SOFTWARE

MLNX_OFED (OpenFabrics Enterprise Distribution)

High-Performance Server and Storage Connectivity Software for Field-Proven RDMA and Transport Offload Hardware Solutions

Clustering using commodity servers and storage systems is seeing widespread deployments in large and growing markets such as high performance computing, data warehousing, online transaction processing, financial services and large scale web 2.0 deployments. To enable distributed computing transparently and with maximum efficiency, applications in these markets require the highest I/O bandwidth and lowest possible latency. These requirements are compounded with the need to support a large interoperable ecosystem of networking, virtualization, storage, and other applications and interfaces. The OFED from OpenFabrics Alliance (www.openfabrics.org) has been hardened through collaborative development and testing by major high performance I/O vendors. Mellanox OFED (MLNX_OFED) is a Mellanox tested and packaged version of OFED and supports two interconnect types using the same RDMA (remote DMA) and kernel bypass APIs called OFED verbs – InfiniBand and Ethernet. 10/20/40Gb/s InfiniBand and RoCE (based on the RDMA over Converged Ethernet standard) over 10/40GigE are supported with OFED by Mellanox to enable OEMs and System Integrators to meet the needs end users in the said markets.

Delivering Converged and Higher I/O Service Levels

Networking in data center environments comprise of server-to-server messaging (IPC), server to LAN and server to SAN traffic. Applications in that run in the data center, especially those belonging to the target markets, expect different APIs or interfaces optimized for their behavior to enable highest performance.

- Efficient & high performance HPC clustering: In HPC markets, MPI (message passing interface) is widely used. OFED delivers multiple optimized verbs-level features both in the kernel and user spaces to enable MPI middleware to gain the maximum latency and throughput performance. For example, bandwidth results in excess of 3200MB/s and application latency of 1 microsecond have been achieved, besides enabling scaling to large clusters while improving on memory usage and latency related efficiencies.
- Lowest Latency for financial services applications: MLNX_OFED with InfiniBand and RoCE delivers the lowest latency for financial applications that use the common OFED verbs interface.
- Lowest Latency and bandwidth for clustered database applications: The Reliable Datagram Services (RDS) protocol included in MLNX_OFED provides an efficient sockets-based interface for Oracle RAC database applications. UDAPL is another user level RDMA interface used by applications such as IBM DB2 pureScale.
- Web 2.0 and other traditional sockets-based applications: MLNX_OFED includes the field proven implementation of IP-over-IB, enabling IP-based applications to work seamlessly over InfiniBand. It also includes the IBTA (www.InfiniBandTA.org) defined Sockets Direct Protocol (SDP) enabling traditional TCP/IP sockets-based applications to capitalize on the RDMA and transport offload capabilities of InfiniBand and RoCE. Standard UDP/TCP/IP sockets interfaces are supported over L2 NIC (Ethernet) driver implementation for applications that are not sensitive to lowest latency.

BENEFITS

- Virtual Protocol Interconnect allows Mellanox ConnectX and ConnectX-2 devices to run both InfiniBand and 10GigE traffic simultaneously on two ports
- Single software stack that operates across all available Mellanox InfiniBand and Ethernet devices and configurations such as mem-free, QDR/DDR/SDR, 10 /40 GigE, and PCI Express modes
- Support for HPC applications for scientific research, oil and gas exploration, car crash tests, bench marking etc. E.g., Fluent, LSDYNA
- Support for Data Center applications such as Oracle 11g/10g RAC, IBM DB2, Financial services applications such as IBM WebSphere LLM, Red Hat MRG, NYSE Data Fabric
- Support for high-performance block storage applications utilizing RDMA benefits

SOCKETS LAYER

- SDP and IP-over-IB component enable TCP/IP and sockets-based applications to interface seamlessly to and benefit from InfiniBand transport
- L2 NIC UDP/TCP/IP interface

ACCESS LAYER

- Supports the OpenFabrics defined Verbs API at the user and kernel levels. User level verbs allow protocols such as MPI and UDAPL to interface to Mellanox InfiniBand and 10/40GigE RoCE hardware. Kernel levels verbs allow protocols like SDP, iSER, SRP to interface to Mellanox InfiniBand and 10/40GigE RoCE hardware.

SCSI MID LAYER

- The SCSI Mid Layer interface enables SCSI based block storage and management applications to interface with the SRP Initiator component and the Mellanox InfiniBand hardware

- Storage applications: To enable traditional SCSI and iSCSI-based storage applications to enjoy similar RDMA performance benefits, MLNX_OFED includes the SCSI over RDMA Protocol (SRP) initiator and target, and the iSCSI RDMA Protocol (iSER) that interoperate with various target components available in the industry. SRP solutions over InfiniBand have been proven to deliver impressive 910MB/s (random read) and 725MB/s (random write) I/O storage performance with 1MB blocks. iSER can be implemented over both InfiniBand or RoCE.

