Mellanox HPC-X™ Software Toolkit Release Notes

Rev 1.9

NOTE:
THIS HARDWARE, SOFTWARE OR TEST SUITE PRODUCT ("PRODUCT(S)") AND ITS RELATED DOCUMENTATION ARE PROVIDED BY MELLANOX TECHNOLOGIES "AS-IS" WITH ALL FAULTS OF ANY KIND AND SOLELY FOR THE PURPOSE OF AIDING THE CUSTOMER IN TESTING APPLICATIONS THAT USE THE PRODUCTS IN DESIGNATED SOLUTIONS. THE CUSTOMER'S MANUFACTURING TEST ENVIRONMENT HAS NOT MET THE STANDARDS SET BY MELLANOX TECHNOLOGIES TO FULLY QUALIFY THE PRODUCT(S) AND/OR THE SYSTEM USING IT. THEREFORE, MELLANOX TECHNOLOGIES CANNOT AND DOES NOT GUARANTEE OR WARRANT THAT THE PRODUCTS WILL OPERATE WITH THE HIGHEST QUALITY. ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT ARE DISCLAIMED. IN NO EVENT SHALL MELLANOX BE LIABLE TO CUSTOMER OR ANY THIRD PARTIES FOR ANY DIRECT, INDIRECT, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES OF ANY KIND (INCLUDING, BUT NOT LIMITED TO, PAYMENT FOR PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY FROM THE USE OF THE PRODUCT(S) AND RELATED DOCUMENTATION EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Mellanox Technologies
350 Oakmead Parkway Suite 100
Sunnyvale, CA 94085
U.S.A.
www.mellanox.com
Tel: (408) 970-3400
Fax: (408) 970-3403

Table of Contents

List Of Tables

Release Update History

Table 1 - Release Update History

Release	Date	Description
Rev 1.9	July 17, 2017	Added a UCX known issue (the last one in the Known Issues table), please see Section 3, “Known Issues”, on page 9
	July 03, 2017	Initial version of this HPC-X version.

1 Overview

These are the release notes for the Mellanox HPC-X™ Rev 1.9. The Mellanox HPC-X™ Software Toolkit is a comprehensive software package that includes Open MPI, OpenSHMEM, PGAS, UPC, MXM, UCX, FCA tool suite for high performance computing environments. HPC-X provides enhancements to significantly increase the scalability and performance of message communications in the network. HPC-X™ enables you to rapidly deploy and deliver maximum application performance without the complexity and costs of licensed third-party tools and libraries.

1.1 HPC-X™ Requirements

The platform and requirements for HPC-X are detailed in the following table:

Platform	Drivers and HCAs
OFED / MLNX_OFED	• OFED 1.5.3 • MLNX_OFED 1.5.3-x.x.x, 3.3-x.x.x
HCAs	• ConnectX®-5 / ConnectX®-5 Ex • ConnectX®-4 / ConnectX®-4 Lx • ConnectX®-3 / ConnectX®-3 Pro • ConnectX®-2 • Connect-IB®

1.2 Important Notes

When HPC-X is launched in an environment without resource manager (slurm, pbs, ...) installed, or from a compute node, it will use Open MPI default rsh/ssh based launcher which does not propagate the library path to the compute nodes.

In such case, pass the `LD_LIBRARY_PATH` variable as following:

```
% mpirun -x LD_LIBRARY_PATH -np 2 $HPCX_MPI_TESTS_DIR/examples/hello_c
```

2 Changes and New Features in This Release

HPC-X™ Rev 1.9 provides the following changes and new features:

