radix appear in the tables and should be kept separate. Electrical egress counts the lane pairs leaving the package at the host SerDes rate. This is the package-perimeter question, reported here as the share of usable copper shoreline consumed by the scale-up plane and the headroom that remains. Optical rollup counts the fibers leaving the faceplate, normalized to a 200 G-per-wavelength payload. Electrical egress determines whether the package can be served in copper at all; optical rollup determines how much faceplate and rear surface the served bandwidth consumes.

The rollup scheme trades wavelength density against interoperability and is chosen by traffic plane. DR carries one wavelength per fiber. OCI uses the OIF 4+4 arrangement: eight wavelengths per fiber at 53G each. The 8+8 scheme uses sixteen-wavelength DWDM. Scale-up traffic stays on standards-aligned DR or OCI where multi-vendor interoperability matters. A Scale-DSA plane can use denser 8+8 when density matters more than interoperability.

The fit reading is the front/rear stoplight at the stated tray height. Where one tier exceeds a single tray, the table reports trays per tier instead. The reference enclosure is a no-midplane canister with a one-OU front faceplate and a six-slot rear EB zone, both of which scale with tray height. In that design, both front and rear surfaces are EB faceplates. Class H and Class V are the tray-orientation classes defined in Platform Form Factor Evolution; in this chapter, they are used as provisioning targets.

8.3. Class-H (MHS/MGX Trays): Reach Scenarios

The AI baseline architecture is dedicated compute and switch trays, rail-optimized so the scale-up domain is composed across several switches. For today's MHS/MGX trays and racks, this class of horizontally mounted trays is termed ‘Class-H’. Scale-up

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

currently escapes in copper across this class, but against a fixed ceiling: the bandwidth a die drives off its edge is set by shoreline density — on the order of a terabit per second of bidirectional traffic per millimeter of usable edge — and a reticle-class die offers only so much edge. A switch lives hardest against that ceiling, because a switch is by function all I/O. A 102.4T merchant switch already carries more scale-up radix than one reticle die edge can escape in copper, which is why current parts disaggregate their SerDes into chiplets wrapped around the packet die — and why a faster-SerDes variant ships in a smaller package than a slower one: the edge, not the logic, sets the size. Copper still closes this generation at 200G for in-rack reach; what it cannot do is absorb the next doubling. At 204.8T on the same reticle the escape demand doubles again while the edge does not, leaving two exits — raise the lane rate to 448G, where passive copper no longer reaches across a rack, or leave copper. Co-packaged optics already ships at exactly this tier. The switch, not the accelerator, typically is the first endpoint photonics carries.

Accelerators have not yet reached that boundary. Their scale-up rides a large multi-die package with edge to spare: the R-/MI-class draws on only a small fraction of its available shoreline, the TPU-class less. So copper escape is comfortable today, and the headroom is real. That headroom is not a reprieve from the change but a measure of how far behind the switch these parts sit; as rack density rises, the same ceiling reaches them. On a separate axis, the low-bandwidth TPU-class is faceplate-bound by its laser floor: a single optical source dominates its small budget, so its faceplate is set by the source, not the data rollup.

Table M: Class-H Reference Provisioning — Reach Scenarios

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Endpoint	Scale-up Tb/s/dir	Copper edge @200G	Scheme	Data fibers	ELSFP	Fit front / rear
R-/MI tray (R200 / MI400, 4 XPU)	57.6	holds · +58.8	DR	576	9	GREEN / AMBER
			OCI	288	9	GREEN / GREEN
			8+8	36	2	GREEN / GREEN
R-/MI tray, next-gen (R300 / MI500, 4 XPU)	160	holds · +11.5	DR	1600	25	AMBER / AMBER
			OCI	800	25	GREEN / AMBER
			8+8	100	4	GREEN / GREEN
TPU tray (TPU v7, 4 XPU)	19.2	holds · +72.3	DR	192	3	GREEN / GREEN
			OCI	96	3	GREEN / GREEN
			8+8	12	1	GREEN / GREEN
Switch 102.4T (the binder)	102.4	26 · 4× short	DR	1024	16	GREEN / AMBER

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Endpoint	Scale-up Tb/s/dir	Copper edge @200G	Scheme	Data fibers	ELSFP	Fit front / rear
			OCI	512	16	GREEN / GREEN
			8+8	64	2	GREEN / GREEN

Copper edge @200G — each accelerator's usable package edge serves its full scale-up demand with the headroom shown (Tb/s/dir of spare edge, ≈ 1 Tb/s per usable mm at 200G SerDes; a SerDes generation doubles per-mm density). The 102.4T switch serves only 26 Tb/s/dir of its bisection, so the balance escapes optically (N× short). Scale-up, data fibers, and ELSFP are per tray (4 XPU); copper-edge and headroom are per package. Compute full-ORU; switch full-ORU. Scale-DSA plane (8+8) paired per endpoint.

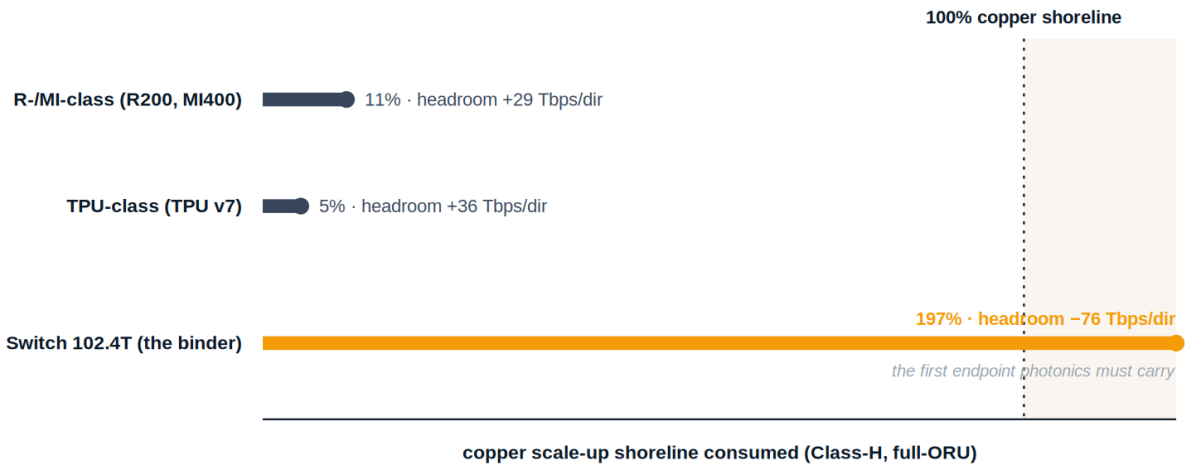


Figure 8.2. *Compute sits below the copper shoreline; the switch is first to exhaust it.*

8.4. Class-V (Reach): Interconnect Reach Binds

The first vertical scenario is the reach-bound case: an eight-accelerator canister at the density where the binding constraint is no longer escape area but the distance the scale-up link spans. The accelerator that escaped comfortably at its own package edge cannot make the inter-tray reach across the canister at 224G, much less at 448G, where copper does not close past roughly half a meter. The per-endpoint photonic bill is

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

modest, between nine and a few hundred fibers depending on rollup, and indifferent to span: the same rollup that serves a package-edge link serves a cross-canister one. Reach is the constraint photonics removes first.

Table N: Class-V Reference Provisioning — Reach-Bound Case

Endpoint	Scale-up Tb/s/dir	Copper edge @200G	Scheme	Data fibers	ELSFP	Fit front / rear
R-/MI tray (R200 / MI400, 4 XPU)	57.6	holds · +58.8	DR	576	9	GREEN / AMBER
			OCI	288	9	GREEN / GREEN
			8+8	36	2	GREEN / GREEN
R-/MI canister (R200 / MI400, 8 XPU)	115.2	holds · +58.8	DR	1152	18	AMBER / AMBER
			OCI	576	18	GREEN / AMBER
			8+8	72	3	GREEN / GREEN
R-/MI tray, next-gen (R300 / MI500, 4 XPU)	160	holds · +11.5	DR	1600	25	AMBER / AMBER

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Endpoint	Scale-up Tb/s/dir	Copper edge @200G	Scheme	Data fibers	ELSFP	Fit front / rear
			OCI	800	25	GREEN / AMBER
			8+8	100	4	GREEN / GREEN
TPU tray (TPU v7, 4 XPU)	19.2	holds · +72.3	DR	192	3	GREEN / GREEN
			OCI	96	3	GREEN / GREEN
			8+8	12	1	GREEN / GREEN
V700 tray (4 XPU + rail-switch)	egress 57.6	switch moves 115.2	DR	576	9	GREEN / AMBER
			OCI	288	9	GREEN / GREEN
			8+8	36	2	GREEN / GREEN
Switch 102.4T	102.4	26 · 4× short	DR	1024	16	GREEN / AMBER
			OCI	512	16	GREEN / GREEN
			8+8	64	2	GREEN / GREEN
Switch 204.8T	204.8	26 · 8× short	DR	2048	32	AMBER / AMBER
			OCI	1024	32	GREEN / AMBER

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Endpoint	Scale-up Tb/s/dir	Copper edge @200G	Scheme	Data fibers	ELSFP	Fit front / rear
			8+8	128	4	GREEN / GREEN

Per-XPU bill (×8 at the canister). The copper figure is the package-edge escape, which holds; the binder is reach over the canister span, which the optical rollup is invariant to. Fit shown at half-height (compute-friendly).

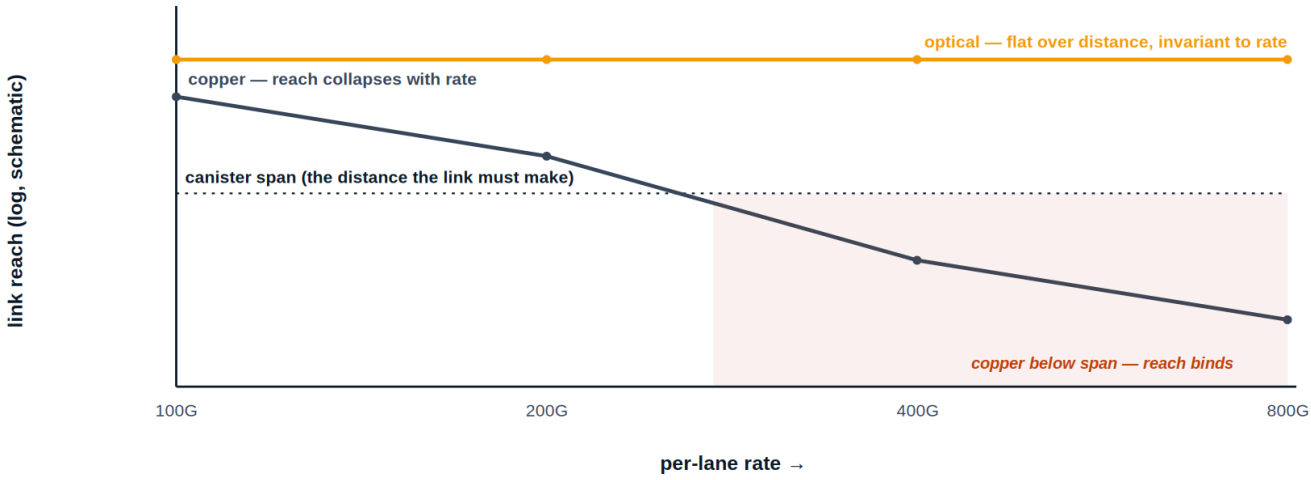


Figure 8.3. *Reach as the binding constraint: copper reach collapsing with lane rate, the optical rollup flat over distance.*

8.5. Class-V (Throughput): Escape Density Binds

The throughput scenarios move the binder from reach to escape density. The decisive fact is geometric: copper IO escape from the perimeter does not grow with the tier so as an example, a switch's usable copper shoreline carries about 26 Tbps/dir whether it is a 204.8T or a 409.6T part. Again, the shoreline length is fixed. Areal Packaged Optics (APO) is a surface resource, and it scales with the package area rather than its edge. The first table makes the gap concrete: the same package that escapes about 26 Tbps/dir in copper carries on the order of 1,200 Tbps/dir of APO I/O, roughly forty-six times its copper shoreline and several times what the tier itself requires. In radix terms,

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

that is thousands of lanes against the 1,024 the domain needs. This is the perimeter-to-area transition that Table F tracks qualitatively in its escape-regime column, with beachfront saturating and APO becoming mandatory as radix climbs, and it is the reason the switch is escapable at all. That immutability has a companion on the routing side: because the port rate doubles with the tier, the switch's addressable scale-up domain holds near 128 across 102.4T at 800G, 204.8T at 1.6T, and 409.6T at 3.2T. The capacity and per-endpoint bandwidth double while the domain itself sits still.

Table O: Perimeter-versus-Area Escape at the Switch

Endpoint	Scale-up need (Tb/s/dir)	Copper perimeter escape (Tb/s/dir)	APO area ceiling (Tb/s/dir)	APO radix @200G	Area unlock
R200 / MI400	14.4	52 (holds)	5,866	29,330	113× cu · 407× need
R300 / MI500	40	54 (holds)	6,248	31,240	116× cu · 156× need
TPU v7	4.8	42 (holds)	1,914	9,568	46× cu · 399× need
Switch 102.4T	102.4	26 (4× short)	1,191	5,956	46× cu · 11.6× need
Switch 204.8T	204.8	26 (8× short)	1,191	5,956	46× cu · 5.8× need

Scale-up radix is the electrical radix the domain needs at its lane rate (= the Table F radix for the tier): 1,024 lanes/dir for either switch (200G or 400G), 72 for the R-class XPU. Copper escape and APO ceiling are per-direction (Tbps/dir). APO radix is that ceiling expressed as electrical radix at the lane rate — the radix area-extraction could feed, far above the need (the unlock factor). The two switch rows are the same 1,024-radix domain on Table G's lane-rate axis — the escalation is lane rate, not domain size.

Given that escape, the rollup decides the faceplate, and the connector inventory follows. The list below counts the front data connectors (SN-MT32) and the rear blind-mate

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

connectors (EB-192) for each rollup — two placement options for the same data — alongside the laser modules. The **XPO (integrated)** row is listed for reference: per the XPO MSA, a 12.8T module presents eight SN-MT16 front ports, so the integrated baseline is eight modules landing 16 four-port adapters at 102.4T and sixteen modules landing 32 at 204.8T, filling roughly half an OU and one OU of front faceplate respectively; the 409.6T (N+2) tier and the 400G-SerDes generation, where only CPO and APO remain in scope, are tabulated in the appendix. Tray height sets the ratio rather than a pass/fail: at the 28 mm tray pitch, matching how NVIDIA® NVSwitch™ cards are built for rail switching, the switch tier is composed across several trays (three for DR, two for OCI, one for the DWDM 8+8 at 204.8T), and a 48 mm tray roughly halves the discrete counts. Compute, an order of magnitude lower in egress, fits one 28 mm tray across every scheme, so the canister packs a uniform 28 mm pitch with compute on a single tray, the switch tier spread across several.

Table P: Class-V Throughput — Connector and Laser Bill by Rollup

Endpoint	Scheme	Data fibers	Front adapters (4-port SN-MT)	EB-192 (rear)	ELSFP	Trays/tier (28 mm)
Switch 102.4T (28 mm tray)	DR	1024	8	6	16	2
	OCI	512	4	3	16	1
	8+8	64	1	1	2	1
	XPO (integ.)	1024	16	n/a	n/a	0.5 OU

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Endpoint	Scheme	Data fibers	Front adapters (4-port SN-MT)	EB-192 (rear)	ELSFP	Trays/tier (28 mm)
Switch 204.8T (28 mm tray)	DR	2048	16	11	32	3
	OCI	1024	8	6	32	2
	8+8	128	1	1	4	1
	XPO (integ.)	2048	32	n/a	n/a	1.0 OU
R-/MI tray (R200 / MI400, 4 XPU)	DR	576	5	3	9	1
	OCI	288	3	2	9	1
	8+8	36	1	1	2	1
R-/MI tray, next-gen (R300 / MI500, 4 XPU)	DR	1600	13	9	25	1
	OCI	800	7	5	25	1
	8+8	100	1	1	4	1

The front column counts 4-port SN-MT adapters: the discrete (CPO) rows land SN-MT32 (128 fibers per adapter), the XPO baseline SN-MT16 (64 fibers per adapter — eight ports and two adapters per 12.8T module as seen in public announcements). Front and rear are alternative landings for the same data, not additive. Trays/tier counts the 28 mm trays needed to compose the tier (rail-switching): DR is rear-EB-bound and needs the most, 8+8 DWDM seats a full tier on one tray. The XPO baseline is effectively an integrated DR rollup — its 2,048 / 4,096 fibers match the CPO DR count — co-packaged into 16 / 32 modules at 1.0 / 2.0 OU, no discrete optical trays. OCI here is the OIF 4+4 profile — eight wavelengths per fiber at 53G — matching Table K's per-port fiber basis.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table Q: Tray-Height Lever — 204.8T Switch Trays per Tier

Tray height	DR	OCI	8+8
28 mm (primary)	3	2	1
48 mm	2	2	1

Trays/tier for the 204.8T switch by tray height: the 28 mm pitch composes the tier across three trays (DR), two (OCI), or one (8+8 DWDM); a 48 mm tray trims DR to two. The denser the rollup, the fewer trays — DWDM and integration both reach one.

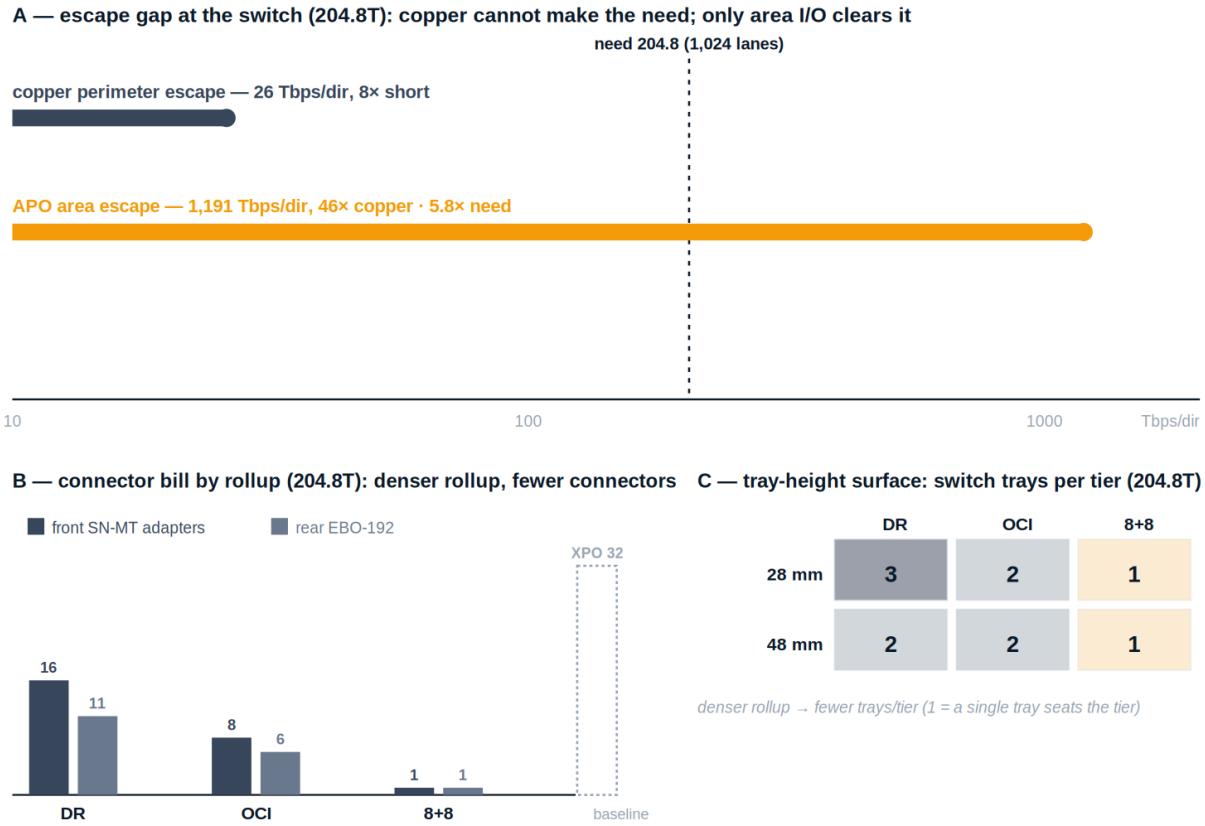


Figure 8.4. *Throughput: the perimeter-versus-area escape gap, the connector bill by rollup, and the tray-height surface.*

8.6. The Wall

Across the scenarios, the binding constraint is not fixed. It migrates as density rises.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

In today's compute trays, scale-up traffic can still escape over copper, and the accelerator is not yet the limiting endpoint. The switch is. It reaches its copper shoreline first: the package edge runs out of practical egress before the logic runs out of usefulness. The next doubling, to the 204.8T / 1.6T generation on a comparable edge budget, is where copper can no longer carry the required escape.

When the design pushes into a canister, reach becomes the first binding constraint across the span, and photonics removes that limit. Push density further, however, and escape binds again at the switch. Tray height then determines the trays-per-tier ratio. At a 28 mm pitch, the tier spreads across multiple trays unless the optical rollup is dense enough; only DWDM with integration holds the tier to one tray.

Each provisioned system marks where its own trajectory reaches the most restrictive of the five constraints, and which constraints are relaxed by copper flyover versus photonic flyover. Read against Figure 4.2, the scenarios show the two axes meeting at the wall: reach extension buys bandwidth up to copper's domain ceiling, but only area-integrated optical escape moves the ceiling itself.

A socketed near-package escape has two independent knobs, and they do different work. Lane rate sets the module count: serving 204.8T of scale-up takes thirty-two 6.4T modules at 224G/lane, sixteen at 448G. The medium, whether NPO fiber at the socket or copper flyover to a faceplate pluggable, sets where the channel ends and which of the five constraints it spends.

Table R: Socketed near-package escape — convergence for 204.8T scale-up

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Lane rate	NPO (E-O at socket, fiber)	Copper flyover (twinax to faceplate)
224G	32 OEs, die-edge terminus. Perimeter binds; corner placement splits SI into two channel classes. No convergent layout.	32 twinax launches, faceplate terminus. Egress lifts to the connector stack-up; SI trimmed in cable. Converges, cable-heavy, faceplate retimer.
448G	16 OEs, uniform edge ring, shortest channel, no corners. Converges.	16 launches, half the cable; ~25 cm copper reach, retimer likely. Converges, reach-tight.

Lane rate sets the module count; the medium sets the channel terminus and the binding constraint. Only the 224G NPO cell has no uniform, closing edge layout.

Beachfront is medium-set, not socket-set. Assumed values, since the connector geometry is not yet public: ~16 mm copper connector, ~18 mm external-laser OE, ~22 mm internal-laser OE. The rate fixes the count; the medium fixes the width. The assumed 2 mm copper-to-external-laser margin adds roughly four within-reach modules on a reticle-class shoreline; at substrate scale the verdicts are unchanged.

Only one cell fails to converge. At 224G the NPO ring holds the escape in-plane at the package edge, where thirty-two modules exceed the perimeter and the corner placements fiber cannot trim back into one channel. Copper flyover at the same rate carries the same thirty-two-count pressure, yet it lifts egress off the edge into the connector stack-up (§ The Flyover Cable Layer) and equalizes the channels in trimmable cable, converging with a faceplate retimer at the copper reach ceiling (§ The Copper Reach Constraint). Doubling the rate to 448G halves the count to sixteen; the NPO ring then seats uniformly with the shortest channel in the matrix, and copper flyover converges on half the cable at a quarter of the reach.

This is the wall read from the connector. At the socket the copper-or-optical question decides only which of the five constraints binds, while whether the edge can be served at all is a question of rate. Neither knob substitutes for the area escape that removes the edge from the account entirely (Figure 4.2, Table O).

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

That composite result is the chapter's setup for what follows. At the 204.8T / 1.6T switch generation, the horizontal tray class reaches a density where it can no longer hold the domain the roadmaps require. The link reach cannot be closed, the package escape cannot leave the switch, or the faceplate cannot carry the optical rollup; often, more than one of these fails at the same time.

That is the trigger for a denser organization of the same trays. This chapter establishes the constraint and identifies the point on the trajectory where it bites. The architectural answer to that constraint, the vertical canister with its depth budget and midplane-free EB faceplates, is defined and built in § Platform Form Factor Evolution.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

9. Platform Form Factor Evolution

Hyperscale AI platforms evolve through distinct form factors, each striking a different balance among density, serviceability, and interconnect. The OCP Open Accelerator Infrastructure (OAI) and the HGX architecture that followed are generation 1, the baseline this section builds on rather than revisits. Two generations follow. Generation 2 is the MHS/MGX rack we call Class-H, which we extend here with state-of-the-art interconnect. Generation 3 is Class-V, which we introduce. One passive optical infrastructure serves all three, scalable and composable, adapting to each form factor's mechanical constraints. The three-phase evolution below is the rollout timeline for these classes, a separate axis from the generations named here.

9.1. Tray Orientation Classes

Table S: Platform Form Factors

Characteristic	Class H (Horizontal)	Class V (Vertical / Canister)
Orientation	Trays slide horizontally, blind-mate at rear	trays insert vertically into canister chassis
Examples	Current MHS, Oberon-class, Helios-class	Kyber-class vertical tray architectures
Density	Moderate (72 XPUs / rack typical)	High (288+ XPUs / rack target); 288 V350 / 576 V700-class XPU per rack at 1:4 CPU:GPU ratio; ~1,728 XPU in 3-rack V700 photonic scale-up domain (see § Proposed Modifications to ORv3/MHS)
Fiber routing	Tray rear to rack fiber management	tray edge through canister raceway to front ODF

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Characteristic	Class H (Horizontal)	Class V (Vertical / Canister)
Faceplate position	Front of tray (existing OSFP bay location)	Front of canister or rack (shared across trays)
Key advantage	Proven serviceability model, multiple fabric planes without major tray change	3–4× density; eliminates copper midplane; supports optional distributed rail switching

Both orientation classes provide the same external interface capability at the Optical Faceplate: connectors and front pluggable modules. The rack-level fiber infrastructure (inter-rack cabling, top-of-rack management) is identical regardless of what is behind the faceplate. This means a data center can deploy a mix of Class H and Class V racks using the same fiber plant.



9.1.1. Switch Card Placement by Orientation Class

The compute and switch reference platforms described in § Reference Platform Scope integrate differently depending on tray orientation class, particularly with respect to switch card (NVSwitch™, UALink™ switch, or fabric switch) placement:

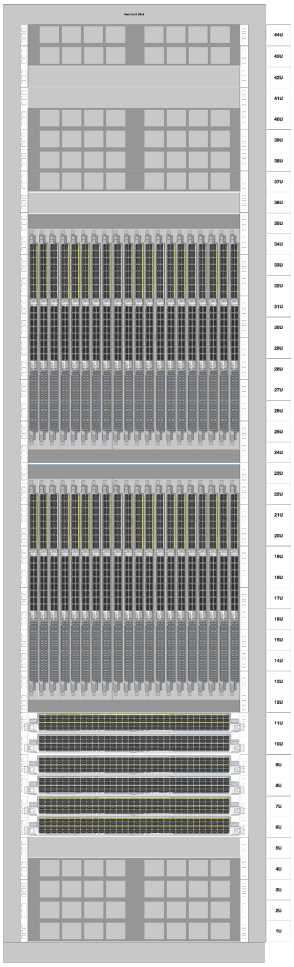
Class H (Horizontal — MHS/Oberon style): Switch trays occupy dedicated positions in the rack, typically at the top or interspersed with compute trays, matching the current MHS and Oberon architecture. Both compute and switch trays face forward, with fiber connectivity at the front faceplate. The ODF/Shuffle sits behind the faceplate and routes fibers between compute and switch tray positions. Because this class already reserves networking tray volume, the principal photonic gain is not a new physical placement of switching silicon; it is multiple fabric planes, cleaner scale-up/scale-out/Scale-DSA separation, reduced copper reach pressure, and generation-invariant passive

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

infrastructure. This is a natural extension of today's layout where fabric-switch trays are separate 1U/2U units in the same rack as compute trays. Fiber routing runs vertically between tray positions through the rack's fiber management channels.

