Architectural Specification: Agile AIDC Optical Path (AAOP) Software Pattern

Document Control & Versioning

- **Document Title:** Architectural Specification: Agile AIDC Optical Path (AAOP) Software Pattern
- **Document ID:** SPEC-AAOP-2026-V1.0
- **Current Version:** v1.0.0
- **Release Date:** June 16, 2026
- **Document Status:** Released for Discussion
- **Author:** Prof. Weiqiang Sun
- **Affiliation:** Shanghai Jiao Tong University (SJTU)
- **Contact Information:** sunwq@sjtu.edu.cn

Revision History

- **v1.0.0 (June 16, 2026) | W. Sun (SJTU):** Official Baseline Release.

Document Distribution & Approvals

- **Target Audience / Distribution:** Industrial Collaborative Entities.
- **Consensus Mandate:** Modifications to core atomic transaction definitions or OOB control channel specifications require a formal review cycle. As per systemic architectural standards, existing research sections and baseline execution models must not be deprecated or removed during future updates.

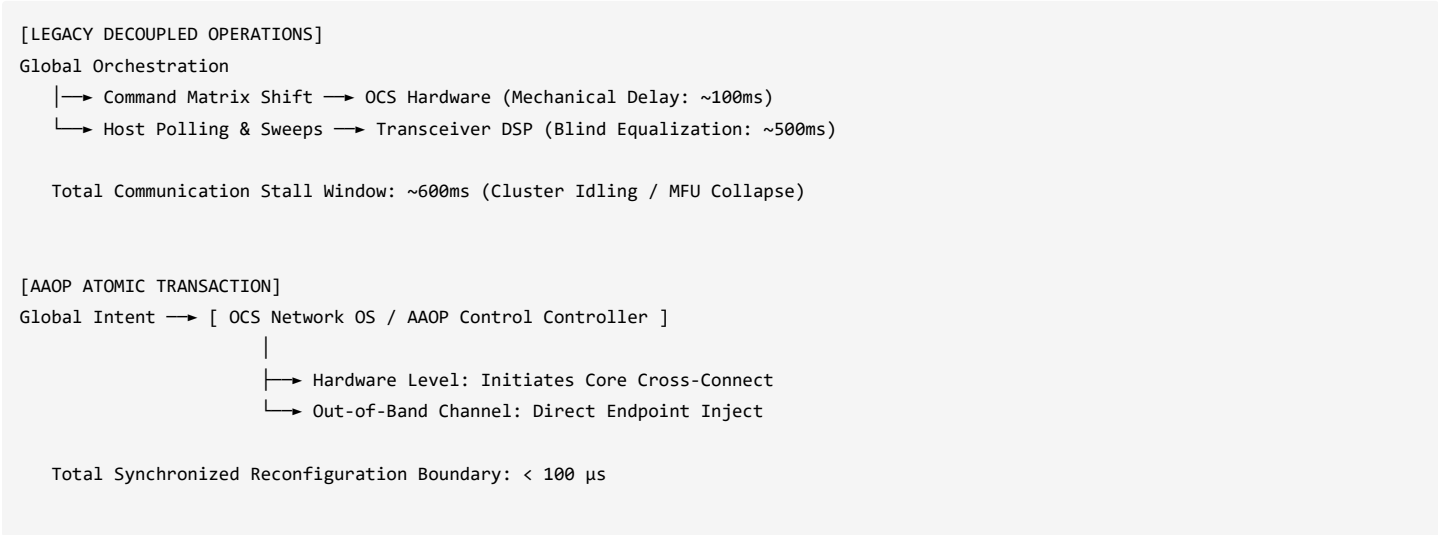
1. Executive Summary & Conceptual Shift

The integration of Optical Circuit Switches (OCS) or Optical Cross-Connects (OXC) into large-scale AI distributed training clusters introduces a severe **Time-Scale Barrier**. While advanced graphics processing pipelines mutate computation stages and collective routing topologies on sub-millisecond schedules, conventional optical layers operate as passive, isolated physical patch-panels.

Reconfiguring an optical topology historically mandates independent decoupled operations: driving a slow mechanical mirror grid, and asynchronously relying on host OS network layers to poll links, perform blind clock-data recovery (CDR), and spend hundreds of milliseconds executing digital equalizer training.