Table 2 - Changes and New Features

Category	Description
HPC-X Content	Updated the following communications libraries and acceleration packages versions: • OpenMPI version 2.1.2a1 • SHArP version 1.3.1 • HCOLL version 3.8.1652 • MXM version 3.6.3103 • UCX version 1.2.2947
UCX	Point-to-point communication API, with tag matching, remote memory access, and atomic operations. This can be used to implement MPI, PGAS, and Big Data libraries and applications- IB transport
	A cleaner API with lower software overhead which provides better performance especially for small messages.
	Support for multitude of InifiniBand transports and Mellanox offloads to optimize data transfer performance: • RDMA • DC • Out-of-order • HW tag matching offload • Registration cache • ODP
	Shared memory communications for optimal intra-node data transfer: • SysV • posix • knem • cma, • xpmem
MXM	Enabled Adaptive Routing for all the transport layers (UD/RC/DC).
	Memory registration optimization.
SHARP	Improved the Out-of-the-box performance of SHARP.
Shared memory	Improved the intranode performance of allreduce and barrier.
Configuration	Changed many default parameter setting in order to achieve best out-of-the-box experience for several applications including - CP2K, miniDFT, VASP, DL-POLY, Amber, Fluent, GAMES-UK, and LS-DYNA.
FCA	As of HPC-X v1.9, FCA v2.5 is no longer included in the HPC-X package
	Improved AlltoAllv algorithm.
	Improved large data allreduce.
	Improved UCX BCOL.

Table 2 - Changes and New Features

Category	Description
OS architecture	Added support for ARM architecture.

3 Known Issues

The following is a list of general limitations and known issues of the various components of this HPC-X release.

Table 3 - Known Issues (Sheet 1 of 2)

Internal Ref.	Issue
-	**Description:** MXM over Ethernet does not function for MTUs which are higher than 1024B when using firmware version 2.11.0500
	Workaround: N/A
	Keywords: MXM over Ethernet
-	**Description:** While running, MXM may show excessive log message.
	Workaround: To minimize the volume of log messages, use: `-x MXM_LOG_LEVEL=fatal` i.e. `% mpirun -x MXM_LOG_LEVEL=fatal ...`
	Keywords: Logs
-	**Description:** A mixed configuration of active ports (one InfiniBand and the other Ethernet) on a single HCA is not supported.

	Workaround: In such case, specify the port you would like to use with:
	`"-x MXM_RDMA_PORTS"`
	or
	`"-x MXM_IB_PORTS"`

	Keywords: Port Configuration
-	**Description:** When stack size is set to "unlimited", some application may suffer from performance degradation.
	Workaround: Make sure that `'ulimit -s unlimited'` is not set before running MXM.
	Keywords: Performance
-	**Description:** MXM v3.4 and v3.5 require that the `max_op_vl` value in OpenSM to be set as >=3.
	Workaround: Set the MXM environment parameter `MXM_OOB_FIRST_SL` to 0 from the command line: `$mpirun -x MXM_OOB_FIRST_SL=0 …`
	Keywords: OpenSM Configuration
-	**Description:** `MXM_IB_USE_GRH` must be set to `"yes"` when one of the following is used: 1. Socket Direct 2. Multi-Host 3. SR-IOV
	Workaround: N/A
	Keywords: EMXM parameters

Table 3 - Known Issues (Sheet 2 of 2)

Internal Ref.	Issue
-	**Description:** Currently, the UPC Barrier does not utilize FCA Barrier, so while GASNET_FCA_ENABLE_BARRIER option that enables/disabled the FCA barrier does affect various UPC collectives, it does not affect UPC Barrier.
	Workaround: N/A
	Keywords: UPC Barrier
-	**Description:** UCX may not work properly with RoCE when running on a large scale.
	Workaround: N/A
	Keywords: UCX
-	**Description:** Using UCX on ARM hosts may result in hangs due to a known issue in OMPI when running on ARM.
	Workaround: N/A
	Keywords: UCX
-	**Description:** As UCX embedded in the HPC-X is compiled with AVX support, UCX cannot be run on hosts without AVX support. In case the AVX is not available, recompile the UCX that is available in the HPC-X with the option: `--with-avx=no`
	Workaround: N/A
	Keywords: UCX

4 Bug Fixes History

Table 4 lists the bugs fixed in this release.

