ADAPTER CARD
PRODUCT BRIEF

ConnectX®-5 VPI Socket Direct™

100Gb/s InfiniBand & Ethernet Adapter Card

ConnectX®-5 †

Intelligent RDMA-enabled network adapter card with advanced application offload capabilities supporting 100Gb/s for servers without x16 PCIe slots

ConnectX-5 Socket Direct with Virtual Protocol Interconnect® supports two ports of 100Gb/s InfiniBand and Ethernet connectivity, very low latency, and very high message rate, OVS and NVMe over Fabric offloads, providing the highest performance and most flexible solution for the most demanding applications and markets: Machine Learning, Data Analytics, and more.

SOCKET DIRECT

ConnectX-5 Socket Direct provides 100Gb/s port speed even to servers without x16 PCIe slots by splitting the 16-lane PCIe bus into two 8-lane buses, one of which is accessible through a PCIe x8 edge connector and the other through a parallel x8 Auxiliary PCIe Connection Card, connected by a dedicated harness.

Moreover, the card brings improved performance to dual-socket servers by enabling direct access from each CPU in a dual-socket server to the network through its dedicated PCIe x8 interface. In such a configuration, Socket Direct also brings lower latency and lower CPU utilization. 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.

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.

Mellanox Multi-Host™ technology, which was first introduced with ConnectX-4, is enabled in the Mellanox Socket Direct card, allowing multiple hosts to be connected into a single adapter by separating the PCIe interface into multiple and independent interfaces.

HPC ENVIRONMENTS

ConnectX-5 delivers high bandwidth, low latency, and high computation efficiency for high performance, data intensive and scalable compute and storage platforms. ConnectX-5 offers 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 VPI utilizes both IBTA RDMA (Remote Data Memory Access) and RoCE (RDMA over Converged Ethernet) technologies, delivering low-latency and high performance. ConnectX-5 enhances RDMA network capabilities by completing the Switch Adaptive-Routing capabilities and supporting data delivered out-of-order, while maintaining ordered completion semantics, providing multipath reliability and efficient support for all network topologies including DragonFly and DragonFly+.

ConnectX-5 also supports Burst Buffer offload for background checkpointing without interfering in the main CPU operations, and the innovative transport service Dynamic Connected Transport (DCT) to ensure extreme scalability for compute and storage systems.

HIGHLIGHTS

NEW FEATURES

- Socket Direct, enabling 100Gb/s for servers without x16 PCIe slots
- 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

BENEFITS

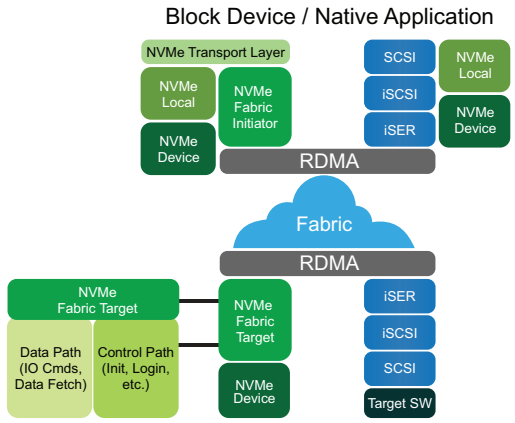
- Up to 100Gb/s connectivity per port
- Industry-leading throughput, low latency, low CPU utilization and high message rate
- Low latency for dual-socket servers in environments with multiple network flows
- Maximizes data center ROI with Multi-Host technology
- Innovative rack design for storage and Machine Learning based on Host Chaining technology
- Smart interconnect for x86, Power, Arm, and GPU-based compute and storage platforms
- Advanced storage capabilities including NVMe over Fabric offloads
- Intelligent network adapter supporting flexible pipeline programmability
- Cutting-edge performance in virtualized networks including Network Function Virtualization (NFV)
- Enabler for efficient service chaining capabilities
- Efficient I/O consolidation, lowering data center costs and complexity

† For illustration only. Actual products may vary.

