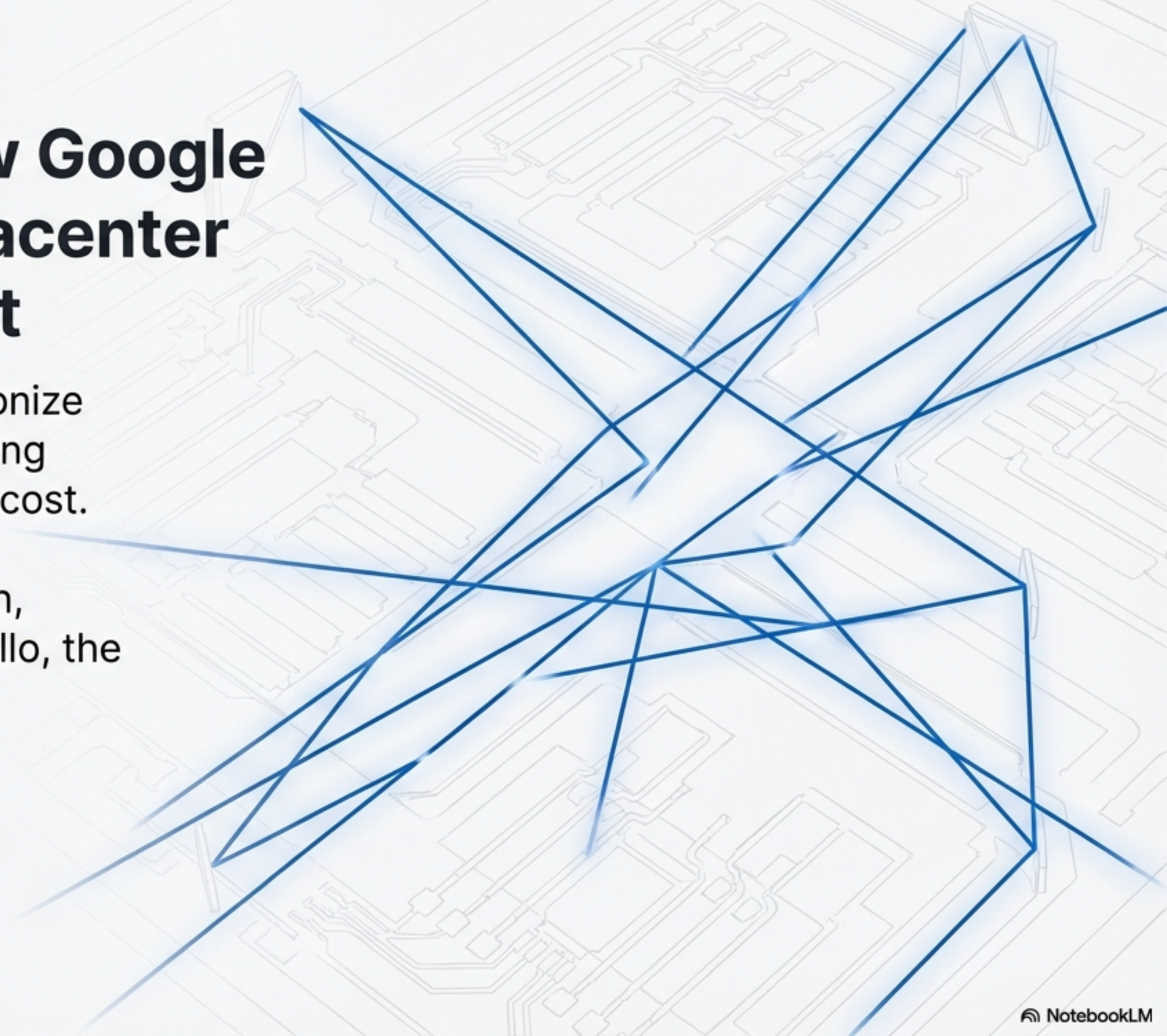
Mission Apollo: How Google Reinvented the Datacenter Backbone with Light

A decade-long journey to productionize optical circuit switching, transforming network scalability, efficiency, and cost.

This presentation details the design, implementation, and impact of Apollo, the world's first large-scale production deployment of OCS for datacenter networking.



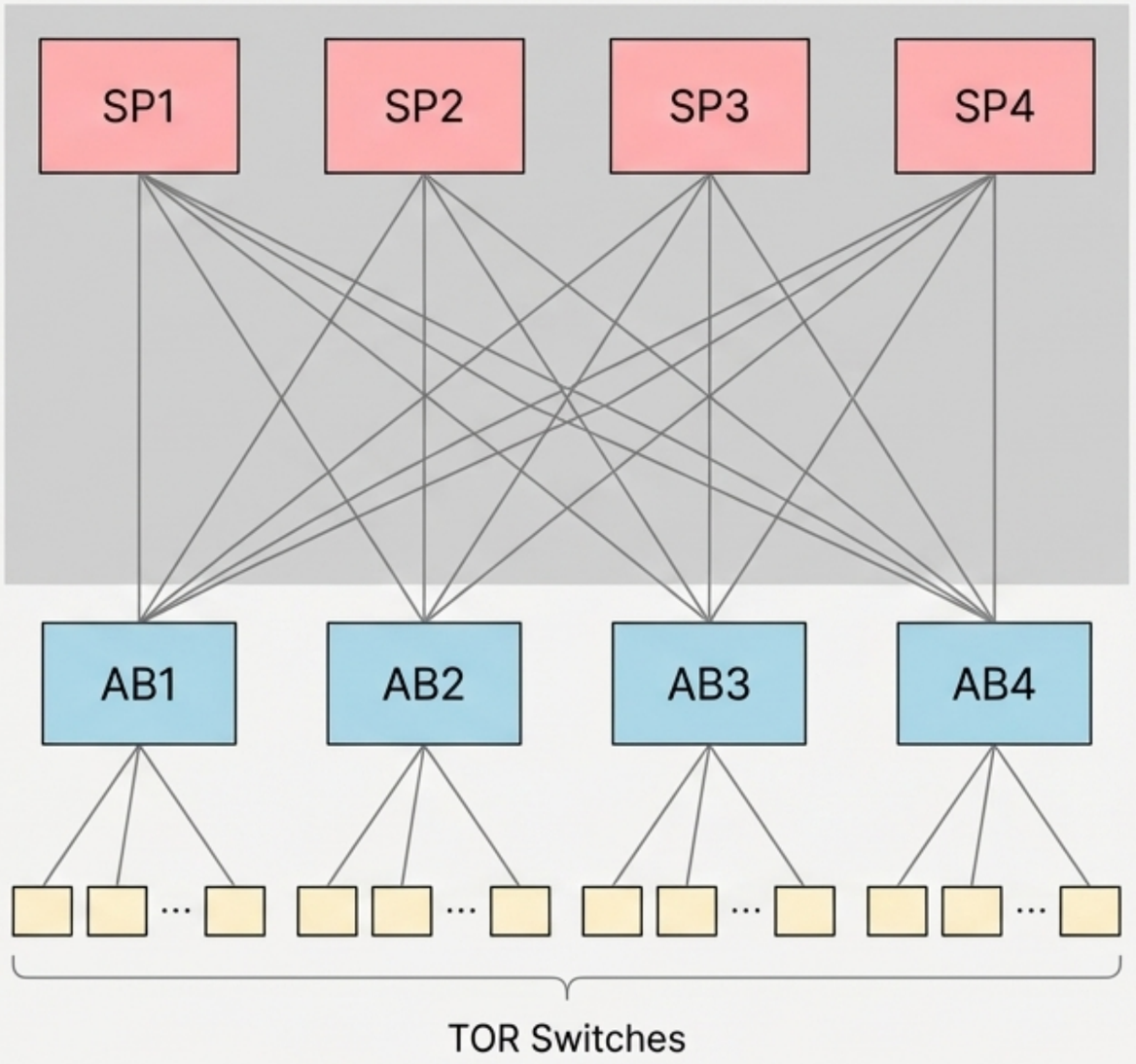
Traditional datacenter networks, built on electrical switches, are hitting a wall of complexity and cost.

