ADAPTER CARD
PRODUCT BRIEF

†

ConnectX®-5 VPI Adapter Card for Open Compute Project (OCP)

Intelligent RDMA-enabled network adapter card with advanced application offload capabilities for High-Performance Computing, Machine Learning, Cloud, and Storage platforms

ConnectX-5 with Virtual Protocol Interconnect® (VPI) supports 100Gb/s InfiniBand and Ethernet connectivity, super-low latency and very high message rate, plus NVMe over Fabric offloads, providing the highest performance and most flexible solution for Open Compute Project servers and storage appliances, while supporting the most demanding applications and markets: Machine Learning, Data Analytics, and more.

HPC Environments

ConnectX-5 VPI for Open Compute Project (OCP) NIC utilizes both IBTA RDMA (Remote Data Memory Access) and RoCE (RDMA over Converged Ethernet) technologies, delivering high bandwidth, low latency, and high computation efficiency for high performance, data intensive and scalable compute and storage platforms. ConnectX-5 offers significant enhancements to HPC infrastructures by providing MPI and SHMEM/PGAS and Rendezvous Tag Matching offload, hardware support for out-of-order RDMA Write and Read operations, as well as additional Network Atomic and PCIe Atomic operations support.

ConnectX-5 complements switch adaptive-routing capabilities, and supports out-of-order data delivery, while maintaining in-order completion semantics. Additionally ConnectX-5 NIC provides multipath reliability and efficient support for many network topologies, such as DragonFly+.

ConnectX-5 also supports GPUDirect® for enhanced Machine Learning applications, Burst Buffer offload for background checkpointing without interfering in the main CPU operations, and the innovative Dynamic Connected Transport (DCT) service that ensures extreme scalability for compute and storage systems.

Storage Environments

NVMe storage devices are gaining popularity, offering very fast storage access. The evolving NVMe over Fabric (NVMe-oF) protocol leverages the RDMA connectivity for remote access. ConnectX-5 offers further enhacements by providing NVMe-oF target offloads, enabling very efficient NVMe storage access with no CPU intervention, and thus improved performance and lower latency.

As with the earlier generations of ConnectX adapters, standard block and file access protocols can leverage RoCE for high-performance storage access. A consolidated compute and storage network achieves significant cost-performance advantages over multi-fabric networks.

HIGHLIGHTS

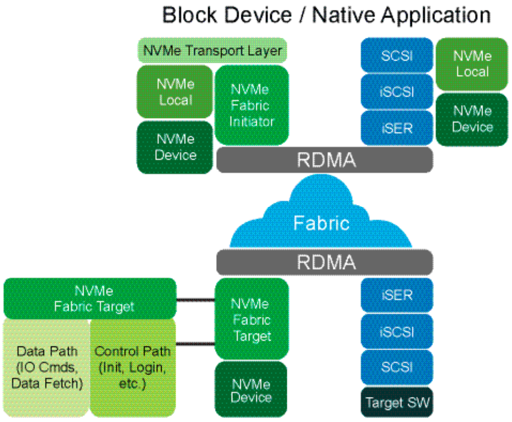
FEATURES

- Tag matching and rendezvous offloads
- Adaptive routing on reliable transport
- Burst buffer offloads for background checkpointing
- NVMe over Fabric (NVMe-oF) offloads
- Back-end switch elimination by host chaining
- Enhanced vSwitch/vRouter offloads
- Flexible pipeline
- RoCE for overlay networks
- RoHS-compliant
- ODCC-compatible

BENEFITS

- Up to 100Gb/s connectivity per port
- Open Compute Project form factor; OCP Specification 2.0 type 1
- Industry-leading throughput, low latency & CPU utilization, and high message rate
- Innovative rack design for storage and Machine Learning
- Smart interconnect for x86, Power, Arm, and GPU-based compute and storage platforms
- Advanced storage capabilities including NVMe over Fabric offloads
- Support for flexible pipeline programmability
- Cutting-edge performance in virtualized networks including NFV
- Enablor for efficient service chaining
- Efficient I/O consolidation, lowering data center costs and complexity

[†] For illustration only. Actual products may vary.



Cloud and Web2.0 Environments

Cloud and Web2.0 customers who are developing their platforms on Software Defined Network (SDN) environments, are leveraging their servers' Operating System Virtual-Switching capabilities to enable maximum flexibility.

Open V-Switch (OVS) is an example of a virtual switch that allows Virtual Machines to communicate with each other and with the outside world. Virtual switch traditionally resides in the hypervisor and switching is based on twelve-tuple matching on flows. The virtual switch or virtual router software-based solution is CPU intensive, affecting system performance and preventing fully utilizing available bandwidth.