Table 4 - Bug Fixes History (Sheet 1 of 2)

Internal Ref.	Issue
884482	**Description:** Fixed internal HCOLL datatype mapping.
	Keywords: HCOLL, FCA
	Discovered in Release: 1.7.405
	Fixed in Release: 1.7.406
884508	**Description:** Fixed internal HCOLL datatype lower bound calculation.
	Keywords: HCOLL, FCA
	Discovered in Release: 1.7.405
	Fixed in Release: 1.7.406
884490	**Description:** Fixed allgather unpacking issues.
	Keywords: HCOLL, FCA
	Discovered in Release: 1.7.405
	Fixed in Release: 1.7.406
885009	**Description:** Fixed wrong answer in alltoallv.
	Keywords: HCOLL, FCA
	Discovered in Release: 1.7.405
	Fixed in Release: 1.7.406
882193	**Description:** Fixed mcast group leak in HCOLL.
	Keywords: HCOLL, FCA
	Discovered in Release: 1.7.405
	Fixed in Release: 1.7.406
-	**Description:** Added IN_PLACE support for alltoall, alltoallv, and allgatherv.
	Keywords: HCOLL, FCA
	Discovered in Release: 1.7.405
	Fixed in Release: 1.7.406
-	**Description:** Fixed an issue related to multi-threaded MPI_Bcast.
	Keywords: HCOLL, FCA
	Discovered in Release: 1.7.405
	Fixed in Release: 1.7.406
Salesforce: 316541	**Description:** Fixed a memory barrier issue in MPI_Barrier on Power PPC systems.
	Keywords: HCOLL, FCA
	Discovered in Release: 1.7.405
	Fixed in Release: 1.7.406

Table 4 - Bug Fixes History (Sheet 2 of 2)

Internal Ref.	Issue
Salesforce: 316547	**Description:** Fixed multi-threaded MPI_COMM_DUP and MPI_COMM_SPLIT hanging issues.
	Keywords: HCOLL, FCA
	Discovered in Release: 1.7.405
	Fixed in Release: 1.7.406
894346	**Description:** Fixed Quantum Espresso hanging issues.
	Keywords: HCOLL, FCA
	Discovered in Release: 1.7.405
	Fixed in Release: 1.7.406
898283	**Description:** Fixed an issue which caused CP2K applications to hang when HCOLL was enabled.
	Keywords: HCOLL, FCA
	Discovered in Release: 1.7.405
	Fixed in Release: 1.7.406
906155	**Description:** Fixed an issue which caused VASP applications to hang in MPI_Allreduce.
	Keywords: HCOLL, FCA
	Discovered in Release: 1.6
	Fixed in Release: 1.7.406

5 Change Log History

5.1 HPC-X Toolkit Change Log History

Table 5 - HPC-X Toolkit Change Log History

Category	Description
Rev 1.8.2	
MXM	Updated MXM version to 3.6.2098 which includes memory registration optimization.
Rev 1.8	
Cross Channel (CC)	Added Cross Channel (CC) AlltoAllv
	Added CC zcpy Ring Bcas
SHARP	Added SHARP non-blocking collectives
Shared memory POWER	Added shared memory POWER optimizations for allreduce
	Added shared memory POWER optimizations for Barrier
Mixed data types	Added support for mixed data types
Non-contiguous Bcast	Added support for non-contiguous Bcast with UMR or SGE in CC
UMR	Added UMR support in CC bcol
Unified Communication - X Framework (UCX)	A new acceleration library, integrated into the Open MPI (as a pml layer) and available as part of HPC-X. It is an open source communication library designed to achieve the highest performance for HPC applications.
HPC-X Content	Updated the following communications libraries and acceleration packages versions: • HCOLL updated to v3.7. Open MPI updated to v2.10
FCA	FCA 2.x is no longer the default FCA used in HPC-X. As of HPC-X v1.8, FCA 3.x (HCOLL) is the default FCA used and it replaces FCA v2.x.
Bug Fixes	See Section 4, “Bug Fixes History”, on page 11
Rev 1.7	
MXM	Updated MXM version to 3.6
FCA Collective	Added Cross-Channel based Allgather, Bcast, 8-byte Allreduce.
FCA	Added MPI datatype support.
	Added optimizations for PPC platforms.
	Added support for multiple SHArP leaders on a single host.
	Added support for collecting SHArP usage statistics.
	Exposed cross-channel non-blocking collectives to the MPI level.
Rev 1.6	

