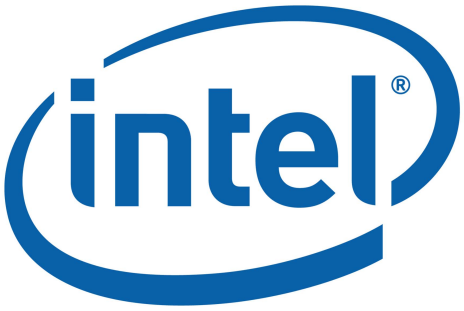


# Intel® Ethernet Switch FM10000

## Specification Update

**Networking Division (ND)**

***May 2016***

Revision 2.0
334304-001

# Revision History

| Revision | Date | Comments |
|---|---|---|
| 2.0 | May 12, 2016 | Initial Release (Intel public). |

# 1. Introduction

This document applies to the Intel® Ethernet Switch FM10000 (FM10000).

This document is an update to a published specification, the *Intel® Ethernet Switch FM10000 Datasheet*. It is intended for use by system manufacturers and software developers. All product documents are subject to frequent revision and new order numbers may apply. New documents may be added. Be sure you have the latest information before finalizing your design.

## 1.1 Product Code and Device Identification

### Product Code: FM10000

The following tables and drawings describe the various identifying markings on each device package:

**Table 1-1 Markings**

| Device | Stepping | Top Marking | S-Spec[1] | SKU Register | FuseBox for Scaling VDDS/VDDF | Description |
|---|---|---|---|---|---|---|
| FM10840 | B0 | EZFM10840 | S LLFX | Available | Available | Ethernet Switch (6x100/9x40/24x25/ 36x10) with 8x4 or 4x8 PCIe data ports. |
| FM10420 | B0 | EZFM10420 | S LLFY | Available | Available | Ethernet Switch (2x100/2x40/8x25/ 8x10) with 4x4 or 2x8 PCIe data ports. |

1. For Tray, Tape, Reel data, see Table 1-3.

**Table 1-2 Device ID**

| Device ID Code | Device ID | Vendor ID | Revision ID | Class Code |
|---|---|---|---|---|
| FM10000 Default Device | 15a4 | 8086 | 0x0 | 0x02 |
| FM10000 Virtual Function Device | 15a5 | 8086 | 0x0 | 0x02 |

**Table 1-3 MM Numbers**

| Product | Tray MM# | Tape and Reel MM# | Reserved |
|---|---|---|---|
| EZFM10840 | 946069 | N/A | |
| EZFM10420 | 946070 | N/A | |

## 1.2 Marking Diagrams


**Figure 1-1 FM10000 Example with Identifying Marks**

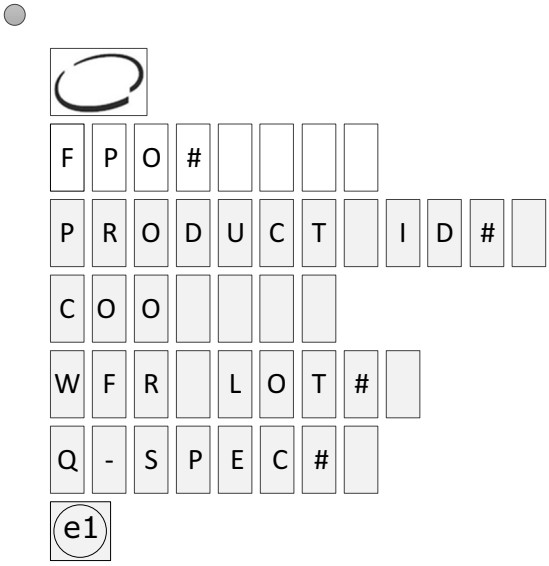
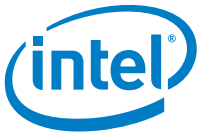



**Figure 1-2 FM10000 Marking Diagram**

- LINE 1: Swirl Logo
- LINE 2: FPO Number (Could be up to 8 characters)
- LINE 3: Product ID Number
- LINE 4: Country of Origin (Manufacturer assembly site)
- LINE 5: Wafer Lot and Wafer ID Number
- LINE 6: Q-spec Number
- LINE 7: Pb free symbol



## 1.3 Nomenclature Used in This Document

This document uses specific terms, codes, and abbreviations to describe changes, errata, sightings and/or clarifications that apply to silicon/steppings. See Table 1-4 for a description.

**Table 1-4 Nomenclature**

| Name | Description |
|---|---|
| Specification Clarifications | Greater detail or further highlights concerning a specification's impact to a complex design situation. These clarifications will be incorporated in the next release of the specifications. |
| Specification Changes | Modifications to the current published specifications. These changes will be incorporated in the next release of the specifications. |
| Errata | Design defects or errors. Errata may cause device behavior to deviate from published specifications. Hardware and software designed to be used with any given stepping must assume that all errata documented for that stepping are present on all devices. |
| Software Clarifications | Applies to Intel drivers, EEPROM loads. |
| Documentation Changes | Typos, errors, or omissions from the current published specifications. These changes will be incorporated in the next release of the specifications. |
| A0, B0, etc. | Stepping to which the status applies. |
| Doc | Document change or update that will be implemented. |
| Fixed | This erratum has been fixed. |
| Fix Planned | This erratum is intended to be fixed in a future stepping of the component. |
| NoFix | There are no plans to fix this erratum. |
| Fixed in NVM | This erratum has been fixed in NVM X.XX. |
| Fix Planned in NVM | This erratum is intended to be fixed in a future NVM version. |
| Eval | Plans to fix this erratum are under evaluation. |

# 2. Hardware Clarifications, Changes, Updates and Errata

See Section 1.3 for an explanation of terms, codes, and abbreviations.

