

High-Bandwidth Low-Latency AI Compute Network Engineering Specifications

- Official Technical Overview & Hardware Datasheet

EXECUTIVE SUMMARY

The High-Bandwidth Low-Latency AI Compute Network Engineering Specifications platform represents a purpose-built, carrier-grade hardware solution designed to meet the unprecedented demands of large-scale artificial intelligence training and inference clusters. As AI models grow beyond trillions of parameters, the underlying network fabric must evolve from a best-effort transport layer into a deterministic, ultra-low-latency switching and routing engine. This document provides a comprehensive technical overview of the chassis, line cards, forwarding ASICs, and thermal management systems that enable non-blocking, 800GE-ready interconnect for GPU-accelerated compute nodes. The platform is engineered for hyperscale data center operators, AI cloud service providers, and national research laboratories requiring absolute performance predictability and wire-speed programmability.



ARCHITECTURE & CHASSIS DESIGN

The system is built upon a modular, 4RU chassis architecture that supports up to 32 hot-swappable line card slots and dual redundant switch fabric modules. At the heart of the platform lies a custom 25.6 Tbps forwarding ASIC, fabricated on a 5nm process, which delivers deterministic latency of under 450 nanoseconds port-to-port. The backplane utilizes a passive orthogonal midplane design, eliminating active components in the data path to improve mean time between failures (MTBF) and signal integrity. Each line card supports 36 x 800GE QSFP-DD800 or 72 x 400GE QSFP112 ports, providing a total system capacity of 51.2 Tbps in a single chassis. The chassis is designed for front-to-back airflow with variable-speed fan trays and supports 1+1 redundant -48V DC or 3+1 AC power supplies.

HARDWARE FEATURES

The platform integrates a comprehensive set of hardware features optimized for AI workload acceleration. Key capabilities include:

- Cut-through switching with dynamic buffer allocation to minimize head-of-line blocking for collective communication operations (AllReduce, AllGather).
- Hardware-accelerated RoCEv2 (RDMA over Converged Ethernet) with congestion control offload for lossless fabric behavior.
- Programmable telemetry engine supporting in-band network telemetry (INT) at line rate, enabling microsecond-level visibility into fabric congestion.
- Precision Time Protocol (PTP) hardware timestamping with sub-100ns accuracy for distributed AI training synchronization.
- Advanced load balancing using dynamic flowlets and packet spraying to eliminate hash polarization across spine-leaf topologies.
- Secure boot, TPM 2.0, and MACsec encryption on every port for zero-trust fabric security.

COMPLIANCE & STANDARDS

The system complies with the following industry standards and regulatory directives:

- IEEE 802.3cd, 802.3ck, 802.3db (800GE electrical and optical interfaces)
- IEEE 802.1Q, 802.1ad, 802.1Qbb (PFC), 802.1Qaz (ETS)
- RoCEv2 and iWARP specifications per IBTA
- NEBS Level 3 and ETSI EN 300 019 for environmental robustness
- UL 60950-1, IEC 62368-1, and CE marking for safety
- RoHS 3 and REACH compliance for materials

TECHNICAL SPECIFICATIONS

The following table summarizes the core hardware parameters of the High-Bandwidth Low-Latency AI Compute Network Engineering Specifications platform.

Parameter	Specification
Form Factor	4RU Modular Chassis, 32 Line Card Slots
Switching Capacity	51.2 Tbps (full duplex)
Forwarding ASIC	25.6 Tbps, 5nm, cut-through
Port-to-Port Latency	< 450 ns (typical, 64-byte payload)
Line Card Options	36 x 800GE QSFP-DD800 / 72 x 400GE QSFP112 / 144 x 100GE
Switch Fabric Modules	2+1 redundant, hot-swappable

Power Supply	1+1 -48V DC or 3+1 AC, hot-swappable
Airflow	Front-to-back, variable-speed fan trays
Management Interfaces	2 x 10GE SFP+, 1 x Console RJ45, 1 x USB
RoCEv2 Support	Hardware offload, lossless PFC/ETS
Telemetry	INT, sFlow, streaming telemetry at line rate
PTP Accuracy	< 100 ns hardware timestamping
MACsec	Line-rate encryption on all ports
Operating Temperature	0°C to 45°C (continuous)
Dimensions (H x W x D)	177 mm x 448 mm x 900 mm
Weight (fully loaded)	Approx. 120 kg
MTBF	> 350,000 hours
Compliance	NEBS Level 3, UL, CE, RoHS 3

ORDERING OPTIONS

Base chassis configurations are available with or without switch fabric modules, allowing for staged deployment. Line card options include 800GE

OSFP/QSFP-DD800 variants, 400GE QSFP112 high-density cards, and 100GE breakout cards for legacy AI node connectivity. Software licenses are offered in three tiers: Standard (L2/L3 forwarding), Advanced (RoCEv2, INT, PTP), and AI Fabric Suite (collective communication optimization, dynamic load balancing). Redundant power supplies and fan trays are ordered separately to match specific data center power and cooling architectures. All configurations include a standard 3-year hardware warranty with optional 5-year extended support.