Table 5 - HPC-X Toolkit Change Log History

Category	Description
MXM v3.5	See Section 5.3, "MXM Change Log History", on page 16
IB-Router	Allows hosts that are located on different IB subnets to communicate with each other. This support is currently available when using the `'openib btl'` in Open MPI. **Note:** When using `'openib btl'`, RoCE and IB router are mutually exclusive. The Open MPI inside HPC-X 1.6 is not compiled with ib-router support, therefore it supports RoCE out-of-the-box.
FCA v3.5	See Section 5.2, "FCA Change Log History", on page 15
Rev 1.5	
HPC-X Content	Updated the following communications libraries and acceleration packages versions: • Open MPI updated to v1.10 • UPC update to 2.22.0 • MXM updated to v3.4.369 • FCA updated to v3.4.799
MXM v3.4.369	See Section 5.3, "MXM Change Log History", on page 16
FCA v3.4.799	See Section 5.2, "FCA Change Log History", on page 15
Rev 1.4	
FCA v3.3	See Section 5.2, "FCA Change Log History", on page 15
MXM v3.4	See Section 5.3, "MXM Change Log History", on page 16
Rev 1.3	
MLNX_OFED	Added support for OFED Inbox drivers
CPU Architecture	Added support for PPC architecture
LID Mask Control (LMC)	Added support for multiple LIDs usage when the LMC in the fabric is higher than zero. MXM will use multiple LIDs to distribute traffic across multiple links and achieve better resource utilization.
Performance	Performance improvements for all transport layers.
Adaptive Routing	Enhanced support for Adaptive Routing for the UD transport layer. For further information, please refer to the HPC-X User Manual section *"Adaptive Routing for UD Transport"*.
UD zero copy	UD zero copy support on receiver side to achieve better bandwidth utilization and reduce CPU usage.

5.2 FCA Change Log History

Table 6 - FCA Change Log History

Category	Description
Rev 3.5	
FCA Collective	Added MPI Allgatherv and MPI reduce
FCA	Added support for SHArP (including SHArP allreduce, reduce and barrier)
	Enhanced scalability for CORE-Direct based collectives
	Added support for complex data types
Rev 3.4	
General	UCX support
	Communicator caching scheme with eviction: improves jobstart and communicator creation time
Collectives	Collectives: Added Alltoallv and Alltoall small message algorithms.
Rev 3.3	
General	Ported to PowerPC
	Thread safety added
Collectives	Improved large message allreduce algorithm (Enabled by default)
	Beta version of network topology awareness (Enabled by default)
Rev 3.0	
Collectives	Offload collectives communication from MPI process onto Mellanox interconnect hardware
	Efficient collectives communication flow optimized to job and topology
MPI collectives	Significantly reduce MPI collectives runtime
MPI-3	Native support for MPI-3
Blocking and Non-blocking collectives	Support for blocking and nonblocking collectives
HCOLL	Supports hierarchical communication algorithms (HCOLL)
Collective algorithm	Supports multiple optimizations within a single collective algorithm
Performance	Increase CPU availability and efficiency for increased application performance
MPI libraries	Seamless integration with MPI libraries and job schedulers
Rev 2.5	
Multicast Group	Added MCG (Multicast Group) cleanup tool
Performance	Performance improvements
Rev 2.2	
Performance	Performance improvements

Table 6 - FCA Change Log History

Category	Description
Dynamic offloading rules	Enabled dynamic offloading rules configuration based on the data type and reduce operations
Mixed MTU	Added support for mixed MTU
Rev 2.1.1	
AMD/Interlagos CPUs	Added support for AMD/Interlagos CPUs
Rev 2.1	
Core-Direct®	Added support for Mellanox Core-Direct® technology (enables offloading collective operations to the HCA.)
Non-contiguous data layouts	Added support for non-contiguous data layouts
PGI compilers	Added support for PGI compilers