### Table 2-1 Summary of Specification Clarifications

| Specification Clarification | Status |
|---|---|
| 1. Designs Incorporating Six or More PCIe End Points (PEPs) | N/A |
| 2. Driving LVPECL Receivers with HSCL | N/A |

### Table 2-2 Summary of Specification Changes

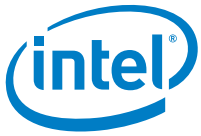
| Specification Change | Status |
|---|---|
| None. | N/A |

### Table 2-3 Summary of Documentation Updates

| Specification Change | Status |
|---|---|
| None. | N/A |

### Table 2-4 Summary of Errata; Errata Include Steppings

| Erratum | Status |
|---|---|
| 1. FM10000 VSSF Sense Pin | A0=Yes, B0=No; Fixed |
| 2. FM10000 PCIe Gen3 Not Working | A0=Yes, B0=Yes; Fixed in NVM 1.12 |
| 3. 10/40 GbE KR/KR4 Rise/Fall Time Violations | A0=Yes, B0=Yes; NoFix |
| 4. 100GBASE-CR4 and 100GBASE-KR4 Max Channel Support Limitation | A0=Yes, B0=Yes; NoFix |
| 5. 10 GbE SFI Rx Fails ITT (Direct Attach) BER Test | A0=Yes, B0=Yes; NoFix |
| 6. PCIe RX Input Impedance with Port Off Below Specification | A0=Yes, B0=Yes; NoFix |
| 7. $I^2C$ Write/Read Does Not Work Reliably | A0=Yes, B0=No; Fixed |
| 8. Some $I^2C$ Timings at 100 KHz Are Out of Specification | A0=Yes, B0=Yes; NoFix |
| 9. 10GBASE-R BER Monitor State Machine Not Compliant with IEEE Std 802.3 | A0=Yes, B0=Yes; NoFix |
| 10. Ethernet Port Signal Interrupt Unreliable if Energy Efficient Ethernet is Enabled | A0=Yes, B0=Yes; NoFix |
| 11. 10 GbE Port Transition from Reset | A0=Yes, B0=Yes; NoFix |



## Table 2-4 Summary of Errata; Errata Include Steppings (Continued)

| Erratum | Status |
|---|---|
| 12. $I^2C$ Master Does Not Check Slave ACK During Read | A0=Yes, B0=Yes; NoFix |
| 13. $I^2C$ Timeout is Larger Than Expected | A0=Yes, B0=Yes; NoFix |
| 14. Mailbox Semaphore Mechanism is Not Operational | A0=Yes, B0=Yes; NoFix |
| 15. WAKE_ERROR_COUNTER in EPL Not Counting Properly in 10 GbE Mode | A0=Yes, B0=Yes; NoFix |
| 16. Ethernet Port Frame Corruption on //T// //LI// | A0=Yes, B0=Yes; NoFix |
| 17. PCIe Tx Voltage Swing May Exceed Specification | A0=Yes, B0=Yes; NoFix |
| 18. UNH test 36.1.1 Synchronization Failure | A0=Yes, B0=Yes; NoFix |
| 19. FM10000 May Exceed 10/40 GbE KR/KR4 V2 Amplitude for High Voltage | A0=Yes, B0=Yes; NoFix |
| 20. FM10000 Exceeds 100 GbE KR4 Tx Steady State Voltage Test | A0=Yes, B0=Yes; NoFix |
| 21. 100 GbE KR4 Tx Effective Jitter Fail (ETUJ) | A0=Yes, B0=Yes; NoFix |
| 22. 100 GbE CR4 Tx Effective Jitter Fail (ETUJ) | A0=Yes, B0=Yes; NoFix |
| 23. DC Electrical Idle Differential Output Voltage Out of Specification | A0=Yes, B0=Yes; NoFix |
| 24. FM10000 Fails UNH 4.1.4 Test: Ethernet Length Not Validated | A0=Yes, B0=Yes; NoFix |
| 25. FM10000 Does Not Pass UNH 31.2.3 and 4.1.3D Tests | A0=Yes, B0=Yes; NoFix |
| 26. PCIe Channel Errors Can Cause DMA Engine Deadlock | A0=Yes, B0=No; Fixed |
| 27. PCIe CEM Loopback Works Only on Lane 0 of Each PEP | A0=Yes, B0=Yes; NoFix |
| 28. I2C_SCL and I2C_SDA Do Not Float when the FM10000 is Powered Off | A0=Yes, B0=Yes; NoFix |
| 29. EPL_TX_FRAME_ERROR_COUNTER May Under-Count | A0=Yes, B0=Yes; NoFix |
| 30. FM10000 Not Compliant at 10 GbE KR Rx MTC1 (Long Channel) Test | A0=Yes, B0=Yes; NoFix |
| 31. FM10000 RX Compliance failures for 25GBase-CR, 100GBase-CR4, 25GBase-KR, and 100GBase-KR4 | A0=Yes, B0=Yes; NoFix |
| 32. Lossless PEP Configuration | A0=Yes, B0=Yes; NoFix |
| 33. PCIe Delayed ACK | A0=Yes, B0=Yes: NoFix |
| 34. PCIe Replay Timeout Exceeded | A0=Yes, B0=Yes: NoFix |

## 2.1 Specification Clarifications

### 1. Designs Incorporating Six or More PCIe End Points (PEPs)

The PCI Express 3.0 Base Specification states that a PCIe device must enter the LTSSM Detect state within 20 ms of the end of Fundamental Reset. If six or more of the FM10000 PEPs come out of reset in the same 11 ms window, and the FM10000 is set to access the SPI NVM in single-pin mode, the 20 ms specification may be violated by up to approximately 3 ms. That could leave the affected FM10000 PEPs in reset state due to inactivity on the PCIe bus.

To stay within the 20 ms specification in designs with six or more PEPs, configure the FM10000 NVM using the FM10000 Boot Image Generator (**rrcBig**) tool to an SPI speed of 50 MHz, and use either dual-pin or quad-pin SPI mode.

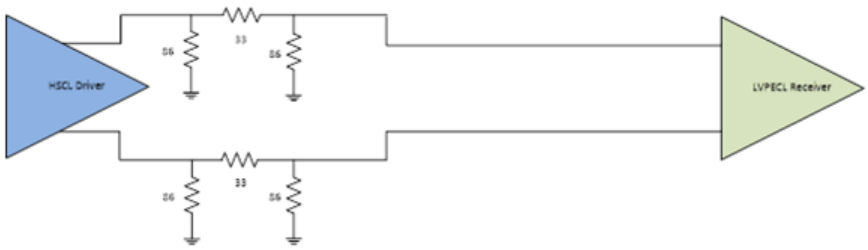
If an NVM configured to single-pin mode is an absolute requirement with six or more PEPs, FM10000 customers MUST avoid resetting six or more PEPs in the same 11 ms window.