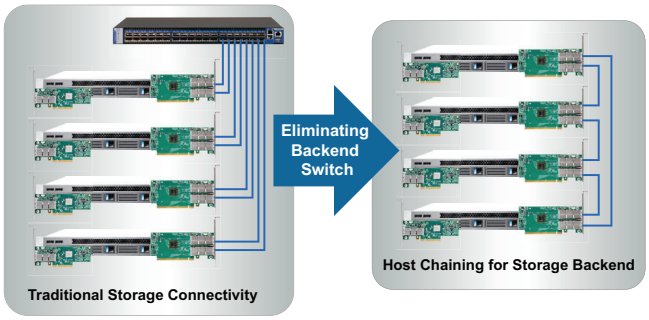
STORAGE ENVIRONMENTS

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.



ConnectX-5 enables an innovative storage rack design, Host Chaining, by which different servers can interconnect directly without involving the Top of the Rack (ToR) switch. With the various new rack design alternatives (Host Chaining and Multi-Host), ConnectX-5 Socket Direct lowers the total cost of ownership (TCO) in the data center by reducing CAPEX (cables, NICs, and switch port expenses), and by reducing OPEX by cutting down on switch port management and overall power usage.



CLOUD AND WEB2.0 ENVIRONMENTS

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

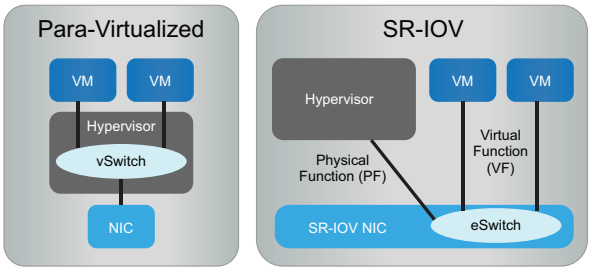
Open V-Switch (OVS) is an example of a virtual switch that allows Virtual Machines to communicate with each other and with the outside world. Virtual switch traditionally resides in the hypervisor and switching is based on multiple-tuple matching on flows. The virtual switch or virtual router software-based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

Mellanox Accelerated Switching And Packet Processing (ASAP^2) Direct technology allows to offload vSwitch/vRouter by handling the data plane in the NIC hardware while maintaining the control plane unmodified. As a results there is 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, which include flexible parser and flexible match-action tables, can be programmed, which enable 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 more Virtual Appliances, Virtual Machines and more tenants on the same hardware.

MANAGEABILITY FOR SOCKET DIRECT

Manageability is supported through a BMC. The Socket Direct PCIe stand-up adapter can be connected to a BMC using MCTP over SMBus or MCTP over PCIe protocols as if it is a standard Mellanox PCIe stand-up adapter. The adapter can be configured per the server's specific manageability solution.

COMPATIBILITY

PCI Express Interface
- PCIe Gen 3.0, 1.1 and 2.0 compatible
- 2.5, 5.0, 8, 16GT/s link rate
- 32 lanes as 2x 16-lanes of PCIe
- Auto-negotiates to x16, x8, x4, x2, or x1 lanes
- PCIe Atomic
- TLP (Transaction Layer Packet) Processing Hints (TPH)
- Advance Error Reporting (AER)
- Process Address Space ID (PASID) Address Translation Services (ATS)
- 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
- RDMA 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 ASAP²
  - 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
- Erasure Coding offload – offloading Reed Solomon calculations
- 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 and Multi-Host, 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
- PLDM for Firmware Update DSP0267
- 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 1149.6

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.

Table 1 - Part Numbers and Descriptions

OPN	Description	Dimensions w/o Bracket
MCX556M-ECAT-S25	ConnectX®-5 VPI adapter card with Multi-Host Socket Direct supporting dual-socket server, FDR/EDR IB (100Gb/s) and 40/50/100GbE, dual-port QSFP28, 2x PCIe3.0 x8, 25cm harness, tall bracket	16.7cm x 6.9cm (low profile) 11.3cm x 4.0cm and 25cm harness
MCX556M-ECAT-S35A	ConnectX®-5 VPI adapter card with Multi-Host Socket Direct supporting dual-socket server, FDR/EDR IB (100Gb/s) and 40/50/100GbE, dual-port QSFP28, 2x PCIe3.0 x8, 35cm harness, active PCI extension card, tall bracket	16.7cm x 6.9cm (low profile) 11.3cm x 4.0cm and 35cm harness

NOTE: All tall-bracket adapters are shipped with the tall bracket mounted and a short bracket as an accessory.

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