Intel® Ethernet Controller X540 Specification Update

LAN Access Division (LAD)

Revision 2.6
May 2013

Legal

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

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

Revisions

Date	Revision	Description
May 2013	2.6	Added Specification Change #5. Added Specification Clarification #6. Added Documentation Update #2. Removed Erratum #4 (replaced by Erratum #37). Added Errata #35, #36, and #37.
February 2013	2.5	Added X540-BT2 Product Code and Device Identification information (tables 1-2 through 1-3). Added Errata #33, #34. Added Documentation Update #1.
December 2012	2.4	Revised section 1.1 - Product Code and Device Identification (added single port sku ordering information). Added Specification Clarification #5. Added Specification Change #4. Added Errata #31 and #32.
July 2012	2.3	Added Specification Clarification #3 and #4. Added Errata #29 and #30. Added Software Clarification #5.
March 2012	2.2	Added Software Clarification #4. Added Erratum #28.
March 2012	2.1	Removed Erratum #23. Added Erratum #27.
January 2012	2.0	Revision change to reflect latest software release. No hardware updates.
January 2012	1.9	Initial public release.

Note: This page intentionally left blank.

1. Introduction

This document applies to the Intel® Ethernet Controller X540 (X540).

This document is an update to a published specification, the *Intel®* Ethernet Controller X540 *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.

References to PCIe* in this document refer to PCIe v2.1 (2.5GT/s and 5GT/s).

1.1 Product Code and Device Identification

Product Code: JLX540AT2 JLX540BT2 and JLX540AT1

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

Table 1-1. Markings

Device	Stepping	Top Marking	Q-Specification	Description
X540	B0	JLX540AT2	SLJEJ	10 GbE, 2-port, 25 x 25 - Lead (Pb) free and RoHS compliant
X540	B0	JLX540AT2	SLJEK	10 GbE, 2-port, 25 x 25 - Lead (Pb) free and RoHS compliant
X540	B0	JLX540AT1		10 GbE, 1-port, 25 x 25 - Lead (Pb) free and RoHS compliant
X540	B0	JLX540BT2		10 GbE, 2-port, 25 x 25 - Lead (Pb) free and RoHS compliant

Table 1-2. Device ID

X540 Device ID Code	Vendor ID	Device ID	Revision ID
Intel® Ethernet Controller X540-AT2/X540-BT2	8086	1528	0
Intel® Ethernet Controller X540-AT1	8086	1560	0
Intel® X540 Virtual Function (Mailbox Communication)	8086	1515	0
Intel® X540 Virtual Function (Microsoft* Hyper-V)	8086	1530	0

Table 1-3. MM Numbers

Product	Tray MM#	Tape and Reel MM#	Reserved
JLX540AT2	917469		
JLX540AT2		917470	
JLX540BT2	920903	920904	
JLX540AT1	924771		