Spiraling Costs: The spine layer becomes prohibitively expensive as network demands grow.

Unsustainable Power: Power consumption for the electrical spine and its associated optical interfaces becomes a major operational burden.

Inflexible Evolution: Upgrading the fabric requires costly and disruptive network-wide rebuilds, often taking months and vacating entire buildings.

The ‘Before” Architecture: The Electrical Spine



Optical Circuit Switching (OCS) long promised a solution, but was deemed impractical for production datacenters.

The Promise (Theoretical Benefits)

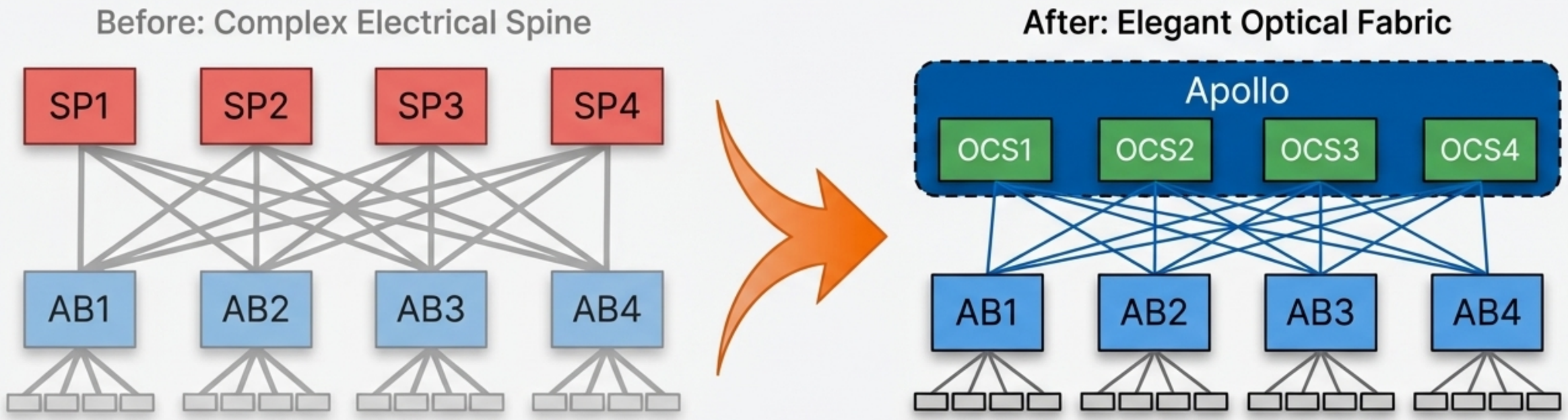
- **Data-Rate Agnostic:** Steers light without processing, allowing optics to evolve independently.
- **Low Latency:** Latency is set by the speed-of-light propagation delay (~5ns per meter in fiber), orders of magnitude lower than EPS.
- **Extreme Energy Efficiency:** Mirrors are capacitive loads; power to maintain connections is exceptionally low.

The 'Conventional Wisdom' (Perceived Barriers)

- Lack of manufacturable, cost-efficient, and reliable OCS hardware at scale.
- Concerns over switching time, signal loss, and operational complexity.

'An explicit goal of this paper is to re-examine and perhaps to reset such conventional wisdom.'

We replaced the entire electrical spine with Apollo, a dynamic optical circuit switching fabric.



The Apollo layer replaces the Spine blocks with cut-through OCSes, fundamentally changing our network architecture. This eliminates the electrical switches and optical interfaces previously used to implement the Spine, resulting in significant cost and power savings. The optical fabric now serves as the backbone for all of our datacenter networks.

At the heart of Apollo is Palomar, our internally-developed 136x136 3D MEMS OCS.

After challenges with vendor solutions, we designed Palomar with three principles: manufacturability, serviceability, and reliability.