The **Agile AIDC Optical Path (AAOP)** software pattern fundamentally replaces this decoupled approach. Under AAOP, **an optical connection route is treated as an inseparable, atomic, software-managed transaction containing both its structural physical glass coordinates and its edge digital link states.**

Legacy Reconfiguration Operations vs. AAOP Atomic Flow



By allowing the OCS Network OS to completely wholesale path realization through a dedicated, hardware-level synchronization layer, AAOP collapses link-establishment latency down to microsecond tiers. This shifts optical interconnect scheduling away from coarse **Task-Level scheduling** (minutes or hours) and enables fine-grained, dynamic **Iteration-Level scheduling** matching the natural execution steps of Large Language Model (LLM) training pipelines.

2. Multi-Tier Layered Control Model

To insulate high-level AI workloads from physical optoelectronic drift while preserving sub-millisecond execution loops, the AAOP control architecture is split into three decoupled operational planes:

Tier 1: Global Cluster Orchestrator (Workload Plane)

- **Functional Scope:** Distributed schedulers (e.g., Slurm, Ray) paired with the AIDC Core Traffic Engineer.
- **Operational Latency:** Seconds to Minutes.
- **Mechanism:** Parses the computational Directed Acyclic Graph (DAG) of the training run. It calculates future scheduling requirements (such as shifting from the rectangular traffic pipelines of 3D Parallelism to the cross-cluster all-to-all exchanges of Expert/MoE Parallelism) and issues high-level, declarative configuration blueprints to the underlying fabric.

Tier 2: AAOP Fabric Controller / OCS NOS (Cohesive Path Plane)

- **Functional Scope:** Core OCS Network Operating System and local runtime managers.
- **Operational Latency:** Milliseconds.
- **Mechanism:** Consumes the macro-intents from Tier 1. It operates as the **Single Source of Truth** for the network state. Tier 2 transforms abstract node requests into physical cross-connect routing layouts, optimizing for localized signal loss profiles, while driving deterministic out-of-band notifications to edge endpoints to prepare them for physical realignment.

Tier 3: Embedded In-Band Data Plane Agents (Hardware Plane)

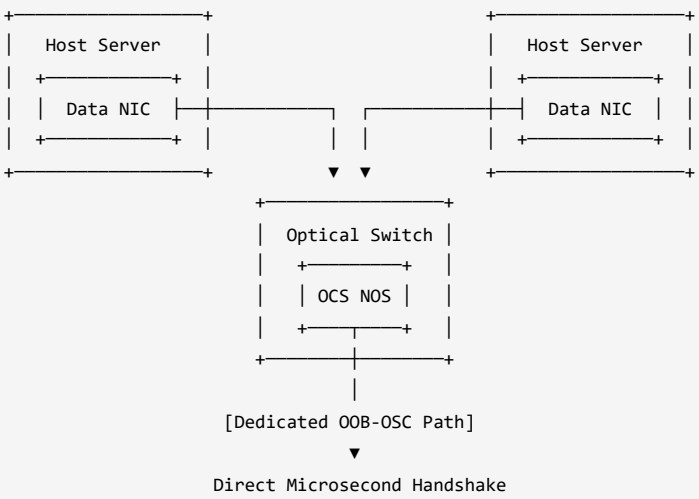
- **Functional Scope:** Transceiver microcode engines, intelligent Network Interface Cards (SmartNICs), and switch register banks.
- **Operational Latency:** Microseconds (μs) to Nanoseconds (ns).
- **Mechanism:** Manages local, automated feedback loops. It executes immediate transceiver parameter realignments—tuning wavelength windows (λ_n), modulating semiconductor amplifiers, executing rapid error masking, and adjusting Variable Optical Attenuators (VOAs) to equalize power levels immediately as incoming laser lines anchor.

3. Core Technical Pillars of AAOP

The realization of an Agile AIDC Optical Path relies on three integrated software and hardware technologies:

Pillar I: The Out-of-Band Optical Supervisory Channel (OOB-OSC)

To safely bypass the overhead of host operating system networking stacks, host bus polling traps, and kernel-space virtualization paths, the AAOP pattern leverages a dedicated, physical supervisory infrastructure.