Figure 9.1. *(Above) ORv3 Rack with Class H trays*

Class V (Vertical/Canister — Kyber style): Class V admits three switch-placement options, defined by whether and where switching silicon occupies the canister.



Front-and-rear (rear-switch mode) pairs a front compute tray with a rear switch tray at V350 depth, back-to-back. A passive fiber coupling zone between them replaces the copper midplane.

On-tray (distributed rail-switch mode) carries rail-switch ASICs on the same V700-class tray as the XPU packages. The compute tray becomes both endpoint and routing element. The rear zone, if used, is a passive optical egress or cartridge boundary rather than an active networking card.

Front-only carries no switching in the canister at either tray depth. A V350 tray runs front-only with its rear position unpopulated by a switch tray; a V700-class tray uses its extra depth for XPUs instead of rail-switch ASICs. Both variants source switching from elsewhere in the rack — typically dedicated Class H switch trays.

Figure 9.2. *(Above) ORv3 Rack with Class V trays for compute and Class H shown for switching in front only placement.*

Figure 9.3 contrasts the two switch-integrated configurations against the front-only baseline. Left ("A") is front-and-rear placement, trays back-to-back on rear optical connectors, similar to NVIDIA's Kyber racks. Right ("B") is the longer V700-class tray enabling on-tray placement. Both are formalized as Class V canister depth designators in § Proposed Modifications to ORv3/MHS.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

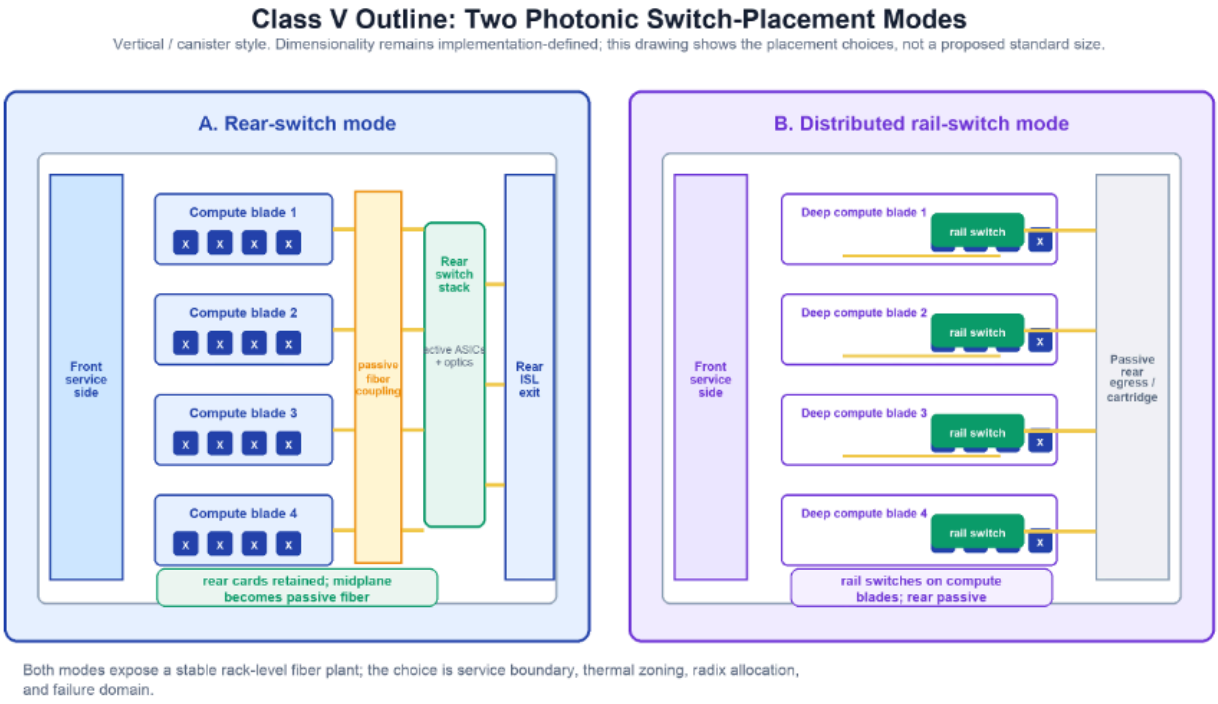


Figure 9.3. *Class V Showing rear-switch mode vs. distributed rail-switch mode; dimensions are illustrative, not normative.*

Both Class H and Class V placement models are accommodated by the same passive optical infrastructure components (Faceplate, ODF, Raceway) and the same Link Profiles, regardless of whether a tray's fiber exits at the front, the rear, or both. The switch reference platform Base Specification should therefore define the optical interface requirements for four deployment patterns: Class H with dedicated switch trays, Class V front-and-rear, Class V on-tray, and Class V front-only. Integrators can then choose based on density, serviceability, thermal zoning, and radix allocation without changing the rack-level fiber plant.

Multiple fabric planes need no new form factor. The front and rear connectors already route to the switches, so one fiber plant carries the scale-up, scale-out, and Scale-DSA planes with no mechanical change. The new classes, and V700 on-tray rail switching in particular, open room to do more, developed in § Scale-DSA: The Third Scaling Axis.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

The platform evolves along two vectors. On the switch, optical integration deepens toward the silicon; on the compute tray, density doubles as photonic escape frees the board. Neither waits for the other, and the three phases below place each in time.

9.2. Optical Integration Continuum at the Switching Silicon

The switch-side vector is an integration continuum. Port density at a switch tray has historically been set by pluggable-cage geometry, a property of the housing, not the silicon. The industry separates the two in stages: pluggable modules on the faceplate, near-package modules that move inward and route fiber to the panel, and co-packaged integration that brings the optical boundary within the package. Each stage shortens the copper channel between silicon and optics, and its energy and signal-integrity cost falls with it (§ Optical Engine Selection and Economics).

Co-packaged integration ends the copper term. The electrical-to-optical transition occurs in millimeters, the channel budget leaves link planning, and what remains is propagation: deterministic, flat across the envelope, indifferent to SerDes generation. The faceplate becomes a fiber-distribution surface, and switch radix binds on silicon bandwidth and waveguide density rather than cage geometry. The high-radix path of § Path to High-Radix Switching becomes realizable at this boundary.

The reference architecture is defined at the fiber boundary, not at any point on the continuum. The interior-facing fiber interface and the rear-facing ISL interface are the tray's contract; how deeply integration proceeds behind them is the industry's to navigate. The contract holds as the boundary moves inward, so the rack and fiber plant carry every stage without redesign.

9.3. Density on the Tray

The compute-side vector is density, and photonics unbinds it. Board area freed from copper fan-out, pluggable cages, and discrete regulators turns four packages into eight

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

in the same envelope (§ Density Doubling Through Board Area Recovery, § The Flyover Cable Layer). The binder then migrates from interconnect to cooling and power, a problem of mechanical and electrical engineering, not physics. The specifics and the open thermal topology live in § Binding Constraints at 8-Package Density.

9.4. Three-Phase Platform Evolution

The photonic architecture envisions a multi-phase platform evolution, with each phase building on the infrastructure of the previous one: Phase 1 opens the switch vector, Phase 2 lands the density vector and the Class-V canister, Phase 3 extends both.

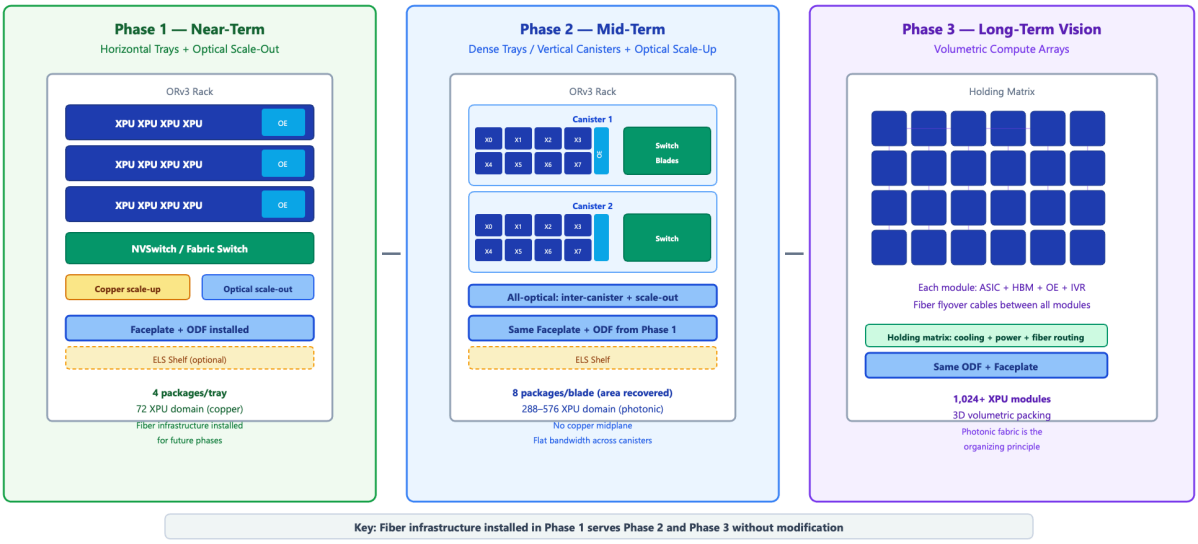


Figure 9.4. *Three-Phase Platform Evolution.*

Phase 1 — Horizontal trays with photonic scale-out (near-term): Current MHS/ORv3 Class H trays with co-packaged or near-package OEs replacing pluggable transceivers. Copper remains for intra-tray scale-up. Photonics extends scale-out reach and reduces transceiver power and complexity. The Optical Faceplate, ODF, and fiber infrastructure are installed in this phase.

Phase 2 — Dense trays and vertical canisters with photonic scale-up (mid-term): Two concurrent density paths. For Class H, photonic interconnect and IVR enable

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

greater than today's 4, for example 8-package compute trays in existing tray envelopes, doubling per-tray compute density (§ Density Doubling Through Board Area Recovery) while supporting multiple fabric planes through existing network-tray positions. The optical flyover cable layer replaces copper trace routing, and the board simplifies to a power distribution and mechanical support structure. For Class V vertical tray architectures, photonics eliminates the copper midplane and enables either rear optical switch trays or deeper front-access compute trays with onboard rail switches. The latter option turns each compute tray into a fabric participant and supports dragonfly-like rack groups within a single logical scale-up tier. Both paths use the fiber infrastructure installed in Phase 1 without modification. Scale-up domains expand to full-rack and multi-rack operation at flat or explicitly engineered oversubscription.

Phase 3 — Volumetric compute arrays (long-term vision): XPU modules in dense three-dimensional arrays connected via photonic flyover cables. Modules are self-contained (ASIC + HBM + OE + IVR), connected to a mechanical support structure (the "holding matrix") that provides cooling, power, and fiber routing. Traditional trays and canisters give way to a volumetric packaging approach where the photonic fabric is the organizing principle. The 8-package tray of Phase 2 is the conceptual precursor: Phase 3 extends the same density logic from two-dimensional tray packing to three-dimensional volumetric packing.

Phase 3 is presented as a directional vision, not a near-term design target. Its inclusion signals that architectural decisions made in Phases 1 and 2 should not foreclose this evolutionary path.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

10. Reliability, Serviceability, and Economics

The operational and economic case for photonic interconnect extends beyond BW and power. Consolidating the optical plant improves link reliability, simplifies field service, and lowers both capital and operating cost; this section makes each case and points to the section where the supporting mechanism is defined. The quantitative targets are left for the ecosystem to calibrate.

10.1. Reliability Improvements

Retries from random bit errors, burst errors, or link flaps increase latency and reduce effective BW. Random bit errors may be caused by marginal signal integrity, noise, thermal drift, or component degradation. Burst errors can result from transient disturbances, connector contamination, laser instability, or fiber impairment. Link flaps, where a link rapidly transitions between up and down states, can have an even larger performance impact because they may trigger retraining, traffic rerouting, or higher-level recovery procedures. Because local compute is synchronous, reduced link BW can slow the entire AI cluster. Reliability targets for photonic scale-up systems should therefore be defined not in terms of raw bit error rate alone, but in terms of workload-visible impact.

Reliability targets scale with line rate. Bit error rate is a per-bit metric; uncorrected-error frequency and link-flap event rates scale linearly with throughput at constant BER. Holding workload-visible error tolerance constant as per-lane rates progress from today's 100 Gb/s generation toward 224 Gb/s and 448 Gb/s therefore requires proportional improvement in pre-FEC channel BER (today on the order of 10^{-4} to 10^{-6} depending on profile and FEC scheme), achieved through coordinated advances across the physical, algorithmic, and protocol layers rather than through any single mechanism.

The architecture's direct contribution to this scaling is at the physical layer. A controlled-loss optical plant consolidated at the Optical Faceplate and ODF (§ Optical Faceplate) reduces the connector and routing losses that accumulate across pluggable

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

systems, and the optional External Laser Source shelf (§ Fiber Raceway and Optical Distribution Frame) provides a serviceable home for high-quality sources whose wavelength stability, RIN, and linewidth bound achievable SNR that are independent of OE replacement cycles. Algorithmic levers including equalization, FEC matched to channel statistics, link training, and protocol-layer recovery, remain the domain of SerDes and PHY vendors and are accommodated at the Link Profile boundary (§ Link Profile Framework).

Production CPO deployments have demonstrated 5× reliability improvement over pluggable optics (2.6M hour vs 0.55M hour MTBF).[1] This improvement comes from eliminating the pluggable connector cycling, reducing the number of components in the optical path, and integrating the OE in a controlled, hermetic package environment.

At cluster scale (100,000+ links), this reliability improvement translates directly to reduced operational disruption. A 5× MTBF improvement means 5× fewer link failures per year, 5× fewer maintenance interventions, and 5× fewer workload interruptions from link-related faults. Combined with the front-accessible service model and single-link fault-isolation properties described in § Serviceability Through Architecture, reducing the maintenance burden proportionally.

The Class V rail-switch option should be evaluated against the same failure-mode baseline as rear switch-card designs. Switch ASIC failure already exists in the dense-rack architecture; moving the ASIC from a rear card to a front-access compute tray changes replacement workflow and blast radius, not the existence of the failure class. Because the copper midplane is removed, the highest-blast-radius passive electrical component is replaced by per-link fiber paths, connectorized breakpoints, and replaceable tray-level active elements.

A natural extension for the community is a shared specification of per-element loss budgets, source-stability requirements, and component aging models indexed against the candidate Link Profiles. Common metrics here would let operators compare

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

reliability outcomes across implementations on a consistent basis, paralleling the role that COM plays for placement selection.

10.2. Serviceability Through Architecture

The Optical Faceplate and ODF architecture described in (§ Optical Faceplate, § Optical Distribution Frame) provides a serviceability model that improves on both pluggable optics and copper midplane approaches:

Table T: Serviceability

Service Action	Copper Midplane	Pluggable Optics	This Architecture
tray replacement	Extract from midplane (high risk, canister-wide impact)	Pull tray, reseat transceivers	Disconnect at ODF, pull tray, reconnect (front access, single-link scope)
Link fault isolation	Difficult (trace buried in PCB)	Swap transceiver	Test at ODF break point (no tray extraction)
Topology change	Impossible (fixed PCB traces)	Re-cable at top of rack	Re-patch at ODF (front access)
Failure blast radius	Entire canister (e.g., 36 - 72 XPU packages)	Single transceiver	Single link or single OE
Light source service	N/A	N/A (integrated)	ELSFP hot-swap (front access, no tray impact)

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Cold-plate accommodation of fiber egress is a platform integration constraint; see § Platform Integration Requirements Coverage for the design consideration.

10.3. Economic Benefits

The economic case for this architecture is distributed across the document where each mechanism is defined, rather than asserted as a separate benefits list. Each item below identifies the canonical section for:

CAPEX — Midplane elimination and supply-chain breadth (§ Fiber Raceway and Optical Distribution Frame): Replacing a custom 40+ layer controlled-impedance copper midplane PCB with passive fiber and a sheet-metal raceway reduces per-rack material cost, broadens the supply chain from a few qualified PCB fabs to the global fiber and connector industry, and eliminates the generational midplane redesign cycle.

CAPEX — Density improvement (§ Platform Form Factor Evolution): Higher XPU density per rack reduces the per-XPU share of rack, power delivery, cooling, and floor-space cost. The Class V canister target of 288 XPUs per rack against a Class H baseline of 72 is illustrative of the 3–4× range that current implementations are pursuing.

CAPEX — Scales of economy: Best-in-class requirements narrow the supplier field, so volume per qualified part rises and unit cost falls. Competition on each remaining requirement keeps the technical bar moving, and customers gain both supply-chain choice and the benefit of that volume.

OPEX — Interconnect energy (§ COM and Placement Selection): Effective pJ/bit is the integrating metric for placement-class selection. Movement from pluggable optics (15–20 pJ/bit) toward co-packaged classes (toward ~5 pJ/bit) is what recovers the per-rack network power quantified in § Power Efficiency at Scale.

OPEX — Generation portability (§ Integration by Reference): The fiber plant, raceway, ODF, and faceplate installed for one generation serve the next. The upgrade cycle

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

replaces trays, not infrastructure, amortizing the passive plant across multiple compute generations.

OPEX — Maintenance burden (§ Reliability Improvements, § Serviceability Through Architecture): The reliability and serviceability gains developed earlier in this section jointly reduce the frequency, duration, and skill level required for maintenance interventions.

10.4. Diagnostic and Fault Isolation: COM-to-RAS Runtime Closure

Robust diagnostics are another important design consideration. Fast and accurate diagnostics help identify what needs to be repaired and can significantly shorten the deployment and repair cycle. Diagnostics also provide valuable insight into the root cause of a failure. In addition to locating failed components, diagnostics can identify link elements that have degrading performance.

The architecture should therefore expose sufficient telemetry to quickly determine whether a failure originates in the electrical interface, OE, fiber path, ODF connection, external laser source, or remote endpoint. For example, minimum diagnostic indicators should include optical transmit and receive power, laser bias/current/temperature, wavelength, preFEC BER, FEC corrected and uncorrected error counts, link flap history, loss of lock (LOL) and loss of signal (LOS).

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

11. System-Level Impact on AI Workloads

11.1. MFU Improvement Path

Model FLOPS Utilization (MFU) is the ratio of observed floating-point throughput to theoretical peak. Today's best-reported MFU for large model training is 38–43%. The losses come from three sources, all interconnect-related:

- Communication overhead: Gradient synchronization, parameter sharing, and activation transfer consume XPU cycles waiting for data. Larger scale-up domains reduce the frequency of these stalls.
- Activation recomputation: When scale-up BW is insufficient, systems recompute activations rather than store and retrieve them. Photonic BW extension eliminates the need for recomputation in many cases.
- Topology-induced serialization: When the scale-up domain is smaller than the model's parallelism requirement, pipeline stages serialize. Larger domains enable wider parallelism.

The sensitivity of training throughput to scale-up domain boundaries is especially pronounced for Mixture-of-Experts (MoE) architectures, which are becoming the dominant design pattern for frontier models. MoE training requires all-to-all communication during expert routing, a traffic pattern that is acutely sensitive to domain boundary crossings. Published simulation results[20] of 4.7T-parameter MoE training on 32,768 GPUs demonstrated 1.55–2.71× training time improvement when a single 512-GPU optically-connected scale-up pod replaced four separate 128-GPU electrically-connected pods, with the largest gains observed at high expert counts (256 experts) where all-to-all communication dominates total training time. These results suggest that photonic HBD extension produces its greatest workload-level returns precisely where the industry is heading architecturally.

By extending the scale-up domain from 72 to 1,024–2,304+ XPUs at full BW, the architecture addresses all three sources. The target MFU of 55–65% represents a

[20]Published simulation results, IEEE Symposium on High-Performance Interconnects (Hot Interconnects), August 2025. 4.7T-parameter MoE model on 32,768 GPUs.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

30–70% improvement in effective utilization, getting more useful work from the same hardware investment.

11.2. Power Efficiency at Scale

The power analysis from § Rack-level interconnect energy savings illustrates the scale of the opportunity. In a dense rack architecture with >=144 XPU packages (at confirmed rack power levels of 600 kW or more), the electrical network can consume a modeled ≈37% of total system power at the pluggable baseline, depending on topology choices. Photonic interconnect at 2.3–5 pJ/bit (APO; compared to 15–25 pJ/bit electrical) reduces the network's share of system power to a modeled ≈15%, recovering on the order of 100 kW per rack[21] for additional compute or reduced facility power. At data center scale (thousands of racks), this represents megawatts of power savings.[22]

11.3. Photonic HBD: Extending the High Bandwidth Domain

Photonic interconnect enables a regime where physical density and logical disaggregation are independent design knobs, not competing ones. The density side was developed in § Platform Form Factor Evolution. This section takes up the logical-disaggregation consequence: how far the high-bandwidth domain can extend once distance stops tapering bandwidth.

The OCP OSAI whitepaper defines the High Bandwidth Domain (HBD) as the set of resources reachable at full memory-semantic BW.[23] Today, the HBD is bounded by copper reach to a single rack or fraction of a rack. With photonic interconnect, the HBD extends naturally:

Table U: Photonic HBD

[21] illustrative model result
[22] NVIDIA 800V HVDC architecture blog: supporting 1MW+ IT racks.
[23] OSAI Whitepaper v1.0.0 (OCP, Jan 2025): 576 accelerators per rack as geometric scaling target.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Configuration	Interconnect	Domain Size	Racks	Reach
Copper baseline	Electrical	72 XPUs	1–2	~2m (224G)
CPO	Optical scale-out	72–576	1–8	>500m
APO	Full optical	576–1,024+	1–16+	>500m

Placement regimes here map onto the platform roadmap in § Three-Phase Platform Evolution: CPO anchors photonic scale-out; APO denotes the area-domain regime that dense scale-up and volumetric configurations require.

The photonic fabric provides uniform BW independent of distance within the rack: any XPU can access any memory pool, any switch port, or any other XPU at full BW, regardless of physical position. This link-layer property is what permits the HBD to extend coherently across canister and tray boundaries rather than tapering at them. In distributed rail-switch Class V systems, this HBD can be realized as a dragonfly-like single-tier logical fabric rather than as a separate rear-card networking tier. A rack becomes a local group with dense tray-to-tray links; selected rail-switch ports become global optical links to other racks in the same HBD. The fabric is all-to-all at the logical level while the physical topology remains radix-efficient: local and global link counts are sized from the XPU injection BW, rail-switch radix, wavelength-scaled fiber capacity, and the allowed oversubscription target.

Memory disaggregation becomes architecturally natural in the photonic HBD. XPU-to-remote-memory links are physically identical to XPU-to-XPU links—they use the same fiber infrastructure, the same raceway, the same ODF. The system designer simply assigns some tray positions to memory trays instead of compute trays. Protocols such as CXL®.mem can provide cache-coherent access to these disaggregated memory pools, with the photonic link providing the reach and BW that copper cannot. Appendix E quantifies the resulting access-time and pooling gains.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

The same substitutability principle generalizes: any tray position may be assigned to compute, memory, NIC, or mixed role classes provided the link contracts of the placement-class taxonomy (§ Placement Class Taxonomy) are satisfied. Role composition becomes a deployment-time choice rather than a design-time commitment.

11.4. Data-Center Roll-Up at the Scale Boundary

This subsection rolls up power, domain size, and the binding physical constraint across publicly-announced scaling trajectory of accelerator platforms, with the goal of making the data-center-level consequence of the architecture explicit. The roll-up uses NVIDIA NVL-class platforms as concrete dimensional anchors because they are the only publicly detailed reference platforms at these scales as of early 2026. The architecture is protocol-agnostic at the Link Profile contract layer (§ Link Profile Framework): the same passive fiber plant and channel budgets serve NVLink™, UALink™, UEC/SUE, PCIe®, and any other electrical protocol that meets the Link Profile envelope. The 1,024-accelerator domain ceiling defined by UALink™ 1.0 [7] sits at the same scale point as the Feynman-class NVL1152 entry below.

Three-tier system power, anchored at the Class V dense-rack operating point (144 packages, on the order of 1.0–1.5 PB/s aggregate scale-up BW, 600 kW rack envelope per § Power Efficiency at Scale and Rack-level interconnect energy savings), partitions into compute (XPU plus HBM, the workload that earns the budget), switching (the electrical fabric, which at flat-bisection targets consumes a modeled ≈37% of total system power), and the photonic backbone (which scales the domain at under 2.3-5 pJ per bit). Substituting the photonic backbone for pluggable optics in the scale-up path compresses network share from ≈37% to ≈15%, recovering the per-rack power quantified above. At the smaller Class H per-rack operating points, the same accounting yields proportionally smaller absolute recoveries; at Feynman-class multi-rack scale, the binding constraint shifts to photonic radix rather than power share, as detailed below.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table V maps the architectural framework across four publicly-announced rack scale anchors spanning the 2024–2028 trajectory, ordered by package count. Each anchor binds on a different physical axis, and one set of Link Profile contracts and one passive fiber plant span all four. Scale-up bandwidth is quoted per GPU package and per-rack aggregates are package bandwidth × package count (GPU scale-up only); package counts are physical packages, not die-counted.

Table V: Four-anchor scaling rollup.