Port Count

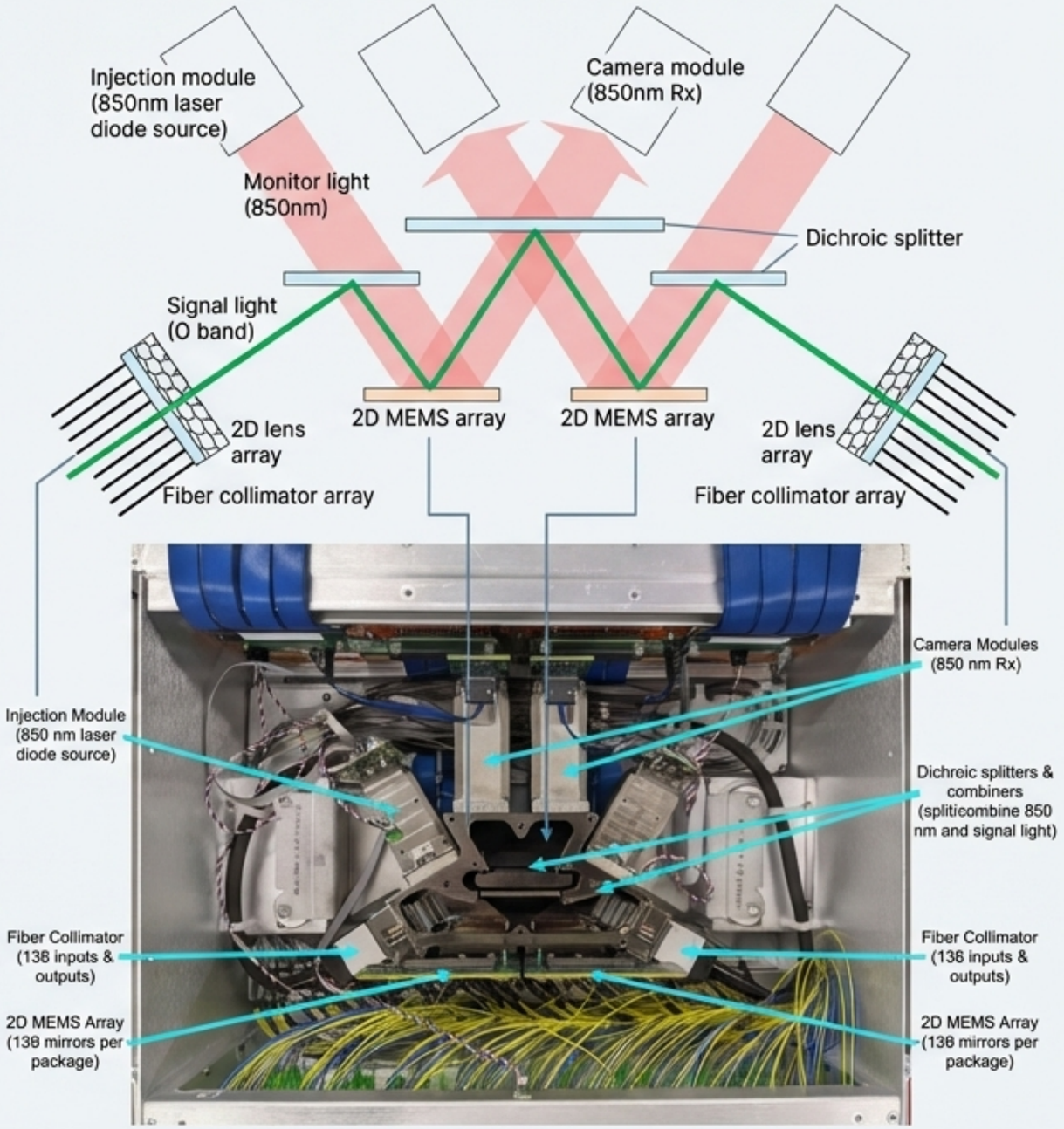
136x136 non-blocking connectivity.

Core Technology

Based on two 2D MEMS mirror arrays that steer light beams.

Key Innovation

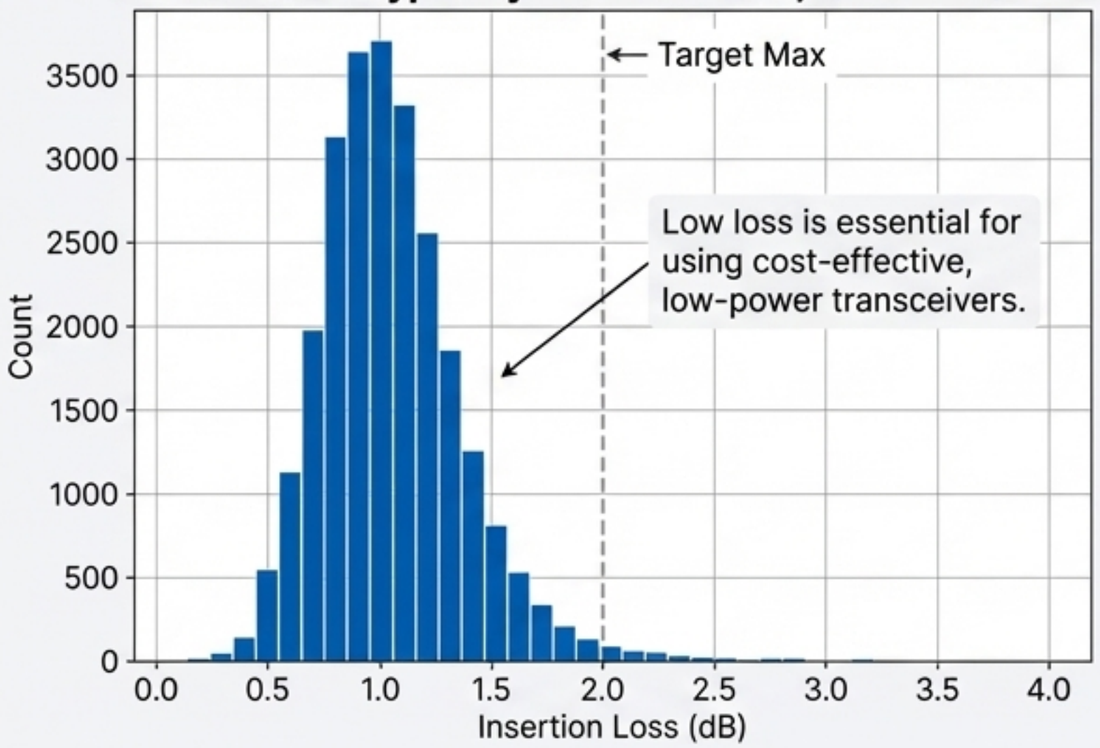
A camera-based feedback system simplifies mirror control. A single camera image per MEMS array is used to optimize alignment, a critical choice for enabling a low-cost, manufacturable solution at scale.



