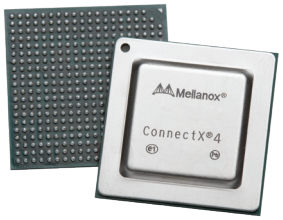
ADAPTER SILICON
PRODUCT BRIEF



ConnectX®·4 †

ConnectX®-4 VPI IC

100Gb/s InfiniBand & Ethernet Adapter IC

Single/Dual-Port Adapter supporting 10/25/40/50/56/100Gb/s with Virtual Protocol Interconnect® and Multi-Host Technology

ConnectX®-4 adapter with Virtual Protocol Interconnect (VPI), supporting 10/25/40/50/56/100Gb/s InfiniBand and Ethernet connectivity, provides the highest performance and most flexible solution for high-performance, Web 2.0, Cloud, data analytics, database, and storage platforms.

With the exponential increase and usage of data, the advantages in scientific algorithms, and the creation of new applications, the demand for the highest interconnect throughput, lowest latency and sophisticated data acceleration engines continues to increase. ConnectX-4 enables data centers to leverage the world's leading interconnect adapter for increasing their applications while reducing operational and capital expenses.

ConnectX-4 provides an unmatched combination of 100Gb/s bandwidth, sub 700ns latency and 150 million messages per second. It includes native hardware support for RDMA over InfiniBand and Ethernet, Ethernet stateless offload engines, GPUDirect®, and the Multi-Host Technology.

The Multi-Host Technology enables the next generation scalable data center design to achieve maximum CAPEX and OPEX savings without compromising on network performance.

HPC ENVIRONMENTS

ConnectX-4 delivers high bandwidth, low latency, and high computation efficiency for the High Performance Computing clusters. Collective operation 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 the CORE-Direct capabilities by removing the restriction on the data length for which data reductions are supported.

MELLANOX MULTI-HOST™ TECHNOLOGY

Mellanox's ConnectX-4 Multi-Host technology enables connecting multiple hosts into a single interconnect adapter by separating the ConnectX-4 PCIe interface into multiple and independent interfaces. Each interface can be connected to a separate host with no performance degradation. ConnectX-4 offers four fully-independent PCIe buses, lowering total cost of ownership in the data center by reducing CAPEX requirements from four cables, NICs, and switch ports to only one of each, and by reducing OPEX by cutting down on switch port management and overall power usage.

Each host can be active or inactive at any time, independent of the other hosts, and receives bandwidth of its own. Bandwidth is split between the hosts, either evenly (default) or based on configurable differentiated Quality of Service (QoS), depending on the data center's needs.

HIGHLIGHTS

BENEFITS

- 10/25/40/50/56/100Gb/s connectivity for servers and storage
- Industry-leading throughput and low latency performance for HPC, Web access and storage
- Maximizing data centers' return on investment (ROI) with Multi-Host technology
- 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

FEATURES

- EDR 100Gb/s InfiniBand or 100Gb/s Ethernet per port
- Single and dual-port options available
- 10/25/40/50/56/100Gb/s speeds
- 150M messages/second
- Multi-Host technology
- Connectivity to up-to 4 independent hosts
- Hardware offloads for NVGRE and VXLAN encapsulated traffic
- CPU offloading of transport operations
- Application offloading
- Mellanox PeerDirect™ communication acceleration
- End-to-end QoS and congestion control
- Hardware-based I/O virtualization
- Accelerated Switching and Packet Processing (ASAP²)
- Erasure Coding offload
- T10-DIF Signature Handover
- Ethernet encapsulation (EoIB)
- Typical power of EDR/100GbE 1-port – 11.2W (ATIS score)
- RoHS2-R6

©2017 Mellanox Technologies. All rights reserved.

† For illustration only. Actual products may vary.

Multi-Host technology features uncompromising independent host management, with full independent NC-SI/MCTP support to each host and to the NIC. IT managers can remotely control the configuration and power state of each host individually, such that management of one host does not affect host traffic performance or the management of the other hosts, guaranteeing host security and isolation. To further lower the total cost of ownership, ConnectX-4 supports management of the multiple hosts using a single BMC, with independent NC-SI/MCTP management channels for each of the managed hosts.

Multi-Host also supports a heterogeneous data center architecture; the various hosts connected to the single adapter can be x86, Power, GPU, or Arm, thereby removing any limitations in passing data or communicating between CPUs.

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. ConnectX-4's SR-IOV capability and Multi-Host technology are mutually exclusive, and each host in a Multi-Host server can leverage an individual SR-IOV implementation.