System	AMD MI450 / Helios	NVIDIA GB200 NVL72	Kyber 2-canister	Kyber 4-canister
Class / form factor	Class H · ORW	Class H · Oberon	Class V · canister	Class V · canister
GPU packages	72	72	144	288
Domain size (GPUs)	72	72	144	288
Switch tier	1	1	2	2
Scale-up fabric / medium	UALink™ / IF, copper	NVLink™, copper	NVLink™, copper	NVLink™, copper
Scale-up BW per GPU	3.6 TB/s	1.8 TB/s	3.6 TB/s	3.6 TB/s
Per-rack scale-up aggregate	~260 TB/s	~130 TB/s	~520 TB/s	~1.04 PB/s
Memory per GPU	432 GB HBM4	192 GB HBM3E	~512 GB HBM4E	~512 GB HBM4E
Memory BW per GPU	19.6 TB/s	8 TB/s	~26 TB/s	~26 TB/s

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

System	AMD MI450 / Helios	NVIDIA GB200 NVL72	Kyber 2-canister	Kyber 4-canister
Rack power	~120 kW+ (est.)	~120–140 kW	~400–500 kW (est.)	800 kW – 1 MW
Scale-up energy regime	Copper, ~10–15 pJ/bit; stressed	Copper, ~5–10 pJ/bit; sufficient	Pluggable optics required at flat bisection, 15–25 pJ/bit	Photonic-class; copper midplane infeasible
Network share, electrical	~11–16%	~4–9%	~31%	~31%
Network share, photonic 2.3-5 pJ/bit	~5–7%	~3–4%	~15%	~15%
Power recovered (elec→photonic)	~15–30 kW	~5–10 kW	~65–130 kW (est.)	~130–260 kW (est.)
Binding physical constraint	Copper reach at next-gen rates	None	Copper midplane failure at flat bisection	Escape radix — fiber per switch ASIC
Photonic adoption driver	Reach + retimer-latency elimination; scale-out pluggable → CPO	Generation-portable fiber plant; retimer-latency elimination	Feasibility within rack power envelope	Photonic radix scaling

Columns 3–4 model a two-die Rubin Ultra in the Kyber canister — R300-class I/O and memory (HBM4E) on R200-class cores — rather than the four-die R300 (144 packages / 576 dies). We adopt this given foundry- and packaging-community commentary on challenges still to be resolved at four dies: HBM4 validation timelines, ~4 kW four-die thermal envelopes, and large-die yield and CoWoS-L packaging. Memory reflects roughly half the four-die

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

R300 target (1 TB HBM4E, 16-stack), itself an aspirational capacity. Package counts are physical accelerator packages (18 trays × 4 = 72 per canister), not die-counted. The four-canister rack (288 packages) is a 2× architectural-maximum extrapolation beyond the announced single Kyber rack (144 packages); Class V power and recovered-power figures are estimates on this basis.

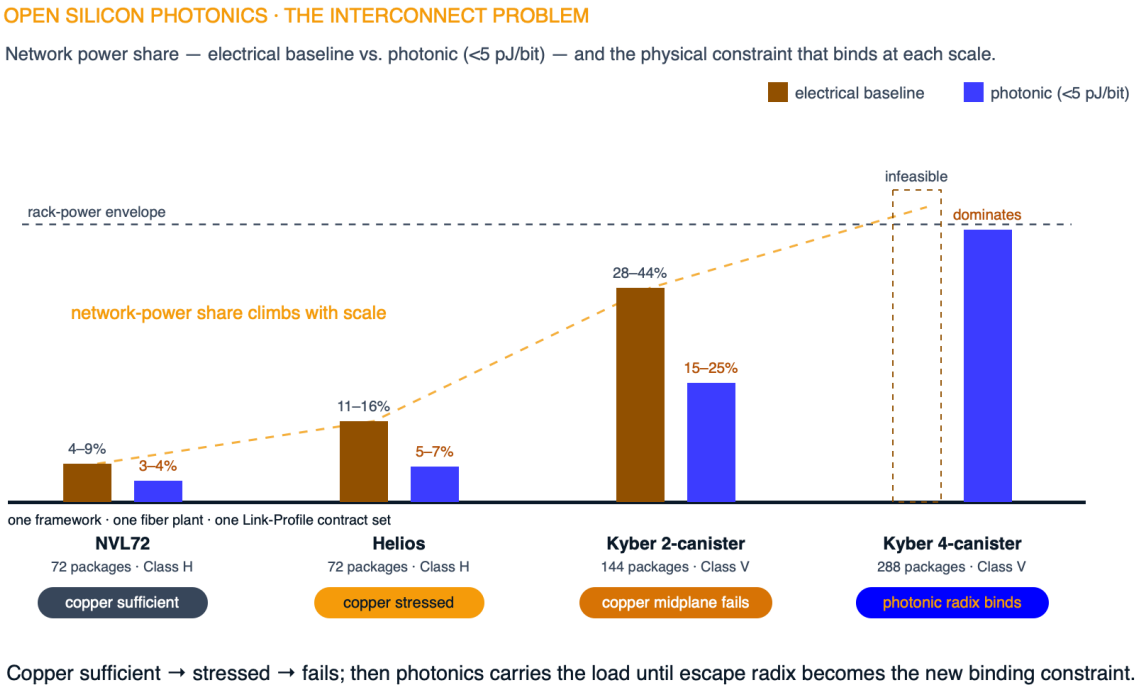


Figure 11.1. *Binding-constraints across the four scale anchors.*

A small-multiples sample strip of four panels (MI450/Helios, GB200 NVL72, Kyber 2-canister, Kyber 4-canister). Each panel plots network power share as paired bars, electrical baseline versus photonic at 2.3-5 pJ/bit, with the binding physical constraint named beneath: copper-stressed, copper-sufficient, copper-fails, photonic-radix-binds. One continuous baseline rule spans all four panels to mark the single framework. The figure shows at a glance the constraint migration that Table V carries row by row

Per-GPU BW values follow the public NVLink™ scale-up convention (bidirectional per GPU/die). The four anchors are sized against the most aggressive publicly-disclosed requirement set in each scale category as of early 2026. The architecture closes against the worst-case demand at each scale point; less demanding platforms, including AMD MI-class on UALink™, Intel Gaudi-class, and custom hyperscaler accelerators, sit inside the envelope by construction and do not require separate columns. The architecture's

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Link Profile contracts (§ Link Profile Framework) are protocol-agnostic and serve NVLink™, UALink™, UEC/SUE, PCIe®, and any other electrical protocol that meets the envelope. Network-share percentages at the electrical baseline are derived from per-rack aggregate BW and a 15–20 pJ/bit pluggable-optics estimate; with APO at 2.3-5 pJ/bit, the share is scaled accordingly. The Class V single-rack column carries the network-share figures from § Power Efficiency at Scale. See § The Density-Bandwidth Dilemma for the document-wide package-level convention

Per-rack BW density rises approximately tenfold across the four anchors, and absolute power recovery scales with it. Class H operates in the regime where copper still meets requirements and the photonic adoption case rests on generation-portable fiber plant and elimination of retimer-stage latency. Class V at single-rack density is where the copper midplane fails outright and photonic interconnect is required for flat bisection; this is also where absolute kW recovery is largest. At Feynman-class multi-rack scale, the binding constraint shifts again, from copper reach to photonic radix: fiber escape density per switch ASIC package becomes the dominant physical limit, and OCI Phase 2 timing relative to platform ramp becomes the dominant roadmap variable. A single architectural framework composed of channel budgets, placement classes, and Link Profile contracts, spans all four regimes.

Across the four anchors, the architectural delta resolves to two foregrounded quantities and one structural observation. Domain scale grows from 72 to 1,024–1,152 XPUs, a factor of approximately fourteen to sixteen, with the HBD spanning eight dense racks coherently rather than tapering at canister or rack boundaries. Network power share moves from ≈37% to ≈15% at the Class V single-rack operating point, with absolute recovery on the order of 100 kW per rack at this density and proportionally smaller absolute recoveries at lower per-rack BW points. The binding physical constraint migrates monotonically across the scaling axis from copper-sufficient, copper-stressed, copper-fails, through photonic-radix-binds, and each transition is addressed by the same framework rather than a new one. Workload-level useful-work improvements

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

(MFU, training time, inference throughput) are downstream of these structural changes and are workload-dependent; the MoE training and MFU gains developed earlier in this section are illustrative of the same mechanism.

Copper bounds the rack. Photonic radix bounds the campus. The open contract bounds the cost.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

12. OCP Platform Integration

MHS scope expansion. The Modular Hardware System (MHS) project has historically focused on CPU and PCIe®-class I/O; high-BW network I/O (200/400G Ethernet and InfiniBand) has been the domain of OCP NIC, which sits outside MHS by design. AI accelerator and switch platforms now require 200/400G interconnect at densities and placement classes that fit neither pattern — these signals generally do not attach directly to CPUs, and they exceed what OCP NIC was scoped to address. The architecture in this paper is positioned to extend MHS to encompass additional Scale-Up and Scale-Out elements including GPU/XPU and switch compute classes, with the proposed MHS-IAI workstream as the formal vehicle for that extension. The platform integration requirements below and the modifications in § Proposed Modifications to ORv3/MHS are scoped accordingly.

Table W: Platform Integration Requirements Coverage

Requirement	Architecture treatment	Reference
Power	Bus-bar interfaces, ELS module provisioning, FRU power coordination.	§ Passive Optical Infrastructure Components · § Proposed Modifications to ORv3/MHS
Cooling	Cold-plate accommodation, fluid coupling, OE thermal envelope, photonic thermal contract. Cold-plate geometry should accommodate both lateral (edge-coupled) and vertical (surface-coupled) optical egress without compromising thermal solution coverage; vertically-coupled FAUs require	§ Front-Panel Module Thermal Envelope · § Rear Thermal Envelope · § Photonics

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Requirement	Architecture treatment	Reference
	interference-free mounting (apertures, channels, or discontinuous plate geometry). **Crucially, the cold-plate profile must accommodate the potentially elevated Z-height of substrate-mounted vertical interconnects—such as OpenCPX and Co-Packaged Copper (CPC) flyover connectors sitting directly on top of a BGA—to eliminate physical interference and tolerance stack-up risks under monolithic cold plates.** Edge-coupled FAUs and copper cable connectors may sit outside or beneath the cold plate where connector tolerance to clamping force and local thermal gradient permits.	Thermal Supply · § Optical Engine Classes
Telemetry	OCP RAS API endpoint conformance, OIF ELSFP CMIS conformance, OCP Hardware Management Specifications alignment, link-quality reporting, alarm taxonomy.	§ Envelope and Contract · § Diagnostic and Fault Isolation
I/O	Optical Faceplate, ODF, Raceway, Rear Interface Zone, blind-mate provisions.	§ Passive Optical Infrastructure Components
Timing	Reference-clock and timestamp path preservation; OCP TAP coordination for photonic-native reference sources.	§ Envelope and Contract · § Proposed Working Group Cooperation
Firmware	OIF CMIS and OCP MHS and OCP OPF specifications alignment. OE class qualification,	§ Optical Engine Classes · §

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Requirement	Architecture treatment	Reference
	firmware update path, signed image requirements, rollback behavior (further specification required).	Proposed Working Group Cooperation
Security	Eye-safety interlocks for ELS; broader link-layer security (management-plane authentication, tamper detection, credential provisioning) requires further specification.	§ Passive Optical Infrastructure Components · § Proposed Working Group Cooperation
Environmental	Vibration, contamination, EMI, operating-temperature range.	§ Reliability Improvements

Firmware lifecycle. OE class qualification establishes firmware acceptance criteria at design time, but the full firmware lifecycle — secure update path, signed image enforcement, version reporting through the OCP RAS API, and rollback behavior under fault — requires further specification. This work composes with existing OCP firmware specifications applied to the photonic OE as a managed component class.

Security. Eye-safety interlocks for the ELS modules are addressed in the Optical Services Zone treatment (§ Optical Distribution Frame). Broader link-layer security — management-plane authentication, optical-link tamper detection, and secure provisioning of placement-class qualification credentials — requires coordination with OCP security working groups and Optical Distribution Frame (ODF / Fiber Shuffle)s not specified by this architecture.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

12.1. Proposed Modifications to ORv3/MHS

As described in § Integration by Reference, this architecture extends existing OCP specifications by reference rather than re-specifying them. The following summarizes what changes and what does not for each reference platform:

Unchanged across both platforms: Power bus bars, liquid cooling manifolds and quick-disconnects, rack mechanical dimensions, management networking, tray mechanical form factors (1U/2U/4U), and tray power/cooling interfaces. All existing MHS, ORv3, and OSAI definitions are retained.

Relative to OCP MGX Accelerated Computing Rack and Trays Specification (NVIDIA contribution): The rack envelope, tray form factors, power delivery, and cooling provisions defined in OCP MGX [15] for accelerated computing platforms are preserved without modification. Photonic-specific additions appear at the tray faceplate and behind it as the Optical Faceplate, ODF (with optional ELS module population), Raceway, and optional Rear Interface Zone — none of which alter the MGX mechanical or thermal contracts. The HPM Configuration Progression in § Reference Platform Scope illustrates the configuration tiers at which MGX provisions are invoked.

Added for the Compute (XPU) Reference Platform: Optical Faceplate specification at tray/canister front panel (modular connector grid for heterogeneous OE link classes), ODF/Shuffle specification (positioned behind faceplate, topology routing and fault isolation, with optional ELS module population), Raceway specification (fiber channels from OE to ODF), optional rack-level ELS shelf position reservation (for SiPh-ELS class deployments at densities exceeding tray-level zone capacity), and link monitoring integration for optical power telemetry. The compute platform extends MHS tray definitions and OSAI compute tray functional requirements. A Rear Optical Interface Zone position reservation is also added optionally, providing blind-mate scale-up egress, alternative shuffle topologies, and tray-level hot-swap.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Added for the Switch Reference Platform: Optical Faceplate specification (potentially different connector mix and density than compute, optimized for high-port-count switch ASICs), ODF/Shuffle specification (with optional ELS module population), Raceway specification, and optional rack-level ELS shelf position reservation. The switch platform extends MHS switch tray definitions and OSAI HBD switch tray functional requirements. Switch-specific considerations include higher port count per tray, more uniform OE class requirements, and alignment with existing CPO switch deployment experience. A Rear Optical Interface Zone position reservation is also added optionally; § Case Study: The Inter-Canister Tapering Problem describes one specific instantiation for Class V rear-mounted switch placements.

Added for the Class V Canister Reference Configuration: The Class V canister envelope reuses ORv3 by reference and adds one new mechanical entity (the canister); ORv3 rails, bus bars, manifolds, management, and MHS conventions on KOZ, datum, chassis-hook interfaces, and connector placement are inherited unchanged. The Optical Faceplate, ODF, Raceway, and ELS provisions defined above for the compute reference platform apply identically, with raceway routing oriented vertically along the canister front edge. Each canister hosts two stacked rows of 18 tray-slot positions; up to two canisters per rack occupy contiguous OU windows separated by an inter-canister fiber routing zone. Slot pitch is nominally 26–28 mm.

Two card depth designators are defined for Class V canister populations. V350 places two back-to-back trays per slot position (a front-loaded tray and a rear-loaded tray at ~350 mm depth each, separated by a passive photonic coupling zone replacing the copper midplane), yielding up to 72 trays per canister; the Kyber-style population is 36 compute front + 36 switch rear, giving 72 compute trays and 72 switch trays per rack at two canisters per rack. V350 fits the existing MHS Class L envelope at L350 (within the current 250–365 mm depth range). V700-class places one deeper tray per slot position (~700–800 mm front-loaded), with on-tray rail-switching absorbed onto the deeper compute tray and the rear of the canister reserved for passive fiber egress; the canister

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

presents only a front service surface, matching the Class H serviceability vector. V700-class yields 36 compute trays per canister, 72 per rack, and exceeds the current MHS Class L 365 mm depth maximum — therefore requiring a new MHS HPM class with extended depth envelope (~800 mm), proposed here as a joint work item with the OCP MHS workgroup. Both designators inherit MHS conventions on KOZ, datum, chassis-hook interfaces, and connector placement; the numeric depths are illustrative pending Design Specification iteration.

Per-tray content follows the HPM Configuration Progression in § Reference Platform Scope: V350 compute trays at the Dense-class tier (4 XPU + 1 CPU per tray; rail-switches occupy rear positions of the same slot) and V700-class trays at the Ultra-dense tier (8 XPU + 2 CPU per tray with on-tray rail-switch ASIC) at a nominal 1:4 CPU:GPU ratio. Per-XPU footprint follows a 3D CPO ↔ NPO area equivalence: R-Ultra-class packages with 3D CPO and R200-class packages with NPO both converge on roughly 25–30 k mm² per XPU including FAU perimeter and cold-plate border, with the OpenCPX MSA socket envelope as the worst-case sizing reference. CPU↔GPU links use short-reach copper (NVLink™-C2C class) by default and do not consume OE-class allocation or FAU perimeter on the CPU package; the per-tray OE budget is therefore GPU-driven at this baseline, which halves potential per-tray fiber count and aligns the OE-class allocation with scale-up domain participants. At V350 the per-tray area headroom preserves the option for an optical CPU where the integrator requires CPU participation in an optical fabric; at V700-class the tray content density (8 XPU + 2 CPU + on-tray rail-switch ASIC against ~315 k mm² PCB) binds the CPU to copper-only.

Per-rack XPU count: V350 = 288 XPU/rack (72 compute trays × 4 XPU), V700-class = 576 XPU/rack (72 compute trays × 8 XPU). The V700 × 3-rack photonic scale-up domain reaches approximately 1,728 XPU; inter-rack BW is flattened by photonic reach and high-radix rail-switching rather than tapered as it would be with copper interconnect — this is the inter-canister tapering argument of § Case Study: The Inter-Canister

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Tapering Problem raised one level. XPU count uses a vendor-neutral per-package convention; vendors using multi-die packages should multiply by per-package die count to align with vendor marketing nomenclature.

12.2. Standards Alignment

The architecture described in this white paper consumes specifications from multiple standards-development organizations and contributes new specifications only where existing work does not cover the relevant scope. Figure 12.1 maps each layer of the hyperscaler build flow to its primary standards owner and identifies boundary cases where shared ownership is required, including optical management, CPO thermal integration, photonic KGD, and rack-level fiber serviceability.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

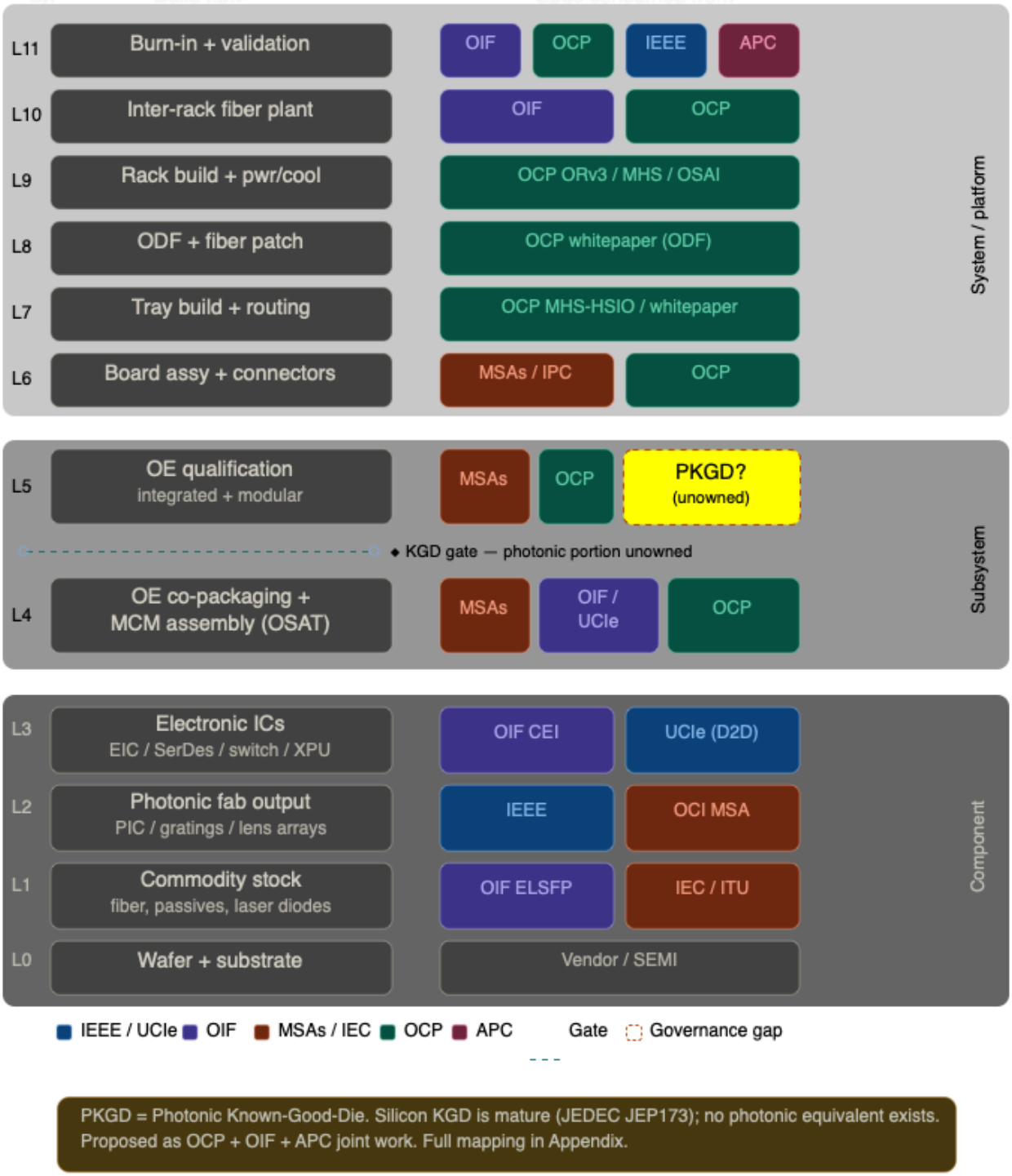


Figure 12.1. *SDO Governance Map across the hyperscaler build flow.*

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

The stack decomposes naturally into three tiers. At the component layer, IEEE 802.3 defines optical physical-layer performance, OIF defines electrical interface implementation agreements (CEI, EEI Framework, ELSFP), and MSAs (OCI, OSFP, SFF-TA) define component-level interoperability contracts. At the subsystem layer, OE co-packaging and module qualification combine MSA form factor specifications with OIF and UCIe® electrical interface definitions and OCP thermal and mechanical requirements. At the system layer, OCP specifications (ORv3, MHS, OSAI, and the proposed MHS-IAI workstream) own platform integration, tray architecture, rack-level build, and inter-rack fiber infrastructure. The OCP RAS API specification (v0.9) provides the management and telemetry conformance layer that the envelope-and-contract framework in § Envelope and Contract incorporates as the link-layer run-time observability dimension; OCP RAS API conformance is referenced as part of the management-and-telemetry contract dimension applied to every placement class.

The Optical Faceplate defined in § Optical Faceplate serves as the architectural decomposition point between these scopes. Specifications below the faceplate govern the OE, its packaging, its electrical interfaces, and its field-replaceable unit qualification, all of which are defined by MSAs, OIF, IEEE, and UCIe® and consumed as reference by this architecture. Specifications above the faceplate govern how the platform consumes the resulting optical subsystem, tray mechanical integration, fiber management, topology implementation, and rack-level assembly, all of which are within OCP's scope. This division preserves multi-vendor competition at the component and subsystem tiers while enabling open platform integration specifications where existing standards do not reach.

One layer on this map is not yet fully owned by any existing organization. Silicon known good die (KGD) methodology is mature, governed by JEDEC JEP173 and supported by well-established wafer probe and statistical yield practices. No equivalent methodology exists for photonic chiplets. Pre-attach verification of a photonic chiplet requires optical performance characterization (insertion loss, extinction ratio, responsivity, wavelength

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

accuracy), electrical signal integrity at the photonic interface, and optical coupling integrity at the grating or edge coupler, three axes that silicon KGD does not cover. Without a Photonic Known-Good-Die (PKGD) methodology, the financial risk of committing photonic chiplets to multi-die co-packaged assemblies falls entirely on the integrator, who will come to depend on the discovery of defective die through post-assembly tests. This white paper proposes PKGD as a joint work item between OCP, OIF, and APC, drawing on OCP's system integration requirements, OIF's optical compliance testing heritage, and APC's test methodology and plugfest competency.

The full governance map, including cross-cutting compliance gates, sub-layer qualification distinctions for different OE classes (3D CPO, 2D CPO, NPO, S-NPO/CPX, M-NPO, and FPO, with XPO as an FPO subset), and detailed SDO document references, appears in Appendix B.

IOWN Global Forum work which includes the AI Interconnect Functional Architecture (March 2026) and the Open All-Photonics Network (APN), operates above the substrate-physical-contract layer this whitepaper specifies, defining functional-architecture and network capability abstractions for AI cluster networking and inter-site optical transport. The OCP-IOWN AI Computing Continuum cooperation framework (February 2026) formalizes this division of labor: OCP-driven open hardware specifications composing with IOWN-driven communications architecture across centralized through edge deployments. The substrate-level physical contracts defined here are intended to compose into that framework as the open hardware building block.

The architecture aligns with and references the following external specifications. Detailed mapping would be developed in the Base Specification:

- OIF: CEI-224G (SerDes electrical), EEI Framework Document (next-generation energy-efficient interface study, covering 448G+ applications), ELSFP-02.0 (external laser form factor), Co-Packaging Framework Document.
- IEEE: 802.3dj (200G/lane Ethernet PHY), 802.3 Beyond 400G Study Group.

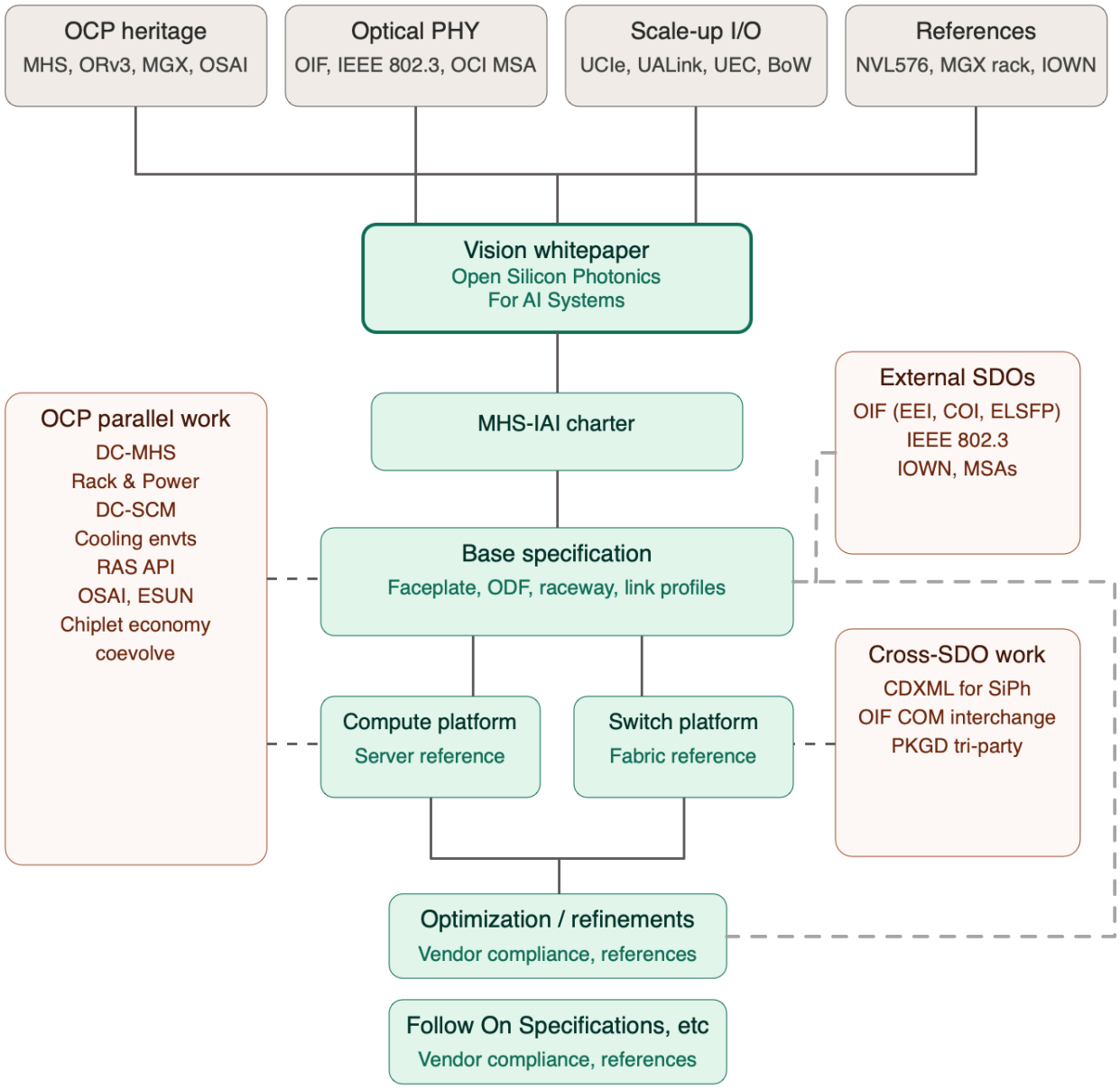
This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

- UCIe®: Universal Chiplet Interconnect Express® (die-to-die interface relevant to OE integration on interposer).
- CXL®: Compute Express Link (memory disaggregation and resource pooling over photonic links).
- UALink™: Ultra Accelerator Link 1.0 (1,024-XPU scale-up domain specification).
- UEC: Ultra Ethernet Consortium 1.0 (optimized Ethernet for AI networking).
- OCP: ORv3, MHS, OSAI, Cooling Environment specifications.
- MSAs (Multi-Source Agreements): OCI MSA (optical compute interconnect PHY for AI scale-up), OpenCPX MSA (socket contract accepting OE or co-packaged copper bundle implementations), XPO MSA (MSA-defined subset of faceplate-pluggable OEs for optical format conversion), CW-WDM MSA (external CW laser source specifications for co-packaged and remote laser architectures), CWDM4 MSA (the 4x25 Gb/s datacenter optics MSA), EB MSA (open blind-mate expanded-beam connector specification).