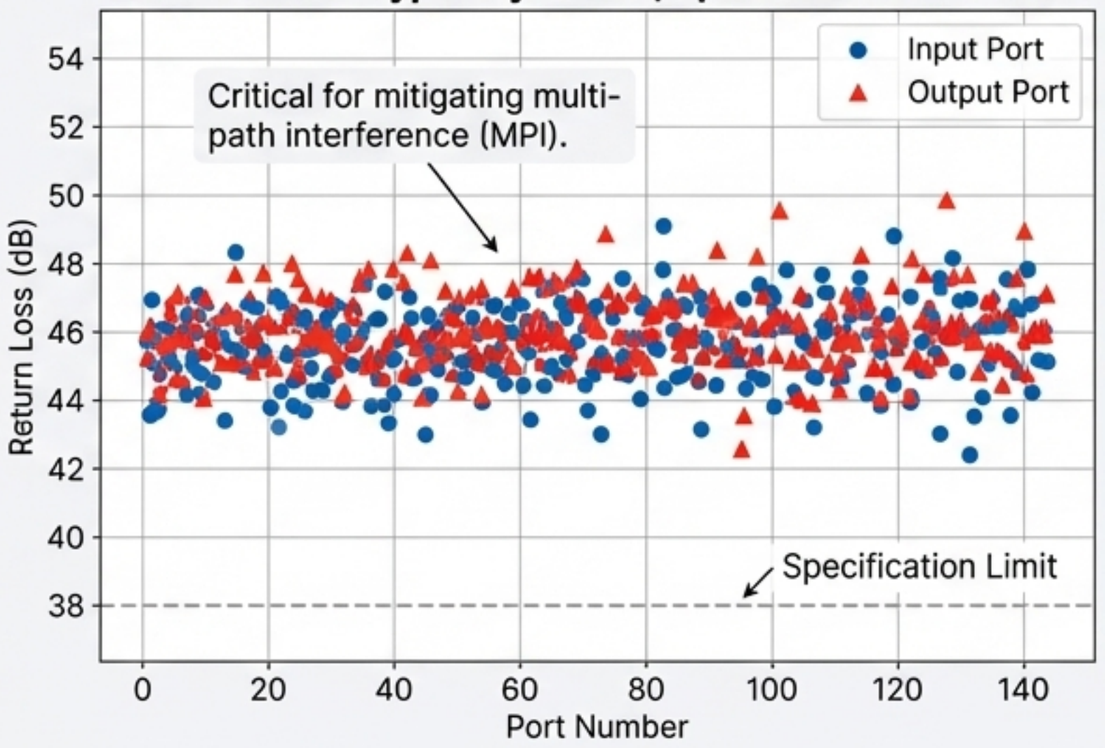
Palomar delivers exceptionally low signal loss and reflections, enabling robust high-speed links.

Meeting aggressive optical performance targets was critical. The stringent return loss requirement stems from the use of bidirectional links, where reflections directly degrade signal-to-noise ratio.

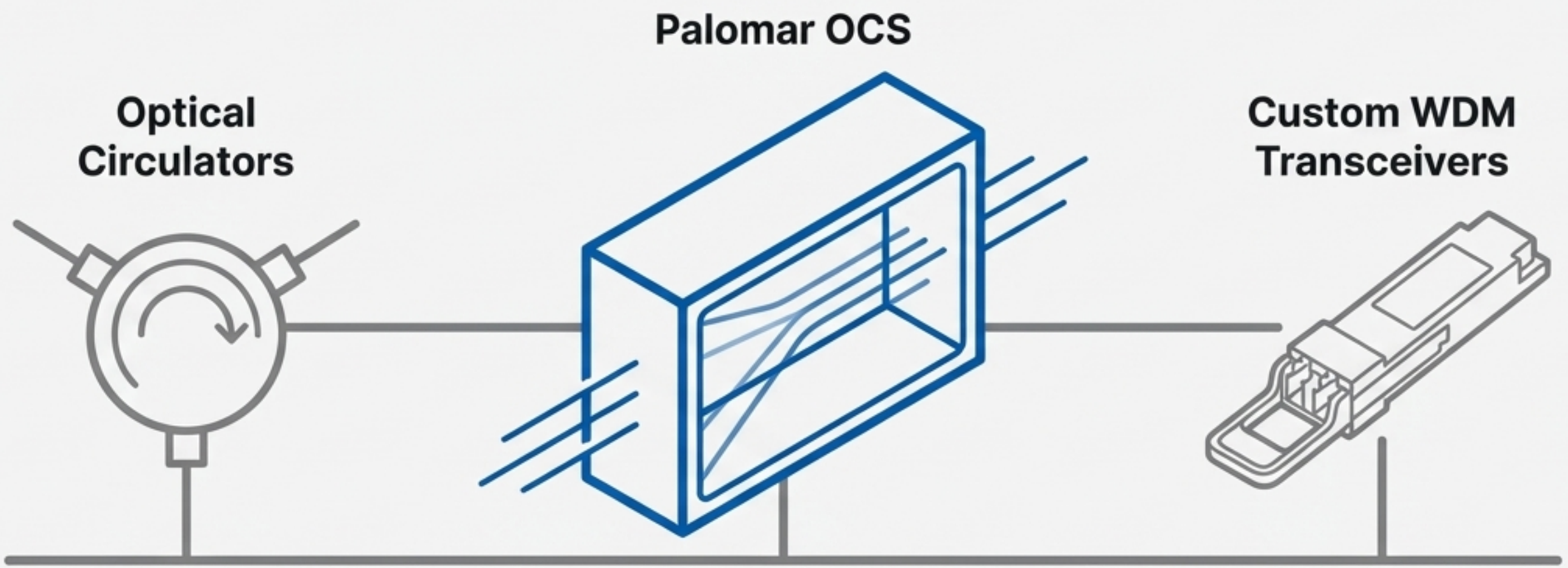
Insertion Loss: Typically <2dB for all 18,496 connections



Return Loss: Typically -46dB, Specification <-38dB



Apollo’s success relies on three critical, co-designed hardware components working in concert.