OVERLAY NETWORKS

In order to better scale Ethernet 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, and enable the traditional offloads to be performed on the encapsulated traffic for these and other tunneling protocols (GENEVE, MPLS, QinQ, and so on). With ConnectX-4, data center operators can achieve native performance in the new network architecture.

ASAP^2 ™

Mellanox ConnectX-4 VPI offers Accelerated Switching And Packet Processing (ASAP^2) technology to perform offload activities in the hypervisor, including data path, packet parsing, VxLAN and NVGRE encapsulation/decapsulation, and more.

ASAP^2 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. ASAP^2 has two formats: ASAP^2 Flex™ and ASAP^2 Direct™. One example of a virtual switch that ASAP^2 can offload is OpenVSwitch (OVS).

RDMA

ConnectX-4 VPI utilizes both IBTA RDMA (Remote Data Memory Access) and RoCE (RDMA over Converged Ethernet) technologies, delivering low-latency and high performance. ConnectX-4 leverages data center bridging (DCB) capabilities and advanced congestion control hardware mechanisms to provide 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.

STORAGE ACCELERATION

Storage applications will see improved performance with the higher bandwidth EDR delivers. Moreover, 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.

DISTRIBUTED RAID

ConnectX-4 delivers advanced Erasure Coding offloading capability, enabling distributed Redundant Array (RAID) 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.

STANDARD & MULTI-HOST MANAGEMENT

Mellanox's host management technology for standard and multi-host platforms 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.

SOFTWARE SUPPORT

All Mellanox adapter cards are supported by Windows, Linux distributions, VMware, FreeBSD, and Citrix XENServer. ConnectX-4 support various management interfaces and has a rich set of tool for configuration and management across operating systems.

©2017 Mellanox Technologies. All rights reserved.

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

Operating Systems/Distributions*
- RHEL/CentOS
- Windows
- FreeBSD
- VMware
- OpenFabrics Enterprise Distribution (OFED)
- OpenFabrics Windows Distribution (WinOF)

Connectivity
- Interoperable with 10/25/40/50/56/100Gb Ethernet switches
- 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 / 56GbE / 50GbE / 40GbE / 25GbE / 10GbE
- 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)

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)
- Registration free RDMA memory access

Multi-Host
- Up to 4 separate PCIe interfaces to 4 independent hosts
- Two PCIe x8 to two hosts or four PCIe x4 to four hosts
- Independent NC-SI SMBus interfaces
- Independent stand-by and wake-on-LAN signals

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 supports

Virtualization
- SR-IOV: Up to 256 Virtual Functions
- SR-IOV: Up to 16 Physical Functions per port
- Virtualization hierarchies (e.g., NPAR and Multi-Host)
  - 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, Open SHMEM
- TCP/UDP, 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
- PLDM for Monitor and Control DSP0248
- PLDM for Firmware Update DSP0267
- SDN management interface for managing the eSwitch
- I^2C interface for device control and configuration
- General Purpose I/O pins
- SPI interface to Flash
- JTAG IEEE 1149.1 and IEEE 1149.6

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

©2017 Mellanox Technologies. All rights reserved.

Table 1 - Part Numbers and Descriptions

OPN	Description
MT27704A0-FDCF-FV	ConnectX-4 VPI, 1-port IC, FDR/56GbE, PCIe 3.0 x16, 8GT/s (RoHS2-R6)
MT27708A0-FDCF-FV	ConnectX-4 VPI, 2-port IC, FDR/56GbE, PCIe 3.0 x16, 8GT/s (RoHS2-R6)
MT27704A0-FDCF-EV	ConnectX-4 VPI, 1-port IC, EDR/100GbE, PCIe 3.0 x16, 8GT/s (RoHS2-R6)
MT27708A0-FDCF-EV	ConnectX-4 VPI, 2-port IC, EDR/100GbE, PCIe 3.0 x16, 8GT/s (RoHS2-R6)
MT27708A0-FDCF-FVM	ConnectX-4 VPI, 2-port IC, FDR/56GbE, Multi-Host, PCIe 3.0 x16, 8GT/s (RoHS2-R6)
MT27708A0-FDCF-EVM	ConnectX-4 VPI, 2-port IC, EDR/100GbE, Multi-Host, PCIe 3.0 x16, 8GT/s (RoHS2-R6)

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

© Copyright 2017. Mellanox Technologies. All rights reserved.
Mellanox, Mellanox logo, ConnectX, CORE-Direct and GPUDirect are registered trademarks of Mellanox Technologies, Ltd.
Mellanox Multi-Host, Mellanox PeerDirect, ASAP², ASAP² Flex™ and ASAP² Direct™ are trademarks of Mellanox Technologies, Ltd. All other trademarks are property of their respective owners.

53445PB
Rev 2.5