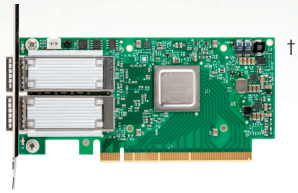
INFINIBAND/ETHERNET (VPI) ADAPTER CARDS

PRODUCT BRIEF

ConnectX®-4 VPI

Single/Dual-Port Adapter Cards supporting 100Gb/s with Virtual Protocol Interconnect®



ConnectX-4 adapter cards with Virtual Protocol Interconnect (VPI), supporting EDR 100Gb/s InfiniBand and 100Gb/s Ethernet connectivity, provide the highest performance and most flexible solution for high-performance, Web 2.0, Cloud, data analytics, database, and storage platforms.

With the exponential growth of data being shared and stored by applications and social networks, the need for high-speed and high performance compute and storage data centers is skyrocketing.

ConnectX®-4 provides exceptional high performance for the most demanding data centers, public and private clouds, Web2.0 and Big Data applications, as well as High-Performance Computing (HPC) and Storage systems, enabling today's corporations to meet the demands of the data explosion.

ConnectX®-4 provides an unmatched combination of 100Gb/s bandwidth in a single port, the lowest available latency, and specific hardware offloads, addressing both today's and the next generation's compute and storage data center demands.

100Gb/s Virtual Protocol Interconnect (VPI) Adapter

ConnectX-4 offers the highest throughput VPI adapter, supporting EDR 100Gb/s InfiniBand and 100Gb/s Ethernet and enabling any standard networking, clustering, or storage to operate seamlessly over any converged network leveraging a consolidated software stack.

I/O Virtualization

ConnectX-4 SR-IOV technology provides dedicated adapter resources and guaranteed isolation and protection for virtual machines (VMs) within the server. I/O virtualization with ConnectX-4 gives data center administrators better server utilization while reducing cost, power, and cable complexity, allowing more Virtual Machines and more tenants on the same hardware.

Overlay Networks

In order to better scale their networks, data center operators often create overlay networks that carry traffic from individual virtual machines over logical tunnels in encapsulated formats such as NVGRE and VXLAN. While this solves network scalability issues, it hides the TCP packet from the hardware offloading engines, placing higher loads on the host CPU. ConnectX-4 effectively addresses this by providing advanced NVGRE and VXLAN hardware offloading engines that encapsulate and de-capsulate the overlay protocol headers, enabling the traditional offloads to be performed on the encapsulated traffic. With ConnectX-4, data center operators can achieve native performance in the new network architecture.

HPC Environments

ConnectX-4 delivers high bandwidth, low latency, and high computation efficiency for the High Performance Computing clusters. Collective communication is a communication pattern in HPC in which all members of a group of processes participate and share data.

CORE-Direct® (Collective Offload Resource Engine) provides advanced capabilities for implementing MPI and SHMEM collective operations. It enhances collective communication scalability and minimizes the CPU overhead for such operations, while providing asynchronous and high-performance collective communication capabilities. It also enhances application scalability by reducing the exposure of the collective communication to the effects of system noise (the bad effect of system activity on running jobs). ConnectX-4 enhances

HIGHLIGHTS

BENEFITS

- Highest performing silicon for applications requiring high bandwidth, low latency and high message rate
- World-class cluster, network, and storage performance
- Smart interconnect for x86, Power, ARM, and GPU-based compute and storage platforms
- Cutting-edge performance in virtualized overlay networks (VXLAN and NVGRE)
- Efficient I/O consolidation, lowering data center costs and complexity
- Virtualization acceleration
- Power efficiency
- Scalability to tens-of-thousands of nodes

KEY FEATURES

- EDR 100Gb/s InfiniBand or 100Gb/s Ethernet per port
- 1/10/20/25/40/50/56/100Gb/s speeds
- 150M messages/second
- Single and dual-port options available
- Erasure Coding offload
- Accelerated Switching and Packet Processing (ASAP²)
- T10-DIF Signature Handover
- Virtual Protocol Interconnect (VPI)
- CPU offloading of transport operations
- Application offloading
- Mellanox PeerDirect™ communication acceleration
- Hardware offloads for NVGRE and VXLAN encapsulated traffic
- End-to-end QoS and congestion control
- Hardware-based I/O virtualization
- Ethernet encapsulation (EoIB)
- RoHS-R6

† For illustration only. Actual products may vary.

the CORE-Direct capabilities by removing the restriction on the data length for which data reductions are supported.

ASAP²™

Mellanox ConnectX-4 EN offers Accelerated Switching And Packet Processing (ASAP²) technology to perform offload activities in the hypervisor, including data path, packet parsing, VxLAN and NVGRE encapsulation/decapsulation, and more.