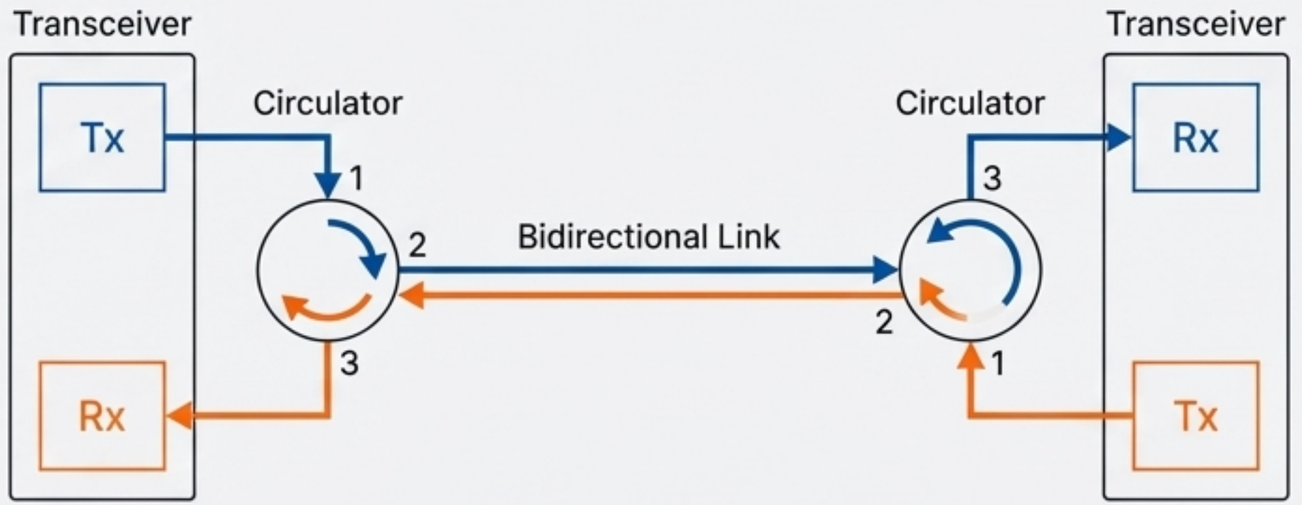
Realizing a cost-effective, large-scale optical switching layer required holistic system design. We developed three essential hardware pillars that were optimized to work together.

Optical circulators enable bidirectional traffic on a single fiber, doubling OCS port count and halving cabling costs.

The circulator is a three-port non-reciprocal device with cyclic connectivity (Port 1 → Port 2, Port 2 → Port 3). By coupling a circulator to each transceiver, we convert a standard duplex link into a bidirectional one. This seemingly simple component has a massive impact:

- **Doubles OCS Radix:** A 136-port OCS effectively becomes a 272-port bidirectional switch.
- **Halves Fiber Plant:** The number of required fiber cables is cut in half.

