Release Notes

Mellanox IB Gold Distribution

Rev 1.8.1

Mellanox IB Gold Distribution Release Notes

Document Number:

Mellanox Technologies, Inc.
2900 Stender Way
Santa Clara, CA 95054
U.S.A.
www.Mellanox.com

Tel: (408) 970-3400
Fax: (408) 970-3403

Mellanox Technologies Ltd
PO Box 586 Hermon Building
Yokneam 20692
Israel

Tel: +972-4-909-7200
Fax: +972-4-959-3245

1 Overview

These are the release notes of Mellanox IB Gold Distribution (IBGD), Rev 1.8.1. The IBGD is a SW package of several modules that together comprise a complete InfiniBand solution for High Performance Computing (HPC), Enterprise Data Center (EDC), and storage.

> Note: If you plan to upgrade IB Gold Distribution on your cluster, please upgrade *all* its nodes to this new version.

This document is organized into the following sections:

1.1 Mellanox IB Gold Distribution Rev 1.8.1 Contents

Table 1 - IB Gold Contents

Component Title	Component Name / Details
IBGD Firmware	• FW-23108 InfiniHost Firmware revision 3.3.5 • FW-25208 InfiniHost III Ex (InfiniHost Mode) Firmware revision 4.7.400 • FW-25218 InfiniHost III Ex (MemFree Mode) Firmware revision 5.1.0 • FW-25204 InfiniHost III Lx Firmware revision 1.0.700 • FW-43132 InfiniScale Firmware revision 5.3.0 • FW-47396 InfiniScale III Firmware revision 0.8.0
Open Subnet Manager	• OpenSM 1.8.0-1
InfiniBand Administration Management	• IBADM 1.8.1 - this includes: - Mellanox Software Tools: MST 4.3.1
OSU MPI	• MVAPICH 0.9.5_mlx1.0.2 (Mellanox's Edition) - this includes: - Cluster Benchmarks: Presta 1.2, Pallas 2.2.1, B/L OSU 1.0
IB	• ib-1.8.1 - this includes: - VAPI (4.1.1) and Access Layer (AL 1.8.1) - IPoIB 1.8.1 - Socket Direct Protocol: SDP 1.8.1 - SCSI RDMA Protocol Initiator: SRP Initiator (Host) 1.8.1 - Connection Manager: CM 1.8.1 - DAPL 1.8.0: * User Direct Access Programming Layer: uDAPL 1.8.0 * Kernel-level Direct Access Programming Layer: kDAPL 1.8.0 * Direct Access Transport: DAT 1.8.0
Parallel remote shell program	• pdsh-2.3-1
IBGranite Fabric Verification Suite	• ibgfvs-1.0.0beta
Additional Packages	• Placed under IBGD-1.8.1/SOURCES but not installed: - Cable Testing & SerDes Configuration Tool (Eye Opening 0.2.0) - Mellanox Embedded Management Tools mlxpart-ppc8xx-0.2.0-rc4

1.2 Supported Platforms and Operating Systems

The following table lists all supported platforms and operating systems by the tools and modules included in this IBGD package.

Table 2 - Supported Platforms and Operating Systems

Architecture	Operating System	Kernel
X86	RedHat Enterprise Linux 4.0 up2	2.6.9-22.ELsmp
	SuSE SLES 9.0	2.6.5-7.111.5-smp
	SuSE 9.3 Pro	2.6.11.4-20a-smp
AMD64 (Opteron)	RedHat Enterprise Linux AS 4.0 up2	2.6.9-22.ELsmp
	SuSE 9.3 Pro	2.6.11.4-20a-smp
	SuSE SLES 9.0	2.6.5-7.111.19-smp
Intel EM64T [2]	RedHat Enterprise Linux AS 4.0 up2	2.6.9-22.EL
	SuSE 9.3 Pro	2.6.11.4-20a-smp
	SuSE SLES 9.0 RC5	2.6.5-7.111.19-smp

1.3 Supported InfiniHost Firmware Versions

Table 3 - Supported Firmware

Firmware Name	Supported Versions	Details / Notes
fw-23108	3.3.2 and later	MT23108 InfiniHost firmware version
fw-25208	4.6.2 and later	MT25208 InfiniHost III Ex - **InfiniHost Mode** Note: IB port operation at DDR is supported from version 4.7.0 and on
fw-25218	5.0.1 and later	MT25208 InfiniHost III Ex - **MemFree Mode** Note: IB port operation at DDR is supported from version 5.1.0 and on
fw-25204	1.0.1 and later	MT25204 InfiniHost III Lx Note: Supports IB port operation at DDR