ASAP² allows offloading by handling the data plane in the NIC hardware using SR-IOV, while maintaining the control plane used in today's software-based solutions unmodified. As a result, there is significantly higher performance without the associated CPU load. ASAP² has two formats: ASAP² Flex™ and ASAP² Direct™.

One example of a virtual switch that ASAP² can offload is OpenVSwitch (OVS).

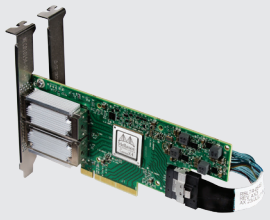
RDMA and RoCE

ConnectX-4, utilizing IBTA RDMA (Remote Data Memory Access) and RoCE (RDMA over Converged Ethernet) technology, delivers low-latency and high-performance over InfiniBand and Ethernet networks. Leveraging data center bridging (DCB) capabilities as well as ConnectX-4 advanced congestion control hardware mechanisms, RoCE provides efficient low-latency RDMA services over Layer 2 and Layer 3 networks.

Mellanox PeerDirect™

PeerDirect communication provides high efficiency RDMA access by eliminating unnecessary internal data copies between components on the PCIe bus (for example, from GPU to CPU), and therefore significantly reduces application run time. ConnectX-4 advanced acceleration technology enables higher cluster efficiency and scalability to tens of thousands of nodes.

INNOVATIVE OPTION - 100Gb/s OVER TWO PCIe x8 SLOTS



Mellanox offers an alternate ConnectX-4 Multi-Host Socket Direct™ card to enable 100Gb/s transmission rate also for servers without x16 PCIe slots. The adapter's 16-lane PCIe bus is split into two 8-lane buses, with one bus accessible through a PCIe x8 edge connector and the other bus through an x8 parallel connector to an Auxiliary PCIe Connection Card. The two cards should be installed into two adjacent PCIe x8 slots and connected using a dedicated harness.

The new card brings improved performance to dual-socket servers. Mellanox Multi-Host Socket Direct™ leverages Mellanox Multi-Host® technology by enabling direct access from each CPU in a dual-socket server to the network through its dedicated PCIe x8 interface.

Multi-Host Socket Direct also brings lower latency and lower CPU utilization to dual socket servers. The direct connection from each CPU to the network means the Interconnect can bypass a QPI (UPI) and the other CPU, optimizing performance and improving latency. CPU utilization is improved as each CPU handles only its own traffic and not traffic from the other CPU.

Multi-Host Socket Direct also enables GPUDirect® RDMA for all CPU/GPU pairs by ensuring that all GPUs are linked to CPUs close to the adapter card, and enables Intel® DDIO on both sockets by creating a direct connection between the sockets and the adapter card.

Storage Acceleration

Storage applications will see improved performance with the higher bandwidth EDR delivers. Moreover, standard block and file access protocols can leverage RoCE and InfiniBand RDMA for high-performance storage access. A consolidated compute and storage network achieves significant cost-performance advantages over multi-fabric networks.

Distributed RAID

ConnectX-4 delivers advanced Erasure Coding offloading capability, enabling distributed RAID (Redundant Array of Inexpensive Disks), a data storage technology that combines multiple disk drive components into a logical unit for the purposes of data redundancy and performance improvement. The ConnectX-4 family's Reed-Solomon capability introduces redundant block calculations, which, together with RDMA, achieves high performance and reliable storage access.

Signature Handover

ConnectX-4 supports hardware checking of T10 Data Integrity Field / Protection Information (T10-DIF/PI), reducing the CPU overhead and accelerating delivery of data to the application. Signature handover is handled by the adapter on ingress and/or egress packets, reducing the load on the CPU at the Initiator and/or Target machines.

Software Support

All Mellanox adapter cards are supported by Windows, Linux distributions, VMware, FreeBSD, and Citrix XENServer. ConnectX-4 VPI adapters support OpenFabrics-based RDMA protocols and software and are compatible with configuration and management tools from OEMs and operating system vendors.

COMPATIBILITY

PCI EXPRESS INTERFACE
- PCIe Gen 3.0 compliant, 1.1 and 2.0 compatible
- 2.5, 5.0, or 8.0GT/s link rate x16
- Auto-negotiates to
- x16, x8, x4, x2, or x1
- Support for MSI/MSI-X mechanisms

CONNECTIVITY
- Interoperable with InfiniBand or 1/10/25/40/50/100Gb Ethernet switches
- Passive copper cable with ESD protection
- Powered connectors for optical and active cable support

OPERATING SYSTEMS/DISTRIBUTIONS*
- RHEL/CentOS
- Windows
- FreeBSD
- VMware
- OpenFabrics Enterprise Distribution (OFED)
- OpenFabrics Windows Distribution (WinOF)

FEATURE SUMMARY*

INFINIBAND
- 2 ports 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

ENHANCED FEATURES
- Hardware-based reliable transport
- Collective operations offloads
- Vector collective operations offloads
- Mellanox 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) – registration free RDMA memory access