From Duplex to Bidirectional

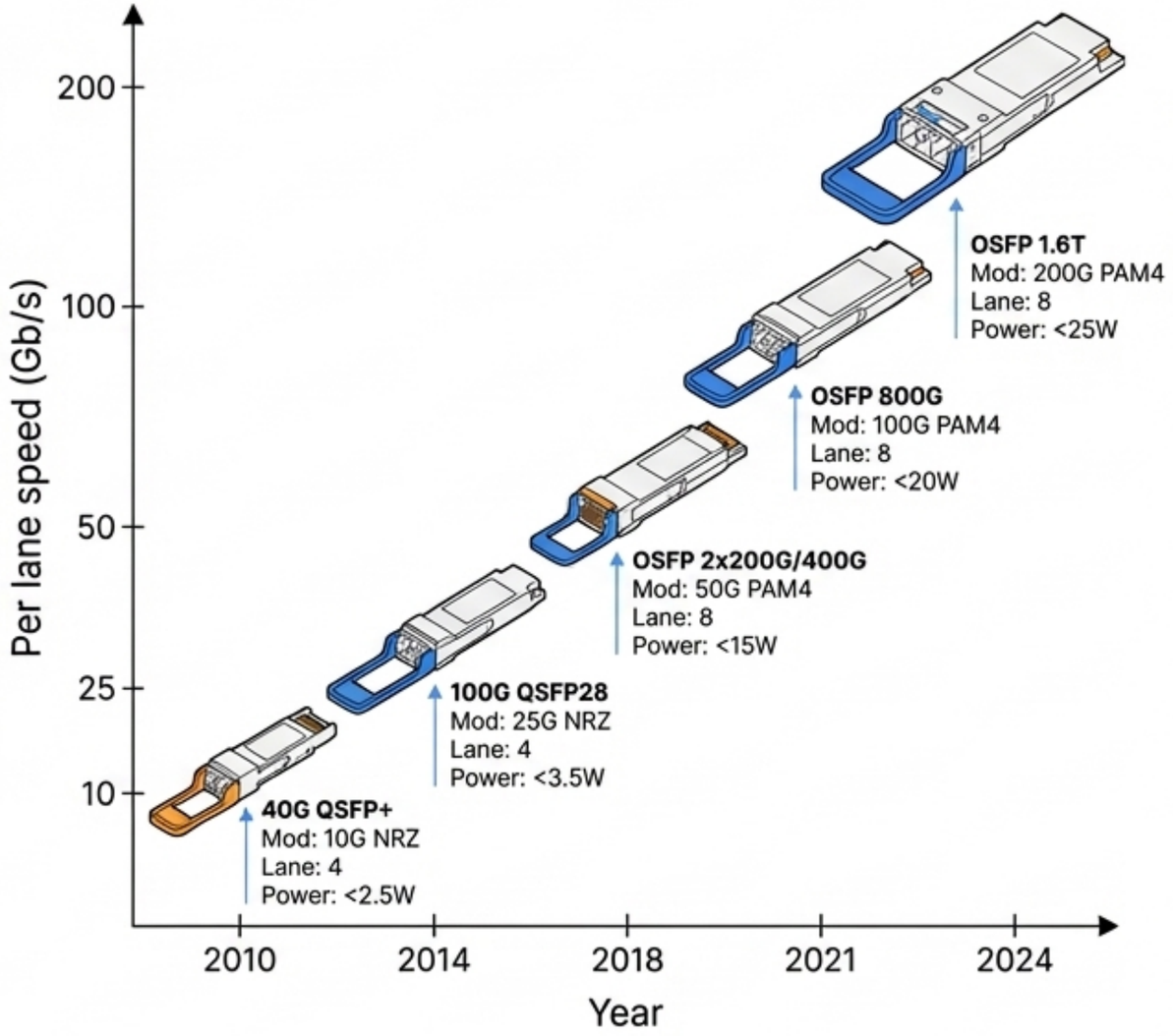


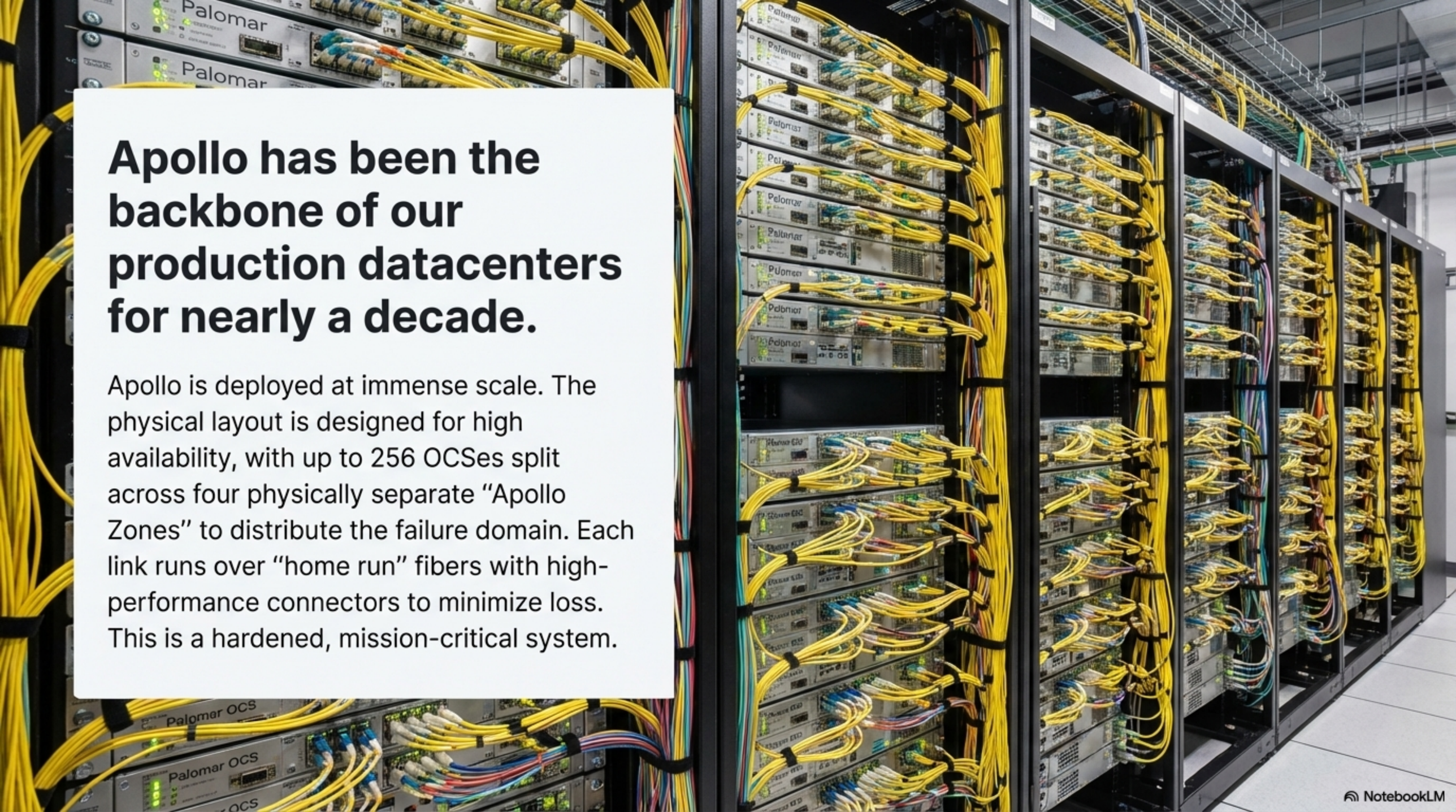
We co-designed four generations of WDM transceivers to master the physics of the Apollo fabric.

Standard optics were insufficient for the unique challenges of an OCS-based bidirectional link. Over a decade, we developed custom WDM transceivers with key innovations:

- **Mastering Impairments:** Migrated to Externally Modulated Lasers (EMLs) and developed DSP-based algorithms to mitigate Multi-Path Interference (MPI).
- **Backward Compatibility:** Maintained the same CWDM4 wavelength grid across all generations, allowing a 40G Aggregation Block to interoperate seamlessly with a 400G block.
- **Scaling Performance:** Continuously increased per-lane speeds from 10G to 100G+, evolving from NRZ to PAM4 modulation.

WDM Transceiver Evolution





Apollo has been the backbone of our production datacenters for nearly a decade.

Apollo is deployed at immense scale. The physical layout is designed for high availability, with up to 256 OCSes split across four physically separate “Apollo Zones” to distribute the failure domain. Each link runs over “home run” fibers with high-performance connectors to minimize loss. This is a hardened, mission-critical system.

Apollo enables us to evolve the fabric seamlessly, adding capacity and new technology without disruption.

The OCS fabric is data-rate agnostic, making it a long-lived asset. This unlocks two powerful capabilities:

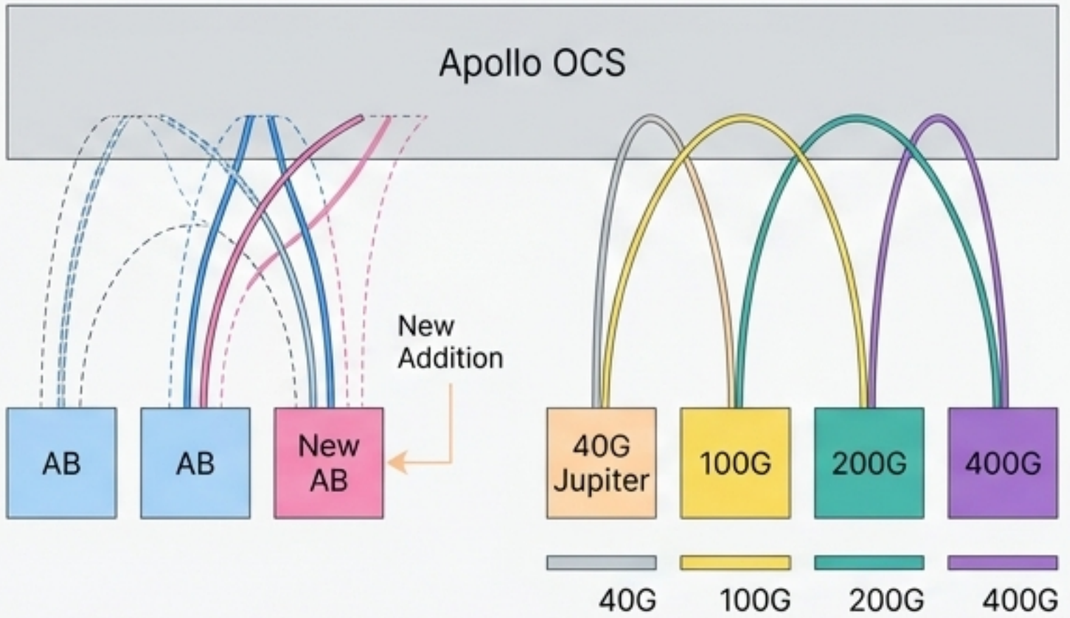
Pay-as-you-grow Expansion

New Aggregation Blocks (ABs) can be added incrementally. The OCS automates the complex “re-striping” of network connectivity, a task unmanageable with manual patch panels.