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

13. Proposed Deliverable Structure: Informative Vision, Normative Contracts, and Validation Profiles

This white paper is the first deliverable in a proposed series aligned with OCP's Modular Contribution Architecture.



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

***Figure 13.1**. Relationship between this white paper, future efforts, and external groups.*

The contribution strategy follows the Base / Design / Product specification hierarchy to maximize community participation and vendor differentiation:

Table X: Deliverables

Deliverable	Content	IP Level	Status
White Paper (this document)	Vision, goals, architectural concepts, degrees of freedom, measurable targets	No IP — conceptual alignment	Draft v1.0 — open for collaborator input
Base Specification	Interface contracts for both Compute and Switch reference platforms: Faceplate grid, ODF requirements, Raceway envelope, Link Profile definitions, RAS requirements. Integration by reference of MHS, ORv3, OSAI, Cooling	Light IP — coarse alignment, goals, use cases	To be developed collaboratively
Design Specification(s)	Prescriptive implementations: Class H compute tray, Class V canister compute, CPO switch tray, specific OE	Medium IP — fine alignment, design-specific	Community contributors extend Base Spec

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Deliverable	Content	IP Level	Status
	integration guides per platform; Class V distributed rail-switch compute tray option (V350 Kyber-style back-to-back and V700-class front-only canister populations per § Proposed Modifications to ORv3/MHS; V350 = Dense-class 4 XPU + 1 CPU/tray, V700-class = Ultra-dense 8 XPU + 2 CPU/tray + on-tray rail-switch)		
Product Specification(s)	Exact BOM, design files, compliance test plans for specific implementations	Medium/High IP — implementation-specific	Vendor contributions

The Base Specification is the critical next deliverable. It would formalize this white paper's architectural concepts into testable interface contracts for both the Compute (XPU) and Switch reference platforms. Each platform section of the Base Spec would extend applicable OCP contributions (MHS, ORv3, OSAI, Cooling Environment) by reference, adding only photonic-specific interface definitions. This integration by reference approach ensures that the photonic architecture inherits the maturity and

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

community validation of existing OCP specifications. The Base Spec should be developed collaboratively through the relevant OCP projects, with broad ecosystem input.

13.1. Compliance as an Ecosystem Accelerator

An open specification becomes deployable only when conformance can be demonstrated repeatably, independently, and without bilateral negotiation between every pair of suppliers and integrators. The Base Specification should therefore include compliance requirements as first-class normative content, not as a certification afterthought. Each interface contract should identify the required evidence of conformance, the applicable test method, the reporting format, and the lifecycle conditions under which conformance remains valid, including firmware updates, component substitutions, environmental limits, and field-service events.

This model doubles as an ecosystem accelerator. OCP has precedent in programs such as S.A.F.E., where approved assessors, community-developed test cases, and public-facing result artifacts reduce duplicated effort and increase trust across the supply chain. A similar model for open photonic AI interconnect could allow vendors, labs, system integrators, and operators to contribute tests, qualify implementations, and publish conformance summaries while protecting proprietary raw data and margin details. The goal is not to centralize implementation control, but to make qualification portable: a component, subsystem, or platform that satisfies the open contract should be recognizable as such across the ecosystem.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

14. Ecosystem Collaboration and Next Steps

The AI industry is at an inflection point where interconnect, not compute, is the binding constraint on system performance. The technology building blocks for photonic interconnect exist. What the ecosystem lacks is the open systems architecture that integrates these building blocks into standardized, interoperable platforms.
This white paper provides a starting point. It is explicitly incomplete and designed to be refined, challenged, and extended by the community. The following areas are identified for collaborative development:

- This architecture does not address *all* challenges required for production deployment.
- Significant companion work is needed in areas including: thermal management of OEs co-packaged with high-power XPU die, advanced packaging integration (2.5D/3D architectures, BoW, UCIe® interoperability), security of optical links and management planes, and foundry-level test and validation methodologies for photonic components at volume.
- Existing Standards Development Organizations (OIF, IEEE, CXL® Consortium) partially address these challenges; this effort should coordinate with them rather than proceed in isolation.

Alongside these technical challenges, ecosystem-level exchange formats are a precondition for vendor-neutral integration at scale. Three categories follow from the envelope-and-contract methodology in § Envelope and Contract: standardized component characterization metadata, vendor-neutral system-level link evaluation, and secure instance-level identity with attestation. Specification of these formats is properly the work of the appropriate standards bodies; this document identifies them as preconditions for the ecosystem this architecture targets rather than prescribing their form.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

14.1. Proposed Working Group Cooperation

Link Architecture (with OIF): refine Link Profile definitions, develop compliance test methodology, define interoperability requirements between OE classes, and ideal WDM modes, grids, and channels for CWDM and DWDM configurations. Evaluate folding IEEE DR to bidirectional to meet OCI's fiber count at IEEE-PHY maturity, a standards-anchored path to higher per-fiber density.

Laser Integration and RAS: Develop ELS module and rack-aggregation specifications, define laser redundancy and hot-swap requirements, quantify reliability targets for different OE classes. Develop OIF CMIS enhancements and/or define manageability frameworks with the OCP Server MHS and Hardware Management communities for optical link monitoring and fault reporting.

Platform Mechanical Integration: Develop Optical Faceplate, ODF, Structured Optical Cabling Plane, and Raceway specifications for both Class H and Class V platforms, including connector adapter standards, color/keying conventions for laser-delivery versus data-transport waveguides, and plane-level replacement requirements. Evaluate provisions for future Optical Circuit Switching (OCS) integration within the ODF position. Coordinate with the OCP MHS workgroup on a new MHS HPM class with extended depth envelope (~800 mm) to accommodate the V700-class designator for Class V canister populations per § Proposed Modifications to ORv3/MHS.

Thermal and Packaging: Address thermal management challenges specific to co-packaged and near-package optics, including thermal interaction between OEs and XPU die, cold plate design for photonic components, and packaging roadmap alignment with Bunch of Wires (BoW), UCIe®, and advanced 2.5D/3D integration. This work complements ongoing efforts at OIF and IEEE.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Compliance, Test, Validation, and Certification: Develop a four-tier test and measurement regime to enable open-architecture photonic ramp from prototype to high-volume manufacturing. Foundry tier: establish Photonic Known-Good-Die (PKGD) methodology to address the L5a gap identified in § Standards Alignment , the wafer-level optical characterization (insertion loss, extinction ratio, responsivity, wavelength accuracy), electrical signal integrity at the photonic interface, and optical coupling integrity at grating or edge coupler, jointly with OIF and APC and referencing JEDEC JEP173 as the silicon-KGD precedent. Component tier: extend per-OE-class qualification through interoperability test event frameworks paralleling OIF's CEI plugfest heritage, aligned with the qualification standards already in scope (Telcordia GR-468, JEDEC JC14.3 for silicon photonics, AEC for higher-reliability applications) and complementing OCI MSA digital diagnostics and OCP RAS API endpoint conformance.

Assembly tier: define post-attach verification methodology for co-packaged and near-package OEs, characterize installed links against the §C envelope (per-topology insertion loss, COM closure, FEC margin), and operationalize the envelope-compression mechanism that converts planning ranges into deployed-context specification values. System and field tier: certify run-time link observability against the COM–RAS coupling defined in § Diagnostic and Fault Isolation, including CPER-format event reporting, BER/PRBS checker conformance per OCI v1.0, alarm taxonomy, and fault isolation at ODF breakpoints. The regime treats the test-and-measurement ecosystem (instrument suppliers, characterization labs, plugfest hosts) as a named ecosystem-participant class and establishes the data-flow path by which deployed characterization compresses §C planning ranges into standards-body specification values over time.

Optical Connector and Fiber-Attach Development: Advance mated-transition performance at each optical boundary the architecture names: PIC fiber attach, tray-internal and faceplate terminations, blind-mate rear interfaces, and ELS delivery. At

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

the PIC attach the ask is a conjunction, not a loss figure alone: detachable, passively aligned, reflow-survivable interfaces holding the 1 dB per-facet planning value of Appendix C.2, with a stretch class toward 0.2 dB where physical-contact alignment physics permits. Faceplate and tray-internal terminations carry the same floor under density, mating-cycle, and field-cleaning constraints. Blind-mate rear interfaces form a separate declared class. Expanded-beam collimation buys mate float, contamination tolerance, and facet power-density headroom at an intrinsic IL premium, and the class is judged on that trade. The same headroom makes expanded-beam interfaces natural candidates for high-power ELS delivery. Every transition improved returns budget directly through the connector-count accounting of § Optical Link Composition: at N mated pairs, closing typical IL from 0.5 dB to 0.2 dB returns 0.3 × N dB to the link.

Precision Time and Reference Distribution (with OCP TAP): Coordinate with the OCP Time Appliance Project on photonic-native timing source form factors complementing the current PCIe®-attached Time Card, supporting reference-clock and timestamp distribution to fabric-attached photonic links. This work establishes the management/timing plane traversal pattern referenced in § Envelope and Contract.

Firmware Lifecycle for Photonic Components: Coordinate with OCP firmware specifications to define the firmware update path, signed image requirements, version reporting, and rollback behavior for OE-class managed components.

Security for Photonic Interconnect: Coordinate with OCP security working groups to address management-plane authentication, optical-link tamper detection, and provisioning of placement-class qualification credentials for photonic components.

System Integration and Workload Optimization: Develop system-level performance benchmarks, coordinate multi-vendor interoperability validation, and quantify MFU improvement through photonic HBD extension across representative AI workloads.

14.2. Invitation to Collaborate

We invite the following stakeholders to join this effort:

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

- Accelerator vendors (GPU, TPU, FPGA, custom ASIC designers)
- OE and photonic IC providers (SiPh, VCSEL, micro-LED, coherent)
- Laser source providers
- Connector and fiber cable manufacturers
- Advanced packaging and foundry providers (2.5D/3D integration, photonic IC fabrication)
- Test, measurement, and certification organizations
- Switch silicon vendors
- System integrators and OEMs
- Hyperscaler operators and cloud service providers
- Standards body representatives (OIF, IEEE, UCIe®, CXL® Consortium, UALink™, UEC)

Starting reference design contributions. The workstream invites donor contributions of reference designs for the Compute (XPU) and Switch reference platforms as concrete starting points from which the Base Specification can be derived. Donor contributions ground the envelope-and-contract framework in proven mechanical, thermal, and integration choices, and accelerate specification development. Any qualifying design generalizes into the open contract layer; donor IP flows outward through the OCP contribution model on equal terms with all other ecosystem participants.

> *The technology is ready. The systems are not. The Open Compute Project community is uniquely positioned to close this gap through collaborative specification development. This white paper is the starting point for that conversation.*

For more information or to participate:

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Please reach out to: OCP Server MHS project for the anticipated Interconnect for AI (IAI).

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

15. Conclusion

This whitepaper shows the path towards making silicon photonic infrastructure for AI buildable. We began with distance and light. The argument of this white paper has been that the conversation between our machines need not stay imprisoned in copper, and that the way to free it is not a single invention but a discipline — four of them, working together. A channel budget that maps the interconnect parameter space by measurement rather than by lore. A placement-class taxonomy that treats each integration boundary as a first-class decision made in the open, not a convention quietly inherited. A COM-driven selection framework that lets transparent analysis, rather than habit, decide where light should take over from copper. And an envelope-and-contract methodology that says what every implementation provides while leaving how it is built to the people who build it. Together these disciplines carry the High Bandwidth Domain past the edge of the rack and onto a passive fiber plant that outlives the SerDes generation that lit it — a generation-invariant substrate carrying scale-up, scale-out, and Scale-DSA traffic on one shared infrastructure.

What makes the path traversable is the discipline of separating specification from implementation — knowing the difference between describing a thing and dictating it. The architecture publishes the channel budgets, the placement classes, the contract dimensions, and the boundary transitions; the ecosystem populates the envelope with the implementations that meet the contracts. New realizations enter the architecture by meeting the contract, not by being predicted by the document.

This is a beginning, not an ending. The contracts drawn here are component-level, and the richest gains lie just past them. An ecosystem free to co-design across the ASIC SerDes, the OE, the channel, and the layers of manageability and RAS that tie them together reaches the system-level and cross-component opportunities no single component can see on its own. Co-design will reshape what lives behind the contracts many times over. The contracts themselves are meant to endure.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

The path is shown. The next step is to derive OCP open specifications from this whitepaper, starting with OCP AI Systems specifications as the base, and move the silicon photonics community towards an open ecosystem. The technology to let our machines speak in light is within reach. Whether they come to speak a common language is a choice we make in the open, together.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Appendix A: Lower PHY Architecture

The architecture's Link Profile contracts (§ Link Profile Framework) are defined at the aggregate port rate, not the per-lane electrical rate. A given aggregate port may be implemented through multiple SerDes lane-rate and lane-count combinations, consistent with the OCI multi-source agreement convention of 4× or 8× aggregation. At each generation, the architecture supports two characteristic implementation paths: a "fast-narrow" path, which uses the leading-edge per-lane SerDes rate with the fewest lanes, and a "slow-wide" path, which reuses prior-generation SerDes IP with a higher lane count. The preferred variant is the one that closes the channel budget at the lowest effective pJ/bit. When the fast-narrow variant closes the budget, it is preferred because it reduces fiber count, aggregate connector loss, and routing complexity. When it does not, the slow-wide variant becomes architecturally necessary, or the placement class should move inward.

The following table characterizes the parameter space across three SerDes generations. Numbers are illustrative across a representative chip I/O range; exact ports per ASIC depend on the published chip I/O target and the duplex counting convention used by the vendor.

Table Y: Bandwidth Matrix

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Generation	Platform class	Aggregate I/O per package	Per-lane SerDes (fast-narrow / slow-wide)	Per-port aggregate (example)	Optical link count per package	Per-fiber BW (16λ BiDi)	32-36 fiber tray-face aggregate
1	XPU scale-up	3.6 TB/s (28.8 Tb/s)†	224G / 112G	800G (4×224G or 8×112G)	~36	3.2 Tb/s	~114 Tb/s
1	Switch	51.2 T	224G / 112G	800G (4×224G or 8×112G)	64	3.2 Tb/s	~114 Tb/s
2	XPU scale-up	7.2 TB/s (57.6 Tb/s)†	448G / 224G	1.6T (4×448G or 8×224G)	~36	6.4 Tb/s	~228 Tb/s
2	Switch	102.4 T	448G / 224G	1.6T (4×448G or 8×224G)	64	6.4 Tb/s	~228 Tb/s
3*	XPU scale-up	14.4 TB/s (115.2 Tb/s)†‡	896G / 448G	3.2T (4×896G or 8×448G)	~36	12.8 Tb/s	~460 Tb/s
3*	Switch	204.8 T	896G / 448G	3.2T (4×896G	64	12.8 Tb/s	~460 Tb/s

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Generation	Platform class	Aggregate I/O per package	Per-lane SerDes (fast-narrow / slow-wide)	Per-port aggregate (example)	Optical link count per package	Per-fiber BW (16λ BiDi)	32-36 fiber tray-face aggregate
				or 8×448G)			

*Gen 3 illustrative; slow-wide path likely dominant until 896G SerDes matures. Modulation format may bifurcate by generation (PAM4, PAM6, PAM8, or future schemes); the aggregate-port contract is modulation-agnostic. †XPU scale-up aggregates consistent with publicly disclosed XPU roadmaps as illustrative scaling targets; ‡14.4 TB/s is the illustrative continuation of the 2× per-generation cadence rather than a roadmap-anchored figure.

Reach across generations: copper at the per-lane rate compresses from ~1–3 m (Gen 1) to <~0.5–1 m (Gen 2) to <~0.25–0.5 m (Gen 3); fiber reach is >500 m across all generations.

Two observations follow from the table. First, per-port aggregate and per-lane SerDes are identical across XPU and Switch platform classes at each generation — the architecture's contract framework applies uniformly, and only port count per package differs. Second, the 32-36 fiber tray-face aggregate is identical across the two platform classes per generation, which is the structural intent of the common passive infrastructure: a tray-face built for a switch in any given generation carries the equivalent of multiple XPU packages' scale-up I/O at the same BW, and vice versa.

The optical link count reports the number of independent Link Profile instances per package required to carry the aggregate I/O. For Switch packages this equals the chip's

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

switching radix at the optical interface; fabric radix at the network level is derived through topology (single-tier, folded Clos, multi-stage). For XPU packages this equals the count of scale-up optical links, which may align one-to-one with the protocol's link unit (e.g., NVLink™-class links) or may bundle multiple protocol links per optical link depending on implementation. Network fabric or scale-up domain size is a topology-dependent consequence of these counts, not equivalent to them.

Within a given chip I/O budget at each generation, optical link count and per-port aggregate are conjugate parameters: their product is fixed by the table above, but the split between them is the architect's choice. The same Gen 2 chip I/O at 102.4 T can be partitioned as 64 × 1.6 T (favoring per-port BW), as 128 × 800 G (favoring optical link count), or as 256 × 400 G; each lies on the same chip I/O contour but lands in different system territory. Workloads emphasizing large collective transfers (e.g., AllReduce in pre-training) typically benefit from fewer-and-fatter links because per-transfer serialization is reduced; workloads emphasizing destination diversity (e.g., MoE expert routing, sharded KV cache, all-to-all patterns) typically benefit from more-and-thinner links because fabric reach is exercised more often than per-link throughput. Connector and fiber-attach consequences follow accordingly: higher optical link count at lower per-port aggregate increases package-perimeter total attach (more independent link instances at the same per-port FAU contract); lower link count at higher per-port aggregate concentrates BW into fewer fibers at higher per-fiber BW (more wavelengths or higher modulation per fiber for the same Profile). Tray-face dimensioning at the Optical Faceplate and ODF ranges across this space. The architecture accommodates both directions through the envelope-and-contract methodology: per-port aggregate is the contract dimension; the split between link counts and per-port aggregates is the implementation choice within the supported generation.

Rail-switch sizing uses the same arithmetic. A compute tray with onboard rail switches allocates total switch I/O among four buckets: XPU injection BW, intra-rack local links, inter-rack global links, and any independent fabric planes such as Fabric A/Fabric B or

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Scale-DSA. The design task is therefore not simply to maximize fiber count; it is to choose the rail size, optical link count, and per-fiber wavelength density that close the target scale-up domain with acceptable hop count, BW taper, and failure-domain separation.

The arithmetic resolves to per-tray bandwidth and laser sizing across the configuration tiers of § Reference Platform Scope. Table Z instantiates four cases that span the Dense-class and Ultra-dense-class envelopes, with and without optical CPU participation, and reports the implied fiber count and ELS optical aggregate at the Gen 1 design point.

Table Z: Per-Tray Optical BW and ELS Sizing by Configuration Case

Case	Tray content	CPU↔GPU	Per-tray BW (Gen 1)	Gen 2	Gen 3	Fibers per direction (Gen 1)	ELS optical aggregate per tray (Gen 1 @ +10 dBm/λ)
1	4 GPU + 1 CPU (V350, Dense)	Copper	**7.2 TB/s** (4 × 1.8)	14.4 TB/s	28.8 TB/s	18	~1.4 W (~10 W wall-plug @ 15% WPE)
2	4 GPU + 1 CPU (V350, Dense, optical CPU)	Optical (+4 × 1.8 = +7.2 TB/s)	**14.4 TB/s**	21.6 TB/s	36.0 TB/s	36	~2.9 W (~20 W wall-plug)
3	8 GPU + 2 CPU (V700,	Copper	**14.4 TB/s**	28.8 TB/s	57.6 TB/s	36	~2.9 W (~20 W wall-plug)

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Case	Tray content	CPU↔GPU	Per-tray BW (Gen 1)	Gen 2	Gen 3	Fibers per direction (Gen 1)	ELS optical aggregate per tray (Gen 1 @ +10 dBm/λ)
	Ultra-dense)		(8 × 1.8)				
4	8 GPU + 2 CPU (V700, Ultra-dense, optical CPU)†	Optical (+8 × 1.8 = +14.4 TB/s)	**28.8 TB/s**	43.2 TB/s	72.0 TB/s	72	~5.8 W (~40 W wall-plug)

Per-tray BW is BiDi aggregate (GPU scale-up + CPU↔GPU C2C when CPU is optical); fiber count is per direction at 16λ BiDi 3.2 Tb/s/fiber; ELS aggregate assumes 8 Tx-side wavelengths per fiber at +10 dBm/λ delivered to the OE, mid-envelope of the OIF ELSFP-02.0 range (Table AG). Wall-plug shown at illustrative 15% WPE; production ELS modules vary. CPU↔GPU C2C link bandwidth is taken at 1.8 TB/s per CPU–GPU pair (NVLink™-C2C class). †Case 4 reflects an aspirational ceiling enabled by smaller-footprint CPU classes (e.g., ARM-class) freeing area on the V700 PCB; not supported by V700 as currently defined in § Reference Platform Scope and § Proposed Modifications to ORv3/MHS , included here to size the worst-case envelope for the Optical Faceplate and ELS provisioning.

Cases 2 and 3 land at the same per-tray BW (14.4 TB/s at Gen 1) by different routes: an optical-CPU V350 and a copper-CPU V700 produce identical laser-sizing pressure on the ELS and identical fiber-count pressure on the Optical Faceplate. The passive fiber plant and ELS provisioning size against the BW envelope; the configuration choice is downstream. This is the envelope-and-contract methodology of § Envelope and Contract applied to configuration-agnosticism in that the passive infrastructure does not

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

need to know which tray configuration sits behind the faceplate, only the bandwidth envelope it is to carry.

Fiber Count Derivation

The aggregate port rate determines port count for a given chip I/O; per-port lane count and WDM density determine fiber count per port. The derivation chain proceeds in five steps:

Chip I/O ÷ per-port aggregate → ports per ASIC

Lane count per port × (Tx+Rx multiplier) ÷ WDM channels per fiber → fibers per port

Ports per ASIC × fibers per port → fibers per ASIC

Fibers per ASIC × XPUs per tray → fibers per HPM

Fibers per HPM × HPMs per tray → fibers per tray-face

The Tx+Rx multiplier is 1 for BiDi WDM (Tx and Rx share fibers via wavelength separation in opposite directions) and 2 for non-BiDi configurations (Tx and Rx occupy separate fibers). WDM channels per fiber ranges from 1 (DR-style single-wavelength) through 4 (FR4) to 16+ (DWDM BiDi). The two multipliers compound: a profile that is both BiDi and dense-WDM minimizes fibers per port; a profile that is single-wavelength and non-BiDi maximizes it.

Worked example: 102.4 T chip I/O at 1.6 T per port = 64 ports per ASIC. Fibers per ASIC depends on profile and path selection:

Table AA: Fibers per ASIC by Profile (Worked Example)

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Optical configuration	WDM ch / fiber	Tx+Rx	Fibers / port	Fibers/ASIC, 64-port (UALink™)	Fibers/ASIC, 72-port (NVLink™)
8λ BiDi (OCI 4+4 class)	8	1	1	64	72
FR4-class (4λ/fiber)	4	2	4	256	288
2λ (FR-class)	2	2	8	512	576
Single-λ DR8 (slow-wide)	1	2	16	1,024	1,152

All rows realize a 1.6 Tb/s port as 8 wavelengths at 200 Gb/s; the dense row shares Tx and Rx bidirectionally on one fiber (OCI 4+4 class), the rest use separate Tx/Rx fibers. The per-ASIC totals are sized to coincide with familiar scale-up domain magnitudes — power-of-two (UALink™-class) and ×72 (NVLink™/NVL72-class) — so the fiber cost of trading WDM density for fiber count reads against domains the audience already sizes systems around. These are fibers per switch ASIC, not domain populations.