- **Topology Mapping:** Every high-throughput compute link ($400\text{G}/800\text{G}/1.6\text{T}$) from a GPU accelerator node or an Electrical Packet Switch (EPS) leaf is hard-bundled with an auxiliary low-speed control fiber line or a dedicated, structurally isolated out-of-band signaling wavelength channel.
- **Hardware Profile:** Operating at low signaling speeds ($1\ \text{Gbps}$ to $10\ \text{Gbps}$) with cheap, ultra-reliable components, this control channel bypasses high-power burst-mode processing requirements. It serves as an uncompromised, always-on communications path connecting the switch control processor directly to edge transceivers.

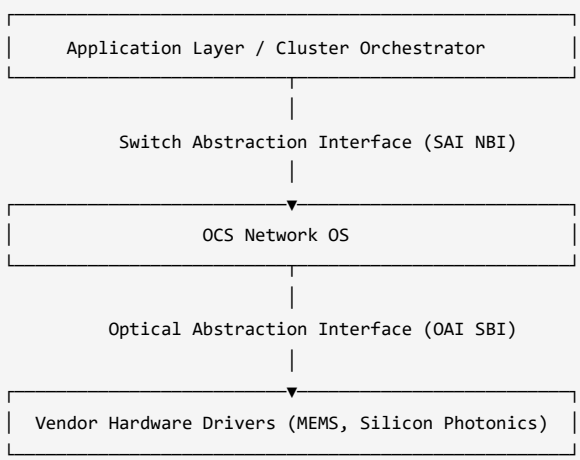
Pillar II: The AAOP Control Protocol Stack

A specialized, lightweight, non-IP software protocol suite executes over the OOB-OSC to coordinate path convergence:

- **Instant Neighbor Verification:** The moment a physical channel change clears inside the switch core, a hard-coded handshake frame fires down the OOB-OSC, identifying port pairing indices instantly without broadcasting.
- **Impairment Mitigation Monitoring:** Endpoint transceivers and the core switch continuously exchange localized optical telemetry via the control plane, dynamically steering VOAs to correct for insertion loss anomalies or mechanical micro-vibrations without resetting the digital link.
- **Equalization Coefficient Pre-Loading ("Warm Lock"):** Standard Digital Signal Processors (DSPs) spend hundreds of milliseconds executing blind convergence sweeps to re-stabilize Feed-Forward (FFE) and Decision Feedback Equalization (DFE) tap weights when incoming light resets. The AAOP protocol circumvents this. Because the OCS operating system maintains a strict map of the source-to-destination pipeline, it pushes pre-calibrated, baseline equalization coefficients directly into the receiving transceiver's memory register over the control link. The receiver achieves an immediate physical lock on incoming data without scanning.

Pillar III: Standardized Interface Abstractions (SAI/OAI Extensions)

To prevent proprietary lock-in and enable cross-vendor interoperability across multi-tier clusters, the software boundary is divided into open APIs:



- **Northbound Switch Abstraction Interface (SAI) Extensions:** Extends Open Compute Project (OCP) paradigms to permit global AI engines to handle optical reconfigurations without dealing with underlying physical optics or microcode details.
- **Southbound Optical Abstraction Interface (OAI):** Establishes a decoupled driver layer between the core network operating system and underlying physical actuators (e.g., 3D MEMS mirrors, Silicon Photonics Mach-Zehnder arrays, or tunable filters). This guarantees that as solid-state optical switches emerge, the AAOP protocol stack and orchestration patterns remain unchanged.

4. Performance & Structural Projections

By offloading path transactions from the compute data path to the specialized AAOP control engine, synchronization overhead transitions from macro-seconds to micro-seconds:

Reconfiguration Milestone	Legacy Network Route	Optimized In-Band	AAOP Solution
Control Signal Propagation	5 – 20 ms	0.5 – 1 ms	**< 10** μs
Control Protocol RTT	10 – 50 ms	1 – 2 ms	**< 20 – 50** μs
Topology Notification	10 – 100 ms	0.5 – 2 ms	**< 20** μs
Matrix Command Dispatch	5 – 20 ms	0.5 – 2 ms	**< 10** μs
Post-Switch Validation	20 – 100 ms	1 – 5 ms	**< 50** μs

Estimates calculated on a standard model scaling up to 10,000 GPU acceleration endpoints with a maximum local fiber routing length of 1 km.

