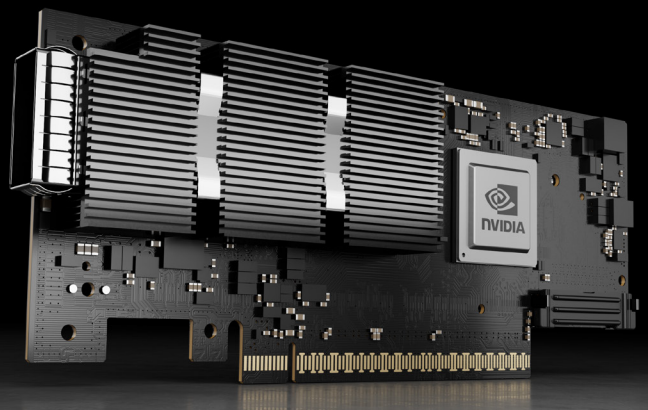
DATASHEET



NVIDIA CONNECTX INFINBAND ADAPTER PORTFOLIO

Smart, efficient, scalable

Complex workloads demand ultra-fast processing of high-resolution simulations, extreme-size datasets, and highly parallelized algorithms. As these computing requirements continue to grow, the NVIDIA® Quantum InfiniBand platform—the world's only fully offloadable, In-Network Computing acceleration technology—provides the dramatic leap in performance needed to achieve unmatched performance in high-performance computing (HPC), AI, and hyperscale cloud infrastructures with less cost and complexity.

NVIDIA ConnectX® smart host channel adapters (HCAs), with best-in-class performance and efficiency, are the ideal solution for HPC clusters that demand high bandwidth, high message rate, and low latency to achieve the highest server efficiency and application productivity. With remote direct memory access (RDMA) traffic consolidation and hardware acceleration for virtualization, ConnectX provides optimal I/O services to achieve the maximum return-on-investment (ROI) for data centers, high-scale storage systems, and cloud computing.

World-Class Performance and Scale

ConnectX adapter cards provide best-in-class performance and efficient computing through NVIDIA **Magnum IO™** and its advanced acceleration and offload capabilities. Network protocol processing and data movement are completed in the adapter without CPU intervention. Application acceleration, support for the NVIDIA Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™, and GPU communication acceleration via NVIDIA GPUDirect® technology bring further performance improvements. These technologies enable higher cluster efficiency and scalability to hundreds-of-thousands of nodes.

Accelerated Storage

The NVIDIA Quantum InfiniBand platform is being leveraged by NVIDIA's storage partners to meet the scaling, bandwidth, and latency requirements needed to exploit the potential of their storage solutions. Standard block and file access protocols leveraging InfiniBand RDMA result in high-performance storage access. ConnectX adapters support SRP, iSER, NFS RDMA, SMB Direct, SCSI and iSCSI, as well as NVMe over Fabrics (NVMe-oF) storage protocols. ConnectX adapters also offer a flexible signature handover mechanism based on the advanced T10 Data Integrity Field (DIF) implementation.

KEY BENEFITS

- World-class cluster performance leveraging In-Network Computing engines
- Efficient use of compute resources
- High bandwidth and low-latency services
- Smart interconnect for x86, Power, Arm, and GPU-based compute and storage platforms
- Virtualization acceleration
- RDMA and TCP/IP for I/O consolidation
- SR-IOV technology: VM protection and QoS
- Compliant with OCP 2.0 and 3.0 NIC specifications, depending upon model, as well as Open Data Center Committee (ODCC) compatible

TARGET APPLICATIONS

- High performance computing
- Machine learning, artificial intelligence, and data analysis platforms
- Clustered databases and high-throughput data warehousing
- Latency-sensitive financial analysis and high frequency trading
- Embedded systems leveraging high performance and low latency
- Performance storage applications, such as backup, restore, and mirroring
- Compute and storage platforms

Enabling I/O Virtualization

ConnectX adapters provide comprehensive support for virtualized data centers with single root I/O virtualization (SR-IOV) allowing dedicated adapter resources and guaranteed isolation and protection for virtual machines (VM) within the server. I/O virtualization on InfiniBand gives data center managers better server utilization and LAN and SAN unification while reducing cost, power, and cable complexity.

Connect Multiple Hosts to Your Adapter

NVIDIA Multi-Host™ connects multiple compute or storage hosts into a single interconnect adapter, separating the adapter PCIe interface into multiple and independent PCIe interfaces with no performance degradation. The technology is provided on selected adapters and enables designing and building new scale-out heterogeneous compute and storage racks with direct connectivity between compute elements, storage elements, and the network. Resulting in better power and performance management, while achieving maximum data processing and data transfer.