1.2 Marking Diagram

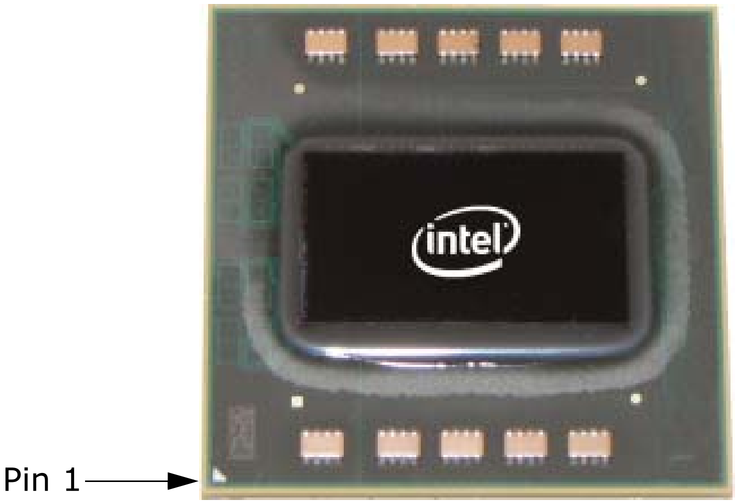
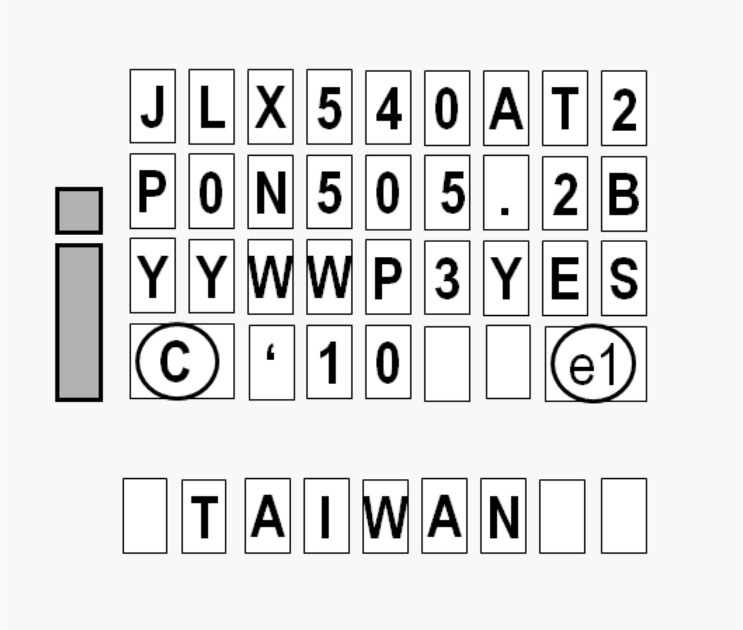


Figure 1-1. Example With Identifying Marks



- LINE1: Product code
- LINE2: Wafer lot# concatenated with Assembler vendor code
- LINE3: Assy YYWW followed by Q-spec# (no "Q") and ES for eng sample
- LINE4: Copyright, ' YY, Pb-free mark
- LINE 5: Country of origin (COO)

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 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.
Sightings	Observed issues that are believed to be errata, but have not been completely confirmed or root caused. The intention of documenting sightings is to proactively inform users of behaviors or issues that have been observed. Sightings may evolve to errata or may be removed as non-issues after investigation completes.
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.
Documentation Changes	Typos, errors, or omissions from the current published specifications. These changes will be incorporated in the next release of the specifications.
A1, B1, etc.	Stepping to which the status applies.
Doc	Document change or update that will be implemented.
Fix	This erratum is intended to be fixed in a future stepping of the component.
Fixed	This erratum has been previously fixed.
NoFix	There are no plans to fix this erratum.
Eval	Plans to fix this erratum are under evaluation.
Red Change Bar/ or Bold	This Item is either new or modified from the previous version of the document.

2. Hardware Sightings, Clarifications, Changes, Updates and Errata

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

Table 2-1. Summary of Hardware Clarifications, Changes and Errata; Errata Include Steppings

Specification Clarifications	Status
1. PCIe Completion Timeout Value Must Be Properly Set	N/A
2. Master Disable (Datasheet Section 5.2.5.3.2)	N/A
3. MCTP/DMTF Standard Compliance	N/A
4. PCIe: No Snoop is Enabled By Default	N/A
5. IPv6 Extended Headers are Parsed by the X540	N/A
6. Selecting a Rx Pool Using VLAN Filters	N/A
Specification Changes	**Status**
1. PBA Number Module — Word Address 0x15-0x16	N/A
2. Updates to PXE/iSCSI EEPROM Words	N/A
3. NC-SI Pull-Down Resistor Value Change	N/A
4. RXMTRL.UDPT Initial Value	N/A
5. The Flow Director FDIRErr(0) Bit In The Rx Descriptor Is Valid Only If The FLM Bit Is Set	N/A
Documentation Updates	**Status**
1. X540-AT2 and X540-BT2 Current Consumption	N/A
2. NC-SI_CRS_DV Pull-up/Pull-down Requirement	N/A
Errata	**Status**
1. Flow Director: Length Error Bit Not Updated on a Remove Operation	B-Step; NoFix
2. Flow Director: Filter Might Lose the Length-Error Attribute in Perfect-Match Mode	B-Step; NoFix
3. Flow Director: L4 Packet Type Gives Wrong Indication	B-Step; NoFix
4. Replaced by Erratum #37.	N/A
5. No Length Error on VLAN Packets With Bad Type/Length Field	B-Step; NoFix
6. GPRC and GORCL/H Also Count Missed Packets	B-Step; NoFix
7. FCoE: In Order to Read DMA-Rx FCoE Context, CSRs Need to Add a Dummy Write	B-Step; NoFix
8. In 100 Mb/s Link Mode, CSR Access to DMA-Rx Might Reach an Internal Timeout	B-Step; NoFix
9. MACSec: When PN=0b, a Packet is Not Dropped	B-Step; NoFix
10. MACSec Statistics: LSECRXUC, LSECRXNUSA and LSECRXUNSA Statistics Counters Not implemented According to Specification	B-Step; NoFix
11. Cause of an Interrupt Might Never be Cleared	B-Step; NoFix
12. The X540 Doesn't Meet the Timing Requirements for PAUSE Operation in 1 GbE Speed	B-Step; NoFix

