Release Notes

MT25208 InfiniHost™ III Ex Firmware

InfiniHost-compatible Mode: PCI DevID 25208 (Decimal)

fw-25208 Rev 4.8.200

MT25208 InfiniHost(TM) III Ex InfiniHost-mode Firmware fw-25208 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 for the MT25208 InfiniHost™ III Ex firmware, fw-25208 Rev 4.8.200. It is appropriate for the MT25208 device functioning as an MT23108 InfiniHost-compatible device. For the most updated list of HCA Adapter Cards supported by this firmware, visit the firmware download pages via http://www.mellanox.com.

Note: After burning new firmware to an HCA board, reboot the machine so that the new firmware can take effect.

The document consists of the following sections:

2 Major New Features

Note: The following changes are with respect to Rev 4.7.600 of the firmware.

- Support for programming the PCI Express maximum payload size via the firmware INI file. See the parameter **`emt_max_payload_size_supported`** in the file `fw-25208-defaults.ref`
- Support for **`log_ge_per_port`** parameter which is used for assigning gather engines to queues. See Section 2.4, "HCA QoS Control," in the *InfiniHost™ Programmers Reference Manual, Rev. 1.19, Document No. 2111PM*
- Added the INI parameter **`watchdog_enable`** to enable the IB flow control watchdog timer. See the file `fw-25208-defaults.ref`
- Support for 400 MHz DDR SDRAM DIMMs
- Changes to parity error detection mechanism as follows:
  - Added the INI parameter **`pqcp`** to reduce harmless QP Context array parity errors
  - Added a FW mechanism for correcting TPT (Translation & Protection Table) parity errors

3 Bug Fixes

The following table describes known issues that were fixed in Rev 4.8.200 with respect to Rev 4.8.000 of the firmware.

Table 1 - Rev 4.8.200 Bug Fixes

	Issue	Description	Discovered in	Fixed in
1.	Small message rate degradation	Fixed	4.8.000	4.8.200

The following table describes known issues that were fixed in Rev 4.8.000 with respect to Rev 4.7.600 of the firmware.

Table 2 - Rev 4.8.000 Bug Fixes

	Issue	Description	Discovered in	Fixed in
1.	QP context corruption	• HCA might stall under stress (ID:36209) • Upon SuspendQP in conjunction with QP Error (ID:38876)	4.7.600	4.8.000
2.	Corrupted Read/Atomic response	Possible corruption of Read or Atomic responses (ID:36353, 37014, 37026, 39022)	4.7.600	4.8.000
3.	Bringing IB port up on DDR device	DDR IB link is not going up after being put in port DISABLE state. (ID:36586)	4.7.600	4.8.000
4.	LinkSpeedEnable	LinkSpeedEnable is not applied correctly. (ID: 36592)	4.7.600	4.8.000
5.	relaxed_ordering_en bit is not PCI Express compliant	PCI Express bit relaxed_ordering_en is R/W instead of Read Only	4.7.600	4.8.000
6.	Vendor Specific MADs	Wrong Vendor Key (V_KEY) authentication and mismatch behavior. (ID:36127) Wrong Revision ID in General Info MAD. (ID: 38025)	4.7.600	4.8.000
7.	Serdes Parameters	Serdes Parameters are not applied correctly to port 2 (ID:37782-3)	4.7.600	4.8.000
8.	RDMA_READ and ATOMIC Dead-Lock	Internal leak may occur when doing RDMA_READs and ATOMIC on Send Queues. (ID: 27761)	4.7.600	4.8.000
9.	Fatal error in large-stress condition	FW may fail (FW_INTERNAL_ERROR) upon large stress (ID: 38845,39393)	4.7.600	4.8.000
10.	QUERY_DEBUG_MSG	msg_hdr_size is wrong (ID:38428)	4.7.600	4.8.000
11.	SYSTEM_DISABLE command may get stuck if IB port runs at DDR	Fixed (ID : 36757,39325)	4.7.600	4.8.000
12.	Potential FW deadlock	FW may deadlock upon heavy stress (ID : 39393)	4.7.600	4.8.000
13.	HCA may hang upon SW reset	If the HCA has a pending In-Bound read and a SW reset occurs simultaneously, the HCA may hang (ID: 38053,39660,39661)	4.7.600	4.8.000
14.	Wrong link state returned to GetPort-Info query on DDR Auto-negotiation	While a link is in the DDR Auto-negotiation process, it may report a POLLING state for the physical link and an INITIALIZE state for the logical link while it should report POLLING and Link-Down, respectively	4.7.600	4.8.000