**Note:** In all cases, when creating the FM10000 NVM using **rrcBig**, customers should ensure that the NVM is configured to a SPI speed of 50 MHz, as this gives the fastest PCIe reset de-assertion handling.

### 2. Driving LVPECL Receivers with HSCL

Intel strongly recommends that customers drive the LVPECL receivers on the FM10000 with LVPECL transmitters. FM10000 validation was performed with LVPECL driving LVPECL.

Some customers may decide to use HSCL transmitters in place of LVPECL transmitters. Using an HSCL driver may be possible using a circuit similar to the diagram below. Be aware that Intel has not validated this circuit with the FM10000.



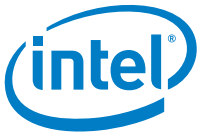

Customers should review the Ruby Rapids and Red Gorge reference design schematics for other potential circuits.

## 2.2 Specification Changes

None.

## 2.3 Documentation Updates

None.



## 2.4 Errata

### 1. FM10000 VSSF Sense Pin

Status: A0=Yes, B0=No; Fixed

### 2. FM10000 PCIe Gen3 Not Working

Problem:

PCIe Gen3 FW development and validation is not complete. FM10000 PCIe Gen3 does not work properly.

Implication:

FM10000 PCIe Gen3 does not work.

Workaround:

FM10000 customers should use PCIe Gen1 and Gen2.

Status: A0=Yes, B0=Yes; Fixed in NVM 1.12

FM10000 NVM version 1.12 or higher allows PCIe Gen3 to work.

### 3. 10/40 GbE KR/KR4 Rise/Fall Time Violations

Problem:

The FM10000 does not meet Rise & Fall time 10/40 GbE KR/KR4.

Typically, $T_{rise}/T_{fall}$ = 17.4 ps to 23.2 ps, where 24 ps ≤ *Limit* ≤ 47 ps.

Implication:

Fast Rise & Fall time could result in higher EMI. IEEE testing could also be affected.

Workaround:

None.

It is always best practice to follow the FM10000 schematic and layout guides to achieve the best possible EMI results.

Status: A0=Yes, B0=Yes; NoFix

## 4. 100GBASE-CR4 and 100GBASE-KR4 Max Channel Support Limitation

### Problem:

As per IEEE 802.3bj specification, maximum channel loss for 100GBASE-CR4 and 100GBASE-KR4 applications is 35 dB at 12.89 GHz with RS-FEC. The FM10000 does not have a complete FEC implementation that can correct bit errors. Therefore, the FM10000 cannot support the max channel loss of 35 dB. Maximum supported loss by the FM10000 for 100GBASE-CR4 and 100GBASE-KR4 applications is 30 dB at 12.89 GHz.

### Implication:

For 100GBASE-CR4, it is up to customers how they allocate this 30 dB budget to the cable (for 100GBASE-CR4) and PCB portions of the channel. For example, if the intention is for a design to take the full 6.81 dB budget allocated by the IEEE 802.3bj specification for the PCB portion at each end of the cable, only 16.38 dB budget remains for the cable. In this case, the customer is limited to cables that have ≤ 16.38 dB loss at 12.89 GHz. If customers can save a few dB of loss from the PCB portion, they can use that in their cable loss budget, which may potentially allow them to use higher loss cables.

For 100GBASE-KR4, as long as the total insertion loss is ≤ 30 dB at 12.89 GHz, the loss can be distributed in any number of ways among the circuit boards in the KR4 channel path.

For both 100GBASE-CR4 and 100GBASE-KR4, if the customer does not own or control all of the boards (and/or cables) in the full channel path, they must determine how much of the full insertion loss budget is consumed by the other boards in the same channel. Provided they consume less than 30 dB, the difference between the sum of the insertion losses of the other boards and 30 dB is the channel budget that remains for the FM10000 circuit board.

For more details, refer to *Intel® Ethernet Switch FM10000 100GBASE-CR4 Cable Specifications Application Note* (Doc# 560651)

### Workaround:

There is no silicon workaround for this issue. Customers may choose lower loss cables and/or may design lower loss PCB channel to work around this limitation.

An alternative is to add a re-timer with FEC support.

### Status: A0=Yes, B0=Yes; NoFix

## 5. 10 GbE SFI Rx Fails ITT (Direct Attach) BER Test

### Problem:

The FM10000 fails Rx SFI ITT (Interference Tolerance Test) at high temperature.

SFI Interference Tolerance Test tests the Rx ability to cope with an SFP+ direct-attach cable utilizing a reference Tx. Part of *Appendix E, -SFP+ Direct Attach Cable Specifications "10GSFP+CU"*.

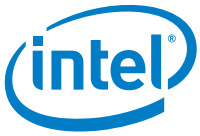
### Implication:

Interoperability/Operability issues might occur using high-loss SFP channels.

### Workaround:

None.

### Status: A0=Yes, B0=Yes; NoFix



## 6. PCIe RX Input Impedance with Port Off Below Specification

### Problem:

When the FM10000 is powered down, the measured port input impedance can be less than 20 KΩ (Specification Value impedance >20 KΩ for voltage higher then 200 mV).

It is found that when the FM10000 is powered down, the measured RX impedance can be less than 20 KΩ with input signal common mode offsets of 200-500 mV. In those cases, input impedances as low as 250 Ω can be measured. Common mode offsets of 200 mV meet specification.

### Implication:

This may create interoperability issues if a PCIe device is directly connected to the FM10000 while it is powered down, and the partner detect signal has an output common mode value of ~ 500 mV (false Tx detect).

### Workaround:

None.

### Status: A0=Yes, B0=Yes; NoFix

## 7. $I^2C$ Write/Read Does Not Work Reliably

### Status: A0=Yes, B0=No; Fixed

## 8. Some $I^2C$ Timings at 100 KHz Are Out of Specification

### Problem:

The following $I^2C$ timings are out of specification for 100 KHz operation when the FM10000 is operating as a master, and configured for 100 KHz operation by setting I2C_CFG.*Divider* to 52 decimal (0x34h).