Table 2-1. Summary of Hardware Clarifications, Changes and Errata; Errata Include Steppings

Errata	Status
13. The X540 Doesn't Meet the Timing Requirements for PAUSE Operation in 100 Mb/s	B-Step; NoFix
14. NC-SI Additional Multicast Packets Might Be Forwarded to the MC	B-Step; NoFix
15. SMBus: Unread Packets Received On One Port Might Cause Loss of Ability To Receive on Other Port	B-Step; NoFix
16. NC-SI: Packet Loss When the MC Sends Packets to Both Ports and One Port Has Link Down	B-Step; NoFix
17. FCoE: Exhausted Receive Context is not Invalidated if Last Buffer Size is Equal to User Buffer Size	B-Step; NoFix
18. Gen1 Tx Compliance Pattern Test Mode (Wrong Disparity)	B-Step; NoFix
19. LEDs Cannot Be Configured To Blink in LED_ON Mode	B-Step; NoFix
20. NVM Missing or Blank	B-Step; NoFix
21. Management Component Transport Protocol (MCTP) in D3 State	B-Step; NoFix
22. LED Does Not Blink In Invert Mode	B-Step; NoFix
23. In Certain Configurations, LPLU (at S5) Can Link at 1 GbE	B-Step; NoFix
24. External POR Assertion	B-Step; NoFix
25. The Allow Link Down (ALD) Feature Doesn't Work While Using Function Swap	B-Step; NoFix
26. PCIe Gen2 TX Common Return Loss	B-Step; NoFix
27. Double Image Policy Flow Is Not Applicable to PHY Image Module	B-Step; NoFix
28. Flow Director Filters Configuration Issue	B-Step; NoFix
29. PCIe Compliance Pattern is Not Transmitted When Connected to a x4/x2/x1 Slot	B-Step; NoFix
30. PF's MSI TLP Might Contain the Wrong Requester ID when a VF Uses MSI-X	B-Step; NoFix
31. PCIe Rx Termination During Power Up	B-Step; NoFix
32. EICR Bit 23 Can Be Read As Set Even If Not Initialized	B-Step; NoFix
33. NC-SI: Get NC-SI Pass-through Statistics Response Might Contain Incorrect Packet Counts	B-Step; NoFix
34. IPv4 Checksum Error Might Be Reported For Multicast Frames Over 12 KB	B-Step; NoFix
35. Flow Director: Collision Indication Can Be Cleared By Adding A New Filter	B-Step; NoFix
36. RXMEMWRAP Register Content Is Inaccurate	B-Step; NoFix
37. Flow Director Statistics Inaccuracy	B-Step; NoFix

2.1 Specification Clarifications

1. PCIe Completion Timeout Value Must Be Properly Set

The X540 Completion Timeout Value[3:0] must be properly set by the system BIOS in the Intel X540 PCIe Configuration Space Device Control 2 register (0xC8; W). Failure to do so can cause unexpected completion timeouts.