The 16× span across the profile × path matrix defines the dimensional envelope the passive infrastructure is to accommodate. The § Optical Faceplate , § Optical Distribution Frame (ODF / Shuffle) , § Fiber Raceway, and § Fiber Attach Envelope are each to be designed to tolerate the upper bound permitted by the supported profiles, even when implementations populate at the lower end. This is the envelope-and-contract methodology applied to fiber count: the architecture publishes the envelope; specific implementations select the profile and path within it.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Appendix B: Photonic Ecosystem Reference

The chart on the following page maps the photonic ecosystem build flow as a governance reference. Each row (L0–L11) is a layer of system construction, from wafer and substrate at the bottom to burn-in and validation at the top. For each layer the chart identifies the specifications that an integrator consumes, the standards bodies or MSAs that own them, and the compliance gates that qualify the layer for release into the next. Three governance tiers along the right edge identify primary ownership at each scale: OCP for system and platform integration, MSA/OIF/OCP shared for subsystem qualification, and IEEE/OIF/IEC at the component level. The PKGD (Photonic Known-Good-Die) layer at L5a is highlighted as currently unowned, identifying the methodology gap discussed in § Standards Alignment as a candidate joint work item across OCP, OIF, and APC.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

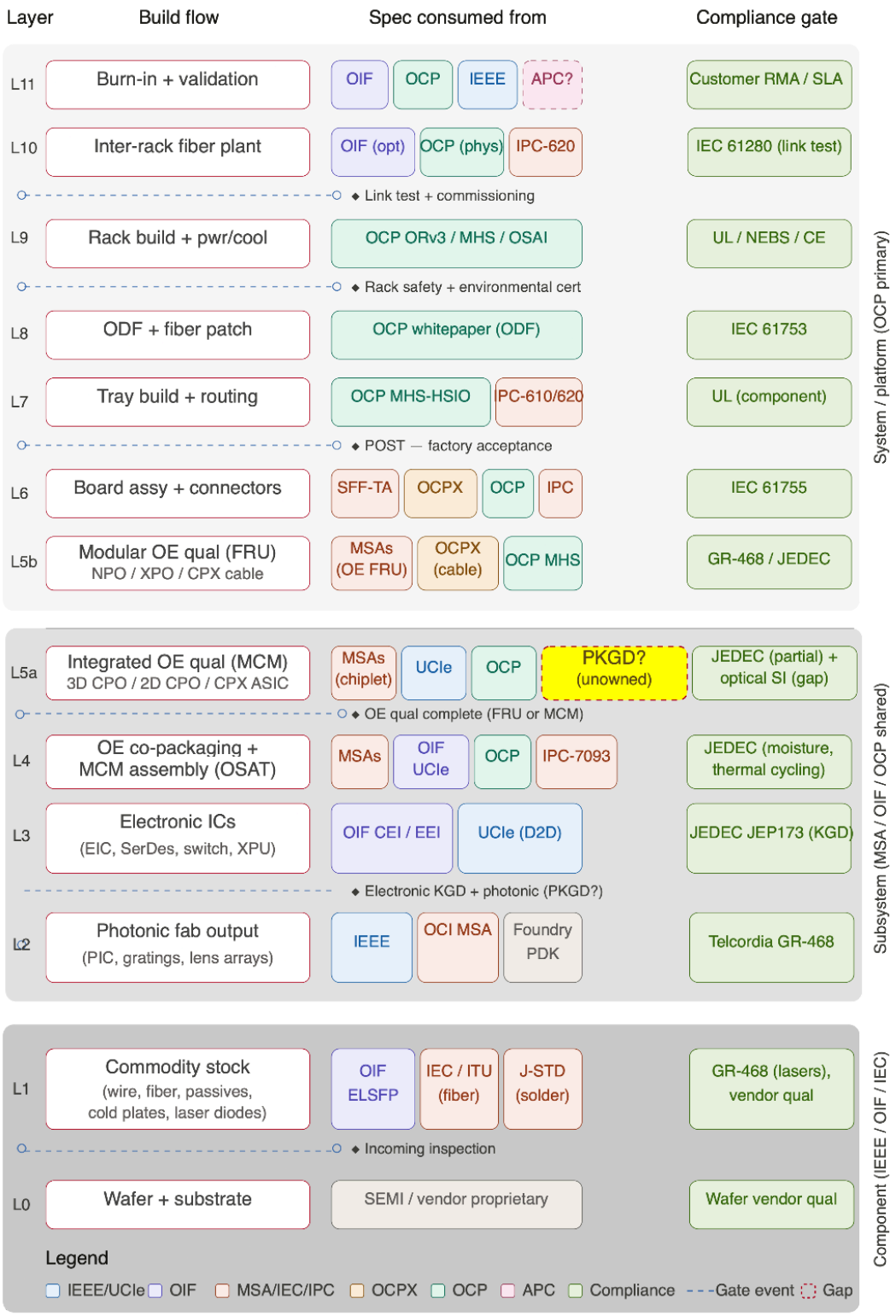


Figure B.1. *Photonic ecosystem build-flow governance reference (full-detail variant).*

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

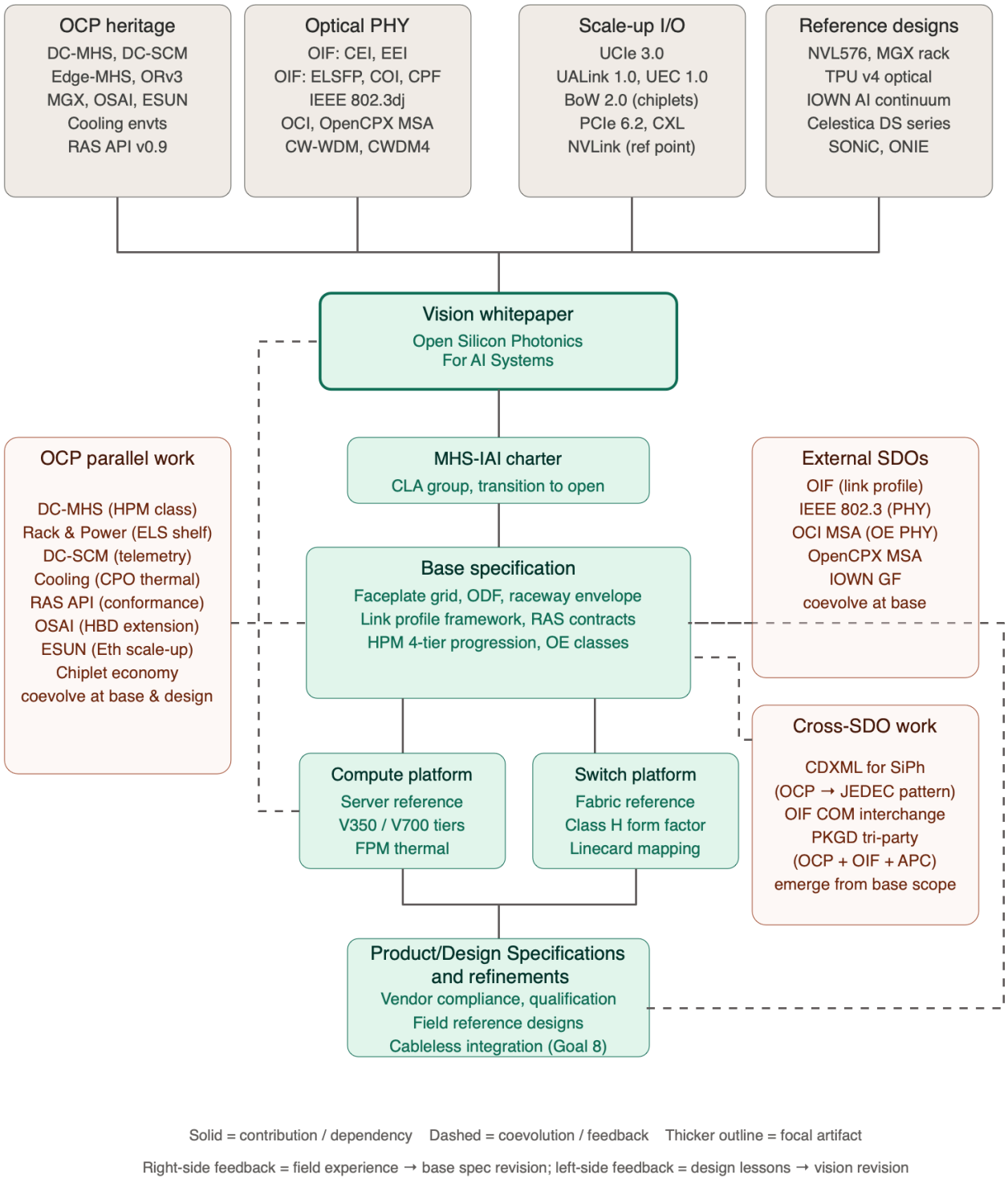


Figure B.2. *Relationship between this effort, the OCP, and outside SDOs and future efforts.*

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Appendix C: Photonic Link Budget Reference

C.1 Scope and Method

Purpose and structure

§ Optical Link Composition establishes the framework: six architectural loss contributors compose the budget from transmitter to photodetector, six topology classes parameterize the contributor counts, and the architecture commits to the framework while deferring specific dB allocations and SNR-margin closure here.

This appendix operationalizes that framework along a deliberate path. C.2 derives each loss mechanism from underlying physics so the parameter spaces are intelligible without recourse to vendor disclosure. C.3 maps mechanisms onto the six topology classes and closes the per-topology loss accounting. C.4 establishes transmitter output and receiver sensitivity envelopes against which the topology budgets close. C.5 closes SNR margin per modulation column against the C.4 envelopes. C.6 instantiates the framework against four worked reference cases at 224G and 448G (DR and OCI 4+4 BiDi). C.7 decomposes the tolerance envelope and identifies the qualification levers that recover the envelope under deployment-context discipline. C.8 surfaces the items where vendor disclosure or standards-body characterization would compress the planning ranges into specification values.

Scope

In scope: per-contributor insertion loss derivation from physical-layer principles (waveguide-mode overlap, modulator topology, fiber attenuation, connector alignment tolerance, mux/demux BW physics); per-topology loss accounting for the six topology classes enumerated in § Optical Link Composition ; transmitter output and receiver sensitivity envelopes consistent with the modulation columns identified below; statistical envelope treatment of polarization-dependent loss, wavelength-dependent loss, temperature variation, component aging, production tolerances, and connector-pair repeat-mate variability.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Out of scope: the electrical channel (Tables I and J in § Channel Budget Reference cover the bump-to-bump electrical envelope; this appendix begins at the OE optical transmit boundary and ends at the photodetector); FEC architecture and coding gain (these are upper-PHY and standards-determined; the appendix uses representative FEC margins to close the SNR budget without specifying which code an implementation runs); specific vendor component values where the public state of the art does not yet support a defensible first-principles bound — those are carried as ranges and called out at the relevant point in C.2 and C.3.

Modulation columns

The architecture is waveform-neutral at the link-profile level (§ Link Profile Framework); the appendix treats three modulation columns to make the design space legible.

PAM4 IM/DD is the primary column. The optical channel sustains the 56 GBaud and 112 GBaud symbol rates within available SiPh modulator BW, and the architecture's link profiles are framed around PAM4 as the buildable baseline. PAM4 IM/DD is built on more than a decade of standardized component characterization (IEEE 802.3, OIF CEI-LR, OCI v1.0). Closed-form sensitivity expressions and component-IL bounds are well established and the appendix derives them.

PAM6 (PAM8 and other higher orders) enters this appendix only on the copper side of the OE. OIF's CEI-448G-LR work is exploring PAM6 as a copper-channel alternative at 448 Gb/s, trading symbol rate (~86.5 GHz Nyquist) for modulation complexity. The architecture's position is that the OE absorbs PAM6 at the electrical interface where the host SerDes presents it, and re-modulates as PAM4 on the optical channel; the optical PHY presented at the fiber boundary is PAM4 regardless of the electrical-side modulation. There is no industry movement toward PAM6 optical modulation: PAM4 represents a decade of hard-won SiPh modulator, DSP, and characterization work, and the optical community's center of gravity remains there. PAM6 therefore appears as a

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

boundary condition on the OE electrical interface but does not appear as a separate column in the optical SNR analysis.

Coherent-lite is the third column. Coherent detection with reduced-complexity DSP recovers SNR through phase-and-amplitude signaling and quadrature processing rather than direct intensity discrimination. The advantage trades DSP energy for receiver performance: a sensitivity gain in the range of 5–15 dB at a pJ/bit cost driven by the DSP architecture. The width of that range is dependent on vendor implementation of the TX and RX stages — modulator format, local oscillator integration, equalizer depth, and FEC interaction all move the operating point. Closed-form sensitivity expressions exist for the underlying receiver physics, but the public state of the art does not yet support a defensible single-point estimate for short-reach intra-datacenter and scale-up applications. The receiver framework itself is publicly documented in OIF work, notably the 800LR Implementation Agreement (an 800G adaptation built on the 400ZR coherent DSP architecture); the open item is the short-reach, scale-up application of that framework, not the framework itself. The appendix carries coherent-lite as a sensitivity-bounded column rather than a point estimate.

Notation and margin envelope

All optical powers are expressed in dBm referenced to 1 mW. All losses are in dB (positive = attenuation). Receiver sensitivity is expressed in dBm at the relevant FEC threshold (typically 2.4×10^{-4} pre-FEC BER for KP4 or equivalent); the appendix uses 1.12× FEC overhead consistent with the main document's channel budget reference framing. OSNR is expressed in a 0.1 nm reference BW unless noted. Wavelength references follow O-band single-mode convention (G.652D SMF, 1310 nm region) unless explicitly stated.

Six envelope contributors — polarization-dependent loss, wavelength-dependent loss, temperature variation of component IL, component aging, production tolerances, and

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

connector-pair repeat-mate variability — are reserved as a statistical margin at the photodetector rather than allocated as point losses against individual contributors. The appendix carries 3 dB as the total envelope, applied to all per-topology budgets in C.3 and decomposed against qualification levers in C.7.

C.2 Physical Layer Losses

The six architectural contributors enumerated in § Optical Link Composition distribute across five physical mechanism families. Grouping by physics rather than by spatial location is what makes the parameter spaces tractable: lateral-misalignment loss at a connector pair and at a PIC-to-fiber edge coupler is the same physics applied at different boundaries, and the per-channel insertion loss of a WDM demux and a WDM mux is the same physics applied at different ends of the same fiber. This section derives each family from first principles, identifies the lower bound the physics imposes, and identifies the boundary at which vendor or standards data populates the deployed value.

Modal/coupling losses at waveguide-to-fiber transitions. Coupling between dissimilar single-mode guides — laser-to-PIC, PIC-to-fiber, fiber-to-photodetector — is governed by the modal overlap integral between the two guided modes. For two Gaussian-approximated modes of 1/e amplitude radii w_1 and w_2 perfectly aligned, the intrinsic coupling efficiency is $\eta = 2 w_1 w_2 / (w_1^2 + w_2^2)$, reaching unity only at $w_1 = w_2$. SiPh waveguide modes are sub-micron (typical mode field radius 0.15–0.25 μm); SMF G.652D modes are $w \approx 4.6$ μm (MFD ≈ 9.2 μm) at 1310 nm. The intrinsic mismatch should be addressed by a mode-converting structure: an inverted taper, a grating coupler, or a 3-D printed beam-shaping element. State-of-the-art edge-coupled designs achieve ≈ 0.5 dB per facet; grating couplers run 1.5–3 dB per facet with polarization dependence as an additional contributor; emerging photonic-wire-bond/3-D printed couplers report sub-0.5 dB.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Lateral misalignment compounds the intrinsic loss. For two identical Gaussian modes of mode-field radius w aligned to within transverse offset d, coupling efficiency degrades as $\eta(d) = \exp(-d^2/w^2)$ — Marcuse's standard form. The 1 dB threshold sits at $d \approx 0.5\ w$; for SMF this is ±2.3 μm and for SiPh waveguide modes it is ±0.1 μm. Mode-conversion structures relax the deployed tolerance to the SMF scale, so alignment tolerance dominates assembly yield rather than the deployed coupling loss. Angular misalignment $\eta(\theta) = \exp[-(\pi\ n\ w\ \theta\ /\ \lambda)^2]$ is well-controlled by ferrule geometry and contributes negligibly below ~0.5°. Physics-bounded lower limit: ≈ 0.5 dB per facet for mature edge-coupled SiPh-to-SMF attach. Practical envelope: 0.5–2 dB per transition. Vendor data populates: actual per-facet IL and its statistical distribution. Architecture's planning value: 1 dB per facet for SiPh-ELS designs with mature edge-coupled fiber attach.

Modulator and active-element insertion loss. The IM/DD modulator imposes an insertion loss governed by topology. Mach-Zehnder modulators recombine two phase-modulated arms by interference; IL decomposes into input/output coupler IL (≈ 0.3 dB each), phase-shifter waveguide loss (carrier-depletion designs run ≈ 1 dB/mm), and chirp/extinction trade-off loss, landing typical carrier-depletion 56 GBaud MZMs at 4–6 dB total IL. Thin-film lithium niobate (TFLN) MZMs reach 2–3 dB at the cost of process compatibility. Ring-resonator modulators are sub-100 μm² resonant structures that modulate by detuning the cavity resonance. Intrinsic IL is governed by the intrinsic-to-coupling Q ratio: critical coupling yields infinite extinction at zero on-resonance IL, but practical operating points trade extinction for IL and run 1–3 dB. Ring modulators carry a ~80 pm/K wavelength drift and require thermal control. Electro-absorption modulators tune the Franz-Keldysh or quantum-confined Stark absorption and run 3–6 dB IL; they are common in III-V integrated designs and uncommon in pure SiPh. Modulator topology is an OE-internal choice, so the IL contribution depends on vendor selection. Comprehensive characterization appears in IEEE Journal of Selected Topics in Quantum Electronics SiPh modulator surveys and in OFC and ECOC technical-program contributions from the major SiPh foundries. Vendor

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

data populates the column. Architecture's planning ranges: 1–3 dB (RRM), 3–6 dB (MZM and EAM).

WDM mux/demux insertion loss. Mux/demux IL appears symmetrically at the two ends of the link, and the two implementation families relevant at 4λ–16λ channel counts have distinct scaling. Arrayed waveguide gratings diffract wavelengths through a free-propagation region into channel-specific output waveguides; AWG IL scales weakly with channel count and lands at 2–4 dB for 4–16 channel devices in SiPh, with per-channel non-uniformity of ±0.5–1 dB across the band. Cascaded ring filters — the natural choice for RRM-based designs — compose per-channel drop filters in series; per-channel IL is low (0.5–1 dB) but compounds along the cascade, so the worst-case channel pays N × per-stage loss. Echelle gratings and thin-film filter cascades are alternative implementations with their own IL/non-uniformity trade-offs. Architecture's planning value: 3 dB per WDM endpoint at 8–16 channels; per-channel non-uniformity captured in the WDL envelope.

Fiber propagation loss. Single-mode G.652D attenuation is approximately 0.32–0.40 dB/km in the O-band (1310 nm) used by all near-term SiPh and OCI implementations, dominated by Rayleigh scattering with a contribution from residual OH absorption. C-band (1550 nm) attenuation is approximately 0.18–0.22 dB/km, but near-term implementations are O-band by SiPh modulator-BW and chromatic-dispersion considerations. At intra-rack distances (≤ 50 m) propagation loss is negligible (~0.02 dB); at inter-rack distances (50–500 m through ODF and shuffle) it remains in the 0.02–0.2 dB range, small relative to connector and OE-internal contributions. Propagation loss becomes consequential only at the inter-pod scale (> 500 m), where the architecture's format-conversion staging (OFC at the pod boundary, § Optical Link Composition) is already the chosen reach-extension mechanism. Macrobending and microbending loss are tolerance contributors; G.657-class bend-insensitive variants relax the minimum bend radius from 30 mm to 7.5 mm or smaller. Closed-form physics;

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

not vendor-determined. Budget allocation = propagation distance × 0.4 dB/km O-band, with bend-induced loss in the production tolerance margin.

Connector pair insertion loss. A connector pair is the physical mating of two single-mode fiber end-faces, alignment-controlled by a ferrule. Three contributors compose connector IL. Lateral misalignment (transverse offset d between fiber cores) is the dominant contributor, governed by the same $\eta(d) = \exp(-d^2/w^2)$ form derived above; for SMF modes of common $w \approx 4.6$ μm, the loss reaches 0.1 dB at $d \approx 0.7$ μm, 0.3 dB at $d \approx 1.2$ μm, 0.5 dB at $d \approx 1.6$ μm. Production SMF connectors (LC, MPO/MTP) achieve 0.1–0.5 dB typical, 0.75 dB worst-case per IEC 61753-1 grade B, held manufacturer-independent by random-mate qualification per IEC 61755-3-3; the alignment-tolerance physics sets the floor. Angular misalignment (tilt θ between core axes) contributes negligibly under typical ferrule control. End-face geometry either physical-contact (PC) or angled physical contact (APC, 8° polish) — eliminates Fresnel reflection at the glass-air interface; flat-polish geometries pay 0.3 dB per face and are unsuitable for high-channel-count parallel connectors. APC is standard for MPO/MTP at scale-up density. Architecture's planning value: 0.5 dB per connector pair typical, 0.75 dB worst-case for tolerance-corner closure. Connector count is the explicit design variable from § Optical Link Composition ; each pair eliminated returns 0.5–0.75 dB directly to the budget.

Polarization-dependent loss, wavelength-dependent loss, and aging envelope. PDL, WDL, and aging are not point losses but statistical envelopes that need to be reserved at the photodetector regardless of which contributor they affect. PDL arises in components whose IL differs between the two principal polarizations: grating couplers (0.5–3 dB PDL), some MZM topologies, AWGs at channel-edge wavelengths. Polarization state at the modulator output is fixed by the SiPh waveguide, so PDL accumulates only across downstream polarization-sensitive components; typical end-to-end allocation 0.5–1 dB. WDL captures per-channel IL non-uniformity across the

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

WDM band — AWG channel-edge attenuation, mux/demux passband ripple, modulator wavelength-dependent IL; typical end-to-end allocation 1–2 dB across 8–16 channels. Aging contributors include laser output degradation (1–2 dB over 5 years at module operating temperature), connector IL drift from repeat mating (~0.1 dB per 500 mates for LC, less for MPO), and PD responsivity drift; typical end-to-end allocation 1–2 dB. Temperature variation of component IL (most significant for unstabilized AWGs and uncompensated ring resonators) and production tolerances compound into the residual margin. C.7 decomposes this envelope into its independent contributors and develops the qualification levers that recover envelope under deployment-context discipline. Architecture's planning envelope: 3 dB total at the photodetector.

Table AB: OE-Internal Loss Mechanism Summary

Mechanism family	Physical basis	Typical contribution	Bounded by
Modal/coupling	Mode-overlap integral + lateral misalignment	0.5–2 dB per transition	Physics + assembly process
Modulator IL	Modulator topology	1–6 dB total (topology-dependent)	Vendor topology choice
WDM mux/demux IL	AWG / ring-filter physics	3 dB per endpoint, ×2 per link	Vendor implementation
Fiber attenuation	Rayleigh scattering + OH absorption	0.3–0.4 dB/km O-band	Closed-form physics
Connector pair IL	Lateral misalignment	0.5 dB typical, 0.75 dB worst-case per pair	Standardized (IEC 61753-1)

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Mechanism family	Physical basis	Typical contribution	Bounded by
PDL / WDL / aging envelope	Statistical accumulation at PD	3 dB total (decomposed in C.7)	Vendor + standards

C.3 maps these mechanisms onto the six topology classes enumerated in § Optical Link Composition and closes the per-topology loss accounting; C.4 derives the transmitter output power and receiver sensitivity envelopes against which the topology budgets should close.

C.3 Per Topology Insertion Loss Accounting

C.3.1 Decomposition

The optical link budget for any topology class composes from three independent terms:

Total link IL = OE-internal (Tx side + Rx side) + topology overhead (connectors + fiber + any format-conversion stage) + reserved envelope (PDL + WDL + aging + tolerance)

OE-internal IL is a function of the OE architecture, not the topology — a SiPh-ELS engine has the same internal budget whether deployed in a direct intra-tray topology or routed through an ODF. Topology overhead is a function of connector count, fiber path length, and the presence or absence of an O→E→O format-conversion stage, with no dependency on what OE class is at either end. The 3 dB envelope from C.1 is reserved at the photodetector regardless. Decomposing this way lets the OE-internal budget be computed once per representative architecture and the topology overhead worked once per topology class; the per-topology link budget is the sum.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

This separation also clarifies where each design lever lives. The OE designer optimizes against the OE-internal term through modulator topology, mux/demux integration, and edge-coupler quality. The system integrator optimizes against the topology overhead through connector-count discipline (§ Optical Link Composition) and topology selection. The standards body and component vendors operate against the envelope through tolerance specification and qualification. The three are independently negotiable, which is the structural reason the architecture commits to a budget framework in § Optical Link Composition rather than to specific dB allocations.

C.3.2 OE Internal Insertion Loss- Two Reference Architectures

Two SiPh-ELS reference architectures bracket the practical range at PAM4 and are framed against the public literature distribution rather than as point estimates. The single-λ MZM architecture is representative of DR-class designs; the 4λ ring-resonator WDM architecture is representative of OCI-class designs. Planning values used below are midpoints within publicly characterized ranges; the budget table in C.3.5 carries the spread as cell ranges so the per-topology total IL reflects literature distribution rather than a single point estimate. Both assume external CW laser source (ELSFP) per the architecture's SiPh-ELS designation.

Single-λ MZM (DR-class). The transmit chain comprises laser-to-PIC edge coupling (≈ 1 dB), an MZM modulator, and PIC-to-fiber edge coupling (≈ 1 dB). No WDM mux is present. MZM IL in carrier-depletion SiPh designs is characterized at 4–6 dB on-chip for PAM4 at 56 GBaud (Wang et al., Photonics Research 6(2), 2018, reports 5.4 dB on-chip at 56 GBaud PAM4); the planning midpoint is 4 dB recognizing that the upper end of the range applies to the highest-extinction operating points. Tx OE-internal IL = 1 + 4 + 1 = 6 dB at the planning midpoint, with a 5–8 dB range carried into the per-topology budget. The receive chain mirrors the transmit chain less the modulator: fiber-to-PIC coupling (≈ 1 dB) plus photodetector coupling (≈ 0.5 dB intrinsic, supported by published Ge-on-Si

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

PDs demonstrating 0.7–1.1 A/W external responsivity in production SiPh foundries, equivalent to ~80–90% external quantum efficiency). Rx OE-internal IL ≈ 1.5 dB.

4λ RRM-WDM (OCI-class). The transmit chain comprises laser-to-PIC edge coupling (≈ 1 dB), an array of four ring-resonator modulators arranged in a cascade on a common bus waveguide, and PIC-to-fiber edge coupling (≈ 1 dB). The ring-modulator cascade is structurally significant: each ring modulates on-resonance for its assigned wavelength and passes the others off-resonance, so the cascade serves as modulator array and WDM multiplexer simultaneously. Published ring modulators in production processes report off-resonance IL <0.2 dB per pass (Li et al., ACS Photonics 9(4), 2022, and related work) and on-resonance modulation IL of 1–3 dB at PAM4 line rates from 50–100 Gb/s. Per-channel transmit IL at the planning midpoint = 1 (laser) + 2 (on-resonance) + 0.6 (three off-resonance passes at 0.2 dB) + 1 (edge) = 4.6 dB, with a 3.6–6.6 dB range carried forward. The receive chain places a similar four-ring cascade as drop filter into channel-specific photodetectors, with per-channel IL = 1 + 2 + 0.6 + 0.5 = 4.1 dB at the midpoint.

Comparison at OCI-class 4λ density:

Table AC: OE-Internal IL Comparison at OCI 4λ Density

OE architecture	Tx OE-internal	Rx OE-internal	Tx + Rx
Single-λ MZM + discrete mux (hypothetical 4λ MZM with separate WDM mux)	1 + 4 + 3 + 1 = 9 dB	1 + 3 + 0.5 = 4.5 dB	13.5 dB
4λ RRM-WDM cascade (representative OCI-class)	1 + 2 + 0.6 + 1 = 4.6 dB	1 + 2 + 0.6 + 0.5 = 4.1 dB	8.7 dB

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

The ~5 dB Tx+Rx budget advantage of the RRM-WDM cascade over an MZM plus discrete mux is the link-budget basis for favoring ring-based OE designs at higher wavelength counts. Individual modulator insertion loss and ring pass-through loss remain vendor-dependent within the ranges shown. The architectural relationship is more stable: ring cascades combine modulation and wavelength multiplexing, while MZM-based designs require separate modulation and mux stages.

The “single-λ MZM + discrete 4λ mux” case is included as an explicit reference point, not as a deployed OCI-density configuration. In practice, production 4λ designs at OCI lane rates are ring-based because the integrated modulation-and-mux structure avoids the extra budget penalty surfaced by this comparison. Stating the baseline explicitly keeps the architectural conclusion grounded in the optical budget rather than in an assumed technology preference.

C.3.3 Topology Overhead

Topology overhead composes from three terms: connector-pair IL, fiber attenuation, and (where applicable) the O→E→O format-conversion stage. Connector-pair IL is taken at the 0.5 dB planning value derived in C.2. Fiber attenuation is taken at 0.4 dB/km O-band consistent with G.652D in the 1310 nm region. Distances in the table below are the architectural typical values from § Optical Link Composition Table A; integrators populate actuals against their specific deployments.