Virtual Protocol Interconnect Support

MLNX_OFED uses multi-layer, yet efficient device driver architecture to enable multiple I/O connectivity options over the same hardware I/O adapter (HCA or NIC). The same OFED stack installed on a server can deliver I/O services over both InfiniBand and Ethernet simultaneously, and ports can be repurposed to meet application and end user needs. For example, one port on the adapter can function as Ethernet in L2 NIC mode or RoCE mode and the other port can operate as InfiniBand. Or, both ports can be repurposed to run as InfiniBand or Ethernet. By doing so, maximum flexibility is provided for how the above described converged and higher I/O service levels are delivered.

High Availability (HA)

MLNX_OFED includes high availability support for message passing, sockets and storage applications. Specifically, the standard Linux channel bonding module is supported over IPoIB enabling failover across ports on the same adapter, or across adapters. Similarly, standard multi-pathing and failover over the SRP Initiator is supported for SCSI applications using standard Linux implementations such as device mapper multipath (dm-multipath) driver. Some vendor-specific fail-over/load-balancing driver models are supported as well.

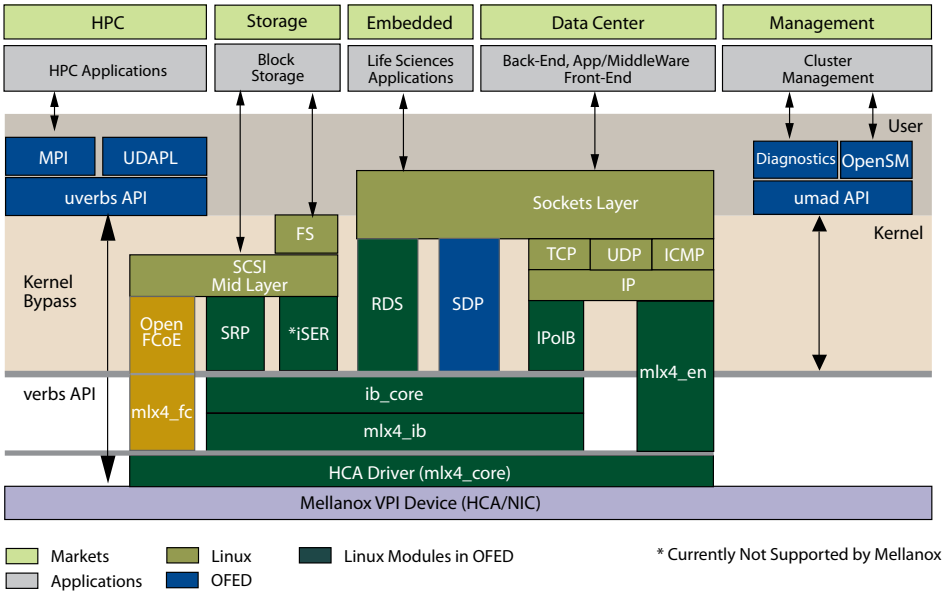
COMPONENTS

- Drivers for InfiniBand, RoCE, L2 NIC, FCoE
- Access Layers and common verbs interface
- VPI (Virtual Protocol Interconnect)
- OSU MVAPICH and Open MPI
- IP-over-IB
- SDP
- SRP Initiator
- iSER Initiator*
- uDAPL
- RDS
- Subnet Manager (OpenSM)
  - Installation, Administration and Diagnostics
- Tools
  - Performance test suites

*Currently not supported by Mellanox

DEVICE SUPPORT

- InfiniHost® Adapter Silicon
- InifiniHost® III Ex /Lx Adapter Silicon
- ConnectX® Adapter Silicon
- ConnectX® -2 Adapter Silicon
- Memory & Memory-free Adapter Cards
- InfiniScale® Switch Silicon
- InifiniScale® III Switch Silicon
- InfiniScale IV® Switch Silicon
- MTS3600 InfiniBand Switch



Note: SRP Target is included in the OFED package but is not shown in this diagram

Platform Supplier Products and Support	
RedHat	RHEL 4 U8, RHEL 5.3, 5.4, 5.5
CentOS	Versions 5.3, 5.4, 5.5
Novell	SLES 10 sp3, SLES11
Intel	x86, IA64, EM64T, EM64T x86_64, PCI-X and PCIe platforms
AMD	AMD64 and AMD64-Ex Opteron, PCI-X and PCIe platforms
IBM	PPC64

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

OPENFABRICS ALLIANCE