1.4 Supported HCA Adapter Cards

Table 4 - Supported HCA Adapter Cards

HCA Card OPN	Code Name	Description
MHX-CEXXX / MHX-CEXXX-T[1] (previously MTPB23108)	Cougar	InfiniHost PCI-X HCA Adapter Card
MHXL-CFXXX / MHXL-CFXXX-T[1] (previously MTLP23108)	Cougar Cub	Low Profile InfiniHost PCI-X HCA Adapter Card
MHEL-CFXXX / MHEL-CFXXX-T[1] (previously MTLP25208) MHEL-CFXXX-TC	Lion Cub SDR Lion Cub SDR Rev C	InfiniHost III Ex HCA Adapter Card
MHGA28-1T / MHGA28-2T	Lion Cub 128/256 DDR	InfiniHost III Ex HCA Adapter Card with 128/256MB local memory. IB ports support double data rate (DDR) operation
MHEA28-XS / MHEA28-XT [2]	Lion Mini SDR	MemFree InfiniHost III Ex HCA Adapter Card
MHGA28-XT	Lion Mini DDR	InfiniHost III Ex MemFree HCA Card
MHES14-XT	Tiger SDR	InfiniHost III Lx PCI Express x4 HCA Card
MHES18-XS / MHES18-XT [2]	Cheetah SDR	InfiniHost III Lx PCI Express x8 HCA Card
MHGS18-XS / MHGS18-XT [2]	Cheetah DDR	InfiniHost III Lx PCI Express x8 HCA Card

1. XXX reflects the size of on-board memory (in MB): 128, 256, or 512. '-T' in the OPN indicates a Tall Bracket card; *no* '-T' indicates a Short Bracket card.
2. XT stands for Tall Bracket; XS stands for Short Bracket.

1.5 Tested Switch Platforms

Table 5 provides a partial list of the switch platforms on which IBGD Rev 1.8.1 was tested.

Table 5 - Tested Switch Platforms - Partial list

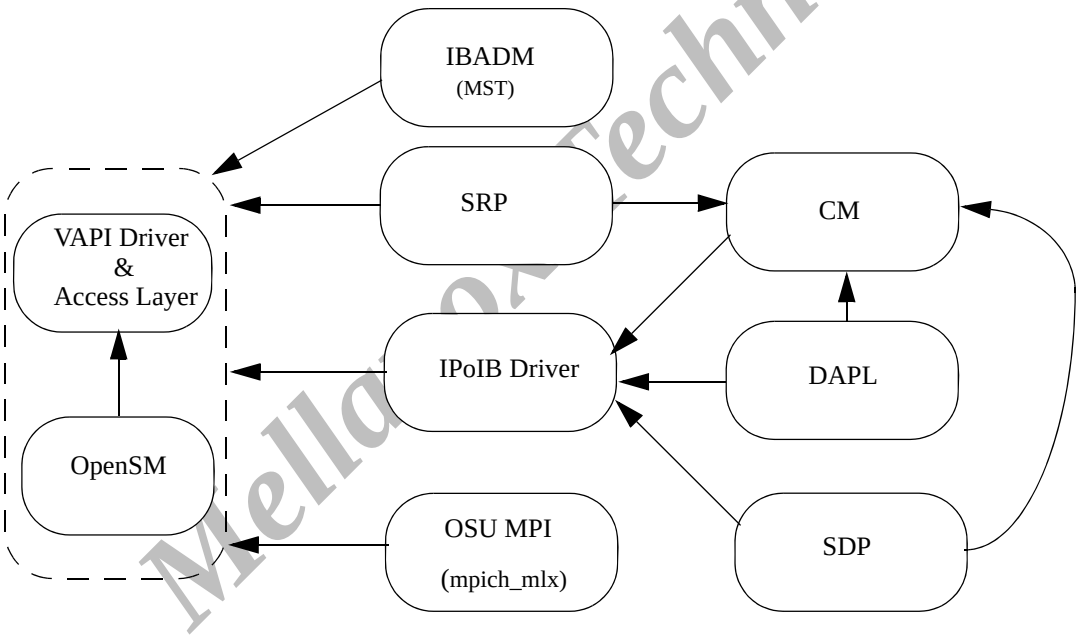
Switch Platform OPN	Make	Description
MTS2400	Mellanox Technologies	24 4X IB SDR port switch platform
MTS9600	Mellanox Technologies	96 4X IB SDR port switch platform
MTS14400	Mellanox Technologies	144 4X IB SDR port switch platform
F-X430060, F-X430062, F-X430061, F-X430063	Flextronics	24 4X port switch platforms (IB @SDR)
F-X430044, F-X430058, F-X430046, F-X430047	Flextronics	24 4X port switch platforms (IB @DDR)

1.6 Inter-Package Dependencies

During the IBGD package installation, some dependencies between the package components must be observed. The following is a summary list of these dependencies. Figure 1,“IBGD Package Inter-Dependencies” summarizes the dependencies in the form of a diagram.