Table AD: Per-Topology Fiber Path Composition

Topology class	Connectors	Fiber path	Connector IL	Fiber IL	Format-conv. stage	Topology overhead
Direct intra-tray (bonded)	0	< 1 m	0.0 dB	< 0.001 dB	—	**0.0 dB**

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Topology class	Connectors	Fiber path	Connector IL	Fiber IL	Format-conv. stage	Topology overhead
Direct intra-tray (connectorized)	1	< 1 m	0.5 dB	< 0.001 dB	—	**0.5 dB**
Point-to-point intra-rack	2	1–5 m	1.0 dB	< 0.002 dB	—	**1.0 dB**
ODF-routed intra-rack	3–4	1–10 m	1.5–2.0 dB	< 0.004 dB	—	**1.5–2.0 dB**
Shuffle-routed inter-rack	4–6	5–50 m	2.0–3.0 dB	0.002–0.02 dB	—	**2.0–3.0 dB**
Inter-pod with OFC	4–6 + O→E→O	5–100 m	2.0–3.0 dB	0.002–0.04 dB	(structurally distinct, see C.3.4)	**see C.3.4**

Two observations follow. First, within the architectural reach envelope, fiber attenuation is not the limiting term. The optical budget is dominated by interface loss: couplers, connector pairs, shuffle points, and distribution points. That is why § Optical Link Composition treats connector count as the explicit design variable.

Second, topology selection moves real margin. The connector-loss span from bonded intra-tray paths to shuffle-routed inter-rack paths is roughly 0 to 3 dB. That span is comparable to the reserved envelope and roughly half a typical OE-internal budget. Each connector pair eliminated returns about 0.5 dB directly to link margin.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

This accounting is consistent with OCI v1.0. OCI defines a 2.5 dB total external insertion-loss budget for its 500 m SMF reference link and attributes that budget mainly to connector losses, with connector count and type determining the populated case (§2.4). The appendix covers a wider topology space: from zero-connector bonded intra-tray paths, through OCI-class inter-rack paths, and out to format-converted inter-pod paths. The conclusion is not that distance is irrelevant, but that at these reaches connectorization, not fiber attenuation, is the dominant architectural variable.

C.3.4 The Inter-Pod Case - Format Conversions as Two Independent Budgets

The inter-pod topology with format conversion (OFC at the pod boundary) is structurally distinct: the O→E→O stage is not a loss to be added to a single budget but a regeneration point that establishes two budget-independent optical hops. Each hop closes its own budget independently.

For the inter-pod path, the structure is:
[Pod-1 OE]—(intra-pod-1 topology)—[OFC ingress]—(electrical regeneration)—[OFC egress]—(intra-pod-2 topology)—[Pod-2 OE]

The OFC's electrical regeneration interior decouples the two halves: Pod-1 OE transmit power, intra-pod-1 connectors and fiber, and OFC-ingress receiver close as one budget; OFC-egress transmit, intra-pod-2 connectors and fiber, and Pod-2 OE receive close as a separate budget. Each half is structurally equivalent to a shuffle-routed inter-rack topology (4–6 connectors, ~5–50 m fiber, ~2.0–3.0 dB topology overhead per half). The latency cost of the regeneration stage is paid in PHY processing and FEC encoding/decoding at the OFC and is captured in the main document's latency characterization rather than the optical loss budget. The architectural point is that format-conversion-induced optical IL does not exist at the OFC; the cost is the electrical regeneration and its associated latency and pJ/bit.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

C.3.5 Per Topology Link Budget - Total IL Closure

Combining OE-internal, topology overhead, and the 3 dB envelope produces the total IL each topology class should close at the photodetector. Total IL is shown as a range reflecting the OE-internal literature distribution and the per-topology connector-count spread; midpoint values are useful for first-order planning but the range is the more honest reflection of the design space.

Table AE: Total Link IL by Topology Class

Topology class	OE-internal (Tx+Rx)	Topology overhead	Envelope	Total link IL
Single-λ MZM (DR-class), Tx+Rx = 6.5–9.5 dB				
Direct intra-tray (bonded)	6.5–9.5 dB	0.0 dB	3.0 dB	**9.5–12.5 dB**
Direct intra-tray (connectorized)	6.5–9.5 dB	0.5 dB	3.0 dB	**10.0–13.0 dB**
Point-to-point intra-rack	6.5–9.5 dB	1.0 dB	3.0 dB	**10.5–13.5 dB**
ODF-routed intra-rack	6.5–9.5 dB	1.5–2.0 dB	3.0 dB	**11.0–14.5 dB**
Shuffle-routed inter-rack	6.5–9.5 dB	2.0–3.0 dB	3.0 dB	**11.5–15.5 dB**
Inter-pod (each half)	6.5–9.5 dB	2.0–3.0 dB	3.0 dB	**11.5–15.5 dB per half**
4λ RRM-WDM (OCI-class), Tx+Rx = 7.2–11.2 dB				

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Topology class	OE-internal (Tx+Rx)	Topology overhead	Envelope	Total link IL
Direct intra-tray (bonded)	7.2–11.2 dB	0.0 dB	3.0 dB	**10.2–14.2 dB**
Direct intra-tray (connectorized)	7.2–11.2 dB	0.5 dB	3.0 dB	**10.7–14.7 dB**
Point-to-point intra-rack	7.2–11.2 dB	1.0 dB	3.0 dB	**11.2–15.2 dB**
ODF-routed intra-rack	7.2–11.2 dB	1.5–2.0 dB	3.0 dB	**11.7–16.2 dB**
Shuffle-routed inter-rack	7.2–11.2 dB	2.0–3.0 dB	3.0 dB	**12.2–17.2 dB**
Inter-pod (each half)	7.2–11.2 dB	2.0–3.0 dB	3.0 dB	**12.2–17.2 dB per half**

The 5–7 dB spread per row reflects the public-literature range of OE-internal designs at PAM4 lane rates; the spread across topology classes at fixed OE class is 2–3 dB. The three-lever framing follows: OE design (5–7 dB), topology selection (2–3 dB), and tolerance envelope (3 dB worst-case, qualification-elastic per C.7) compose multiplicatively, and an integrator's choice of where to operate against each lever determines link viability against the Tx-output / Rx-sensitivity envelope developed in C.4.

C.3.6 Connector-Count Sensitivity

The per-connector IL contribution is ~0.5 dB; cumulative connector IL for the six topology classes ranges from 0 to 3 dB. Two architectural observations follow.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

First, the lowest connector-count topologies (bonded and single-connector intra-tray) recover 2.5–3 dB of link margin compared to the highest-count topology (six-connector shuffle-routed inter-rack). At 224 Gb/s/lane using PAM4, 3 dB of recovered margin is approximately the gap between perimeter-placed (NPO/2D CPO/XPO at faceplate distance) and area-domain placed (3D CPO under-logic) optical-engine integration regimes. The § Optical Link Composition framing that connector count is the explicit design variable is, in budget terms, the same lever that placement-class selection operates: both buy or pay link margin in dB units that close the same budget.

Second, the bonded direct intra-tray topology — zero connector pairs between OE and OE — is the only topology that fully eliminates the connector contribution. Every other topology pays at least 0.5 dB to connectorization. The architectural implication is that for highest-density on-tray XPU-to-XPU links where the OE-to-OE fiber path can be a continuous bonded element, the topology choice is not just a serviceability decision but a 0.5 dB budget recovery that may be the difference between closing a 4λ RRM-WDM Profile B link and not closing it. This is the physical basis for the § Optical Link Composition companion principle: the in-rack fiber path is treated as a single bonded passive element wherever physical access does not require otherwise.

C.3.7 Reference Anchors and Open Items

The C.3 accounting draws on the following public-literature and standards anchors:

Table AF: C.3 Reference Anchors and Public-Literature Citations

Planning value	Range	Anchor
MZM IL at 56 GBaud PAM4	4–6 dB on-chip	Wang et al., *Photonics Research* 6(2), 2018

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Ring modulator IL at PAM4	1–3 dB on-resonance, <0.2 dB off-resonance	Li et al., *ACS Photonics* 9(4), 2022, and related literature
PD external responsivity at 1310/1550 nm	0.7–1.1 A/W (~80–90% QE)	Vivien et al., Michel et al., *Optics Express* (multiple)
Per-facet edge coupler IL (SiPh-to-SMF)	0.5–1.0 dB (mature designs)	OFC technical-program contributions, multiple foundry disclosures
Connector pair IL	0.5 dB typical, 0.75 dB worst-case	IEC 61753-1 grade B
External link IL reference (500 m SMF, intra-DC)	2.5 dB	OCI v1.0 §2.4

The OCI v1.0 reference is the strongest anchor: it is a standards-body characterization rather than a single paper, and it lines up directly with the architecture's topology accounting at shuffle-routed inter-rack scale. Open items where no public anchor yet supports a defensible point estimate: the partition of the 3 dB envelope between PDL/WDL and aging/temperature/tolerance contributors (developed against published reliability frameworks in C.7 but not yet against vendor disclosure for the specific component mix targeted here); coherent-lite receiver sensitivity at short-reach intra-datacenter scale (carried as a 5–15 dB sensitivity-gain range against PAM4 baseline, dependent on vendor TX/RX implementation choices that are not yet publicly characterized).

C.4 develops the transmitter output and receiver sensitivity envelopes against which the totals in C.3.5 close.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

C.4 Transmitter Output and Receiver Sensitivity Envelopes

C.4.1 Closure Inequality

The link closes when source-side optical power overcomes total link IL with adequate margin against sink-side sensitivity. Stating this as a single inequality:

P_TX,out − L_total ≥ P_RX,sens, where P_TX,

out is the modulator output OMA per channel (dBm), measured at the modulator output port before any OE-internal mux or coupling loss; L_total is the total link IL composed in §C.3.5:

OE-internal (Tx+Rx) + topology overhead + tolerance envelope; P_RX,sens

is the receiver sensitivity at the photodetector input (dBm OMA, at the specified pre-FEC BER threshold; KP4 FEC threshold of 2.4×10^{-4} assumed throughout).

Rearranging:

P_TX,out − P_RX,sens ≥ L_total.

The left-hand side is the available link budget, a property of the modulation/format choice and the OE class. The right-hand side is the consumed link budget, a property of the OE design × topology × envelope composition from §C.3.5. C.4 characterizes the left-hand side; C.4.4 composes it against §C.3.5 to identify which combinations close.

This appendix references P_TX,out at the modulator output (rather than at the OE faceplate, the convention used by OCI v1.0 so that OE-internal TX loss appears explicitly inside L_total rather than implicitly inside the TX spec. Faceplate-referenced specifications are recovered by subtracting the OE-internal TXcontribution; the two conventions yield equivalent closure results.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

C.4.2 Transmitter Output Envelope

The modulator output OMA is bounded above by laser power delivered to the modulator and below by modulator insertion loss and bias point. Both bounds are vendor-design choices within published envelopes:

Table AG: Transmitter Envelope Components (Laser Power and Modulator IL)

Parameter	Range	Anchor
ELS power per λ (delivered to OE)	+5 to +13 dBm	OIF ELSFP-02.0; OCI v1.0 §2.5 (Implementation-Specific within this envelope)
MZM IL on-chip (56 GBaud PAM4)	4–6 dB	Wang et al., Photonics Research 6(2), 2018
Ring modulator IL on-resonance (PAM4)	1–3 dB	Li et al., ACS Photonics 9(4), 2022
PAM4 extinction ratio	3.5–4.5 dB	OCI v1.0 Table 2-2

Composing ELS supply minus modulator IL gives the modulator output OMA envelope per channel:

Table AH: Modulator Output OMA Envelope per Channel

Modulation × rate	P_TX,out (per-channel OMA at modulator output, dBm)	Bounding mechanism
PAM4 @ 224G	−1 to +5	MZM-based: ELS minus 4–6 dB IL. Ring-based: ELS minus 1–3 dB IL. Spread tracks laser supply and modulator architecture, not lane rate.
PAM4 @ 448G	−1 to +5	Same source-side envelope. At 448G the binding constraint is modulator BW, not output power.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Modulation × rate	P_TX,out (per-channel OMA at modulator output, dBm)	Bounding mechanism
Coherent-lite	−1 to +5	Same source-side envelope. I/Q modulator structures redistribute power across quadrature arms; the per-channel envelope is preserved.

The source-side launch power is treated as a common planning envelope. Laser supply and modulator IL set the planning output; modulation format mainly changes receiver sensitivity, DSP energy, and required margin. Under this model, differentiation across the three modulation columns is carried on the receiver and signal-processing side, while the launch-power envelope remains common.

Two operational notes. Total launched power per fiber: For WDM transmitters with N wavelengths per fiber, total per-fiber launched power scales as $P_{TX,out} + 10 \cdot \log_{10}(N)$ dB. At 16λ BiDi with +2 dBm per channel, total per-fiber per-direction launch is +14 dBm. Class 1 eye-safety classification under IEC 60825-1 at the optical faceplate constrains total per-fiber aggregate power; the WDM envelopes considered in this architecture sit within this envelope with margin. Squelched-transmitter behavior: Modulator squelching reduces OMA to ≤ −15 dBm (OCI v1.0) while preserving average optical power to maintain heater and bias lock. Operationally important for link bringup and fault containment; does not enter the closure inequality.

C.4.3 Receiver Sensitivity Envelope

The receiver sensitivity is the OMA at the photodetector input required to recover the modulated signal at the specified pre-FEC BER. It is bounded below by photodetector shot noise and TIA thermal noise, and discriminated by the modulation format's intrinsic SNR penalty relative to a noise-floor-limited reference.

Table AI: Receiver Sensitivity Envelope Components

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Parameter	Range	Anchor
PD external responsivity (1310/1550 nm)	0.7–1.1 A/W	Vivien et al., Michel et al., Optics Express (multiple)
PAM4 SNR penalty vs. NRZ at equal symbol rate	~9.5 dB	Equiprobable 4-PAM Q-factor relation
KP4 FEC pre-FEC BER threshold	2.4×10^{-4}	IEEE 802.3 (multiple PMDs)
Coherent-lite sensitivity advantage over IM/DD PAM4	5–15 dB	Hirokawa et al. JLT 39 (2021): analog coherent, ~5 dB; Morsy-Osman et al. Opt. Express 26 (2018) : DSP-free, ~8 dB; Ahmed et al. IEEE JSSC (2020/2023) : DSP-coherent, ~12–15 dB

The resulting per-channel sensitivity envelope at the photodetector input:

Table AJ: Per-Channel Sensitivity Envelope at Photodetector

Modulation × rate	P_RX,sens (per-channel OMA at PD, dBm)	Discriminating physics
PAM4 @ 224G	−9 to −11	BW-limited TIA noise; cross-validates against OCI v1.0 stressed sensitivity of −6.2 dBm at faceplate after back-calculation through 3–5 dB OE-internal Rx loss.
PAM4 @ 448G	−6 to −8	Sensitivity tightens by ~3 dB from 224G; noise BW doubles and noise floor scales as √BW.
Coherent-lite	−14 to −26	LO gain and matched-filter receiver. Spread spans analog coherent (~5 dB advantage, lower energy/bit)

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Modulation × rate	P_RX,sens (per-channel OMA at PD, dBm)	Discriminating physics
		and DSP-assisted coherent (~15 dB advantage, higher energy/bit).

The 12 dB span across the coherent-lite row is the dominant axis of variation in C.4 and the design knob most discriminating for placement-class viability. At the analog (~5 dB) end, coherent-lite closes the same topology set as PAM4 but with margin recovered for aging and tolerance. At the DSP-assisted (~15 dB) end, coherent-lite opens inter-rack and inter-pod topology classes that PAM4 cannot close without format conversion — at a pJ/bit cost that § Power Efficiency at Scale absorbs.

C.4.4 Available Link Budget and Closure Visibility

Composition of §C.4.2 and §C.4.3 produces the available link budget. The closure analysis anchors at the spec-grade operating point — upper-end TX output (~+5 dBm modulator OMA) paired with lower-end RX sensitivity (~−11 dBm at PD for PAM4-224G), corresponding to ELS supply at the upper end of the OIF ELSFP envelope, modulator architectures with low insertion loss, and TIA/PD pairs at published sensitivity floors. This is the operating point where mature designs sit: OCI v1.0 and 200GBASE-DR1 systems close at their specified reaches at this point today. Anchoring closure analysis at the median of the design envelope, or below it, would understate that deployed reality and produce closure flags inconsistent with shipped product. Designing against the median is therefore not conservatism, but simply inaccuracy. Spec-grade is the realistic plan; sub-spec is the unrealistic one.

Table AK: Available Budget at Spec-Grade vs Envelope Range

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Modulation × rate	Available budget at spec-grade (dB)	Envelope range (dB)
PAM4 @ 224G	16	8 to 16
PAM4 @ 448G	13	5 to 13
Coherent-lite, analog (~5 dB advantage)	21	13 to 21
Coherent-lite, DSP-assisted (~15 dB advantage)	31	23 to 31

Cross-referencing the spec-grade column against the §C.3.5 total link IL ranges produces the closure visibility map. Flags: ✓ closes with margin; ◐ closes tight or marginally; ✗ does not close in the format envelope.

Table AL: SNR Closure Visibility Map by Topology and Modulation

Topology class	PAM4 @ 224G	PAM4 @ 448G	Coherent-lite, analog	Coherent-lite, DSP
Direct intra-tray (bonded)	✓	◐[1]	✓	✓
Direct intra-tray (connectorized)	✓	◐	✓	✓
Point-to-point intra-rack	✓	◐[1]	✓	✓
ODF-routed intra-rack	✓	◐[1]	✓	✓
Shuffle-routed inter-rack	◐[1]	✗	✓	✓
Inter-pod (each half)	◐[1]	✗	✓	✓

[1] Flag shown for DR-class L_total median; OCI-class is decision-relevant at these cells. For PAM4-224G the marked cells become firmly ◐ at OCI-class with ~1.3 dB residual margin. For PAM4-448G the marked cells shift toward ✗ at OCI-class median, leaving only direct intra-tray connectorized as a robust closure at this rate.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

[2] Coherent-lite advantage spans analog implementations (~5 dB; Hirokawa et al., Morsy-Osman et al.) and DSP-assisted implementations (~15 dB; Ahmed et al.). The two columns carry distinct pJ/bit costs which § Power Efficiency at Scale absorbs.

Three operational conclusions follow. **One.** At PAM4-224G, the spec-grade envelope closes intra-rack topology classes with adequate margin and becomes tight at shuffle inter-rack and inter-pod (each half), particularly for OCI-class 4λ RRM-WDM OE designs. The OCI v1.0 reference is consistent with the lower-margin observation: its 2.5 dB external-IL specification carries only ~2.7 dB margin against −1 dBm TX and −6.2 dBm stressed RX at the faceplate, which back-calculates to a tight ◐ at OCI-class shuffle inter-rack in the framework above. **Two.** At PAM4-448G, the available budget compresses by ~3 dB while the §C.3.5 IL band is unchanged. Closure contracts to direct intra-tray, with diminishing margin even there for OCI-class OE designs. In this regime the OE-internal design lever (5–7 dB swing between MZM and ring modulator architectures) and the topology lever (2–3 dB swing across the intra-rack class) compound to determine viability; the format alone is insufficient. **Three.** Coherent-lite at the upper bound of its sensitivity advantage opens inter-pod reach without format conversion, at a pJ/bit cost the § Power Efficiency at Scale power accounting absorbs. Its architectural significance is reach extension, not replacement: IM/DD remains the lowest-energy format for short reaches, while coherent-lite extends the addressable topology envelope without O→E→O conversion at pod boundaries.

The three-lever framing from §C.3.6 — OE design (5–7 dB), topology selection (2–3 dB), tolerance envelope (3 dB fixed) — combines with the modulation/format axis introduced in C.4 to span the closure space. An integrator's choice of where to operate against each lever, paired with format selection per link, determines link viability across the architecture's topology range.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

C.4.5 Note on Modulations Beyond PAM4 (Electrical-Side Only)

OIF is actively exploring modulations beyond PAM4 at 448 Gb/s/lane as copper-side waveforms. PAM6 is the leading near-term candidate, reducing the Nyquist requirement from approximately 112 GHz to approximately 86.5 GHz by trading modulation complexity for symbol rate. PAM8 remains under industry exploration as a more aggressive option, reducing Nyquist further to approximately 75 GHz at the cost of additional vertical SNR penalty (approximately 2.5 dB beyond PAM6's penalty over PAM4) and tighter TX linearity, DAC/ADC resolution, and FEC margin requirements. The motivation for either is host SerDes and twinax flyover viability, not the optical link. From the optical-link perspective addressed in this appendix, neither PAM6 nor PAM8 contributes movement to the C.4 envelopes: modulator drivers, photodetectors, and equalization at their symbol rates have not been demonstrated to outperform PAM4 at equal lane rate on the metrics that bound the link budget. The architecture acknowledges both as electrical-channel design variables and is structurally compatible with bifurcated or trifurcated channel-budget rows by modulation format if either matures within the deployment window, but no optical envelope assumptions move on their account.

C.5 SNR Closure by Modulation Column

C.4.4 establishes the closure visibility map, which (topology, modulation, OE-class) cells close, close tight, or do not close at spec-grade operating points. C.5 deepens that analysis along two axes the visibility flags do not capture: the first-principles physics that sets the receiver sensitivity envelope in C.4.3, and the quantitative residual SNR margin (in dB) that determines whether a closing cell closes with operational headroom or only marginally. The closure flag is a coarse summary; the SNR margin is the operational quantity that distinguishes a robust deployment from a fragile one.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

C.5.1 Closure-to-Margin Framework

The closure inequality from §C.4.1 (P_TX,out − L_total ≥ P_RX,sens) defines a binary outcome: the link closes or it does not. The residual SNR margin is the dB-distance the left-hand side exceeds the right:

$$M_{SNR} = P_{TX,\,out} - L_{total} - P_{RX,\,sens}$$

For closure with operational headroom, the architecture's planning convention is M_SNR ≥ 2 dB at spec-grade operating points. This 2 dB margin envelope is what absorbs the per-channel variation across deployed units, the year-to-year drift across the operational envelope that C.7 partitions, and the FEC threshold uncertainty that vendor BER characterizations carry around the nominal 2.4×10^{-4} pre-FEC operating point.

The receiver sensitivity itself emerges from the noise physics of the PIN+TIA receiver chain. For an unamplified Ge-on-Si PD with responsivity R (A/W) followed by a TIA of input-referred RMS noise current σ_n (A) over the signal BW, the sensitivity at a Q-factor threshold Q_th is:

$$P_{RX,\,sens} = \frac{Q_{th} \times \sigma_n}{R}$$

The Q-factor threshold Q_th = 2.95 corresponds to the KP4 FEC BER threshold of 2.4×10^{-4}, after which the FEC code (KP4: t = 15, n = 544, k = 514) recovers BER to better than 10^{-15} for operational data integrity. For PAM4 modulation, the SNR penalty relative to NRZ at equal symbol rate is ~9.5 dB — three logical levels share the eye-opening that NRZ allocates to one, and the inner-eye Q-factor at equiprobable signaling scales as 1/3 the OMA Q-factor of an equivalent NRZ system.

The TIA noise σ_n scales as √BW where BW is the signal BW, approximately equal to the symbol rate. From 56 GBaud (224 Gb/s/lane using PAM4) to 112 GBaud (448 Gb/s/lane using PAM4), BW doubles and σ_n increases by ~3 dB. This is the structural reason P_RX,sens at 448G is ~3 dB worse than at 224G in the C.4.3 envelope: the TIA

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

noise rises with BW while the responsivity and Q-factor threshold are unchanged. The receiver sensitivity envelope of C.4.3 is, in this framing, the noise-physics translation of the modulation column choice onto the closure inequality.

C.5.2 PAM4 IM/DD Margin Map

At spec-grade operating points for TX output at +5 dBm modulator OMA per channel, RX sensitivity at −11 dBm at PD input for 224 Gb/s/lane using PAM4 and −8 dBm for 448 Gb/s/lane using PAM4 — the available budget per channel is 16 dB at 224 Gb/s/lane and 13 dB at 448 Gb/s/lane (per C.4.4). Cross-referencing against the C.3.5 per-topology total IL ranges produces the residual SNR margin per (topology, OE-class) cell, expressed as M_SNR in dB.

Table AM: Residual SNR Margin per Topology × OE Class

Topology class	DR-class @ 224G (M_SNR, dB)	4λ-WDM @ 224G	DR-class @ 448G	4λ-WDM @ 448G
Direct intra-tray (bonded)	+3.5 to +6.5	+1.8 to +5.8	+0.5 to +3.5	−1.2 to +2.8
Direct intra-tray (connectorized)	+3.0 to +6.0	+1.3 to +5.3	+0.0 to +3.0	−1.7 to +2.3
Point-to-point intra-rack	+2.5 to +5.5	+0.8 to +4.8	−0.5 to +2.5	−2.2 to +1.8
ODF-routed intra-rack	+1.5 to +5.0	−0.2 to +4.3	−1.5 to +2.0	−3.2 to +1.3
Shuffle-routed inter-rack	+0.5 to +4.5	−1.2 to +3.8	−2.5 to +1.5	−4.2 to +0.8
Inter-pod (each half)	+0.5 to +4.5	−1.2 to +3.8	−2.5 to +1.5	−4.2 to +0.8

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Cells with M_SNR ≥ +2 dB at the planning midpoint close with operational headroom. Cells with 0 ≤ M_SNR < +2 dB close but operate without margin against the qualification recovery envelope C.7 develops. Cells with M_SNR < 0 do not close in the modulation column.

Three reads follow.

One, at 224 Gb/s/lane using PAM4 the DR-class envelope closes every topology class with positive midpoint margin and operational headroom (M_SNR ≥ +2 dB) at every cell except inter-pod halves. The OCI-class envelope is tighter as the operational headroom is preserved through ODF-routed intra-rack and contracts to marginal-closure or sub-closure at shuffle and inter-pod. The architectural reading: at 224G the OE-internal lever (5 dB DR vs. OCI swing) is the binding constraint at the upper topology range, not the topology lever itself.

Two, at 448 Gb/s/lane using PAM4 the closure envelope contracts substantially. DR-class shuffle and inter-pod don't close at the planning midpoint; OCI-class shuffle and inter-pod don't close anywhere in their literature range. Closure at 448 Gb/s/lane using PAM4 is bounded to intra-rack topologies for both OE classes, with the OE-internal lever again binding at the upper topology range.

Three, the negative-margin cells at 448 Gb/s/lane using PAM4 are exactly the cells C.4.4 flags as ✗. The quantitative dB-margin map clarifies how far below closure these cells sit and therefore how much budget swing, through envelope qualification (C.7), through OE-internal compression, or through topology recovery, would restore them.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

C.5.3 Coherent-lite Margin Envelope

The coherent-lite column carries a 5–15 dB sensitivity advantage relative to PAM4 IM/DD at equal symbol rate, dependent on vendor implementation of the TX (single-ended vs. I/Q modulator, modulator format), the LO (integrated vs. external, with associated phase-noise floor), and the RX DSP (equalizer depth, FEC interaction, carrier-phase recovery architecture). At the analog end (~5 dB advantage; Hirokawa et al. and Morsy-Osman et al. references in C.4.4), the available budget rises to 21 dB at the spec-grade operating point. At the DSP-assisted end (~15 dB advantage; Ahmed et al.), it rises to 31 dB.