- tHigh minimum should have been 4.0 μs, it is actually 3.0 μs (A0/B0 Step)
- tHD;STA minimum should have been 4.0 μs, it is actually 3.0 μs (A0/B0 Step)
- tSU;STO minimum should have been 4.0 μs, it is actually 3.0 μs (A0/B0 Step)
- tSU;STA minimum should have been 4.7 μs, it is actually 3.5 μs (B0 Step)

### Implication:

Slow $I^2C$ devices might not work properly if addressed from the FM10000.

### Workaround:

Select 400 KHz capable devices whenever possible. In the FM10000, select 400 KHz operation by setting I2C_CFG.*Divider* to 10 decimal (0xAh). The FM10000 meets timing constraints at 400 KHz device. If it is not possible to use 400 KHz, then review timing constraints on 100 KHz devices to ensure it can operate with the shorter timing diagrams.

### Status: A0=Yes, B0=Yes; NoFix

## 9. 10GBASE-R BER Monitor State Machine Not Compliant with IEEE Std 802.3

### Problem:

The BER error monitor is not turned off when a 10 GbE port enters low-power mode, and reports high error count as long as the port in this mode.

### Implication:

Customers with any 10 GbE port in low-power mode will see high BER error count if monitoring the BER error counter.

### Workaround:

Software should ignore BER count while the port is in low-power mode, and must also disable link down report on high BER count when Energy Efficient Ethernet is enabled.

### Status: A0=Yes, B0=Yes; NoFix

## 10. Ethernet Port Signal Interrupt Unreliable if Energy Efficient Ethernet is Enabled

### Problem:

FM10000 SERDES_IP.*RxSignalOk* interrupt not working properly when reporting status while the port is in low power mode.

### Implication:

The SERDES_IP.*RxSignalOk* interrupt should not be used when port is in low-power mode.

### Workaround:

Cancel this interrupt source when Energy Efficient Ethernet is used, and poll PCS_1000BASEX_RX_STATUS.*SyncStatus* periodically to detect whether the port is synchronized to the link partner.

### Status: A0=Yes, B0=Yes; NoFix

## 11. 10 GbE Port Transition from Reset

### Problem:

The 10GBASE-R receiver allows a transition from the RX_INIT to the RX_LI state when a link partner sends //LI// while a port is taken out of reset. A compliant 10GBASE-R EEE implementation should execute the RX_INIT->RX_E->RX_LI set of transitions, and increment *CodeErrorCnt* by one. The FM10000 does not increment this counter in this case. This has no impact on the correct functioning of a network.

### Implication:

*CodeErrorCnt* is not incremented during the described transition.

### Workaround:

Ignore *CodeErrorCnt* during reset.

### Status: A0=Yes, B0=Yes; NoFix

## 12. I2C Master Does Not Check Slave ACK During Read

### Problem:

I2C read operations initiated from the FM10000 do not check slave device's ACK after sending out the first byte with the address and R/W bit. If no device matching that particular address is present to assert ACK, the I2C master nevertheless continue with the read operation, and completes the command successfully, returning 0xFF data.

I2C write operations initiated from the FM10000 do check slave device's ACK after sending the address and write command, and terminates with No_device(3) if the device is not present or does not return an ACK.

Reading a non-existent device does not cause the FM10000 I2C master to deadlock.

### Implication:

Cannot use the FM10000 I2C Read to poll for devices.

### Workaround:

Do not use the FM10000 I2C Read to poll for I2C devices.

### Status: A0=Yes, B0=Yes; NoFix

## 13. I2C Timeout is Larger Than Expected

### Problem:

The FM10000 I2C controller implements a timeout which aborts any transaction if the I2C clock is stretched for too long, returning I2C data and clock to tri-state. The value of this timeout is hard-coded at 43 seconds, and thus not programmable by software. This value far exceeds known reasonable clock stretch limits of existing devices on the market. The timeout does not happen as long the target I2C devices do not endlessly stretch clocks.

### Implication:

Very long timeout for FM10000 I2C if attached I2C device stretches out clock for long periods.

### Workaround:

Verify that I2C devices connected to the FM10000 do not stretch clock too long.

### Status: A0=Yes, B0=Yes; NoFix

# 14. Mailbox Semaphore Mechanism is Not Operational

## Problem:

The mailbox control registers cannot be written without modifying the semaphore owner bit. Original definition defined the owner bit stay unchanged if written as zero. Furthermore, writing zero to this bit clears the bit instead.

## Implication:

Unable to use mailbox semaphore mechanism.

## Workaround:

Since the mailbox control register requires it be written from time to time for consuming messages or for changing the enable interrupt setting, the following workaround flow is recommended. Instead of using semaphores, the proposed workaround uses a lockless mechanism. It splits the mailbox memory in two half regions, each half written by one different entity.

Mailboxes are available for PF⇔VF and PF⇔SM for communications between a PF and VF drivers, and between PF drivers and external Switch Managers. Each PCIe end point has its own set of mailboxes. This channel can be used for the PF driver to send command/status updates to the VFs (such as link change, memory parity error, etc.) or for the VF to send requests to the PF (add to VLAN, etc.), same for PF to SM and vice versa.

**Note:** An external Switch Manager is a Switch Manager running on another host than the current directly-attached host. The external Switch Manager receives interrupts through the PCI_INTERRUPT_OUT_SM from this PCIe interface, which gets routed to either the INT_N pin or routed to another PCIe via LSM. If routed to another PCIe, the signal is received via PCI_INTERRUPT_IN signal. This later one results in an MSI-X interrupt posted to the PF or VF.

The mailboxes are stored in PCIE_MBMEM0..2047 (32-bit words) and used this way:

- 64 x 64-byte mailboxes for communication of the PF with VFs (one 64-byte mailbox per VF, first 32 bytes written by the VF, last 32 bytes written by the PF).
- 1 x 4 KB mailbox for communication between PF and SM (first 2 KB written by the SM, last 2 KB written by the PF).