4 Invariant Sector (IS) Changes / Fixes

The following table describes Invariant Sector changes and/or fixes with respect to the last firmware release.

Table 3 - IS Changes / Fixes

	Issue	Description	Discovered in	Fixed in
1.	PCI Express x2	A PCI Express x2 link width is not allowed anymore	4.7.600	4.8.000

5 Known Issues

The following table describes known issues in this firmware release and possible workarounds.

Table 4 - Known Issues

Index	Issue	Description	Current Implemented Workaround in FW	Possible Workaround	Patch Release (fix)	Scheduled Release (fix)
1.	MSIx vectors	Writing to MSIx vectors (Address/Data/Mask) does not take immediate effect. There may be MSIx messages that leave the device according to the old vector	NA	Commit a PCI configuration cycle after the MSIx modification	NA	NA
2.	QUERY_DDR	Query does not return JEDEC vendor ID yet. Scope of status is limited to active / not active	NA	NA	NA	NA
3.	RTR2RTS_QPEE, SQD2RTS_QPEE: changing optional fields rra_max and ra_buf_index is not supported.	The optional fields rra_max and ra_buf_index are not supported in the RTR2RTS_QPEE and SQD2TRS_QPEE commands	Change requests for these fields will not take effect, and no error indication is provided	Mask these optional fields	NA	NA
4.	PCI 2.3 control and status for interrupts	InfiniHost™ III Ex does not support PCI2.3 control and status bits for interrupts	NA	NA	NA	NA
5.	Change of memory bars on a disabled system	Changing memory bars size / addresses between SYS_DIS and SYS_EN may cause the InfiniHost™ III Ex to hang (ID: 24206)	NA	NA	NA	NA
6.	BAR resizing on an enabled system	Changing bar sizes when a system is enabled may cause the InfiniHost™ III Ex to hang (ID: 24208)	NA	NA	NA	NA
7.	SW reset via configuration cycles	SW reset via config cycles may create double PCI Express completions for the configuration transaction	NA	If InfiniHost ™ III Ex boots in memory controller mode, perform power cycle / hot reset after restoring the Flash	NA	NA
8.	SW reset is performed during a configuration transaction	If SW reset is performed while a configuration transaction is outstanding, it may create double PCI- Express completions for the configuration transaction	NA	Do not perform SW reset during configuration cycles	NA	NA

6 Creating a Device Configuration (.ini) File

Mellanox firmware burning tools enable setting and/or changing configuration variables by the use of an optional configuration (.ini) file. This is needed in case the default values of some variables do not suit a user's specific system requirements. This section describes how to create this configuration file.

To begin with, the .ini file is a text file is composed of one or several configuration sections (see Section 6.1 for the format and/or an example). It is recommended to include, under the appropriate sections, only those variables that need to be changed.

A firmware release includes a reference file called fw-25208-defaults.ref. This file contains the list of all variables which can be configured by a configuration (.ini) file. For each variable the reference file includes a short explanation, the [<section>] it should be under, the range of possible values, and a line with the default setting of the variable which is assumed by the firmware release.

To create the .ini file, simply copy the lines with the variables you wish to set, paste them under their appropriate [<section>] headings, and change the setting values as desired.

6.1 Configuration (.ini) File Format

The .ini file is composed of one or more sections with variable settings. Each section in the file starts with its name between square brackets, e.g. [ADAPTER], [HCA], [IB], etc. The section name is followed by one or more lines of configuration settings and comments, as in the .ini file example shown below. Note that comment lines start with a semicolon.