- The IPoIB driver depends on the installation of the IB Gold Distribution stack with OpenSM running.
- IBADM depends on the installation of the IB Gold Distribution stack with OpenSM running.
- CM depends on the installation of the IB Gold Distribution stack with OpenSM running. It also depends on the IPoIB module.
- OSU MPI depends on the installation of the IB Gold Distribution stack with OpenSM running. The MPI module also requires an established network interface (either InfiniBand IPoIB or Ethernet).
- kDAPL depends on the installation of the IB Gold Distribution stack. It also depends on the CM and IPoIB modules.
- uDAPL depends on the installation of the IB Gold Distribution stack with OpenSM running. It also depends on the IPoIB module.
- SDP depends on the installation of the IB Gold Distribution stack with OpenSM running. The SDP module also depends on the IPoIB module.
- SRP Initiator depends on the installation of the IB Gold Distribution stack with OpenSM running.

Figure 1: IBGD Package Inter-Dependencies



2 Main Changes from Previous Release (1.8.0)

2.1 General Package Changes

1. All patches of IBGD 1.8.0 were incorporated in this release.
2. Reduced list of supported platforms and operating systems. See Table 2 on page 5.
3. Initrd support: IB modules can now be inserted into an initrd image. Changes were made to:
   - openibd service: openibd start will not fail if modules are already loaded and IB devices will be created
   - modprobe.conf / modules.conf: options for the ib_client_query module are rewritten in the same line
   - /etc/modprobe.conf is used instead of /etc/modprobe-openib.conf, and /etc/modules.conf is used instead of /etc/modules-openib.conf
4. The OpenSM SLDD daemon added in order to resolve hand-over issue (see “OpenSM Changes” on page 9). Changes:
   - Added the sldd.sh script under /usr/bin directory
   - Updated /etc/opensm.conf file with new parameters to sldd.sh and opensm
   - Updated the IBGD install.sh script to obtain the list of IP addresses of all opensm servers in the IB subnet

2.2 SRP Changes

1. Implemented multiple SCSI hosts to enable all physical paths. This allows two different ports to connect to the same fabric and see the same targets along different paths.
2. Re-implemented the target_bindings parameter to become of the format:
   `<tgt_service_name>.<tgt_service_id>.<tgt_port_dlid>.<tgt#>.<port#>.<hca#>:...`
3. Removed the target_bindings module parameter. The configuration file /etc/mlx_srp_bindings.conf is used instead to pass the binding string to the SRP driver.
4. Removed the partial fail-over feature between ports/HCA(s) in SRP to allow a generic multi paths/power-paths driver(s) to work.
5. Added task management (Abort, LUN reset, Bus reset, Host reset)
6. SRP now avoids scsi mid-layer taking a device offline in case of a cable pull, power cycling targets, taking targets offline/out-of-fabric for repair. Within 60 seconds of such events, SRP will return DID_IMM_RETRY to cause the scsi mid-layer to retry. After 60 seconds, if SRP cannot re-establish connections with offline devices/ targets, SRP will return SCSI_MLQUEUE_HOST_BUSY in the queue command interface.
7. Implemented auto-detect of old targets or new targets joining the fabric. SRP establishes connections with these targets, and for newly added targets it creates new a scsi_host, and asks the scsi mid-layer to scan the new hot-added targets.
8. Support for OpenSM Master hand-over and re-assigned LIDs. SRP now detects the re-assignment of LIDs and re-establishes connections to targets if required.
9. Removed srp_persistent_bind.sh and remove_srp_persistent_bind.sh from modprobe.conf /modules.conf files. Please contact your Mellanox assigned FAE to understand how SRP persistent binding can be achieved.

2.3 MPI Changes

1. MPI now uses shared library by default
2. Optimized AlltoAll flow

2.4 OpenSM Changes

1. Added synchronization of guid2lid files across the cluster to preserve GUID to LID mapping upon an OpenSM Master hand-over. To use this feature you need to provide the list of IP addresses during IBGD installation. See the IBGD Installation Guide (filename: IBGD_Installation_Guide.txt) for details.
   Moreover, using the new feature requires ssh or rsh without a password between all opensm servers. The default in /etc/opensm.conf is ssh.