---

5. Deployment Framework & Boundary Conditions

Asymmetric Failure Modes

Deploying an out-of-band framework requires robust decoupled error isolation. The AAOP engine handles two asymmetric failure states:

1. **Control Path Lost, Data Path Active:** If the OOB-OSC fails but high-speed transport light remains within nominal limits, the AAOP layer sets a structural flag locking the current topology. It signals the global orchestration frame to pause scheduled shifts on that route while protecting active training calculations.
2. **Control Path Active, Data Path Lost:** If the control path reports normal parameters but high-speed light undergoes severe power degradation, the AAOP protocol bypasses standard timeout counters, immediately signaling an atomic path teardown to prevent data corruption or serialization drops.

Technology Transition Plan

- **Near-Term (Mechanical Integration):** The AAOP software protocol interfaces with 3D MEMS matrices ($10\ \text{ms} - 100\ \text{ms}$ physical mirror delays). The software architecture focuses on completely masking the post-switch link-training overhead ($500\ \text{ms}$ saved down to $< 50\mu s$), maximizing Model Flops Utilization (MFU).
- **Long-Term (Solid-State Integration):** As Silicon Photonics Mach-Zehnder Interferometer (MZI) grids or fast tunable laser assemblies scale up, physical switching times drop into the nanosecond domain. The AAOP model architecture preserves its utility here: by pre-loading coefficients and eliminating operating system interaction loops, it ensures that software processing overhead does not replace mechanical mirrors as the next cluster bottleneck.

6. Out-of-Band (OOB) Channel Payload Architecture & Host Implications

The OOB channel serves as a low-latency, deterministic control plane that isolates physical optical setup adjustments from high-speed data compute rails. This section outlines the telemetry, signaling primitives, and control sequences carried over the OOB link, alongside structural implications for the host endpoint system.

I. Real-Time Link Parameter Tuning

- **Payload Characteristics:** Dynamic transmission of transmitter/receiver optical parameters (power attenuation targets, VOA adjustments, laser biasing metrics, and pre-FEC BER profiles).

- **Operational Flow:** Slight variations in insertion loss across the optical cross-connect matrix alter the optical power arriving at the receiver. The switch and transceiver execute microsecond feedback loops over the OOB channel to stabilize the signal-to-noise ratio.
- **Host System Implications:**
- *Hardware Abstraction:* The host's high-speed transceiver firmware must expose its inner register banks (such as CMIS registers) directly to the OOB controller via a low-level SPI or I^2C bridge.
- *CPU Bypass:* The host CPU cannot participate in this tuning loop. The transceiver's on-board microcontroller must read and apply parameters autonomously in the background, keeping physical micro-adjustments completely invisible to the host OS.

II. Dynamic Wavelength Tuning (λ-Steering)

- **Payload Characteristics:** Tuning configurations for lasers, wavelength channel grid assignments, and filtering window alignment matrices.
- **Operational Flow:** In advanced AIDCs utilizing Wavelength Division Multiplexing (WDM) or tunable routing blocks, establishing a path requires matching the laser's physical emission wavelength to the exact port route chosen by the OCS. The OOB channel pre-allocates and sends these wavelength indices before lighting up the data path.
- **Host System Implications:**
- *State Synchronization:* The host network interface card (NIC) must maintain an asynchronous internal state machine that cleanly separates the physical laser channel from the logical MAC address layer.
- *Equalization Mapping:* Because changing wavelengths modifies the physical behavior of light, the host must link its physical parameter tables directly to the active wavelength index, reloading profiles without causing false link-down alerts.

III. Asynchronous State Reporting & Health Telemetry

- **Payload Characteristics:** High-frequency signaling frames covering real-time physical states, including Loss of Signal (LOS), Lock-Loss warnings, thermal anomalies, mechanical mirror vibration metrics, and pre-FEC error trends.
- **Operational Flow:** The OCS core and edge transceivers continuously stream health telemetry over the OOB channel, creating an immediate feedback loop that catches performance dips before they lead to packet dropouts.
- **Host System Implications:**
- *Decoupled Interrupt Framework:* Under AAOP, the host NIC must intercept physical layer errors and shift them to the OOB daemon, preventing the operating system from tearing down routing tables prematurely.
- *Error Masking:* The host operating system must actively mask brief optical dropouts during reconfiguration windows, holding data transmission queues in a hardware-level "pause" state rather than triggering standard link-down timeout behaviors.