Excerpt from fw-25208-defaults.ref:

```
;;;;; VPD support can be Disabled/Enabled

;;;;; Under [ADAPTER] section

;;;;; Boolean parameter. Possible values: true, false .

vpd_enable = true
```

Example of a .ini file:

```
;Begin of .ini file

[ADAPTER]

vpd_enable = false

;This is a comment line

;End of .ini file
```

7 History

Table 5 - History of Fixed Bugs (Sheet 1 of 4)

Issue	Description	Discovered in	Fixed in
Bind Memory Window checks	The Bind Memory Window command does not check that all index bits of the window to be bound are identical to those of an old window (ID: 35035)	4.7.500	4.7.600
Wrong M_Key check	Given MKeyProtectBits<2, if a SubnGet(NodeInfo) with a wrong M_Key is received while the M_Key Lease Period counter is already active (due to a previous M_Key violation), the counter gets stopped instead of continuing with the original count (ID: 33388)	4.7.500	4.7.600
Wrong Client ReRegister bit	Client ReRegister bit may be wrong (ID:33958,35356)	4.7.500	4.7.600
APM with changed SL may result in corruption	A schedule queue corruption may occur upon an APM with a changed SL (ID: 32672)	4.7.500	4.7.600
Wrong behavior upon UC send with first message packet dropped	If the first packet of an Unreliable Connection gets discarded, subsequent UC behavior may be wrong (ID: 34689)	4.7.500	4.7.600
High latency from SDRAM CKE to FastSelfRefresh GNT	Fixed. (ID: 35270)	4.7.500	4.7.600
Wrong APM behavior	Packet from wrong port may be dropped even if MigReq is not expected. Also, a bad path packet may be allowed in MIGRATED state. (ID: 35179)	4.7.500	4.7.600
SERR# (System Error)	The Signaled System Error Status bit is not set after an Error message is sent (ID: 35458, 35703). The enable_serr Control bit has no effect. (ID: 23409, 26881, 35693)	4.7.500	4.7.600
CQ overrun	A CQ overrun may occur upon CQ stress (ID: 34370)	4.7.500	4.7.600
QP Flush during an SQ Drain may result in corruption	Flushing the QP during an SQ Drain may cause a schedule queue corruption (ID: 34879)	4.7.500	4.7.600
Send Work queue corruption	HW may overrun Send WQ in case of multiple RDMA READs or in case of entering Limited State (ID: 35649)	4.7.500	4.7.600
Wrong MAD status	MAD status may be “Invalid combination of Method and Attribute” instead of “Invalid Attribute” (ID: 35226)	4.7.500	4.7.600
ACK delay when sender is back-pressured	This causes a transport timeout to the remote peer	4.7.0	4.7.400
RNR timer is always set to 0x1F	RNR timer for a Sender retry is always configured to 0x1F regardless of the QP set value (ID:31959)	4.7.0	4.7.400
Unreliable-QP context corruption	(ID:31927)	4.7.0	4.7.400
SDRAM INIT operation when exiting Self Refresh not implemented	Now it is implemented	4.7.0	4.7.400
ECC errors handling	ECC errors may not be reported at all, may cause a wrong SERR# assertion, or may be reported with the wrong ColumnAddress. (ID:31614)	4.7.0	4.7.400
Link up in 1X rather than 4X	Signal Detect must happen before RX power-up; otherwise, the port comes up in 1X mode instead of 4X (ID:32401)	4.7.0	4.7.400

Table 5 - History of Fixed Bugs (Sheet 2 of 4)

