NVIDIA MELLANOX INFINIBAND

NDR 400G ARCHITECTURE

The 7th generation of NVIDIA® Mellanox® InfiniBand provides AI developers and scientific researchers the fastest networking performance available to take on the world's most challenging problems.

It provides ultra-low latency and doubles data throughput with NDR 400Gb/s and adds new NVIDIA In-Network Computing engines to provide additional acceleration to deliver the scalability and feature-rich technology needed for supercomputers, artificial intelligence (AI) and hyperscale cloud data centers. NDR InfiniBand continues to enhance and extend NVIDIA In-Network Computing technologies, including pre-configured and programmable compute engines such as NVIDIA® Mellanox® Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™, MPI tag matching, and MPI All-to-All and programmable cores, delivering both the best cost per node and the best return on investment (ROI).

With this new generation of InfiniBand interconnect, NVIDIA continues to set world records for high performance networking: delivering 2X higher bandwidth per port, 3X higher switch silicon port density, 5X higher switch system capacity, and 32X higher AI acceleration power.

NDR InfiniBand technology offers the third generation of SHARP technology, allowing virtually unlimited scalability for large data aggregation through the network, with support for 64 parallel flows — which is 32X higher capacity of the co-processing network compared to the previous generation. MPI All-to-All and MPI tag matching hardware engines, in addition to other features like advanced congestion control, adaptive routing, self-healing networking and more, all provide critical enhancements to high performance computing (HPC) and AI clusters to reach even higher performance than ever.

NDR INFINIBAND OFFERING

The NDR switch ASIC delivers 64 ports of 400 Gb/s InfiniBand speed or 128 ports of 200 Gb/s, the third generation of Scalable Hierarchical Aggregation and Reduction Protocol (SHARPv3), advanced congestion control, adaptive routing and Self-healing Networking technology. Based on the NDR switch ASIC, the NVIDIA offering includes edge and modular switch systems.

Performance

- NDR 400Gb/s bandwidth per port
- 64 NDR 400G ports or 128 NDR200 200G ports in a single switch device
- 2048 NDR ports or 4096 NDR200 ports in a single modular switch
- Over 66.5 billion packets per second (bi-directional) on a single NDR switch device

Breaking the Records

- 2X bandwidth per port
- 3X the switch radix versus HDR InfiniBand
- 32X higher AI acceleration power per switch
- Over 1M nodes in a 4-switch tier Dragonfly+ network, 6.5X higher vs HDR InfiniBand

Key Features

- Full transport offload
- RDMA, GPUDirect RDMA, GPUDirect Storage
- Programmable In-Network Computing engines
- MPI All-to-All hardware acceleration
- MPI Tag Matching hardware acceleration
- Third generation of Scalable Hierarchical Aggregation and Reduction Protocol (SHARPv3)
- Advanced Adaptive Routing, Congestion Control and Quality of Service
- Self-Healing Networking

† For illustration only. Actual products may vary.

The NDR Host Channel Adapter (HCA) ASIC delivers 400 Gb/s data throughput. It supports 32 lanes of PCIe Gen5 or Gen4 for host connectivity. The adapter also supports multiple pre-configured In-Network Computing acceleration engines such as MPI All-to-All and MPI Tag Matching hardware, as well as multiple programmable compute cores.

NDR InfiniBand connectivity is built on the most advanced 100Gb/s per lane SerDes technology. NDR InfiniBand physical connectivity is based on OSPF connectors, on both the switches and HCA endpoints. Each switch OSFP connector holds 2X NDR or 4X NDR200 InfiniBand ports. Each HCA OSFP connector carries either a single NDR InfiniBand or NDR200 port. The NDR cabling offering includes active and passive copper cables, transceivers, and MPOs.

NDR INFINIBAND EDGE SWITCHES

The NDR family of edge switches comprises switches with 64 NDR ports on 32 physical OSFP ports, which can be split to deliver up to 128 NDR200 ports. The offering of compact 1U switches includes air-cooled and liquid-cooled flavors, internally managed and externally managed (aka unmanaged) switches.

The NDR family of switches enables connectivity to other switches or hosts at varying speeds (NDR, NDR200, HDR, HDR100 and EDR) and carries an aggregated throughput of 51.2 terabits per second (Tb/s) bi-directional throughout, with a landmark of more than 66.5 billion packets per second capacity. As an ideal rack-mounted InfiniBand solution, the NDR edge switch allows maximum flexibility as it enables a variety of topologies, including Fat Tree, DragonFly+, multi-dimensional Torus, and more.

NDR INFINIBAND MODULAR SWITCHES

The NDR family of modular switches offers the following configurations:

- 2048 ports of 400G NDR or 4096 ports of 200G NDR200
- 1024 ports of 400G NDR or 2048 ports of 200G NDR200

The large modular switch is based on a non-blocking two-level fat tree “fabric-in-a-rack” occupying dual rack width, and carries a total unidirectional throughput of 819 terabits per second (Tb/s) or bidirectional throughput of 1.64 petabits per second (Pb/s) - 5X over the previous generation of HDR InfiniBand modular switches.

The mid-size modular switch delivers a two-level non-blocking fat tree of 1024 NDR nodes, in a single rack width chassis. It carries a total unidirectional throughput of 409Tb/s or bidirectional throughput of 819Tb/s.

NDR HOST CHANNEL ADAPTERS

NDR Host Channel Adapters (HCAs) are offered in various form factors, delivering single or dual ports at NDR or NDR200 speeds. The popular form factors are standard CEM with 16 lanes of PCIe Gen4 and Gen5. Some flavors come with the option to connect an additional 16-lane auxiliary card, leveraging NVIDIA Mellanox Socket Direct® technology, to achieve 32 lanes of PCIe Gen4. Other form factors include OCP 3.0 with OSFP connectors, OCP 3.0 with QSFP connectors, as well as CEM PCIe x16 with QSFP connectors.

The NDR InfiniBand HCAs deliver advanced In-Network Computing with MPI All-to-All and MPI Tag Matching hardware engines, and other fabric enhancement features such as Quality of Service (QoS), adaptive routing, congestion control, and more.

The NDR HCA also includes multiple programmable compute cores which enable the offloading of pre-processing data algorithms and application control paths, from the CPU or GPU to the network, providing higher performance, scalability and overlapping between compute and communication tasks.

NDR TRANSCEIVERS AND CABLES

NDR connectivity includes transceivers and MPOs, Active Copper Cables (ACCs) and Direct Attached Cables (DACs), and features the following:

> Maximum flexibility to build any topology desired — the twin port transceiver enables plugging in two MPOs per NDR port. The single port transceivers are used at the HCA side when NDR or NDR200 optical connectivity is needed. NVIDIA also offers a variety of MPOs, straight and split and at various lengths, to achieve maximum flexibility and best performance.

> Twin port transceivers with finned or flat OSFP connectors — can be plugged into air-cooled or liquid-cooled switches, respectively.

> Switch to HCA connectivity — both DACs (up to 1.5m), and ACCs (up to 3m) are offered. The 1-to-2 split cable is used to connect an OSFP switch port (that holds two NDR ports) to two individual NDR HCAs. The 1-to-4 split cable is used to connect an OSFP switch port to four HDR200 HCAs.

> Short OSFP to OSFP DAC — Connects two parallel NDR connections between two neighboring switches. This option enables cost-effective topologies, such as DF+ that leverages the co-location of spine switches in a single rack.

Learn more at **www.nvidia.com**