Rapid Tech Refresh

New-generation optics are backward-compatible, allowing us to deploy 400G ABs asmer upgrade, allowing us to deploy 400G ABs alongside older 100G systems, avoiding costly, all-at-once upgrades.

Seamless Fabric Evolution



We dynamically reconfigure network topology to match real-time traffic demands, maximizing efficiency.

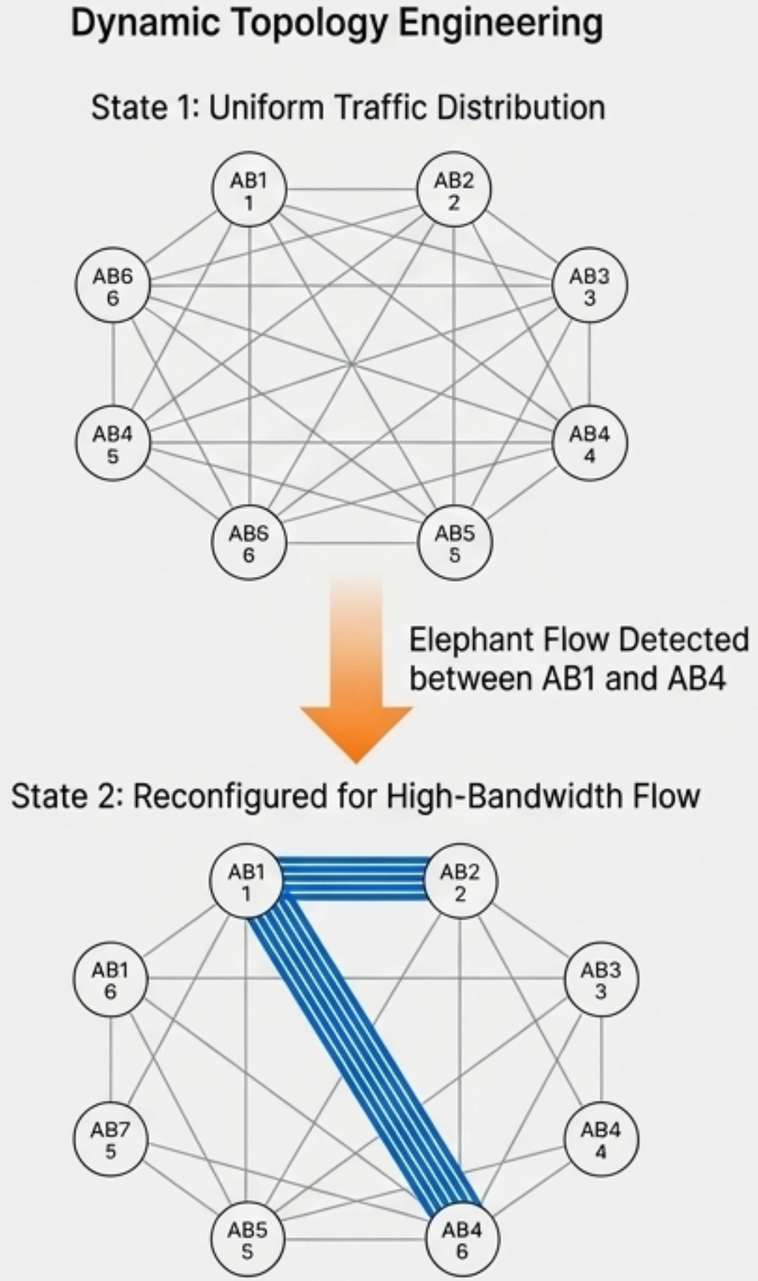
With the electrical spine removed, we use “Topology Engineering” to manage traffic. The OCS allows us to reconfigure inter-AB connectivity on the fly. This capability is especially powerful for:

Handling Elephant Flows

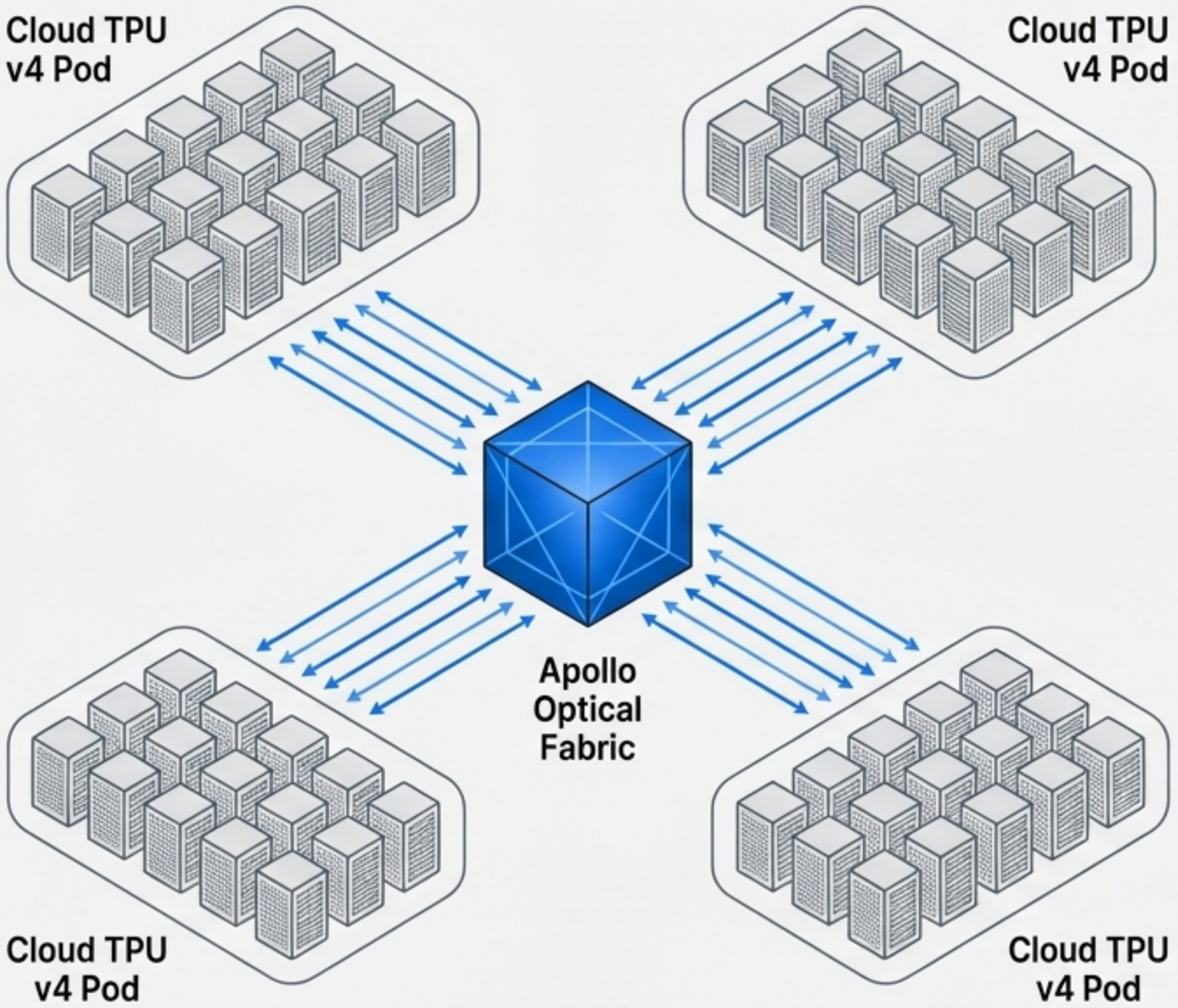
If a large, long-lived data transfer begins between two ABs, we can create more direct optical paths between them to maximize bandwidth.