IV. Predictive Coefficient Pre-Loading ("Warm Lock" Signaling)

- **Payload Characteristics:** Digital Signal Processor (DSP) tap weight vectors, Feed-Forward Equalization (FFE) matrices, and Decision Feedback Equalization (DFE) initial training states.
- **Operational Flow:** Instead of executing a blind, time-consuming scanning loop to align internal equalizer coefficients when light returns to a port, the host reads a pre-calculated tap vector pushed over the OOB channel by the AAOP controller.
- **Host System Implications:**
- *Memory-Mapped Access:* The host NIC DSP must support a memory-mapped interface (MMIO) allowing the OOB microcontroller to overwrite active equalizer registers directly.
- *Firmware Modification:* Standard proprietary transceiver firmware must be modified to accept external, predictive coefficient seeding, moving directly from a dark physical state to a fully locked digital state based on the injected data payload.

References

1. Optical Network Paradigms & Dispersed Bottlenecks

- **Farrington, N., et al. (2010).** "Helios: a hybrid electrical/optical switch architecture for modular data centers." *Proceedings of the ACM SIGCOMM 2010 conference*.
- **Poutievski, L., et al. (2022).** "Jupiter Evolved: Transforming Google's Datacenter Network with Optical Circuit Switches." *Proceedings of the ACM SIGCOMM 2022 Conference*.
- **Glick, M., et al. (2024).** "Optical switching for data centers and advanced computing systems." *Journal of Optical Communications and Networking (JOCN)*, vol. 16, no. 2, pp. 45-58.
- **Zilberman, N., et al. (2016).** "Starlight: Shared Optical Circuits for Software-Defined Data Centers." *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 601-614.
- **Mizuno, K., et al. (2023).** "Scaling Optical-Circuit-Switched Clos Networks to Guarantee Admissible Blocking Probability in Data Centers." *IEEE/OSA Journal of Optical Communications and Networking (JOCN)*, vol. 15, no. 8, pp. 512-524.

2. Multi-Layer Control Planes & Systems Orchestration

- **Ballani, H., et al. (2020).** "Sirius: A Flat, Agile Data Center Network Configured by Information." *Proceedings of the 2020 ACM SIGCOMM Conference*.
- **Microsoft Research / Goyal, A., et al. (2024).** "TE-CCL: Rethinking Machine Learning Collective Communication as a Multi-Commodity Flow Problem." *Proceedings of the ACM SIGCOMM 2024 Conference*, pp. 16–37.
- **Venkatesh, A., et al. (2023).** "Whirlpool: Dynamic Topology Switching for Deep Learning Clusters." *Proceedings of the 28th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '23)*.
- **Porter, R., et al. (2018).** "Software-defined networking control plane for seamless integration of multiple silicon photonic switches in Datacom networks." *Optics Express*, vol. 26, no. 12, pp. 15234-15246.
- **ITU-T Technical Report (10/2024).** "GSTR-ION-2024: Technical report on intelligent optical networks towards 2030." *International Telecommunication Union*.

3. Transceiver Diagnostics & Burst Integration

- **Mellette, W. M., et al. (2017).** "RotorNet: A Scalable, Low-complexity, Optical Datacenter Network." *Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '17)*, pp. 267–280.
- **Mellette, W. M., Forencich, A., et al. (2024).** "Realizing RotorNet: Toward Practical Microsecond-Scale Optical Networking." *Proceedings of the ACM SIGCOMM 2024 Conference*.

4. Machine Learning Workload & Collective Coordination

- **Wang, Y., et al. (2022).** "Blast: Accelerating Deep Learning Training with Dynamic Optical Interconnects." *Proceedings of the 2022 International Conference on Management of Data (SIGMOD)*.
- **Zhu, Z., et al. (2023).** "Understanding Collective Communication Performance in Large-Scale Deep Learning Training." *IEEE Transactions on Parallel and Distributed Systems*, vol. 34, no. 4, pp. 1120-1135.

5. Co-Packaging, Advanced Integration & Emerging Physics

- **Cheng, Q., et al. (2018).** "Co-packaged optics for high-performance computing and data centers." *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 4, pp. 1-11.
- **Soref, R. (2014).** "The past, present, and future of silicon photonics." *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, no. 4, pp. 13-24.