AI Compute Acceleration Module Hardware Architecture Specification - Official Technical Overview & Hardware Datasheet

EXECUTIVE SUMMARY

The AI Compute Acceleration Module represents a paradigm shift in edge and core network infrastructure, purpose-built to meet the escalating computational demands of artificial intelligence and machine learning workloads within telecom environments. This hardware architecture specification defines a carrier-grade acceleration platform that seamlessly integrates into existing network topologies, delivering unprecedented performance density for inference and training tasks at the network edge. Engineered for deployment in space-constrained central offices, edge data centers, and hyperscale facilities, the module provides a robust, scalable foundation for AI-driven network optimization, real-time analytics, and intelligent traffic management. This document serves as the definitive technical reference for system architects, pre-sales engineers, and network planners evaluating high-performance compute acceleration for next-generation telecom infrastructure.

ARCHITECTURE & CHASSIS DESIGN

The AI Compute Acceleration Module is architected around a modular, high-density 2RU chassis designed for standard 19-inch EIA racks. The system features a hybrid compute architecture combining purpose-built AI accelerator ASICs with general-purpose x86 host processors, interconnected via a high-bandwidth PCIe Gen5 fabric. The chassis supports up to eight hot-swappable acceleration modules, each delivering independent computational resources. A passive backplane design ensures maximum signal integrity and minimizes points of failure, while redundant cooling fans and power supplies provide carrier-grade availability. The front-panel accessible module bays allow for in-service expansion and replacement, critical for maintaining continuous network operations.

HARDWARE FEATURES

At the heart of the module lies a custom AI accelerator ASIC fabricated on a 5nm process node, delivering exceptional performance-per-watt for matrix multiplication and convolution operations essential to neural network processing. Each acceleration module integrates 64GB of HBM3 high-bandwidth memory, providing 1.2TB/s of memory bandwidth to feed the compute engines. The host subsystem features dual Intel Xeon Scalable processors with up to 2TB of DDR5 system memory, ensuring efficient data orchestration and control plane operations. Network connectivity is provided through dual 400GbE QSFP-DD ports per module, supporting hardware-accelerated packet processing and RDMA over Converged Ethernet (RoCEv2). The chassis includes an integrated BMC for out-of-band management and a TPM 2.0 module for hardware root-of-trust security.

COMPLIANCE & STANDARDS

The AI Compute Acceleration Module is designed to meet the most stringent international standards for telecom and datacenter equipment. The platform complies with NEBS Level 3 (GR-63-CORE, GR-1089-CORE) for operation in central office environments, ensuring resilience against seismic events, temperature extremes, and electromagnetic interference. Safety certifications

include UL 60950-1, IEC 62368-1, and CSA C22.2 No. 60950-1. EMC compliance is maintained per FCC Part 15 Subpart B Class A, EN 55032 Class A, and EN 55024. The module supports SNMPv3, NETCONF, and RESTCONF management interfaces, ensuring seamless integration into existing OSS/BSS frameworks.

TECHNICAL SPECIFICATIONS

The following table summarizes the key technical parameters of the AI Compute Acceleration Module platform. All specifications are subject to change without notice and represent nominal values under standard operating conditions.

Parameter	Specification
Form Factor	2RU 19-inch EIA rack-mountable chassis
Accelerator ASIC	Custom 5nm AI accelerator, up to 8 modules per chassis
Memory per Module	64GB HBM3, 1.2TB/s bandwidth
Host Processors	Dual Intel Xeon Scalable (4th Gen), up to 2TB DDR5
Network Interfaces	2x 400GbE QSFP-DD per module, RoCEv2 support

PCIe Fabric	PCIe Gen5 x16 per module, 128GB/s bidirectional
Power Supply	1+1 Redundant AC/DC, 2000W per PSU
Cooling	N+1 Redundant hot-swap fan trays
Management	BMC with SNMPv3, NETCONF, RESTCONF, Redfish
Security	TPM 2.0, Secure Boot, Hardware Root-of-Trust
Operating Temperature	0°C to 45°C (continuous), -5°C to 55°C (short-term)
Compliance	NEBS Level 3, UL/IEC 62368-1, FCC/EN Class A

ORDERING OPTIONS

The AI Compute Acceleration Module is available in multiple configurations to suit varying deployment scales and performance requirements. The base chassis (ACM-2000) includes the 2RU enclosure, backplane, redundant power supplies, and fan trays, but requires separate accelerator and host module selection. Accelerator modules (ACM-A100) are available with either 32GB or

64GB HBM3 memory options. Host compute modules (ACM-H200) are offered with single or dual processor configurations. Integrated bundle packages (ACM-B400) combine the chassis with four accelerator modules and dual host modules for rapid deployment. All configurations include a standard three-year hardware warranty with options for extended support and advanced replacement services.