5.3 MXM Change Log History

Table 7 - MXM Change Log History

Category	Description
Rev 3.6	
General	Updated MXM version to 3.6
Rev 3.5	
Performance	Performance improvements
Rev 3.4.369	
Initialization	Job startup performance optimization
Supported Transports	DC enhancements and startup optimizations
Rev 3.4	
Supported Transports	Optimizations for the DC transport at scale
Rev 3.3	
LID Mask Control (LMC)	Added support for multiple LIDs usage when the LMC in the fabric is higher than zero. MXM will use multiple LIDs to distribute traffic across multiple links and achieve better resource utilization.
Adaptive Routing	Enhanced support for Adaptive Routing for the UD transport layer.
UD zero copy	UD zero copy support on receiver side to achieve better bandwidth utilization and reduce CPU usage.
Rev 3.2	

Table 7 - MXM Change Log History

Category	Description
Atomic Operations	Added hardware atomic operations support in the RC and DC transport layers for 32bit and 64bit operands. This feature is set to ON by default. To disable it run: `oshrun -x MXM_CIB_USE_HW_ATOMICS=n ...` **Note:** If hardware atomic operations are disabled, the software atomic is used instead.
MXM API	Added two additional functions (`mxm_ep_wireup()` and `mxm_ep_power-down`) to the MXM API to allow pre-connection establishment for MXM (rather than on-demand). For further information, please refer to the HPC-X User Manual section *"MXM Performance Tuning"*.
Event Interrupt	Added solicited event interrupt for the rendezvous protocol for potential performance improvement. For further information, please refer to the HPC-X User Manual section *"MXM Performance Tuning"*.
Performance	Performance improvements for the DC transport layer.
Adaptive Routing	Added Adaptive Routing for the UD transport layer. For further information, please refer to the HPC-X User Manual section *"Adaptive Routing for UD Transport"*.
Rev 3.0	
Service Level	Service Level support (at Alpha level)
Adaptive Routing	Adaptive Routing support in UD transport layers
Supported Transports	Dynamically Connected Transport (DC) (at GA level)
Performance	Performance optimizations
Rev 2.1	
Supported Transports	Dynamically Connected Transport (DC) (at Beta level)
	RC is currently fully supported
	KNEM support for Intra-node communication
Performance	Performance optimizations
Rev 2.0	
Reliable Connected	Added Reliable Connection (RC) support (at beta level)
MXM Binding	MXM process can be pinned to a specific HCA port. MXM supports the following binding policies: • static - user can specify process-to-port map • cpu affinity based - HCA port will be bound automatically based on process affinity
On-demand connection establishment	Added on-demand connection establishment between the processes
Performance	Performance tuning improvements

Table 7 - MXM Change Log History

Category	Description
Rev 1.5	
MXM over Ethernet	Added Ethernet support
Multi-Rail	Added Multi-Rail support

5.4 HPC-X™ Open MPI/OpenSHMEM Change Log History

Table 8 - HPC-X™ Open MPI/OpenSHMEM Change Log History

Category	Description
Rev 1.8.2	
Acceleration Packages	Added support for new MXM, FCA, HCOLL versions
Job start optimization	Added job start optimization
Performance	Performance improvements
Rev 2.2	
Performance	Added Sandy Bridge performance optimizations.
memheap	Allocated memheap using contiguous memory provided by the HCA.
ptmalloc allocator	Replaced the buddy memheap by the ptmalloc allocator.
multiple pSync arrays	Added the option of using multiple pSync arrays instead of barrier synchronization between collective routines (fcollect, reduction routines)
spml yoda	Optimized small size puts
Performance	Performance optimization
Memory footprint optimizations	Added memory footprint optimizations

5.5 HPC-X™ UPC Change Log History

Table 9 - HPC-X™ UPC Change Log History

Category	Description
Rev 2.18.0	
Acceleration Packages	Added support for new MXM, FCA, HCOLL versions
PMI2 support	Added job start PMI2 support
Rev 2.2	
FCA library	Linking with FCA library instead of using dlopen at runtime.
MPI	Fixed an issue using some of MPIs as job spawner (e.g. MPICH2) Use MPI_-BYTE rather than MPI_CHAR, and use MPI_IN_PLACE.