Cross-referencing against the C.3.5 total IL ranges produces the coherent-lite margin envelope. At the analog end, every topology class closes with M_SNR ≥ +3 dB at OCI-class median, including inter-pod halves; this is the structural significance of analog coherent-lite for the architecture — it opens inter-pod reach without resort to O→E→O conversion at the pod boundary. At the DSP-assisted end, the margin envelope extends well above operational headroom for every topology and OE-class combination considered, with the binding constraint shifting from optical SNR to the pJ/bit floor of the DSP as the § Power Efficiency at Scale power accounting shows.

The architectural reading: coherent-lite does not displace PAM4 IM/DD at short reach where IM/DD's lower pJ/bit makes it the energy-optimal choice. Its role is to extend the topology envelope addressable at lane rates and reach combinations where IM/DD has consumed its margin. C.7's envelope-qualification levers and coherent-lite's sensitivity advantage are complementary as qualification recovers ~1.0–1.5 dB across all formats; coherent-lite recovers 5–15 dB for a specific format at a defined energy cost.
The width of the 5–15 dB range remains the open item C.8 catalogs: The architecture commits to coherent-lite as an envelope-extending modulation column and defers the within-envelope point estimate to vendor disclosure at the time of deployment.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

C.6 Reference Cases by Lane Generation (DR and OCI)

C.6 instantiates the framework against two paired reference cases that anchor the design space at both lane rates. IEEE 802.3dj DR, a single-wavelength parallel single-mode, sets the perimeter-placed standards baseline that the industry transitions out of as scale-up domain density grows. OCI, a four-wavelength bidirectional on a single fiber per direction-pair, per OCI MSA v1.0, sets the area-domain wavelength-density baseline the architecture sizes toward. Each is worked at 224Gb/s and at 448Gb/s per lane, giving four worked cases that bracket the design space the rest of the document references.

C.6.1 Case Construction

All four cases assume shuffle-routed inter-rack topology with the most demanding intra-pod topology class enumerated in § Optical Link Composition, where the topology lever is at its upper end (2.0–3.0 dB overhead) and the envelope is at its 3 dB worst-case planning value. The intent is to stress-test the framework at the binding topology rather than at the easier intra-tray cases that close more readily; cases that close at shuffle-routed inter-rack close with more margin at every shorter topology.

The shuffle-routed inter-rack topology presents 4–6 connector pairs (faceplate ingress, ODF mate, inter-rack harness mate, ODF egress, faceplate egress, plus one optional patch) over 5–50 m of intra-pod fiber. Total topology overhead is 2.0–3.0 dB per C.3.3. The DR cases assume the single-λ MZM OE architecture from C.3.2: Tx OE-internal IL = 1 + 4 + 1 = 6 dB at the planning midpoint (range 5–8 dB), Rx OE-internal IL = 1 + 0.5 = 1.5 dB. Tx+Rx OE-internal = 7.5 dB at midpoint (range 6.5–9.5 dB).

The OCI 4+4 BiDi cases assume the 4λ RRM-WDM OE architecture from C.3.2: Tx OE-internal IL = 4.6 dB per channel at midpoint (range 3.6–6.6 dB), Rx OE-internal IL = 4.1 dB per channel at midpoint. Tx+Rx OE-internal = 8.7 dB at midpoint (range 7.2–11.2 dB). The OCI 4+4 BiDi configuration uses one fiber per direction with four wavelengths

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

multiplexed; the per-channel budget is the closing quantity since each wavelength should close independently against the photodetector for its assigned channel.

C.6.2 Worked Budgets and Closure

The four cases compose:

Table AN: Worked Reference Cases by Lane Generation (DR and OCI)

Case	OE-internal (Tx+Rx)	Topology overhead	Envelope	Total L_total	Available budget	M_SNR at midpoint
1. DR: 200G/λ PAM4 (802.3dj)	7.5 dB (6.5–9.5)	2.5 dB (2.0–3.0)	3.0 dB	**13.0 dB (11.5–15.5)**	16 dB	**+3.0 dB** (range +0.5 to +4.5)
2. DR, next lane: 400G/λ PAM4 (no ratified PHY)	7.5 dB (6.5–9.5)	2.5 dB (2.0–3.0)	3.0 dB	**13.0 dB (11.5–15.5)**	13 dB	**0.0 dB** (range −2.5 to +1.5)
3.OCI Gen 1: 4×53G NRZ (v1.0)	8.7 dB (7.2–11.2)	2.5 dB (2.0–3.0)	3.0 dB	**14.2 dB (12.2–17.2)**	24 dB	**+9.8 dB** (range +6.8 to +11.8)

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Case	OE-internal (Tx+Rx)	Topology overhead	Envelope	Total L_total	Available budget	M_SNR at midpoint
4. OCI Gen 2: 4×100G NRZ (roadmap)	8.7 dB (7.2–11.2)	2.5 dB (2.0–3.0)	3.0 dB	**14.2 dB (12.2–17.2)**	21 dB	**+6.8 dB** (range +3.8 to +8.8)

C.6.3 Architectural Reading

The four cases reveal a coherent structure that maps directly onto the architecture's design choices.

Case 1 (DR, 200G/λ) closes robustly. The midpoint M_SNR of +3.0 dB exceeds the 2 dB operational-headroom threshold, and the full range stays above 0 dB across the literature distribution. This is the regime where current production 200GBASE-DR1 systems operate today; the architecture's framework reproduces this closure without prescribing it. The 16 dB available budget paired with the 13 dB midpoint L_total is the structural reason 224 Gb/s/lane using PAM4 IM/DD is the buildable baseline at shuffle-routed inter-rack scale.

Case 2 (DR, next lane) collapses to marginal. The midpoint M_SNR drops to 0 dB — the link closes at the planning midpoint but with no operational headroom, and ~40% of the literature distribution does not close at all (the 6.5–9.5 dB Tx+Rx OE-internal range tightens the L_total range to 11.5–15.5 dB against a 13 dB budget). At 448 Gb/s/lane using PAM4 in the shuffle-routed inter-rack topology, the architecture is operating against the limit of the format, not against the limit of any single lever. Three responses follow from the lever framing: tighten the OE-internal lever (move to ring-based 4λ

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

designs, accepting the wavelength-count penalty Case 4 surfaces); tighten the topology lever (move from shuffle-routed inter-rack to ODF-routed or point-to-point intra-rack, accepting the smaller scale-up domain); or qualify the envelope (recover 1.0–1.5 dB through C.7 levers, restoring case-2 closure across the literature range).

Case 3 (OCI Gen 1, 4×53G NRZ) closes robustly. Modeled on its own NRZ basis, Gen 1 closes at +9.8 dB midpoint and holds above +6.8 dB across the literature range. The four wavelengths cost about 1.2 dB of OE-internal loss; NRZ's per-symbol sensitivity returns about 8 dB. The net is the widest margin in the set. The earlier PAM4-floor reading understated OCI by that 8 dB. One ceiling holds: at eight wavelengths the eye-safety cap binds the per-fiber aggregate before the receiver does. So OCI banks its margin as reach and yield, not as room to climb power classes.

Case 4 (OCI Gen 2, 4×100G NRZ) closes with headroom. Gen 2 doubles the symbol rate and stays NRZ, trading reach headroom, not detection margin. It closes at +6.8 dB midpoint and holds above +3.8 dB across the range. The 3 dB it gives up against Gen 1 is the baud-rate cost alone; the modulation is unchanged. PAM4-DR, by contrast, tightens to 0.0 dB at its next lane generation, where no ratified optical PHY yet exists. The interoperable NRZ profile is therefore the robust closer and the better-grounded forward point.

Across the four cases, the architecture’s three structural levers: the OE-internal design, topology selection, envelope qualification, together with its modulation-column choice compose to define the closure envelope. No single lever closes DR at its next lane generation (assuming PAM4 etc); the levers compose additively in log scales like dB, and multiplicatively in linear scales, to bring the binding cases into the closing envelope. This is the structural reason the architecture commits to the lever framework rather than to specific point allocations: an integrator selects against the combination

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

that fits a specific system context, with the framework producing the dB economy of each combination.

C.7 Tolerance Envelope Partitioning

The 3 dB tolerance envelope established in §C.3.5 is the architecture's planning convention — the budget margin reserved against contributors that erode link margin over deployed lifetime, across operating conditions, and across unit-to-unit production variation. C.4 anchored closure analysis against this fixed value to keep the framework vendor-neutral. C.7 decomposes the envelope into its independent contributors, sizes each from public anchors, and identifies the qualification levers that recover envelope where deployment context allows. The 3 dB total is the worst-case upper bound; tightening it is an integrator choice with direct consequence for the closure visibility map of §C.4.4.

C.7.1 Envelope Decomposition

Five physically independent contributors compose the envelope. They are independent in the sense that their physical origins are distinct — component aging, environmental sensitivity, polarization symmetry, spectral non-flatness, manufacturing dispersion — and they combine statistically through the link without correlated drift.

Table AO: Tolerance Envelope Contributors and Physical Origins

Contributor	Physical origin	Typical range (dB)	Anchor
Aging (20-year deployment)	Laser power decay, modulator drift, PD responsivity drift, fiber attenuation creep	0.5–1.5	Telcordia GR-468 reliability framework; IEEE 802.3 transceiver aging allocations

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Contributor	Physical origin	Typical range (dB)	Anchor
Temperature drift (0–85 °C operating range)	Wavelength drift on passive filters, modulator threshold drift, TIA gain drift	0.3–0.5	OIF EEI COI WP characterization; vendor MSA datasheets
PDL (polarization-dependent loss)	Passive component asymmetry, fiber polarization mode coupling	0.2–0.5	IEC 61753 connector and component grades
WDL (wavelength-dependent loss)	Non-flat mux/demux passband, FBG ripple, grid-edge effects	0.2–0.5	DWDM filter manufacturer specifications (grid-spacing-dependent)
Production tolerance (1σ to 3σ)	Unit-to-unit variation in coupler efficiency, modulator ER, PD responsivity	0.3–0.5	Foundry process disclosures; IEEE 802.3 component conformance ranges

Two combination models bound the total. Linear sum is the conservative model appropriate when contributors may drift correlatedly or when the link is qualified against worst-case stack-up; the resulting total spans 1.5–3.4 dB across the contributor ranges, which is what justifies 3 dB as the architectural upper bound. Root-sum-square (RSS) is the appropriate model when the contributors are statistically independent and the link is qualified against expected behavior rather than worst-case stack-up; the resulting total spans 0.8–1.8 dB. The architecture uses 3 dB at the linear-sum bound as the vendor-neutral worst-case planning anchor throughout §C.3 and §C.4; qualified deployments operate closer to the RSS bound, leaving 1.0–1.5 dB of envelope as recoverable margin under qualification. This recoverable headroom is the basis for §C.7.2.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

C.7.2 Qualification Levers

The five contributors are not equally elastic under qualification. Three deployment-context choices recover the envelope from the conservative 3 dB upper bound, each compressing a specific contributor with a defined physical mechanism.

Shortened deployment lifetime. Aging is the dominant contributor (0.5–1.5 dB across the envelope) and its allocation scales with the lifetime against which the components are qualified. A 5–10 year deployment, characteristic of hyperscale refresh cycles for AI infrastructure, allows aging allocation to compress to 0.3–0.7 dB. The argument is direct: AI infrastructure refreshes faster than the 20-year telecom heritage of the Telcordia framework assumes, and the envelope is permitted to reflect that.

Controlled environment. Temperature drift contributors are characterized over the 0–85 °C industrial range. Data-center deployments with direct-attach liquid cooling at the OE operate over a narrower range — typically 25–60 °C — compressing temperature contribution to 0.1–0.3 dB. The architecture's preference for direct-attach liquid cooling (§ Design Goals) supports this tightening structurally rather than as an after-the-fact accommodation.

End-to-end pair qualification. Component-level production tolerance assumes worst-case stack-up of independently qualified parts. Qualification of the deployed OE Tx/Rx pair as a system — characterized against the topology in use rather than against unit-level conformance ranges — compresses production-tolerance allocation to 0.1–0.2 dB. The envelope-and-contract methodology (§ Envelope and Contract) supports this directly: the OE interface contract makes the qualified-pair boundary explicit so system-level qualification can attribute against it without re-deriving component-level conformance.

Composing the three levers under a blended combination model — RSS for the small statistically independent contributors and linear addition for the aging term — a hyperscale deployment with 5–10 year refresh, controlled environment, and end-to-end

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

pair qualification operates against an envelope of approximately 1.5–2.0 dB. This recovers 1.0–1.5 dB of margin relative to the 3 dB worst-case. The architecture does not prescribe this tightening; it surfaces the envelope as an explicit design variable that the integrator chooses against deployment context, with the § Envelope and Contract framework providing the qualification surface.

C.7.3 Closure Consequence

The envelope is the most architecturally significant lever in the three-lever framing of §C.3.6 because it is the only lever where qualification recovers margin without changing the OE design or the topology. Recovered envelope flows directly into the §C.4.4 closure visibility map, and the §C.5.2 SNR margin map, and cell-level shifts are concrete.

At PAM4-224G with envelope tightened from 3.0 dB to 1.8 dB (1.2 dB recovery under blended qualification), the §C.3.5 OCI-class shuffle inter-rack L_total drops from 12.2–17.2 dB to 11.0–16.0 dB. The §C.4.4 closure flag at the OCI-class median shifts from firm ◐ (residual margin 1.3 dB) to ✓ (residual margin 2.5 dB). The same recovery shifts inter-pod (each half) at OCI-class from ◐ to ✓, restoring PAM4-224G as a closure-viable format across the upper end of the topology range without requiring coherent-lite or format conversion.

At PAM4-448G the consequence is larger. Envelope tightening from 3.0 dB to 1.8 dB shifts P2P intra-rack and ODF-routed intra-rack at OCI-class from ✗ to ◐, restoring closure viability at intra-rack topologies that the worst-case envelope places out of reach. This is the regime where envelope recovery has its largest architectural consequence: it is the difference between a 448 Gb/s/lane using PAM4 deployment requiring coherent-lite or O→E→O format conversion across the intra-rack topology set, and a 448 Gb/s/lane using PAM4 deployment closing directly with end-to-end qualified OE pairs.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

The architectural conclusion is that the envelope is not the fixed lever in the three-lever framing — it is the qualification-elastic lever. The 3 dB value is the vendor-neutral worst-case planning anchor used throughout §C.3 and §C.4 to keep the framework cross-vendor and unqualified-deployment-safe; the 1.0–1.5 dB of recoverable envelope is what qualified hyperscale deployments unlock. Restating the three-lever framing of §C.3.6 with the sharper terms developed here: OE design (5–7 dB swing, architectural choice), topology (2–3 dB swing, system choice), and envelope (1.0–1.5 dB recoverable, qualification choice). The full qualification-coordinated budget swing across the three levers is 8–11 dB — substantially larger than any single lever's range, because the levers compose additively (in log scales like dB, and multiplicatively in linear scales) when an integrator coordinates them through the § Envelope and Contract envelope-and-contract methodology.

C.8 Open Items and Vendor-Data Boundaries

The appendix's framework is vendor-neutral by construction: per-mechanism IL ranges in C.2 come from public-literature distributions, the per-topology accounting in C.3 composes those ranges without committing to specific vendor values, and the envelope work in C.4–C.7 reserves a planning value (3 dB) that any qualified deployment can tighten under defined conditions. The framework operates on planning ranges where vendor data is incomplete, and the resulting closure flags and SNR margin maps carry the literature distribution explicitly. C.8 surfaces the items where vendor disclosure or standards-body characterization would compress those ranges into specification values and where the architecture's framework would, in consequence, produce tighter closure visibility and sharper SNR margin envelopes.

The open items group into three categories by what data closure would require.
Category 1 — Items where public anchors exist but vendor disclosure would compress the range. The on-resonance modulator IL for production ring resonators (C.2 carries 1–3 dB; foundry disclosure would compress to per-process distributions of perhaps ±0.3 dB around a process-specific midpoint). The edge coupler IL for

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

SiPh-to-SMF attach (C.2 carries 0.5–2 dB; per-foundry production distributions would compress to ±0.2 dB around process midpoints, with the public state of the art at 0.5–1.0 dB for mature processes). The PD responsivity drift over a 5–10 year deployment (C.7 carries 0.5–1.5 dB for the aging envelope, dominated by laser decay; PD-specific drift characterization would partition the aging envelope and let the deployed-environment levers operate on each contributor independently). The MZM IL distribution at production scale for 56 GBaud and 112 GBaud PAM4 operating points (C.2 carries 4–6 dB on-chip; production characterization at 448G symbol rates would either confirm or reset the upper bound).

Category 2 — Items where the public state of the art does not yet support a defensible point estimate. Coherent-lite receiver sensitivity at short-reach intra-datacenter and scale-up applications is the dominant entry; the 5–15 dB sensitivity advantage relative to PAM4 IM/DD is correct in its envelope but the within-envelope point depends on TX/RX vendor implementation choices (modulator format single-ended vs. I/Q, LO integration and its phase-noise floor, DSP equalizer depth and architecture, FEC interaction) that the public state of the art has not yet characterized for short-reach scale-up applications. The coherent-lite pJ/bit floor is the second-tier open item — the energy cost of the sensitivity advantage depends on the same vendor implementation choices and is the lever § Power Efficiency at Scale power accounting manages. The 448 Gb/s/lane using PAM4 modulator BW headroom is a third entry; at 112 GBaud the modulator BW-vs-IL trade-off has been characterized at research scale but not yet at production scale for SiPh-foundry MZM and ring designs.

Category 3 — Items where partitioning the planning envelope against deployed component mix is the closing requirement. The 3 dB envelope of C.7 partitions cleanly across five contributors (aging, temperature drift, PDL, WDL, production tolerance), but the partition itself rests on telecom-heritage allocations (Telcordia GR-468) and on IEC connector grading rather than on AI-deployment-specific

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

component-mix characterization. A deployment specific partition of the SiPh-ELS, RRM-WDM, MZM-DR, and coherent-lite component mixes characterized at intra-DC scale, is the data that would let qualification levers compress the envelope below its current 1.5–2.0 dB qualified-deployment floor. Vendor and operator collaboration through OCP characterization disclosure is the cleanest path; OIF EEI COI WP work is a parallel path with consistent semantics.

The OCI MSA v1.0 specification (§2.4) is the strongest current anchor for Category 1 items at standards-body scale — it characterizes the external link budget at 2.5 dB for the shuffle-routed inter-rack topology class and provides a vendor-neutral reference against which the C.3 accounting reproduces. The architecture's framework is consistent with OCI's standards-derived envelope at the topology classes OCI characterizes; the framework extends OCI's scope at both directions — toward zero-connector bonded intra-tray topologies that OCI does not enumerate, and toward 448 Gb/s/lane rates that exceed OCI v1.0's specification window.

The framework's value to ecosystem participants is the vendor-neutral basis it provides for characterization disclosure: a vendor reporting modulator IL at PAM4 56 GBaud, a foundry reporting edge-coupler IL distributions, or an operator reporting envelope qualification results for a specific deployment-context combination each contributes to a single, composable parameter space the framework references. Closure of the open items above is, in this sense, an OCP/OIF/IEEE/MSA collaboration that the framework is designed to absorb without restructuring.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Appendix D: Architectural Anchors and Sensitivities

The architecture rests on four quantitative claims that recur through the document as the empirical anchors behind strategic conclusions. This appendix derives each claim from physical-layer reasoning, identifies the assumptions on which the derivation rests, and documents the sensitivity of each value to assumption variation. The intent is auditability: a reader who wishes to challenge a strategic conclusion can do so by attacking the underlying numerical anchor rather than by sniping at downstream consequences. The Target Outcomes table carries the headline operational targets with their citation sources; Appendix C carries the link-budget numerical work in full; this appendix occupies the architectural-physics layer that bridges those two.

D.1 Perimeter versus area scaling

Claim. Integrated electrical I/O BW grows linearly with chip side length L, while logic capacity grows with chip area L^2. This regime change motivates Areal Packaged Optics (APO) classification which would be the integration of any electrical pin to any optical I/O over the die area, as a structural break from perimeter-limited (CPO/NPO) Optical Engine placement, and underwrites the approximately 2.65× material-efficiency advantage demonstrated in published 3D-stacked photonic interposer comparisons against 2D NPO and CPO ring placement at equivalent aggregate BW.

Derivation. For a die or package of side length L, perimeter available for shoreline I/O scales as 4L, while logic area scales as L^2. BW-to-compute ratio — the binding metric for AI training systems where parameter and gradient traffic scales with compute — therefore declines as $L/L^2 = 1/L$ for perimeter-limited integration and remains constant for area-integrated integration. The starvation zone widens with package size; this is the structural physics behind shoreline-limited I/O.

Assumption. Single-reticle scope and isotropic waveguide routing: waveguides exit through any face of an integrated photonic interposer rather than only through a

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

designated shoreline. Multi-reticle stitching (demonstrated publicly across reticle-stitch progressions from millimeter-scale single reticles to multi-centimeter stitched envelopes) preserves the L^2 scaling across reticle boundaries through cross-reticle waveguide routing; the perimeter-versus-area asymmetry persists at the multi-reticle envelope. Sensitivity. The 2.65× figure follows from material accounting under standard 2D ring-bus and 3D-stacked area-integrated geometries at equivalent aggregate BW. Substituting ring-resonator and MZM-cascade hybrid placements compresses the advantage by approximately 0.3 to 0.5×; aggressive 3D stacking with multiple photonic layers expands it. The architectural conclusion — that APO is a regime change rather than a placement-continuum point — is robust to factor-of-two variation in the headline ratio because the regime change is in the scaling exponent, not in the constant.

Realizability. The 2D-to-3D transition is where this scaling-exponent change is physically taken: perimeter-bound (2D) co-packaged placement maximizes shoreline density but remains L-limited, while area-domain (3D) co-packaged placement distributes optical escape across the die face and enters the L^2 regime. Beyond the material-efficiency advantage derived above, the enabling capabilities of this transition are demonstrated rather than projected. Area-domain optical escape at fiber counts and aggregate bandwidth consistent with L^2 scaling has been shown on interconnect substrates positioned directly beneath multi-kilowatt compute complexes; the associated under-die thermal contract has been demonstrated to hold at co-packaged power densities on the order of 1.4 W/mm² and under rapid-aggressor and extended 25–105 °C cycling, addressing the modulator-stability and thermal-coupling concerns that have historically gated 3D integration; and detachable, vertically coupled fiber attach at re-insertion loss on the order of 1.5 dB has been demonstrated in high-volume advanced-packaging flows without active alignment, satisfying the serviceability (Goal 7) and manufacturability (Goal 10) contracts together. These capabilities are realizable across multiple merchant foundries and OSATs rather than a captive process, which is the manufacturability basis the open model requires. The architecture therefore treats

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

the area-domain regime (Goal 9) as a reachable forward target whose enabling capabilities exist today, while leaving the specific coupling, modulation, and packaging choices open to any implementation that satisfies the published contract.

D.2 Single-reticle perimeter ceiling

Claim. The practical aggregate BW ceiling for single-reticle perimeter-limited integration is approximately 51.2 T at current SerDes generations. This is the binding constraint that makes 3D CPO or multi-reticle integration architecturally necessary for next-generation systems; lane-rate increases postpone but do not eliminate the ceiling. Derivation. At the 26 mm reticle limit and a four-side shoreline of approximately 104 mm, allowing approximately 50 percent of perimeter for fiber-attach geometry (the remainder reserved for power delivery, cooling, and mechanical mounting) yields approximately 52 mm of usable shoreline. At a 2.0 Tb/s per mm bidirectional shoreline density target consistent with public 224 Gb/s per lane direct-drive OE demonstrations, the perimeter capacity computes to approximately 52 × 2.0 ≈ 104 T raw, derated by approximately 50 percent for routing and density-loss factors to approximately 52 T usable. This brackets the current-generation switching-silicon benchmark of 51.2 T and is consistent with the next-generation 102.4 T target enabled by lane-rate doubling.

Assumption. 224 Gb/s/lane using PAM4 modulation, single-reticle scope, standard fiber-array attach geometry at 250 µm pitch with optional 127 µm tight-pitch variants. Migration to 448 Gb/s/lane and beyond doubles the per-mm ceiling at the cost of tighter signal-integrity budgets that compress effective shoreline (see Appendix C link-budget derivation); the net effect is approximately 1.5 to 1.7× ceiling growth per SerDes generation against the demanded 2× compute growth of frontier AI systems. The ceiling therefore postpones the constraint by approximately one generation per lane-rate doubling without eliminating it.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Sensitivity. The 50 percent perimeter availability assumption is the dominant uncertainty. Aggressive package design can recover up to 60 percent perimeter utilization, lifting the ceiling to approximately 62 T at 224G; pessimistic designs serving high power-delivery densities can fall to 40 percent, compressing to approximately 42 T. The 2.0 Tb/s per mm shoreline density assumes mature direct-drive integration; retimed pluggable approaches reduce effective per-mm density by 30 to 40 percent due to retimer keep-out zones. The architectural conclusion — that single-reticle perimeter-limited integration faces a binding ceiling near current-generation switching radix — is preserved across this range.

D.3 Retimer latency accumulation

Claim. Multi-stage retimed copper paths accumulate 40 to 120 ns of structural latency that topology can reduce but not fully route around.[24]. Non-retimed photonic paths exhibit propagation-dominated latency below 10 ns for intra-rack distances. Latency is therefore a co-equal architectural requirement alongside BW, not a secondary consequence of BW choice.

Derivation. Each retimer stage contributes approximately 5 to 15 ns of latency: CDR lock and recovery at 1 to 3 ns, SerDes deserialization and reserialization with deep equalization at 3 to 8 ns, and FEC encoding and decoding when applied at 1 to 4 ns. The range reflects vendor-published values across current and roadmap retimer products [OIF COI WP §3.2]. Typical multi-stage retimed copper paths in current and proposed scale-up domains traverse 3 to 8 retimer stages from source XPU to destination XPU depending on tray and rack topology; the compound contribution is 40 to 120 ns at the design center, with the upper bound determined by FEC-bearing paths and the deep equalization required for the longest retimed copper segments. Optical

[24] Topology can reduce the number of retimers a path traverses, but any retimer that remains contributes its own stage latency.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

propagation contributes 5 ns per meter [OIF COI WP §3.2]; sub-10 ns is achievable for intra-rack distances of 1 to 2 m, propagation-dominated.
Assumption. Per-stage retimer latency contributions are approximately additive across the multi-stage path. FEC contribution is bypassed in some retimer architectures; the 5 to 15 ns per-stage range includes this variability. Optical FEC contribution, when present in retimed photonic configurations, is comparable to electrical FEC and is absorbed in the per-stage latency.

Sensitivity. The minimum retimer-stage count of 3 and maximum of 8 bracket the operational range across current and roadmap scale-up topologies. Systems with longer copper segments require additional retimer stages and accumulate proportionally; systems eliminating retimers entirely (direct-drive optics) recover the entire latency budget. For AI training workloads with synchronous step boundaries, the elimination of 40 to 120 ns per scale-up traversal directly improves Model FLOPS Utilization by reducing communication overhead and stall cycles; the magnitude is workload-dependent and is the subject of §11 System-Level Impact on AI Workloads.

