Ultra-Low Latency Forwarding Engine Factsheet: RoCEv2 Network Interface Module Hardware Integration Manual

APPLICATION CONTEXT

The exponential growth of artificial intelligence (AI), high-performance computing (HPC), and disaggregated storage workloads has placed unprecedented demands on datacenter network fabrics. Traditional TCP/IP kernel stacks introduce significant CPU overhead, memory bandwidth consumption, and latency jitter that directly impede job completion times and resource utilization efficiency. Remote Direct Memory Access over Converged Ethernet version 2 (RoCEv2) has emerged as the definitive wire protocol for lossless, low-latency fabric transport, enabling direct application-to-application memory access without host CPU intervention.

This Hardware Integration Manual provides systems engineers, network architects, and datacenter operations teams with the authoritative technical reference for deploying the RoCEv2 Network Interface Module (NIM) within standard and customized server platforms. The module is purpose-built to serve as the foundational building block for scale-out AI clusters, NVMe-oF storage targets, and financial trading infrastructures where deterministic microsecond-level latency and line-rate throughput are non-negotiable operational requirements.

The document scope encompasses physical installation prerequisites, firmware and driver integration workflows, protocol offload capabilities, thermal and power design constraints, and validated performance benchmarks. All specifications herein reflect the production release firmware version and are subject to the compliance and certification statements detailed in subsequent sections.

FORWARDING PIPELINE DESIGN

The RoCEv2 NIM implements a dedicated hardware forwarding pipeline that completely bypasses the host operating system kernel for established queue pairs. This architecture ensures that RDMA read, write, and atomic operations are processed entirely within the ASIC datapath, delivering deterministic latency

independent of CPU scheduling or system load.

The core forwarding engine comprises four distinct stages. The ingress parser classifies incoming Ethernet frames, extracts VLAN tags, and identifies RoCEv2 UDP destination port 4791 traffic for prioritized handling. The packet buffer and reorder engine manages out-of-order frame delivery, ensuring that PFC (Priority Flow Control) pause frames are generated and honored with minimal reaction time to maintain a lossless fabric. The RDMA transaction engine maintains queue pair context, handles memory translation via the on-module MMU, and executes reliable connection transport semantics in silicon. Finally, the egress scheduler enforces DCQCN (Data Center Quantized Congestion Notification) and ETS (Enhanced Transmission Selection) policies before frame transmission.

This pipelined design achieves a sustained throughput of 400 Gbps per module with a port-to-port latency of under 1.2 microseconds for 64-byte packets. The module supports up to 16 million queue pairs and 2 million memory regions, providing headroom for the most demanding multi-tenant cloud and enterprise environments.

HARDWARE SPECS SUMMARY

The RoCEv2 NIM is a PCI Express 5.0 x16 add-in card compatible with standard

2U and 4U server chassis. It features dual QSFP112 transceiver cages capable of supporting 400GBASE-DR4, 400GBASE-FR4, and 400GBASE-CR4 copper and optical interconnects. The module integrates a dedicated 16-lane PCIe Gen5 host interface, an on-board 8GB DDR4 cache for queue pair and memory translation context, and a hardware root of trust for secure boot and attestation.

Comprehensive diagnostic capabilities are provided through an out-of-band 1GbE management port, I2C/SMBus interface for platform management, and a comprehensive suite of on-die counters and telemetry registers accessible via NC-SI and MCTP over SMBus. The module is designed for a typical power envelope of 75W and supports both standard and reverse airflow configurations to accommodate diverse server thermal designs.

TECHNICAL SPECIFICATIONS

The following parameter registry summarizes the primary electrical, mechanical, protocol, and environmental characteristics of the RoCEv2 Network Interface Module. All values represent nominal operating conditions at 25 degrees Celsius ambient unless otherwise noted.

Parameter	Specification

Form Factor	PCIe 5.0 x16 Add-in Card, full-height, half-length
Switching Capacity	400 Gbps per port, 800 Gbps aggregate (dual-port)
Power Supply	75W typical, 12V PCIe slot power, optional 6-pin AUX
Host Interface	PCI Express 5.0 x16, 32 GT/s per lane
Network Interfaces	2 x QSFP112 (400GBASE-DR4/FR4/CR4)
On-Board Memory	8GB DDR4 for QP and MTU context
RDMA Queue Pairs	Up to 16 million
Memory Regions	Up to 2 million
Port-to-Port Latency	< 1.2 microseconds (64-byte packets)
Protocol Support	RoCEv2, IPv4, IPv6, VLAN, PFC, ETS, DCQCN, ECN
Management	1GbE OOB, NC-SI, MCTP over SMBus, Redfish, gRPC
Operating Temperature	0 to 55 degrees Celsius (standard airflow)
Storage Temperature	-40 to 85 degrees Celsius
Humidity	5 to 95 percent non-condensing
Regulatory Compliance	FCC, CE, UKCA, IEC 62368-1, RoHS 3

MTBF	> 1,000,000 hours (Telcordia SR-332)

QUALITY OF SERVICE PROTOCOLS

The RoCEv2 NIM supports a comprehensive suite of Quality of Service mechanisms essential for converged datacenter fabrics. Priority-based Flow Control (PFC) operates on up to eight traffic classes, with configurable pause thresholds and hysteresis to prevent head-of-line blocking. Enhanced Transmission Selection (ETS) allocates bandwidth guarantees per traffic class, ensuring that storage, compute, and management traffic coexist without starvation.

Explicit Congestion Notification (ECN) marking is performed in hardware at line rate, with support for DCQCN-compatible quantization and feedback. The module also implements standardized congestion control for RoCEv2 via the Congestion Management (CM) protocol, enabling fabric-wide traffic engineering. Advanced telemetry features include per-queue latency histograms, buffer occupancy watermark reporting, and drop reason counters, all exposed via standard Redfish and gRPC telemetry interfaces.

COMPLIANCE VERIFICATION

The RoCEv2 NIM has been validated against the following industry standards and regulatory frameworks: IEEE 802.3cd (400GBASE Ethernet), IEEE 802.1Qbb (PFC), IEEE 802.1Qaz (ETS), IEEE 802.1Qau (QCN), PCI Express Base Specification 5.0, RoCEv2 specification as defined by the InfiniBand Trade Association, and the NVMe over Fabrics specification. Electromagnetic compatibility compliance includes FCC Part 15 Subpart B Class A, EN 55032 Class A, and EN 55035. Safety certifications include IEC 62368-1, UL 62368-1, and CSA C22.2 No. 62368-1. The module is RoHS 3 compliant and carries the CE and UKCA marks.

SCENARIO VIEW

In a representative AI training cluster deployment, 64 GPU servers are interconnected via a leaf-spine fabric of 400G RoCEv2 NIMs. Each server hosts eight modules, providing 3.2 Tbps of aggregate host bandwidth. The lossless fabric, configured with PFC and DCQCN, ensures that collective operations such as AllReduce complete with minimal tail latency. Storage traffic from a parallel NVMe-oF array traverses the same physical fabric with guaranteed bandwidth, eliminating the need for separate storage and compute networks. The result is a 40 percent reduction in job completion time and a 30 percent improvement in fabric utilization compared to a conventional TCP/IP-based cluster.