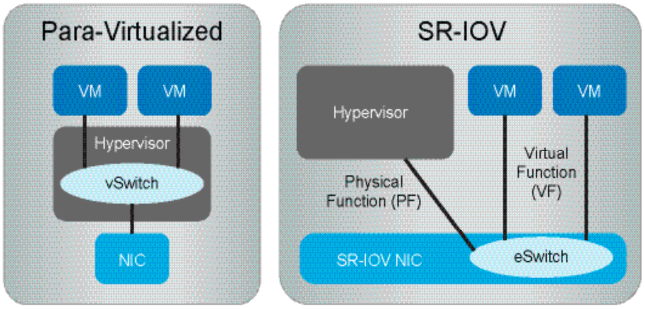
ASAP^2 - Mellanox Accelerated Switch and Packet Processing® technology allows for offloading the vSwitch/vRouter by handling the data plane in the NIC hardware, without modifying the control plane. This results in significantly higher vSwitch/vRouter performance without the associated CPU load.

The vSwitch/vRouter offload functions that are supported by ConnectX-5 include Overlay Networks (for example, VXLAN, NVGRE, MPLS, GENEVE, and NSH) headers' encapsulation and de-capsulation, as well as Stateless offloads of inner packets, packet headers' re-write enabling NAT functionality, and more.

Moreover, the intelligent ConnectX-5 flexible pipeline capabilities, including flexible parser and flexible match-action tables, can be programmed, which enables hardware offloads for future protocols.

ConnectX-5 SR-IOV technology provides dedicated adapter resources and guaranteed isolation and protection for virtual machines (VMs) within the server. Moreover, with ConnectX-5 Network Function Virtualization (NFV), a VM can be used as a virtual appliance. With full data-path operations offloads as well as hairpin hardware capability and service chaining, data can be handled by the Virtual Appliance with minimum CPU utilization.

With these capabilities, data center administrators benefit from better server utilization while reducing cost, power, and cable complexity, allowing for more Virtual Appliances, Virtual Machines and more tenants on the same hardware.



Host Management

Mellanox's host management technology optimizes board management and power, performance and firmware update management via NC-SI, MCTP over SMBus and MCTP over PCIe, as well as PLDM for Monitor and Control DSP0248 and PLDM for Firmware Update DSP0267*.

* Not supported on the UCS M5 generation; Planned support in future UCS generations.

Compatibility*

PCI Express Interface

- PCIe Gen 3.0, 1.1 and 2.0 compatible
- 2.5, 5.0, 8, 16GT/s link rate
- Auto-negotiates to x16, x8, x4, x2, or x1 lanes
- PCIe Atomic
- TLP (Transaction Layer Packet) Processing Hints (TPH)
- PCIe switch Downstream Port Containment (DPC) enablement for PCIe hot-plug
- Access Control Service (ACS) for peer-to-peer secure communication
- Advance Error Reporting (AER)
- Process Address Space ID (PASID) Address Translation Services (ATS)
- IBM CAPI v2 support (Coherent Accelerator Processor Interface)
- Support for MSI/MSI-X mechanisms

Operating Systems/Distributions*

- RHEL/CentOS
- Windows
- FreeBSD
- VMware
- OpenFabrics Enterprise Distribution (OFED)
- OpenFabrics Windows Distribution (WinOF-2)

Connectivity

- Interoperability with InfiniBand switches (up to EDR)
- Interoperability with Ethernet switches (up to 100GbE)
- Passive copper cable with ESD protection
- Powered connectors for optical and active cable support

Features*

InfiniBand

- EDR / FDR / QDR / DDR / SDR
- IBTA Specification 1.3 compliant
- RDMA, Send/Receive semantics
- Hardware-based congestion control
- Atomic operations
- 16 million I/O channels
- 256 to 4Kbyte MTU, 2Gbyte messages
- 8 virtual lanes + VL15

Ethernet

- 100GbE / 50GbE / 40GbE / 25GbE / 10GbE / 1GbE
- IEEE 802.3bj, 802.3bm 100 Gigabit Ethernet
- IEEE 802.3by, Ethernet Consortium 25, 50 Gigabit Ethernet, supporting all FEC modes
- IEEE 802.3ba 40 Gigabit Ethernet
- IEEE 802.3ae 10 Gigabit Ethernet
- IEEE 802.3az Energy Efficient Ethernet (fast wake)
- IEEE 802.3ap based auto-negotiation and KR startup
- IEEE 802.3ad, 802.1AX Link Aggregation
- IEEE 802.1Q, 802.1P VLAN tags and priority
- IEEE 802.1Qau (QCN) – Congestion Notification
- IEEE 802.1Qaz (ETS)
- IEEE 802.1Qbb (PFC)
- IEEE 802.1Qbg
- IEEE 1588v2
- Jumbo frame support (9.6KB)

Enhanced Features

- Hardware-based reliable transport
- Collective operations offloads
- Vector collective operations offloads
- PeerDirect™ RDMA (aka GPUDirect®) communication acceleration
- 64/66 encoding
- Extended Reliable Connected transport (XRC)
- Dynamically Connected transport (DCT)
- Enhanced Atomic operations
- Advanced memory mapping support, allowing user mode registration and remapping of memory (UMR)
- On demand paging (ODP)
- MPI Tag Matching
- Rendezvous protocol offload
- Out-of-order RDMA supporting Adaptive Routing
- Burst buffer offload
- In-Network Memory registration-free RDMA memory access