Issue	Description	Discovered in	Fixed in
MADs:PortInfo Get()	When querying for information about an InfiniHost™ III Ex IB port via its other IB port, the wrong Local port number is returned. Instead of the number of the second port, the one which received the MAD packets, the number of the first port is being returned. (ID: 24177)	4.6.2	4.7.0
A Concurrent Bind and Deallocate for the same Memory Window will prevent closing the Memory Region of this Window	Bind and Deallocate modify the same 'unprotected' variable of the Memory Region. If both operations are attempted simultaneously for the same Memory Window, the variable does not get updated correctly. This prevents closing the Memory Region as the corrupted variable value may indicate that a Memory Window is still bound to it.	4.6.2	4.7.0
Requester ScatterList corruption upon CQ error	A CQ error can cause corruption in the Requester ScatterList Database. As a result QPs may move to error, and the device may stop sending packets (ID: 30670)	4.6.2	4.7.0
IB Tx phase detector should be opened when link state is Config Debounce	(ID: 24332)	4.6.2	4.7.0
FW deadlock in an environment of BIND and HW2SW_MPT	(ID: 29814)	4.6.2	4.7.0
QP deadlocks when doing a RETRY	(ID: 29676)	4.6.2	4.7.0
Big UAR pages	Support for Big UAR pages is not complete (ID: 29496)	4.6.2	4.7.0
Multicast Index miscalculation	Multicast Index miscalculation may cause dropping of multicast packets instead of inserting them. (ID: 29469)	4.6.2	4.7.0
CQ error or QP error together with 2ERR_QPEE may cause CommandIF to hang	(ID: 29431,29737)	4.6.2	4.7.0
FW deadlock when flushing a QP	(ID: 29277)	4.6.2	4.7.0
SRQ deadlocks when QP goes to error	(ID: 29174)	4.6.2	4.7.0
After a Catastrophic Error, HCA start may fail	(ID: 29066)	4.6.2	4.7.0
Port state ACTIVE_DIFFER should be reported as ACTIVE	(ID: 28811)	4.6.2	4.7.0
DIMM Unrecoverable Error not detected	(ID: 28902)	4.6.2	4.7.0
EQC.intr for the Catastrophic Error EQ is hard wired to 0x0	It now can be any legal value (including MSIx) (ID: 28815,28377)	4.6.2	4.7.0
SRQ performance is too low	(ID: 28702)	4.6.2	4.7.0
MSIx vector race when updating MSIx Table	(ID: 26599)	4.6.2	4.7.0
UD starvation	UD messages are not sent because RC ACKs are not arriving (ID: 28374,28427)	4.6.2	4.7.0
SerDes electrical stress may occur if VDDIO > 1.2V	(ID: 28385)	4.6.2	4.7.0

Table 5 - History of Fixed Bugs (Sheet 3 of 4)