The set of registers for this unit are:

- For PF:
  - PCIE_MBMEM[0..2047] — Mailboxes
  - PCIE_MBX[0..63] — VF Mailboxes Control
  - PCIE_GMBX — Global Mailbox Control & Interrupt Status
  - PCIE_MBICR[0..3] — VF Mailboxes Interrupt Cause to MSI-X (PF)
  - PCIE_MBIMR[0..1] — VF Mailboxes Interrupt Mask to MSI-X (PF)
- For VF:
  - PCIE_VFMBMEM[0..15] — VF view of its dedicated mailbox
  - PCIE_VFMBX — VF view of its dedicated mailbox control

Figure 2-1 shows the arrangement:

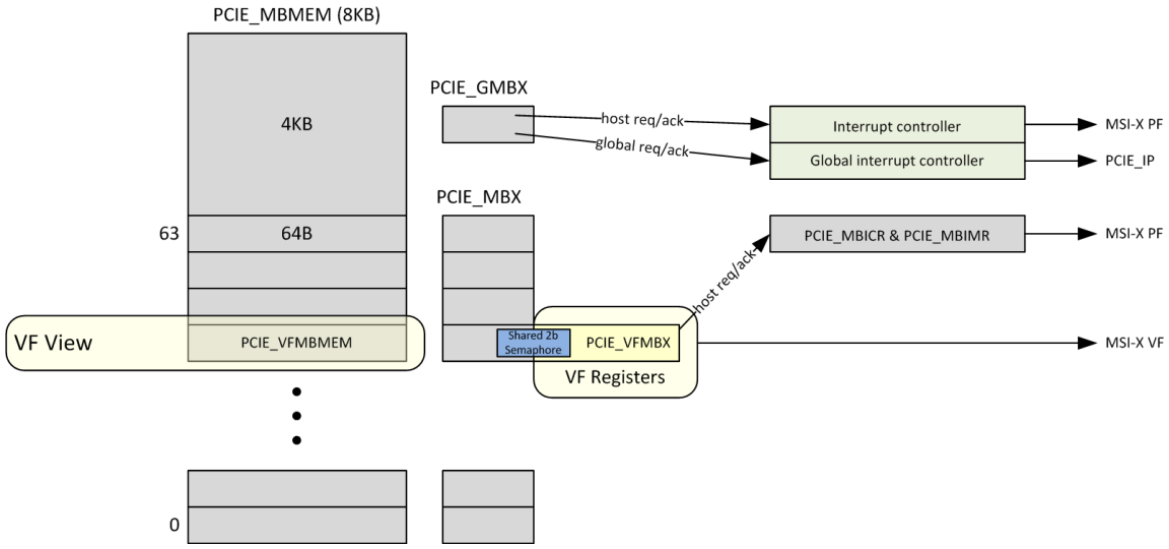



**Figure 2-1 PCIe Mailbox Unit**

The PF writes messages to VF into the last half of the PCIE_MBMEM[64*vf..64*vf+15], and writes messages to SM into the last half of PCIE_MBMEM[1024..2047]. It announces message availability using the PCIE_MBX[vf] and PCIE_GMBX, respectively.

The VF writes messages to PF into the first half of PCIE_VFMBMEM, which points to its mailbox. It announces message availability using the PCIE_VFMBX register which contains fields to route announcement to PF.

The SM writes messages to PF into the first half of PCIE_GMBMEM. It announces message availability using the PCIE_GMBX. When it exits from reset, the SM must clear the contents of its half in the PF/SM mailboxes of all PEPs, just in case it was the owner before resetting.

Following is the process from software:

- Requester:
  1. Writes the message and sends a "request" interrupt to the target.
  2. Waits for an "ack" interrupt from target.

  Requests are announced by setting the MBX.*xxxReq* bit, which results in a mailbox interrupt on the receiver and the request latched in the MBX registers (mirror on MBxICR registers for VF interrupts).

  Acknowledges are detected by reading the MBX registers (mirror on MBxICR registers for VF interrupts).
- Receiver:
  1. Wait for request interrupt.
  2. Read message.
  3. Write back ack.

  Acknowledges are announced by setting the MBX.*xxxAck* bit, which results in a mailbox interrupt on the sender and the acknowledge latched in the MBX registers.

Table 2-5 shows a step-by-step example for PF to VF.

**Table 2-5 PF-to-VF Messaging Flow**

| Step | Event |
|---|---|
| 1 | PF writes message to relevant location in PCIE_MBMEM[n] (in the last half of the corresponding mailbox). |
| 2 | PF sets PCIE_MBX[n].*Req* bit. |
| 3 | Hardware indicates an interrupt to VF #n, setting PCIE_VFMBX.*ReqInterrupt* bit. |
| 4 | VF gets interrupt. |
| 5 | VF reads the message from PCIE_VFMBMEM register. |
| 6 | VF clears PCIE_VFMBX[n].*ReqInterrupt* bit, and sets the *PFAck* bit. |
| 7 | Hardware indicates an interrupt to PF, setting the relevant PCIE_MBX[n].*PFAckInterrupt* bit. |
| 8 | PF clears PCIE_MBX[n].*PFAckInterrupt* bit. |

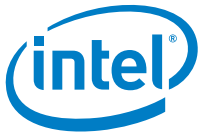
The next table shows a step-by-step example for VF to PF.

**Table 2-6 VF-to-PF Messaging Flow**

| Step | Event |
|---|---|
| 1 | VF writes message to the first half of PCIE_VFMBMEM. |
| 2 | VF sets PCIE_VFMBX.*Req* bit, setting PCIE_MBX[n].*PFReqInterrupt* bit. |
| 3 | Hardware indicates an interrupt to PF, setting the relevant bit in the PCIE_MBICR register. |
| 4 | PF gets interrupt and reads PCIE_MBICR to identify the VF(s). |
| 5 | PF reads the message from relevant location in the PCIE_MBMEM[n] register. |
| 6 | PF clears PCIE_MBX[n].*Req* bit and sets PCIE_MBX[n].*Ack* bit. |
| 7 | Hardware indicates an interrupt to VF, setting the PCIE_VFMBX.*AckInterrupt* bit. |
| 8 | VF clears the PCIE_MBX[n].*PFAckInterrupt* bit. |
| 9 | PF polls PCIE_MBICR until the relevant bit is cleared, and goes back to step 5 for handling mailbox messages from other VFs. |

In any case, the content of the message is hardware independent and is determined by software.