CPU Offloads

- RRDMA over Converged Ethernet (RoCE)
- TCP/UDP/IP stateless offload
- LSO, LRO, checksum offload
- RSS (also on encapsulated packet), TSS, HDS, VLAN and MPLS tag insertion/stripping, Receive flow steering
- Data Plane Development Kit (DPDK) for kernel bypass applications
- Open VSwitch (OVS) offload using ASAP^2
  - Flexible match-action flow tables
  - Tunneling encapsulation/ de-capsulation
- Intelligent interrupt coalescence
- Header rewrite supporting hardware offload of NAT router

Storage Offloads

- NVMe over Fabric offloads for target machine
- T10 DIF – Signature handover operation at wire speed, for ingress and egress traffic
- Storage protocols: SRP, iSER, NFS RDMA, SMB Direct, NVMe-oF

Overlay Networks

- RoCE over Overlay Networks
- Stateless offloads for overlay network tunneling protocols
- Hardware offload of encapsulation and decapsulation of VXLAN, NVGRE, and GENEVE overlay networks

Hardware-Based I/O Virtualization

- Single Root IOV
- Address translation and protection
- VMware NetQueue support
- SR-IOV: Up to 1K Virtual Functions
- SR-IOV: Up to 16 Physical Functions per host
- Virtualization hierarchies (e.g., NPAR, when enabled)
  - Virtualizing Physical Functions on a physical port
  - SR-IOV on every Physical Function
- Configurable and user-programmable QoS
- Guaranteed QoS for VMs

HPC Software Libraries

- Open MPI, IBM PE, OSU MPI (MVAPICH/2), Intel MPI
- Platform MPI, UPC, Open SHMEM

Management and Control**

- NC-SI over MCTP over SMBus and NC-SI over MCTP over PCIe - Baseboard Management Controller interface
- PLDM for Monitor and Control DSP0248
- SDN management interface for managing the eSwitch
- I^{2}C interface for device control and configuration
- General Purpose I/O pins
- SPI interface to Flash
- JTAG IEEE 1149.1 and IEEE 1149.6

Remote Boot

- Remote boot over InfiniBand
- Remote boot over Ethernet
- Remote boot over iSCSI
- Unified Extensible Firmware Interface (UEFI)
- Pre-execution Environment (PXE)

* This section describes hardware features and capabilities. Please refer to the driver and firmware release notes for feature availability.

**Not supported on the UCS M5 generation; Planned support in future UCS generations.

Table 1 - Environment Specifications for ConnectX-5 VPI Adapter Card for OCP

Temperature
Operating: 0°C to 55°C (32°F to 131°F)
Storage: – 40°C to 70°C (– 40°F to 158°F)

Table 2 - Airflow Specifications (LFM) for ConnectX-5 VI Adapter Card for OCP

Airflow Direction	Heatsink to Port		Port to Heatsink	
Mellanox OPN	**Passive Cable**	**Active Cable**	**Passive Cable**	**Active Cable 1.5W**
ConnectX-5 VPI Card for OCP MCX545B-ECAN	300LFM at 55^{0}C	Not supported	400LFM at 55^{0}C	600LFM at 55^{0}C

Mellanox recommends Mellanox cables & modules.
For additional information on tested modules, go to: http://www.mellanox.com/page/firmware_table_ConnectX5IB
a. Select ConnectX-5 Ethernet/InfiniBand
b. Select OPN MCX545B-ECAN
c. Select PSID
d. Select "Release Notes" under Download/Documentation

Table 3 - Cisco-branded Supported Cables and Modules
(For latest updates check the UCS Technical Specs; Also consult the Cisco Compatibility Matrix: https://tmgmatrix.cisco.com)

Cable PID	Description
QSFP-100G-CU3M	100GBASE CR4 Passive Copper Cable, 3M
QSFP-100G-CU5M	100GBASE CR4 Passive Copper Cable, 5M

Optics PID	Description
QSFP-100G-AOC5M	100GBASE QSFP Active Optical Cable, 5m
QSFP-100G-AOC7M	100GBASE QSFP Active Optical Cable, 7m
QSFP-100G-SR4-S	100GBASE SR4 QSFP Transceiver, MPO, 100M over OM4 MMF
QSFP-40/100-SRBD	100G and 40GBASE SR-BiDi QSFP Transceiver, LC, 100m OM4 MMF
QSFP-100G-LR4-S	100GBASE-LR4 QSFP Transceiver, LC, 10km over SMF

Table 4- Ordering Information

Cisco Product ID	Mellanox Part Number	Description	Qualified Cisco Servers
UCSC-O-M5S100GF UCSC-O-M5S100GF=	MCX545B-ECAN	ConnectX®-5 VPI network interface card for OCP with host management, EDR IB (100Gb/s) and 100GbE, single-port QSFP28, PCIe3.0 x16, no bracket, Type-1 Heat Sink	Cisco UCS 4200 Rack Server

*To change the adapter's configuration, including to enable InfiniBand mode, please consult https://docs.mellanox.com/pages/viewpage.action?pageId=12013115.

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com