2.5 VAPI Changes

1. Added the SRQ Limit event for the InfiniHost and Infinihost III Ex HCA devices in memory mode (i.e., with locally attached memory). Activation is enabled only if the firmware used supports this SRQ Limit event. (For InfiniHost, firmware version 3.4.000 or later; for InfiniHost III Ex, firmware version 4.7.400 or later.)
2. Added an option to control the values of the max read request size and the read byte count for PCI Express HCAs (InfiniHost, InfiniHost III Ex/Lx). To use this option, the following lines should be added to /etc/modules.conf:
   ```
   options mod_rhh pci_cap_read_byte_count=<value>
   options mod_rhh pci_cap_max_read_request_size=<value>
   ```
   Valid values for pci_cap_max_read_request_size:
   - 0: 128 bytes max read request size
   - 1: 256 bytes max read request size
   - 2: 512 bytes max read request size
   - 3: 1024 bytes max read request size
   - 4: 2048 bytes max read request size
   - 5: 4096 bytes max read request size (the default)

   Valid values for pci_cap_read_byte_count:
   - 0: 512 bytes
   - 1: 1024 bytes
   - 2: 2048 bytes
   - 3: 4096 bytes (the default)
3. Replaced parameter names to the mlxsys module that controls the amount of physical memory that can be locked:
   - max_lock_pages maximum size of physical memory that can be locked
   - cur_lock_pages current size of locked physical memory

2.6 IBADM Changes

1. The Flint tool (part of the MST package) now auto-detects the size and number of SPI Flash devices attached to the MT25204 InfiniHost III Lx HCA device.

3 Bug Fixes

The following table lists the fixed bugs in this release.

Table 6 - Fixed Bugs

	Title	Details
1.	MPI: Wrong MPI exit status	MPI exit status was reported zero while PALLAS tests were failing. Fixed.
2.	MPI: Incorrect handling of ‘-c’ flag in mpicc/mpif77/mpif90 (Argon bug #10546)	mpicc did not properly handle non-existing or unconnected files when using the ‘-c’ option. Fixed.
3.	MPI: presta com -o 10000 (2 nodes 4 jobs) exits with segmentation fault	This was caused by a missing check for exit status of MPID_SMP_Eagerb_save_short. Fixed.
4.	MPI: Error in MPI_Init flow	The MPI_Init flow tried to use shared memory channel before it was initialized. Fixed.
5.	MPI: mpif90 wrapper filters out the.f95 files	Fixed.
6.	VAPI: iounmap causes a kernel crash	Some kernels crash when iounmap is called for a pointer obtained by ioremap_nocache. Fixed by replacing ioremap_nocache with ioremap.
7.	VAPI: get_user_pages may (rarely) return the bss page for unmapped pages	For an unmapped page, get_user_pages gave the bss page. If later on the HCA used this page and wrote to it, other process(es) crashed. Fixed by changing the call to get_user_pages such that the write parameter depends on the requested permissions for the memory region.
8.	VAPI: Kernel memory allocation may cause a ‘might_sleep’ warning by kernel	Fixed this by changing the kernel memory allocation rule that selects between GFP_ATOMIC or GFP_KERNEL to be identical to the kernel rule in might_sleep() checks.
9.	IPoIB: Connectivity is lost when a remote gid or qpn changes by a ‘recreate’ of the path record.	Fixed.
10.	SDP: Write to Socket in Connection Accepted State returns ECONNRESET	Fixed.
11.	SDP: A partial message is returned though MSG_WAITALL flag is set	Fixed.
12.	SDP: Running TTCP with SDP does not work	Fixed.
13.	OpenSM: SRP Targets started after the SRP Initiator are not discovered	The SRP Initiator was not notified about port state (up/down). Fixed.
14.	IBADM: demonize not found	Fixed.
15.	SRP: SCSI offline device issue	If a SCSI device went offline after its cable was disconnected, it automatically went back online upon cable reconnection and the operation was completed. Fixed.

4 Known Issues

The following table provides major limitations and known issues of the various components of IBGD. Please refer to the component-specific release notes for more details.

Table 7 - Limitations and Known Bugs

	Title of Problem	Details
1.	OSU MPI limitation	A Fortran compiler (for example, gcc-g77) is required for build
2.	SRP limitation	Only the initiator part of the protocol is provided in this package. Please contact Mellanox for the target.
3.	SDP limitation	The SDP ULP cannot run on kernel 2.6.12
4.	AMD 8131 chipset invalid configuration for PCI-X may cause system failure	For details on the invalid configuration, see under http://www.amd.com/us-en/Processors/TechnicalResources/0,,30_182_739_9004,00.html Errata #56 and #58 in the document “AMD-8131™ HyperTransport™ PCI-X Tunnel Revision Guide” IB Gold provides two variables to enable or disable workarounds for these errata. These are: FIX_AMD_8131_ERR56 and FIX_AMD_8131_ERR58 in the configuration file /etc/infiniband/openib.conf. By default, both variables are set to YES to enable the workarounds. A change to either variable requires rebooting the system.
5.	CM limitation	CM support for APM is not fully functional
6.	CM limitation	User-level CM is not supported