Status: A0=Yes, B0=Yes; NoFix



## 15. WAKE_ERROR_COUNTER in EPL Not Counting Properly in 10 GbE Mode

### Problem:

The WAKE_ERROR_COUNTER is used in Energy Efficient Ethernet to count wake up failures after quiescent periods while in low-power mode. This counter should normally be incremented by exactly 1 for every single fault event.

This errata is incorrectly implementing this counter while operating in 10 GbE mode. The counter might be incremented randomly by 2, 3 or 4 for each wake up fault rather than only 1.

The counter operates correctly in 1G mode.

The counter is informative and has no incidence to normal switching functions of the device and would remain to zero if there are no faults while in low power mode.

### Implication:

The WAKE_ERROR_COUNTER in 10 GbE mode can provide incorrect wake up failure counts.

### Workaround:

The software should clear this counter when switching in 10 GbE mode, and divide this counter by 2 to provide an approximate value for the number of faults.

### Status: A0=Yes, B0=Yes; NoFix

## 16. Ethernet Port Frame Corruption on //T// //LI//

### Problem:

The *IEEE Std 802.3-2012, Figure 49-16, Transmit State Machine* specifies that a Terminate (||T||) column can be immediately followed by a low-power idle (||LI||) column. The FM10000 transmitter does not recognize this sequence as a valid sequence during encoding, and replaces the sequence with an error sequence, causing the frame to be received as invalid by the link partner. This situation could occur if an idle period is detected, software enables LPI mode and a frame was just in transmission at the time the software enables the LPI mode.

Similarly, the *IEEE Std 802.3-2012, Figure 49-17 Receive State Machine* specifies that a Terminate (||T||) column can be immediately followed by a low-power idle (||LI||) column. The FM10000 receiver does not recognize this sequence as a valid sequence during decoding, and declares the frame as invalid.

### Implication:

The FM10000 could declare Ethernet frames invalid when a Terminate (||T||) column is immediately followed by a low-power idle (||LI||) column.

### Workaround:

The work around for the Transmit State is for the software to block frame scheduling to the port, wait for any packet in flight to complete (thus ensuring quiet time), enable the LPI mode, then eventually define how to get out of the LPI mode (either by re-enabling frame scheduling or other event).

The work around for the Receive State is the link partner should not enter LPI mode right after a frame, but allow some idle time before entering LPI mode.

### Status: A0=Yes, B0=Yes; NoFix

## 17. PCIe Tx Voltage Swing May Exceed Specification

### Problem:

The FM10000 may exceed maximum p-p Tx voltage swing requirement with a value of 1.27 V compared to a spec of 1.2 V. This only occurs for high voltage corner use cases.

### Implication:

PCIe Tx voltage swing may exceed specified voltage by 0.07 V.

### Workaround:

None.

### Status: A0=Yes, B0=Yes; NoFix

## 18. UNH test 36.1.1 Synchronization Failure

### Problem:

UNH test 36.1.1 Acquire Synchronization, Part A, Sequence 6 FAILED to link.

Sequence 6 checks the ability to synchronize to an unusual pattern, including the reserved code /K28.1/ which contains a comma. The FM10000 does not synchronize to the /K28.1/ reserved pattern.

### Implication:

/K28.1/ is not used by 1000BASECX, and this failure should not cause any interoperability issues.

### Workaround:

None.

### Status: A0=Yes, B0=Yes; NoFix

## 19. FM10000 May Exceed 10/40 GbE KR/KR4 V2 Amplitude for High Voltage

### Problem:

V2 is defined as the average of the middle 4UIs in a Square8 pattern. The IEEE specification allows for 400 mV < v2 < 600 mV (800–1200 mV differential). The test is designated to characterize adaptive link training ability over backplane.

FM10000 V2 can be up to 1260 mV (differential) at high input voltage. Analog supply High voltage was defined as 105% (1.05 V). The Analog voltage change affects the Tx's amplitude linearly.

### Implication:

FM10000 V2 voltage in IEEE test can run 60 mV over specification limit under high voltage (AVDD) conditions,

### Workaround:

None.

### Status: A0=Yes, B0=Yes; NoFix

## 20. FM10000 Exceeds 100 GbE KR4 Tx Steady State Voltage Test

### Problem:

Steady State Voltage (802.3bj Clause 93.8.1.5.2):

The steady state voltage is defined as the sum of the linear fit pulse response (divided by the number of samples per UI). This test is part of the Output WF tests. Specification limit is 0.6 V Max.

FM10000 results are 0.52 V to 0.62 V, and indicate a high amplitude signal steady state voltage at high supply voltage (105% AVDD). The higher Tx Steady State Voltage is linearly affected by the analog voltage, as well as slightly by temperature.

### Implication:

The FM10000 may exceed IEEE 802.3bj Clause 93.8.1.5.2 Steady State Voltage by 0.02 V when analog supply voltage is at 105%.

### Workaround:

None.

### Status: A0=Yes, B0=Yes; NoFix

## 21. 100 GbE KR4 Tx Effective Jitter Fail (ETUJ)

### Problem:

The effective jitter test measures the different components of the uncorrelated jitter.

Three components are defined:

- Effective bounded uncorrelated jitter – Max 0.1 UI.
- Effective Random Jitter – No spec limit.
- Effective Total Uncorrelated Jitter – Max 0.18 UI.

The FM10000 indicates 0.2-0.4UI (8ps - 16ps) for Effective Total Uncorrelated Jitter (ETUJ) across lanes/corners.

### Implication:

Higher Tx jitter can impact interoperability vs. jitter sensitive Rx devices. This might not be seen while operating vs. another FM10000 or other “jitter tolerant” devices.

### Workaround:

None.

### Status: A0=Yes, B0=Yes; NoFix

## 22. 100 GbE CR4 Tx Effective Jitter Fail (ETUJ)

### Problem:

Effective Jitter Test:

The effective jitter test measures the different components of uncorrelated jitter.

Three components are defined:

- Effective bounded uncorrelated jitter – Max 0.1 UI.
- Effective Random Jitter – No spec limit.
- Effective Total Uncorrelated Jitter – Max 0.18 UI.

The FM10000 measures 0.2-0.31UI (8ps - 12ps) for Effective Total Uncorrelated Jitter (ETUJ) across lanes/corners.