ETHERNET
- 100GbE / 56GbE / 50GbE / 40GbE / 25GbE / 10GbE / 1GbE
- IEEE 802.3bj, 802.3bm 100 Gigabit Ethernet
- 25G Ethernet Consortium 25, 50 Gigabit Ethernet
- IEEE 802.3ba 40 Gigabit Ethernet
- IEEE 802.3ae 10 Gigabit Ethernet
- IEEE 802.3az Energy Efficient Ethernet
- IEEE 802.3ap based auto-negotiation and KR startup
- Proprietary Ethernet protocols (20/40GBASE-R2, 50/56GBASE-R4)
- 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)

STORAGE OFFLOADS
- RAID offload - erasure coding (Reed-Solomon) offload
- T10 DIF - Signature handover operation at wire speed, for ingress and egress traffic

OVERLAY NETWORKS
- Stateless offloads for overlay networks and tunneling protocols
- Hardware offload of encapsulation and decapsulation of NVGRE and VXLAN overlay networks

HARDWARE-BASED I/O VIRTUALIZATION
- Single Root IOV
- Multi-function per port
- Address translation and protection
- Multiple queues per virtual machine
- Enhanced QoS for vNICs
- VMware NetQueue support

VIRTUALIZATION
- SR-IOV: Up to 256 Virtual Functions
- SR-IOV: Up to 16 Physical Functions per port
- Virtualization hierarchies (e.g. NPAR)
  - » Virtualizing Physical Functions on a physical port
  - » SR-IOV on every Physical Function
- 1K ingress and egress QoS levels
- Guaranteed QoS for VMs

CPU OFFLOADS
- RDMA over Converged Ethernet (RoCE)
- TCP/UDP/IP stateless offload
- LSO, LRO, checksum offload
- RSS (can be done on encapsulated packet), TSS, HDS, VLAN insertion / stripping, Receive flow steering
- Intelligent interrupt coalescence

REMOTE BOOT
- Remote boot over InfiniBand
- Remote boot over Ethernet
- Remote boot over iSCSI
- PXE and UEFI

PROTOCOL SUPPORT
- OpenMPI, IBM PE, OSU MPI (MVAPICH/2), Intel MPI,
- Platform MPI, UPC, Mellanox SHMEM
- TCP/UDP, EoIB, IPoIB, SDP, RDS, MPLS, VxLAN, NVGRE, GENEVE
- SRP, iSER, NFS RDMA, SMB Direct
- uDAPL

MANAGEMENT AND CONTROL INTERFACES
- NC-SI, MCTP over SMBus and MCTP over PCIe - Baseboard Management Controller interface
- SDN management interface for managing the eSwitch
- I2C interface for device control and configuration
- General Purpose I/O pins
- SPI interface to Flash
- JTAG IEEE 1149.1 and IEEE 1149.61149.6

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

Ordering Part Number	Description**	Dimensions w/o Brackets
MCX455A-ECAT	ConnectX-4 VPI adapter card, EDR IB (100Gb/s) and 100GbE, single-port QSFP28, PCIe3.0 x16, tall bracket, ROHS R6	14.2cm x 6.9cm (low profile)
MCX456A-ECAT	ConnectX-4 VPI adapter card, EDR IB (100Gb/s) and 100GbE, dual-port QSFP28, PCIe3.0 x16, tall bracket, ROHS R6	14.2cm x 6.9cm (low profile)
MCX455A-FCAT	ConnectX-4 VPI adapter card, FDR IB (56Gb/s) and 40/56GbE, single-port QSFP28, PCIe3.0 x16, tall bracket, ROHS R6	14.2cm x 6.9cm (low profile)
MCX456A-FCAT	ConnectX-4 VPI adapter card, FDR IB (56Gb/s) and 40/56GbE, dual-port QSFP28, PCIe3.0 x16, tall bracket, ROHS R6	14.2cm x 6.9cm (low profile)
MCX453A-FCAT	ConnectX-4 VPI adapter card, FDR IB (56Gb/s) and 40/56GbE, single-port QSFP28, PCIe3.0 x8, tall bracket, ROHS R6	14.2cm x 6.9cm (low profile)
MCX454A-FCAT	ConnectX-4 VPI adapter card, FDR IB (56Gb/s) and 40/56GbE, dual-port QSFP28, PCIe3.0 x8, tall bracket, ROHS R6	14.2cm x 6.9cm (low profile)
MCX456M-ECAT	ConnectX-4 VPI adapter card with Multi-Host supporting dual-socket server, EDR IB (100Gb/s) and 100GbE, dual-port QSFP28, dual PCIe3.0 x8, tall bracket, ROHS R6	16.7cm x 6.9cm (low profile) 11.3cm x 4.0cm and 15cm harness

** All listed speeds are the maximum supported and include all lower supported speeds as well.

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