LARCH NETWORKS

Larch Networks accelerates the deployment of high-performance AI infrastructure by transforming raw silicon and hardware into production-ready, highly optimized networking solutions.

With the help of partners we provide full featured fabric, tested and proven.

larch-networks.com

Empowering Open AI Infrastructure: Larch Networks at OCP 2026

OCP 2026 ALTERNATE INFERENCE

Larch Professional Services: Making Silicon Shippable: A De-risked Path to Market

We move beyond standard time-and-materials engagements to deliver fixed-outcome bring-up programs for ODMs, OEMs, and silicon vendors.

- **Turnkey Bring-Up**: We deliver a fully loadable SONiC image, Switch Abstraction Interface (SAI), platform drivers, CLI, and comprehensive regression testing to bridge the gap between reference designs and shippable products.
- **Independent 800G Validation Lab**: Our independent validation facility provides low-friction testing for latency, power, conformance, and interoperability, utilizing tools like Xena and Calnex alongside over 100 automated topological tests.
- **Global Reach for Complex Geographies**: We help local partners deliver complete, supported local solutions in the regions like India, the Gulf, Central Asia, Africa, LatAm. Larch offers complete support for local production, integration and on-the-ground teams.



2 Larch Hardened SONiC: Operable and AI-Optimized

While community SONiC provides a strong foundation, enterprise and NeoCloud environments require an "Operable SONiC" that is fully manageable, visual, and feature-complete.

- **Commercial Support**: Larch provides a hardened, proven SONiC distribution backed by commercial support, custom SDK development, and ongoing maintenance contracts. Special effort is allocated for Marvell Teralynx in NeoCloud AI environments.
- **Closing the Feature Gap**: We provide long-term fix-release support and implement critical on-demand features such as MACsec, EVPN, MPLS, and QinQ/L2PT to meet specific enterprise routing requirements.
- **AI-Native Telemetry**: The platform includes High-Frequency Telemetry and gNMI streaming, delivering microsecond-level queue depths and latency metrics directly to orchestration pipelines to instantly isolate performance drops.

3 Elevating Marvell Teralynx 10 for AI Workloads

Larch Hardened SONiC is actively optimized for Marvell Teralynx 10 (51.2T) scale-out switches, delivering the low-power, low-latency switching foundation required for modern AI clusters. This joint solution provides a highly competitive, power-efficient AI fabric alternative to Broadcom-dominated architectures. By capitalizing on Teralynx silicon, operators can lower Total Cost of Ownership (TCO) and reduce power consumption without sacrificing sub-10ns performance.

- **Adaptive Routing & Auto-Load Balancing (ARS/ALB):** Flowlet routing dynamically distributes massive AI "Elephant Flows" across all 800G and in future 1.6T paths, avoiding hotspots and ensuring optimal lane utilization.
- **Lossless Transport:** Features like Lossless RoCEv2 and Priority Flow Control Watchdog (PFCWD) provide real-time prevention of cluster-wide fabric freezes.
- **Dynamic Buffer Optimization:** Code-driven, real-time Dynamic Explicit Congestion Notification (ECN) tuning optimizes switch buffers specifically for bursty AI workloads.



Available Today, Ready for 1.6T Tomorrow Larch Networks LN-TL64Z800 (64x800G) switch is available to order today, allowing operators to deploy ultra-low latency 800G pipelines immediately. Furthermore, this architecture is future-proof: the exact same Larch SONiC technology, ARS capabilities, and active hardening that power the Teralynx 10 today will seamlessly scale to power the 1.6T pipelines of the Marvell Teralynx 100 Ultra Ethernet devices coming in 2027.

OCP 2026 Demo: Proving ARS Performance at Scale

At OCP, Larch Networks is demonstrating the real-world impact of Local Adaptive Routing and Switching (ARS) on a Delta Teralynx 10 switch running Larch Hardened SONiC.

The demo shows SONiC Local ARS (Adaptive Routing and Switching) on the Marvell Teralynx 10: a BGP ECMP group whose member selection is done per flowlet by the ASIC, based on measured egress port load, instead of by a static 5-tuple hash. Under a skewed load, generated by Xena Z800 Freya tester plain hash ECMP overloads some members and drops traffic while other members sit half idle. With ARS the same offered load is spread evenly, with zero loss and an order-of-magnitude lower latency. The demo also shows how the resulting behavior is observable from outside the box: per-port counters over gNMI (dial-in) and a high-frequency counter stream over OpenTelemetry (push).

Metric (340G Load, 16 Streams, 60s Run)	Static Hash ECMP	Larch ARS (Flowlet Routing)
Load Distribution	Uneven (up to 98.0 Gb/s on loaded members)	Perfect 25.0% split across all 4 members
Packet Loss	4.83% (147,438,000 lost packets)	0.0001% (Effectively zero)
Max Latency	153.2 µs	6.9 µs

The results confirm that while static 5-tuple hashing overloads specific links and drops traffic, ARS dynamically shifts flowlets away from congestion. This results in an up to 30x latency reduction and zero packet reordering, effectively accelerating congested GPU collectives. Furthermore, under a simulated 200 Gb/s member failure, ARS achieves hitless failover with maximum flow gaps of just 0.3 ms, compared to multi-millisecond black-holing under standard ECMP.

LARCH NETWORKS

Tailor-made network hardware & software development

5 Intelligent Orchestration with Dorado Cruz

To complement Larch's robust data plane, the Dorado Cruz AI Fabric Controller provides the intelligence layer for end-to-end network operations.

Cruz translates intent into infrastructure by reading device configurations and automatically emitting fabric resources.

By leveraging the in-band telemetry streamed directly from Larch SONiC, Cruz offers AI-native operations through plain-language queries and predictive alerts.

This orchestration layer ensures that the open AI fabric remains highly visible, manageable, and performant from initial deployment through lifecycle troubleshooting, allowing AI jobs to finish faster and ensuring GPUs are never starved for data.