### Implication:

Higher Tx jitter can impact interoperability vs. jitter sensitive Rx devices. This might not be seen while operating vs. another FM10000 or other "jitter tolerant" devices.

### Workaround:

None.

### Status: A0=Yes, B0=Yes; NoFix

## 23. DC Electrical Idle Differential Output Voltage Out of Specification

### Problem:

The PCIe base r3.1 specification defines the "DC Electrical Idle Differential Output Voltage" for Gen2 and Gen3 to be less than 5 mV.

The FM10000 maximum can be up to 32 mV (the results are higher on slow material).

### Implication:

No known implications.

### Workaround:

None.

### Status: A0=Yes, B0=Yes; NoFix

## 24. FM10000 Fails UNH 4.1.4 Test: Ethernet Length Not Validated

### Problem:

UNH Test 4.1.4: "Frames with length values greater than the length of Data/Pad field should be considered to have invalid length values and that they should be discarded by the DUT."

The FM10000 does not validate Ethernet frame length using the Ethernet Length/Type field. As a result, the FM10000 fails UNH test 4.1.4 "Ethernet length not validated".

### Implication:

The FM10000 validates frame length based on actual frame length, packets with invalid length field will still be forwarded to the destination.

### Workaround:

None.

### Status: A0=Yes, B0=Yes; NoFix

## 25. FM10000 Does Not Pass UNH 31.2.3 and 4.1.3D Tests

### Problem:

The FM10000 accepts pause frames of length up to 192 bytes, but ignores frames of 193 bytes and above. As a result, the FM10000 will not pass the following UNH tests:

- 4.1.3D - "Reception and Transmission of Oversized Frames"
- 31.2.3 - "Receive Oversized MAC Control PAUSE Frames".

### Implication:

The FM10000 will not interpret and react to PAUSE frames greater than 192 bytes. There are no practical reasons for a link partner to send a pause frame greater than 64 bytes, and Intel is not aware of any implementations that generate a PAUSE frame greater than 192 bytes.

### Workaround:

None.

### Status: A0=Yes, B0=Yes; NoFix

## 26. PCIe Channel Errors Can Cause DMA Engine Deadlock

### Status: A0=Yes, B0=No; Fixed

## 27. PCIe CEM Loopback Works Only on Lane 0 of Each PEP

### Problem:

The PCIe Card Electro-Mechanical (CEM) specification states that all lanes should support CEM loopback mode. The FM10000 only supports CEM loopback on lane 0 of each PEP.

### Implication:

The loopback function is normally used for testing purposes, the loopback function is only functional on lane 0 of each PEP.

### Workaround:

None.

### Status: A0=Yes, B0=Yes; NoFix

## 28. I2C_SCL and I2C_SDA Do Not Float when the FM10000 is Powered Off

### Problem:

The FM10000 does not comply with the $I^2C$-Bus 2014 specification, Section 5.1 (Fast-Mode), which mentions:

> "If the power supply to a Fast-Mode device is switched off, the SDA and SCL I/O pins must be floating so that they do not obstruct the bus lines."

With the FM10000 directly attached to $I^2C$ chain with the FM10000 powered off, the I2C_SCL and I2C_SDA pins do not float as required.

### Implication:

If the FM10000 is powered off and directly attached to an $I^2C$ bus, the SCL and SDA lines on the $I^2C$ bus are pulled down to a lower voltage.

### Workaround:

To work around the issue in a live environment where the FM10000 could be off, the user could add a level translator/$I^2C$ bus repeater (such as the PCA9617A) between the FM10000's $I^2C$ pins and the $I^2C$ chain.

### Status: A0=Yes, B0=Yes; NoFix

## 29. EPL_TX_FRAME_ERROR_COUNTER May Under-Count

### Problem:

In switch diagnostic configurations using SAF_MATRIX.*CutThruMode*=0 OR SAF_MATRIX.*IgnoreErr*=1, the EPL_TX_FRAME_ERROR_COUNTER may under-count if two packets are transmitted back-to-back, both packets are marked with error, and the second packet is less than 16 bytes long.

### Implication:

This case can only occur in diagnostic modes. Deployed configurations should set *CutThruMode*=1 and *IgnoreErr*=0, in which case the switch drops the tiny packet with error before it reaches the TX_MAC.

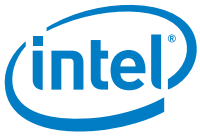
When *IgnoreErr* is enabled to allow diagnostic of link issues by a network engineer, the EPL_TX_FRAME_ERROR_COUNTER may miscount the retransmission of those errors. The ingress counters properly record the number of errors received.

Note that this error condition is very unlikely even in diagnostic modes.

### Workaround:

Customer configurations should set CutThruMode=1 and IgnoreErr=0 as recommended.

### Status: A0=Yes, B0=Yes; NoFix



## 30. FM10000 Not Compliant at 10 GbE KR Rx MTC1 (Long Channel) Test

### Problem:

According to the specification at 802.3ap, Table 72.10 and Annex 69A eq. 69A-6, when using MTC1 in the 10 GbE KR receiver interference tolerance test, the target BER should be equal to or better than $10^{-12}$. Validation results for the FM10000 indicate the BER when using MTC1 in the 10 GbE KR receiver interference tolerance test could be up to $10^{-8}$.

### Implication:

This issue is visible by customers when performing the 10 GbE KR receiver interference tolerance test according to the 802.3ap standard. Validation results for the FM10000 indicate the BER when using MTC1 in the 10 GbE KR receiver interference tolerance test could be up to $10^{-8}$.

This issue does not necessarily affect the capability of the FM10000 to interact with other FM10000 units or with any other link partners. The receiver interference tolerance test defines conditions that are worse than the conditions that would be experienced in real operation against a link partner. Intel testing of the FM10000 10 GbE KR did not show any degradation in the FM10000's capability to interact with other link partners.

### Workaround:

None.

### Status: A0=Yes, B0=Yes; NoFix

## 31. FM10000 RX Compliance failures for 25GBase-CR, 100GBase-CR4, 25GBase-KR, and 100GBase-KR4

### Problem:

The FM10000 has BER failures in IEEE 802.3 RX compliance tests for 25GBase-CR, 100GBase-CR4, 25GBase-KR, and 100GBase-KR4 technologies. The expected BER for those tests is $10^{-12}$.