D.4 Rack-level interconnect energy savings

Claim. Migration from pluggable optics to co-packaged photonic placement classes recovers approximately 50 to 150 kW per rack at the 1.5 PB per second aggregate interconnect BW representative of next-generation dense AI training systems.
Derivation. Aggregate BW of 1.5 PB per second equals 12×10^{15} bits per second. Per-bit energy delta: pluggable optics at 15 to 20 pJ per bit (current production, per Target Outcomes) migrating toward co-packaged classes at ~5 pJ/bit yields a per-link nominal delta of 8 to 20 pJ per bit. Realized delta varies by topology class: bonded intra-tray and pure CPO classes recover the full per-link delta; ODF-routed intra-rack and format-converted inter-pod classes recover less because pluggable, retimer, or format-conversion energy persists at the rack boundary. Across a representative mixed-class realization where 40 to 80 percent of aggregate BW migrates to

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

co-packaged classes and the remainder remains pluggable or format-converted, the realized savings compute to 12 × 10^{15} bit/s × (4 to 16) × 10^{-12} J/bit = 48 to 192 kW per rack. The 50 to 150 kW range cited in § Economic Benefits reflects this mixed-class realization at central assumptions.

Assumption. 1.5 PB per second aggregate represents unidirectional throughput across the rack scale-up domain at next-generation density (consistent with the Class V 288-XPU and beyond density targets in § Platform Form Factor Evolution). Per-bit energy values are at-port averages, not including ELS and management overhead; full-system accounting would adjust by approximately 10 to 15 percent. The mixed-class realization assumption (40 to 80 percent co-packaged) reflects realistic transition deployments where the highest-volume scale-up links migrate first while inter-pod and edge links retain pluggable form factors longer.

Sensitivity. The dominant levers are the migration fraction (which sets the realized delta) and the aggregate BW assumption (which scales linearly). At full CPO migration (100 percent co-packaged), savings recover toward an upper bound near 240 kW at the 20 pJ per bit delta; at 40 percent migration the lower bound near 50 kW applies. The architectural conclusion — that interconnect energy savings are of the same order of magnitude as a typical rack power budget — holds across the assumption range.

D.5 Escape across tiers and SerDes generations — the CPO/APO frontier

The working sections hold the near-term operating point — the 102.4T and 204.8T tiers at 200G SerDes, where discrete pluggable optics still provision the rack. This appendix carries the ladder up to the 409.6T (N+2) tier and the 400G-class SerDes it rides, where the discrete bill stops fitting and only co-packaged (CPO) and area (APO) optics remain in scope. The escape table shows why: copper and CPO are both perimeter resources fixed by edge density, so they fall multiples short and stay short up the ladder; only the area escape grows with the lane rate, and only it clears the tier.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table AP: Escape Ladder — Copper, CPO, and Area Across the Switch Tier Ladder

Switch tier (SerDes)	Scale-up need (Tb/s/dir)	Copper edge (× short)	CPO edge (× short)	APO area ceiling (Tb/s/dir)	APO radix @200G	Frontier
102.4T (224G)	102.4	26 · 3.9× short	83 · 1.2× short	**1,191**	5,956	discrete viable; APO clears
204.8T (224G)	204.8	26 · 7.9× short	83 · 2.5× short	**1,191**	5,956	edge short; APO only
409.6T (448G)	409.6	26 · 15.8× short	83 · 4.9× short	**2,382**	11,910	copper off the table; CPO stays short, APO clears

Per single reticle (104 mm perimeter, ~858 mm² area). Copper escapes 1.0 Tb/s/mm bidi over ~50% usable shoreline (≈ 26 Tb/s/dir); CPO escapes 2.0 Tb/s/mm over the ~80% beachfront optics free once they replace the edge SerDes and interleaved PDN (≈ 83 Tb/s/dir, ~3× copper), per OIF-EEI-COI-WP-01.0. Neither edge density scales with the SerDes generation; only the area ceiling grows with the lane rate (1,191 → 2,382 Tb/s/dir at 224G → 448G). "× short" is need ÷ edge escape per direction. APO radix normalized to a 200G/λ optical payload.

Table AQ: Connector and Laser Bill Across the Ladder — Where Discrete Terminates

Switch tier (SerDes)	Scheme	Data fibers	Front SN-MT32	Rear EB-192	ELSFP	Trays/tier (28 mm)	Discrete fit
102.4T (224G)	DR	1024	8	6	16	2	fits
	OCI	512	4	3	16	1	fits
	8+8	64	1	1	2	1	fits
	XPO (integ.)	1024	16	—	—	0.5 OU	—

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Switch tier (SerDes)	Scheme	Data fibers	Front SN-MT32	Rear EB-192	ELSFP	Trays/tier (28 mm)	Discrete fit
204.8T (224G)	DR	2048	16	11	32	3	tight
	OCI	1024	8	6	32	2	fits
	8+8	128	1	1	4	1	fits
	XPO (integ.)	2048	32	—	—	1.0 OU	—
409.6T (448G)	DR	4096	32	22	64	6	over → CPO/APO
	OCI	2048	16	11	64	4	over → CPO/APO
	8+8	256	2	2	8	1	fits
	XPO (integ.)	4096	64	—	—	2.0 OU	—

D.6 Key claims and evidence

Table AR consolidates the load-bearing quantitative claims with evidence status. Status vocabulary: Public (published data), Modeled (quantitative-kernel result), Derived (arithmetic on stated inputs), Assumption (industry-typical estimate), Target (architecture goal), Forward (pre-demonstration band).

Table AR: Key Claims and Evidence

Claim	Value	Status	Basis
Electrical scale-up link energy	15–25 pJ/bit	Public	OIF-EEI band
Pluggable optics energy	15–20 pJ/bit	Public	OIF-EEI; 1.6T module cross-check

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

CPO energy	5–7 pJ/bit	Public, demonstrated	Vendor disclosures
Area photonic (APO) energy	2.3–5 pJ/bit	Forward	3D SiPho floor to CPO ceiling
Network power share, Blackwell-class pluggable	≈31%	Modeled	§ Rack-level interconnect energy savings
Network power share, Rubin-class pluggable	≈37%	Modeled, counterfactual	Same model, Rubin frontier
Network power share, Class V switchless	≈15%	Modeled, forward	APO band operating point
Power recovered at the Class V point	≈100 kW per 144-package rack; ≈400 kW per 576-XPU domain	Modeled, forward	Same model
MTBF, CPO versus pluggable	2.6M vs 0.55M hours (5×)	Public	Production deployment disclosure
MFU, best reported	38–43%	Public	Llama 3 405B pretraining
MFU target	55–65%	Target	§ Target Outcomes
HBD domain scale	72 → 1,024–1,152 XPU	Target	§ The Scale Boundary
Dense-rack power envelope	600 kW+	Public	Announced rack disclosures
288-package rack	2× beyond announced 144	Assumption	§ Platform Form Factor Evolution

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Retimer bill, 4-package tray	16–32 units at 5–15 W	Assumption	Industry-typical
8-package tray heat	5.6–9.6 kW	Derived	8 × 700–1,200 W

Discrete fit from trays/tier at 28 mm (≤ 2 fits, 3 tight, ≥ 4 over a sane faceplate; a 48 mm tray roughly halves the tray count). At 409.6T the discrete DR bill (4096 fibers, 22 rear EB, 6 trays) is off the faceplate; CPO is faceplate-competitive against the integrated XPO baseline only at the dense 8+8 DWDM rollup, and APO escapes data through the package area — the front then carries lasers only. Front 4-port SN-MT32 = ⌈fibers/128⌉; rear EB-192 = ⌈fibers/192⌉

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Appendix E: Time-Axis Anchors and Sensitivities

The architecture's third value axis is time: the utilization recovered when photonic reach lets a workload address endpoints that copper reaches only by proximity. It rests on two quantitative claims that recur in §5.2 and §11. This appendix derives each from queueing and memory-hierarchy reasoning, states the assumptions each rests on, and documents the sensitivity of each value to assumption variation. The intent matches Appendix D: auditability, so a reader can challenge a strategic conclusion by attacking its numerical anchor. Appendix C carries the link-budget basis and Appendix D the escape-scaling basis; this appendix carries the time basis. A single input is modeled rather than measured: the pooled-tier latency, flagged where it enters.

E.1 The memory-hierarchy access-time cliff

Claim. Inserting a memory-semantic, bounded-latency, composably-reachable sub-microsecond pooled tier between local memory and bulk storage lowers average memory access time by roughly six- to tenfold, set by the storage path it displaces. This is the first-order time-axis lever.

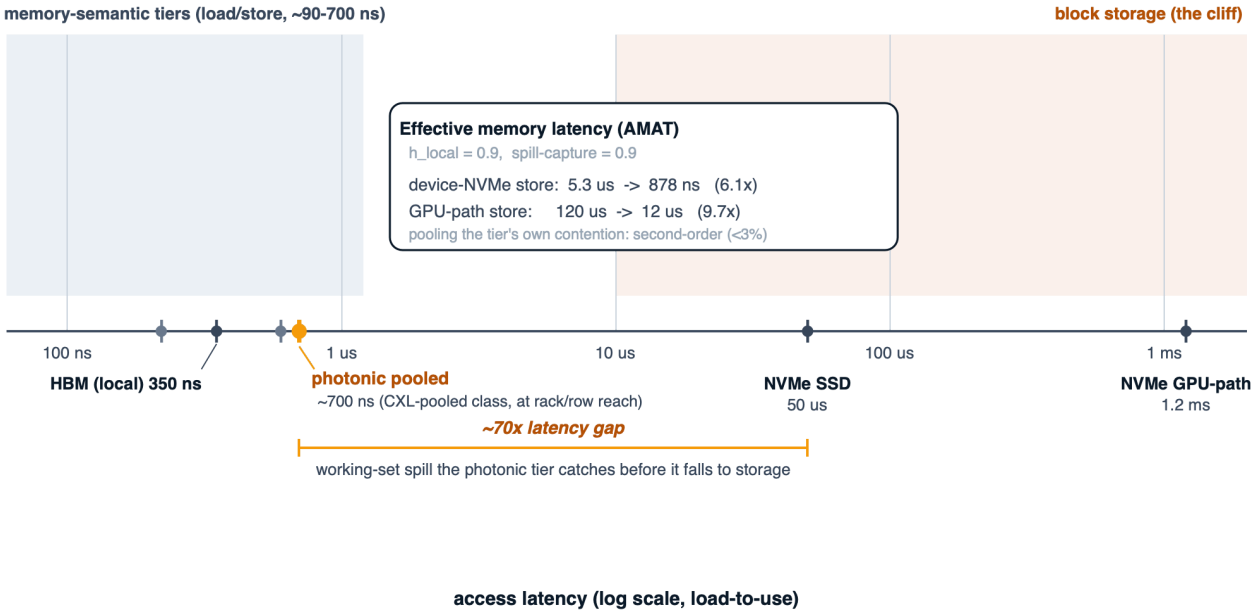
Derivation. Average access time is the load-to-use latency weighted across tiers. A local hit fraction h is served at local memory latency, near 350 ns for HBM. Without the pooled tier, the spill fraction 1 − h falls to bulk storage at tens of microseconds for device NVMe, near 50 µs, or to roughly 1.2 ms on a networked GPU storage path. With the pooled tier, a capture fraction g of that spill is served near 700 ns instead. The lever is the inter-tier latency ratio: the pooled tier sits about 70× faster than device NVMe and about 1,700× faster than the GPU path. At h ≈ 0.9 and g ≈ 0.9, average access time falls from 5.31 µs to 878 ns against device NVMe, a factor of 6.1, and from 120 µs to 12.4 µs against the GPU path, a factor of 9.7.

Assumption. Tier latencies are load-to-use values from published measurements on current accelerator and CXL® memory hierarchies: HBM near 350 ns, CXL®-attached

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

near 220 ns, CXL®-pooled near 600 ns, device NVMe near 50 µs, and GPU-path NVMe near 1.2 ms. The pooled-tier latency near 700 ns is the single modeled value: a CXL®.mem controller term, rack-and-row fiber propagation at about 5 ns/m, and transceiver latency. It is CXL®-pooled-class latency carried to rack and row reach, not latency lower than electrical coherence.

Sensitivity. Two knobs move the result, and the conclusion survives both. Local hit rate: as h rises from 0.8 to 0.99 the gain falls from 7.3× to 2.1× against device NVMe, because a higher local hit rate leaves less spill for the tier to rescue. Spill-capture is the load-bearing knob: as g rises from 0.5 to 1.0 the gain rises from 1.9× to 13.8×, so the tier earns its place only where the working set is reused at the tier's capacity. The modeled 700 ns is not load-bearing; it sits an order of magnitude inside the 70× gap to storage, so a factor-of-two error in it barely moves the ratio. The architectural conclusion holds across the range: the lever is the order-of-magnitude tier gap, not the modeled constant. AI traffic supplies the capture that earns the gain through KV-cache, embedding tables, expert weights, and optimizer state; cold streaming does not benefit.



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Figure E.1. The memory-hierarchy load-to-use ladder: the roughly seventyfold gap from the sub-microsecond pooled tier to bulk storage, and the resulting average-access-time reduction at h ≈ 0.9 and capture ≈ 0.9.

E.2 Pooling the reachable endpoint set

Claim. Pooling the endpoints that photonic reach makes addressable, folding many small allocation domains into one larger high-bandwidth domain, cuts the mean remote-access wait by roughly 65× and the straggler that gates a wide synchronized collective by roughly tenfold, at equal load factor. The payoff concentrates in the tail.

Derivation. Erlang trunking governs the gain: one pool of c servers at load factor ρ blocks and waits less than k pools of c/k servers at the same ρ, because demand variance averages over the larger trunk group. Folding eight 8-endpoint domains into one 64-endpoint high-bandwidth domain at ρ = 0.8 drops the mean wait from 0.286 to 0.0044 service-times, about 65×. For a 512-way collective gated by its slowest participant, the step-gating wait falls from 6.29 to 0.62 service-times, about 10×. A synchronized collective pays the tail, not the mean.

Assumption. The pooled server set is the high-bandwidth domain, scale-up plus Scale-DSA; scale-out is a slower service class held out of the pool. Arrivals are modeled at the Erlang baseline, at load factor near 0.8 and a remote-access fraction of 0.1 to 0.2.

Sensitivity. The trunking gain grows with pool size and concentrates in the tail, so the lever strengthens where it is needed. Mean utilization recovery at ρ ≈ 0.8 is a few percent; the absolute mean is not the claim. Real AI traffic is burstier than the Erlang baseline, which makes pooling more valuable rather than less: a small pool drowns in a burst that a large trunk absorbs. The conclusion that reach buys utilization through the tail holds, and strengthens, under the arrivals real workloads present.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

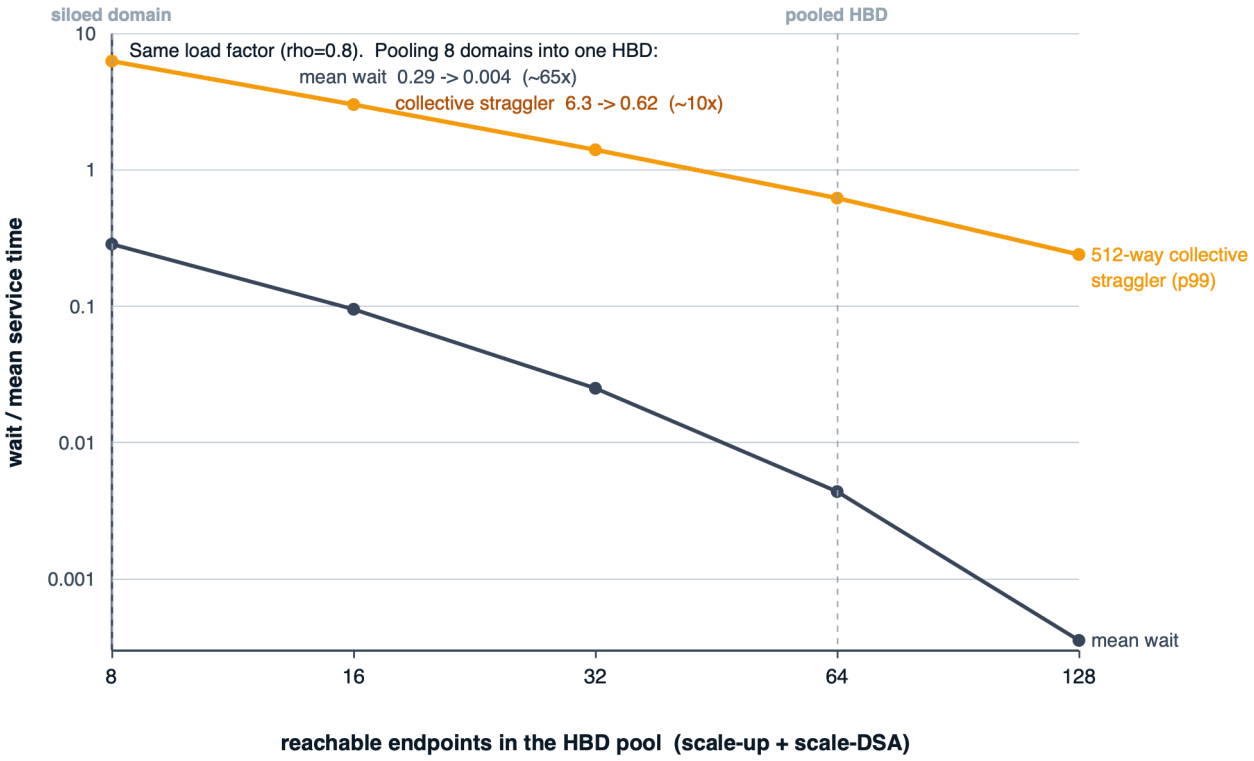


Figure E.2. *Mean remote-access wait gating a 512-way collective*

Mean remote-access wait and the p99 straggler gating a 512-way collective, against pool size at load factor ρ = 0.8. The gain concentrates in the tail, where synchronized collectives are paid.

The third lever, and the ordering. Fire-and-forget issue on a bounded round-trip hides the pooled tier's latency behind concurrency, with the concurrency needed to fully hide it set by Little's Law as the product of request rate and latency; a bounded round-trip, which reach guarantees, makes that concurrency provisionable. This lever is real but smaller. The ordering is fixed: the memory-hierarchy ratio first, pooling second, scheduling third

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

16. Glossary

APO: Areal Packaged Optics.

BiDi: Bidirectional.

BW: Bandwidth

COI: Compute Optics Interface (per OIF EEI/COI work).

COM: Channel Operating Margin.

CPC: Co-Packaged Copper.

CPO: Co-Packaged Optics.

CW: Continuous Wave.

CXL®: Compute Express Link.

DSA / Scale-DSA: Domain-Specific Architecture; scaling axis across heterogeneous compute architectures.

DSP: Digital Signal Processor.

DWDM: Dense Wavelength Division Multiplexing.

EB: Expanded Beam

EB MSA: Expanded Beam Optics Multi-Source Agreement.

EEI: Energy Efficient Interfaces (per OIF EEI/COI work).

ELS/ELSFP: External Laser Source / External Laser Small Form Factor Pluggable.

ESUN: Ethernet Scale-Up Networking.

FAU: Fiber Array Unit.

FEC: Forward Error Correction (RS-FEC, KP4 variants per IEEE 802.3).

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

FEI: Front End Interconnect (per OIF EEI Framework Document, OIF-EEI-FD-01.0) intra-datacenter communications, standards-based interoperability emphasis, typically Ethernet

FPO: Faceplate Pluggable Optics.

FRU: Field Replaceable Unit.

HBD: High Bandwidth Domain.

HBM: High Bandwidth Memory.

HPM: Host Processor Module.

HSIO: High Speed Input Output

IM/DD: Intensity Modulation / Direct Detection.

IOWN GF: Innovative Optical and Wireless Network Global Forum.

ISL: Inter Switch Link.

IVR: Integrated Voltage Regulator.

MFU: Model FLOPS Utilization.

MGX: NVIDIA/OCP Accelerated Computing Rack and Trays specification (modular rack/tray form factor).

MHS: Modular Hardware System.

MHS-IAI: Modular Hardware System-Interconnect for AI[25]

MSA: Multi-Source Agreement.

M-NPO: Mezzanine Near-Packaged Optics.

MZM: Mach-Zehnder Modulator.

[25] Proposed name for OCP Workstream

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

NPO: Near-Packaged Optics.

OCI: Optical Compute Interconnect (MSA).

OCS: Optical Circuit Switching.

ODF: Optical Distribution Frame.

OE: Optical Engine.

OFC: Optical Format Converter.

ORv3: Open Rack v3.

OSAI: Open Systems for AI.

OSNR: Optical Signal-to-Noise Ratio.

PAM4: 4-Level Pulse Amplitude Modulation.

PAM6: 6-Level Pulse Amplitude Modulation.

PAM8: 8-Level Pulse Amplitude Modulation.

PIC: Photonic Integrated Circuit.

PKGD: Photonic Known-Good-Die.

RAS: Reliability, Availability, and Serviceability (per OCP RAS API).

RRM: Ring Resonator Modulator.

SerDes: Serializer/Deserializer.

S-NPO: Socketed Near-Packaged Optics.

SiPh: Silicon Photonics.

SMF: Single-Mode Fiber.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

SOCP: Structured Optical Cabling Plane. Factory-terminated, keyed, labeled, and loss-qualified passive waveguide assembly that carries a defined fabric plane as a replaceable unit; may be realized as ribbon fiber, bonded fiber sheet, molded harness, planar waveguide, or another compliant passive optical format.

TIA: Transimpedance Amplifier.

UALink™: Ultra Accelerator Link.

UEC: Ultra Ethernet Consortium.

VCSEL: Vertical-Cavity Surface-Emitting Laser.

VRM: Voltage Regulator Module.

VSFF: Very Small Form Factor

WDM: Wavelength Division Multiplexing (parent term; includes CWDM, DWDM, BiDi WDM).

WSR: Wavelength-scaled radix. Effective fabric radix within a fixed connector, FAU, or tray-face envelope after accounting for fiber count, wavelengths per fiber, per-wavelength payload rate, and bidirectional directionality.

XPO: eXtra-dense Pluggable Optics.

XPU: Generic term for any accelerator

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

17. References

[1]	Priyadarshi, S., “Unlocking the Potential of Co-Packaged Optics in AI and HPC: Opportunities and Challenges,” IEEE Communications Magazine, vol. 64, no. 2, pp. 24–29, Feb. 2026. DOI: 10.1109/MCOM.001.2500504
[2]	“Accelerating Frontier MoE Training with 3D Integrated Optics,” IEEE Symposium on High-Performance Interconnects (Hot Interconnects), Aug. 2025.
[3]	OIF, “Energy Efficient Interfaces Co-Packaged Optics Interconnect White Paper,” OIF-EEI-COI-WP-01.0, Feb. 2026.
[4]	OIF, “Energy Efficient Interfaces Framework Document,” OIF-EEI-FD-01.0, 2024.
[5]	Advanced Photonics Coalition (APC), https://www.advancedphotonicscoalition.com/
[6]	Open Compute Project, “Open Systems for AI: Blueprint for Scalable Infrastructure,” Version 1.0.0., Jan. 2025.
[7]	UALink™ Consortium, “UALink™ 1.0 Specification,” April 2025.
[8]	Ultra Ethernet Consortium, “UEC 1.0 Specification,” June 2025.
[9]	Optical Compute Interconnect (OCI), OCI MSA, “200G Optical Compute Interconnect (OCI) Optical PHY Specification,” v1.0, 2025.
[10]	IEC 61753-1, “Fibre optic interconnecting devices and passive components — Performance standard — Part 1: General and guidance for performance standards.”
[11]	OCP, “Modular Hardware System (MHS) Edge-MHS HPM Form Factor Base Specification,” v0.5.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

[12]	OCP, “Open Systems for AI — ESUN: Network Operator Requirements Base Specification,” Rev. 1.0.
[13]	IEEE P802.3dj, “Standard for Ethernet — Physical Layers and Management Parameters for 200 Gb/s, 400 Gb/s, 800 Gb/s, and 1.6 Tb/s Operation.”
[14]	IOWN Global Forum, “AI Interconnect Functional Architecture,” March 2026; and OCP–IOWN AI Computing Continuum cooperation framework, February 2026.
[15]	Kartawira, I., DeMerchant, M., Nannapaneni, S., Giacalone, S., Chen, S., and Burch, J., "Optical Circuit Switching for AI and Hyperscale Datacenters: Benefits, Technologies, Current and Future Deployments," Open Compute Project, April 2026.
[16]	OCP, “MGX Accelerated Computing Rack and Trays Specification,” v1.1 (NVIDIA contribution).
[17]	OIF, OIF Energy Efficient Interfaces (EEI) and Co-Packaging Interoperability Demonstration https://www.oiforum.com/wp-content/uploads/OIF_EEI_Interop_Demo_OFC25.pdf
[18]	OIF, “External Laser Small Form-Factor Pluggable (ELSFP) Implementation Agreement,” OIF-ELSFP-02.0, January 2025.
[19]	Arista (and XPO MSA), “XPO: Redefining Pluggable Optics for AI Networking”, https://www.arista.com/assets/data/pdf/Whitepapers/XPO-Whitepaper.pdf

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

18. License

18.1. Creative Commons

OCP encourages participants to share their proposals, specifications and designs with the community. This is to promote openness and encourage continuous and open feedback. It is important to remember that by providing feedback for any such documents, whether in written or verbal form, that the contributor or the contributor's organization grants OCP and its members irrevocable right to use this feedback for any purpose without any further obligation.

It is acknowledged that any such documentation and any ancillary materials that are provided to OCP in connection with this document, including without limitation any white papers, articles, photographs, studies, diagrams, contact information (together, “Materials”) are made available under the Creative Commons Attribution-ShareAlike 4.0 International License found here: https://creativecommons.org/licenses/by-sa/4.0/, or any later version, and without limiting the foregoing, OCP may make the Materials available under such terms.

As a contributor to this document, all members represent that they have the authority to grant the rights and licenses herein. They further represent and warrant that the Materials do not and will not violate the copyrights or misappropriate the trade secret rights of any third party, including without limitation rights in intellectual property. The contributor(s) also represent that, to the extent the Materials include materials protected by copyright or trade secret rights that are owned or created by any third-party, they have obtained permission for its use consistent with the foregoing. They will provide OCP evidence of such permission upon OCP’s request. This document and any "Materials" are published on the respective project's wiki page and are open to the public in accordance with OCP's Bylaws and IP Policy. This can be found at http://www.opencompute.org/participate/legal-documents/. If you have any questions please contact OCP.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

19. About Open Compute Foundation

The Open Compute Project (OCP) brings at-scale innovations and hyperscaler best practices to all, spanning technology domains from the data center to the edge, and the technology stack from silicon, to systems, to site facilities and services. The international OCP Community is made up of organizations and people from hyperscale and tier-2 cloud data center operators, communications providers, colocation providers, diverse enterprises, and technology vendors. The OCP Foundation fosters collaboration within the Community to tackle market challenges in an array of OCP Projects that create open contributions such as specifications, designs, and more. The OCP Solution Provider Program then recognizes the application of those contributions in compliant products, solutions, and data center facilities and services with designations like OCP Inspired, OCP Accepted and OCP Ready. With the tenets of openness, impact, efficiency, scale and sustainability, the OCP engages and educates thousands of engineers every year through many events and webinars. Across many initiatives the OCP Foundation and Community are meeting the market today and shaping the future.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.