Portfolio for Every Data Center

ConnectX adapter cards are available in a variety of form factors to meet specific data center needs, including:

- Standard PCI Express Gen 3.0, Gen 4.0, and Gen 5.0 adapter cards
- Open Compute Project (OCP) cards integrate into the most cost-efficient, energy efficient, and scalable enterprise and hyperscale data centers, delivering leading connectivity for performance-driven server and storage applications. The OCP mezzanine adapter form factor is designed to mate into OCP servers.
  - OCP Spec 2.0 Type 1 and Type 2 mezzanine adapter form factors
  - OCP Spec 3.0 Small Form Factor (SFF) and tall small form factor (TSFF)
- NVIDIA Socket Direct™ cards split the PCI Express bus into two buses, such that each CPU socket gets direct connectivity to the network. With this direct connectivity, traffic can bypass the inter-processors interface, optimizing performance and reducing latency for dual socket servers. Also, each CPU handles only its own traffic, improving CPU utilization. GPUDirect RDMA is also enabled for all CPU/GPU pairs, ensuring that all GPUs are linked to those CPUs that are closest to the adapter card.
  - NVIDIA Socket Direct cards can enable 400Gb/s transmission rates for PCIe Gen 4.0 servers or 200Gb/s for PCIe Gen 3.0 servers, leveraging two PCIe x16 slots. Other models split 16-lane PCIe into two 8-lane buses.
- ConnectX is also available as standalone ASICs.

ConnectX adapters offer the flexibility of connectivity for both InfiniBand and Ethernet protocols on the same adapter card.

Broad Software Support

All ConnectX adapters are supported by a full suite of drivers for major Linux distributions, Microsoft® Windows® Server and VMware vSphere®. Drivers are also available inbox in Linux main distributions, Windows, and vSphere.

ConnectX InfiniBand Adapter Card Portfolio and Specs

Feature	ConnectX-5 Adapter Card	ConnectX-6 Adapter Card	ConnectX-7 Adapter Card
General Specs			
Ports	Single, Dual	Single, Dual	Single, Dual
Port speed	100Gb/s and lower	200Gb/s and lower	400Gb/s and lower
PCI Express (PCIe)	Gen 3.0 x16; Gen 4.0 x16	Gen 3.0/4.0 x16; 32 lanes as 2x Gen 3.0 16-lane PCIe	Gen 4.0/5.0 x16; 32 lanes as 2x Gen 4.0/5.0 16-lane PCIe
Connectors	QSFP28	QSFP56	OSFP, QSFP112
RMDA message rate (million msgs/sec)	200 (ConnectX-5 Ex, Gen4 server); 165 (ConnectX-5, Gen3 server)	215	330-370
RDMA	√	√	√
OOO RDMA (adaptive routing)	√	√	√
Dynamically connected transport	√	√	√
NVIDIA Multi-Host	4 hosts	4 hosts	4 hosts
Storage			
NVMe-oF target offload	√	√	√
T10-DIF/signature handover	√	√	√
Virtualization			
SR-IOV	8 physical functions per port, 512 virtual functions per port	8 physical functions per port, 1K virtual functions per port	16 physical functions per port, 1K virtual functions per port
Congestion control (QCN, ECN)	√	√	√
MPI tag matching offload	√	√	√
Open vSwitch (OVS) offload	√	√	√
VM isolation and protection	√	√	√
Security			
Block-level XTS-AES hardware encryption		√	√
Secure boot			√
Federal Information Processing Standards (FIPS) compliant		√	√
Management			
Hairpin (host chaining)	√	√	√
Host management	√	√	√
Multi-Host isolation and protection	√	√	√
QoS			
Packet pacing	√	√	√
Form Factors			
Standard PCI Express stand-up	√	√	√
OCP	√ OCP 2.0 Type 1 & 2, OCP 3.0 SFF	√ OCP 3.0 SFF	√ OCP 3.0 TSFF/SFF
NVIDIA Socket Direct	√	√	√

For lower InfiniBand port speeds supported, please refer to the product user manual.

ConnectX-5 Ex is a higher throughput and lower latency ConnectX-5 model.

For ordering information and detailed information on feature availability, compliance, and compatibility, please see the product's user manual and the driver/firmware release notes on the public web site. Product datasheets are also available for each ConnectX InfiniBand adapter.

This document focuses on InfiniBand, Ethernet features and specs can be found in the Ethernet product documentation.

Learn more

Learn more at **nvidia.com/en-us/networking/infiniband-adapters**

NVIDIA