The tests that do not comply are:

- 25GBase-CR and 100GBase-CR4 long channel BER:
  The maximum BER seen in this case was approximately $10^{-3}$.
- 25GBase-CR and 100GBase-CR4 short channel BER:
  The maximum BER seen in this case was approximately $10^{-10}$.
- 25GBase-CR and 100GBase-CR4 JTT BER:
  The maximum BER seen in this case was approximately $10^{-3}$.
- 25GBase-KR and 100GBase-KR4 long channel BER:
  The maximum BER seen in this case was approximately $10^{-3}$.
- 25GBase-KR and 100GBase-KR4 JTT BER:
  The maximum BER seen in this case was approximately $110^{-3}$.

### Implication:

The results of this set of tests are more severe than those expected in real-world use.

The tests with BER=$10^{-3}$ are the worst case, but statistically not the dominant value. For example, for the 90% of the cases run for the 25GBase-KR4 and 100GBase-KR4, the BER is <=$10^{-9}$. Intel does not expect results of this test to make any significant impact to real-world use cases.

### Workaround:

None.

### Status: A0=Yes, B0=Yes; NoFix

Fix planned in a future NVM release.

## 32. Lossless PEP Configuration

### Problem:

When a PCIe endpoint RX DMA engine experiences an out-of-descriptors event, it FIFOs packets for up to the time defined by DMA_CTRL.*MaxHoldTime*, at which point a timeout process begins (frame discard). DMA_CTRL.*MaxHoldTime* must be set to a value of less than 100 µs, or the PEP controller does not meet the required pause reaction time. The PEP logic does not generate an Xon pause frame when the descriptor pool exhausts.

As a consequence, when operated as a controller, the FM10000 properly receives and processes pause frames, but cannot be used to implement a lossless fabric if the RX descriptor pool may exhaust. In a port-to-port switch use case, pause is received, processed, and generated correctly for lossless operation.

### Implication:

If HOST A is connected to PEP A on the FM10000, and communicates with HOST B connected to PEP B (or an EPL) on the same FM10000, the following situation can occur during certain types of traffic flow:

1. HOST A receives a burst of data from HOST B and it is very slow to handle it.
2. PEP A is configured to NODROP mode. Therefore, PEP A stops the RX on the lines, and no new data comes in.
3. Meanwhile, HOST A transmits a burst to HOST B and reaches the threshold in the SWITCH buffer because PEP B (or the EPL) RX is slower than PEP A TX.
4. Congestion Management in the switch wants to send PAUSE packet to PEP A but cannot, as the lines are stopped by PEP A (because of NODROP configuration).

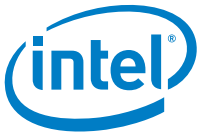
### Workaround:

In a network where PAUSE is enabled, the user should incorporate transport protocols with guaranteed delivery.

Alternatively, configure shapers in the switch scheduler to throttle the rate at which traffic is sent to the host to ensure the RX descriptor pool is not depleted. This solution requires verification that there are no conditions on the Host system preventing the FM10000 driver from servicing burst traffic.

Please contact your local Intel support personnel if there are any questions regarding your FM10000 usage model and the workaround.

### Status: A0=Yes, B0=Yes; NoFix



## 33. PCIe Delayed ACK

### Problem:

The PCIe 3.0 Base Specification requires ACKs to be sent after a ACK LATENCY TIMER TIMEOUT. These timeout values depend on the Gen speed, max payload size, and link widths as per the PCIe specification. A replay is then done by the PCIe link partner if it exceeds this timeout by 3x.

In a busy link, it is expected for an ACK to be delayed by 1 TLP time in addition to the 1x ACK LATENCY TIMER TIMEOUT, as mentioned in the PCIe specification. However, FM10000 ACK sending could be delayed by up to 2 TLPs + 1 DLLP times (about 2x of allowed time).

### Implication:

The delayed ACK is not necessarily an issue, as the PCIe specification allows 3x until the PCIe link partner initiates a REPLAY. If the FM10000 is attached to a PCIe link partner that violates the REPLAY specification, a REPLAY is sent earlier than anticipated. Since REPLAYs should be correctable errors (that can also be masked at the root complex level), they should have a minimal effect on performance.

### Workaround:

None.

**Note:** The REPLAY timings specified in the PCIe 3.0 Base Specification are more generous at higher speeds, so Gen3 operation is recommended. Disabling ACK aggregation would be a possible workaround, but a specific corner case could be reached, as described in Errata #34, "PCIe Replay Timeout Exceeded".

### Status: A0=Yes, B0=Yes: NoFix

## 34. PCIe Replay Timeout Exceeded

### Problem:

The PCIe 3.0 Base Specification requires ACKs to be sent after a ACK LATENCY TIMER TIMEOUT. These timeout values depend on the Gen speed, max payload size, and link widths as per the PCIe specification. A replay is then done by the PCIe link partner if it exceeds this timeout by 3x.

When the FM10000 encounters the specific scenario of the ACK counter ending and an arrival of a TLP, the ACK scheduling is not executed. This ACK is saved internally and is triggered on next Rx TLP only. If the next Rx TLP does not arrive or does not arrive fast enough, this causes a violation of the 3x ACK LATENCY TIMER TIMEOUT on the line, and REPLAY will happen. This could occur at any PCIe speed.

### Implication:

An unnecessary REPLAY is issued by the PCIe link partner. Since REPLAYs should be correctable errors (that can also be masked at the root complex level), they should have a minimal effect on performance.

### Workaround:

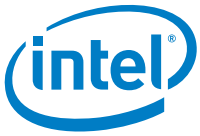
None.

### Status: A0=Yes, B0=Yes: NoFix

# 3. Software Clarifications

**Table 3-1 Summary of Software Clarifications**

| Software Clarification | Status |
|---|---|
| None | N/A |



***NOTE:*** ***This page intentionally left blank.***

# LEGAL

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

The products and services described may contain defects or errors which may cause deviations from published specifications.

Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or by visiting www.intel.com/design/literature.htm.

Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.

* Other names and brands may be claimed as the property of others.