Issue	Description	Discovered in	Fixed in
A SendQ connected to an SRQ may get a wrong RETRY_EXCEEDED	MPT window count is corrupted when BIND is used excessively (ID: 29953)	4.6.2	4.7.0
Fast Self Refresh feature	Fast Self Refresh feature is not functional. (ID: 30014)	4.6.2	4.7.0
DIMM timing parameters	DIMM timing parameters are not configured correctly	4.6.2	4.7.0
Consumer Index corruption in a Completion Queue	When using Increment_CI doorbells, to increment the CI in more than 1, CI may advance wrongly, casing a false CQ overrun, or not detecting a real overrun (ID: 27893)	4.6.1	4.6.2
Memory Region WindowCount corruption	When deallocating a window and trying to bind it simultaneously, the Region entry WindowsCount may be corrupted. (ID: 26829)	4.6.1	4.6.2
Access to VPD with partial ByteEnables	A configuration access to VPD with partial ByteEnables may cause VPD corruption (ID: 27690)	4.5.3	4.6.1
MAD with bad methods	a MAD with an unsupported method should be dropped (ID: 27472)	4.5.3	4.6.1
QP in retry may halt sending	A QP that is executing a message retry may hang and stop sending it (ID: 27252)	4.5.3	4.6.1
Q_Key source changes from QPC / WQE	If the Q_Key source changes from QPC /WQE, InfiniHost™ generates a packet with a wrong Q_key (ID: 21987)	4.5.3	4.6.1
Successful TimeOut-Driven-APM may cause QP context corruption	When an APM occurs as a result of TimeOut, QP context may be corrupted (ID: 26632)	4.5.3	4.6.1
Closing-QP commands get stuck	The commands 2RST_QPEE and RST2ERR_QPEE get stuck if a bad NACK is sent simultaneously (ID: 26243)	4.5.3	4.6.1
Validation of duplicate RDMA_READ/Atomic	Duplicate RDMA_READs/Atomics are not validated against the original requests (ID: 26247)	4.5.3	4.6.1
APM EQE due to TimeOut	An APM that resulted from a TimeOut does not generate an “APM succeeded” EQE (ID: 25948)	4.5.3	4.6.1
Binding Memory Windows across a 4GB boundary	There is an error in binding a Memory Window across a 4GB boundary (ID: 25958)	4.5.3	4.6.1
PCI express DeviceControl.unsupported_error_report_enable is R/W (Invariant Sector)	When flash is corrupted, bit 3 in device control (unsupported error reporting enable) is Read Only. It must be read/write (ID: 25570)	4.5.0	4.5.3
MAD with wrong BaseVersion or ClassVersion	MAD with wrong BaseVersion or ClassVersion should return status 0x1 (ID: 25888)	4.5.0	4.5.3
PortInfo.ResponseTime	The returned PortInfo.ResponseTime is too short (ID: 10597)	4.5.0	4.5.3
MAD with wrong AttributeModifier	MAD with wrong AttributeModifier should cause the response-MAD to have status INVALID_ATTR (ID: 25875)	4.5.0	4.5.3
MTTs in addresses > 4GB	If MTTs are in address > 4GB (e.g. in HIDE DDR), the device may get stuck (ID: 25877)	4.5.0	4.5.3
FW Debug version Data Section corruption	In Debug version of FW, FW Linker can cause a corruption in the Data Section of iRISC (ID: 25774)	4.5.0	4.5.3
PCI express DeviceControl.unsupported_error_report_enable is R/W	When flash is corrupted, bit 3 in device control (unsupported error reporting enable) was Read Only. It must be read/write (ID: 25570)	4.5.0	4.5.3

Table 5 - History of Fixed Bugs (Sheet 4 of 4)

Issue	Description	Discovered in	Fixed in
Link phy error threshold	link phy error threshold was set to a value larger by 1 than the one accepted in portInfo SET() (ID: 25854)	4.5.0	4.5.3
QP state mis-calculation	QPSTATE is shifted in a bit when QPC is non-cacheable. This may cause an unsolicited ACK to be dropped (ID: 25814)	4.5.0	4.5.3
local_ca_ack_delay	QUERY_DEV_LIM.local_ca_ack_delay was 16, which implies 266ms. Changed to 15. Driver can choose to return a lower number for in a non-stress case (ID: 25584)	4.5.0	4.5.3
CQE with error counters	The error counters sq_num_wrfe and rq_num_wrfe miss some increments (ID: 24463)	4.5.0	4.5.3
2err_qp/2rst_qpee starvation	If a CQE with Error is scheduled, and a heavy back-pressure is applied by UpLink, QP may be put to end of queue, and wait a long time till next chance to generate the CQE w/ Error (ID: 25044)	4.5.0	4.5.3
EQE loss	In stress, an ARM doorbell that should discover a non-empty CQ may be bypassed by a SET_CI doorbell. Thus CQ will seem empty, and an EQE will not be generated (ID: 24911)	4.5.0	4.5.3
MSIX memory mapped tables endianess	MSIX vectors and pending bits have wrong endianess (ID: 22794)	4.5.0	4.5.3
PCI compliancy and init changes (Invariant Sector)	Revision number changed to 21	4.0.1	4.5.0

Table 6 - Invariant Sector Fixed Bugs History

Issue	Description	Discovered in	Fixed in
PCI memory space must be disabled by default	PCI memory space must be disabled by default (ID: 27341)	4.5.3	4.6.1
SW reset may block PCIe message generation	In some cases, a SW reset may block the generation of future PCIe messages by the device (ID: 21721,27015)	4.5.3	4.6.1