Improving Efficiency

This allows us to achieve equivalent network throughput with fewer total links or increased throughput with the same number of links.



Apollo: The Interconnect for Large-Scale AI



The Apollo fabric is an ideal interconnect for the predictable, high-bandwidth demands of large-scale Machine Learning.

Large ML training models are communication-intensive and an ideal fit for circuit switching. Google's Cloud TPU v4 Pods, connecting 4,096 chips, leverage this.

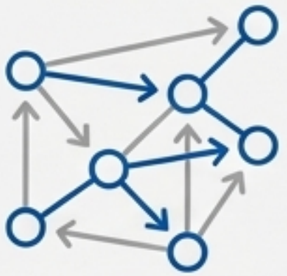
- **Predictable Workloads:** ML systems feature repeating communication patterns perfect for scheduled topology shifts.
- **Low Latency:** The OCS provides near-speed-of-light latency (<100ns for the optics), far superior to multi-hop electrical switches for tightly-coupled communication.
- **High Reliability:** The reliability of the optical fabric is a key enabler for ML clusters where a single link failure can degrade performance.

Apollo proved that optical circuit switching at scale is not only possible, but transformative.

After a decade in production, Apollo has delivered integer-factor improvements and fundamentally reset industry assumptions.

Lowest cost networks at scale with pay-as-you-grow capability.

Fastest adoption of latest-generation optics and networking silicon.



Dynamic network reconfiguration for maximum efficiency and performance

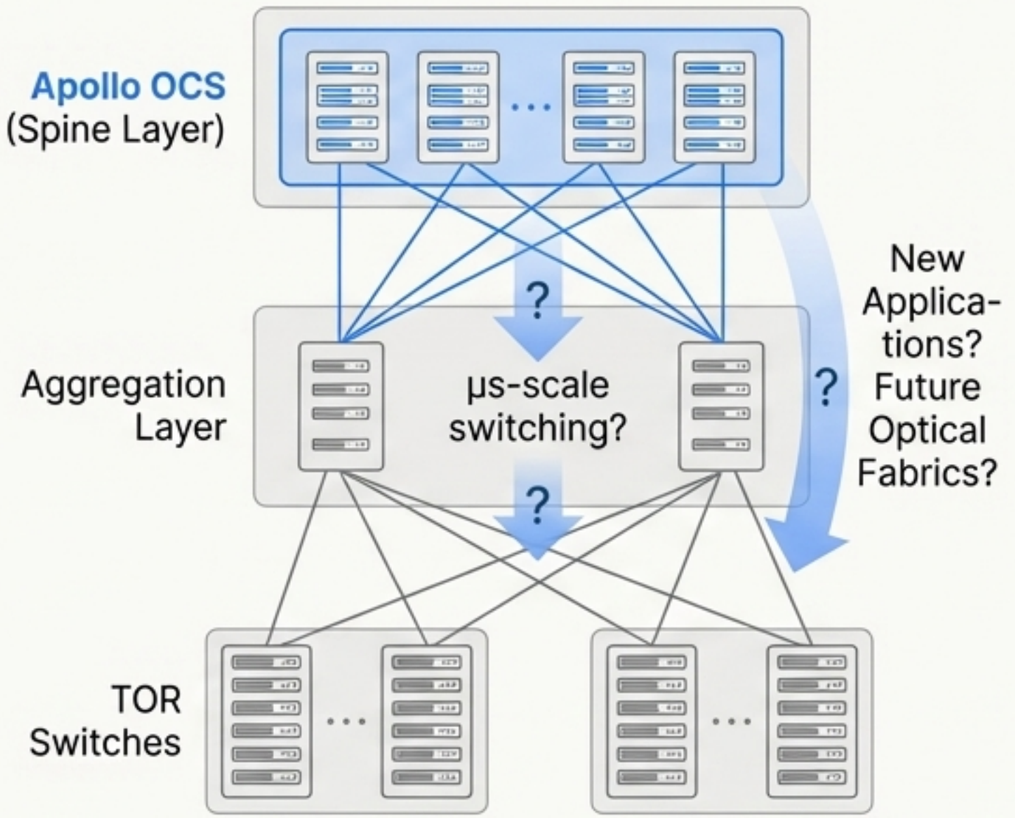
The Mission's Success: We successfully landed a "moonshot" technology in a production environment, demonstrating that large-scale OCS is a practical, deployed, and highly beneficial technology for modern datacenters.

Our mission continues: we are pushing optics deeper into the network to unlock the next wave of innovation.

This work is the initial step in moving the entire datacenter architecture into the optical domain. Future hardware evolution will focus on:

- **Larger Port Count OCS:** To enable further scale-out and more flexible topologies.
- **Faster Switching Speeds:** Microsecond-order switching to support more bursty traffic flows deeper in the network.
- **Lower Loss and Higher Reliability:** To enable new architectures, like cascading multiple OCSes into a larger optical fabric.

The Next Frontier: Optics Deeper in the Stack



A plethora of high-impact future work will follow as we continue to build the datacenter of the future on a foundation of light.