The X540 complies with the PCIe 2.0 specification for the completion timeout mechanism and programmable timeout values. The PCIe 2.0 specification provides programmable timeout ranges between 50 µs to 64 s with a default time range of 50 µs - 50 ms. The X540 defaults to a range of 16 ms - 32 ms.

The completion timeout value must be programmed correctly in PCIe configuration space (in Device Control 2 register); the value must be set above the expected maximum latency for completions in the system in which the X540 is installed. This ensures that the X540 receives the completions for the requests it sends out, avoiding a completion timeout scenario. Failure to properly set the completion timeout value can result in the device timing out prior to a completion returning.

By default, the X540 does not resend the request upon a completion timeout; however, it can be programmed to do so. In this case after the completion timeout occurs, the device assumes the original completion is lost, and resends the original request. In this condition, if the completion for the original request arrives at the X540, this results in two completions arriving for the same request, which might cause unpredictable system behavior. NVM images provided by Intel set the resend feature to off and it is recommended to not enable it.

For details on completion timeout operation, refer to the *Intel® Ethernet Controller X540 Datasheet*.

2. Master Disable (Datasheet Section 5.2.5.3.2)

The driver might time out if the *PCIe Master Enable Status* bit is not cleared within a given time. Examples that delay the clearing of the *PCIe Master Enable Status* bit include flow control, link down, or DMA completions not making it back to the DMA block. In these instances, the driver should check that the Device Status register *Transaction Pending* bit (bit 5) in the PCI config space is clear before proceeding. Also, the driver should flush the transmit data path and initiate two consecutive software resets with a delay larger than 1 µs between them.

The recommended method to flush the transmit data path is as follows:

1. Inhibit data transmission by setting the HLREG0.LPBK bit and clearing the RXCTRL.RXEN bit. This configuration avoids transmission even if flow control or link down events are resumed.
2. Set the GCR_EXT.Buffers_Clear_Func bit for 20 µs to flush internal buffers.
3. Clear the HLREG0.LPBK bit and the GCR_EXT.Buffers_Clear_Func bit.

The *Intel® Ethernet Controller X540 Datasheet* will reflect this specification clarification in the next release.

3. MCTP/DMTF Standard Compliance

The X540 MCTP protocol implementation is based on DMTF DSP0236, DSP0237 and DSP0239 Standards.

The X540 NC-SI over MCTP implementation is described on Section 11.6.4 of the The *Intel® Ethernet Controller X540 Datasheet*.

4. PCIe: No Snoop is Enabled By Default

The X540 enables the No Snoop feature by default after power on. No Snoop feature must be disabled during Rx flow software initialization if there is no intention to use it. To disable No Snoop, the CTRL_EXT.NS_DIS bit should be set to 1b.

5. IPv6 Extended Headers are Parsed by the X540

IPv6 extended headers are parsed by the x540, enabling TCP layer header recognition. As such, the IPv6 extended header fields are not taken into account for the queue classification by a flow director filter. Note that this rule does not apply for security headers and fragmentation headers. Packets with fragmentation headers miss this filter. Packets with security extended headers are parsed only up to these headers and therefore can match only filters that do not require fields from the L4 protocol.

6. Selecting a Rx Pool Using VLAN Filters

Rx Pool selection is described in 82599/x540 DS section 7.10.3.2. Note that pools are first selected by MAC Address filtering, and then by VLAN filtering. If the application is aiming to map packets to pools exclusively by their VLAN tags, it needs to replicate all incoming packets to all the different pools by their MAC Address.

In order to achieve the packet replication, PFVTCTL. Rpl_En should be set and the relevant MAC Address filtering bits should be set:

- MPSAR, PFUTA, MTA and VFTA tables.
- Relevant bits in PFVML2FLT registers – ROMPE, ROPE, BAM and MPE.

Pool selection by VLAN is then controlled by the PFVLVF and PFVLVFB registers.

2.2 Specification Changes

1. PBA Number Module — Word Address 0x15-0x16

Change:

The nine-digit Printed Board Assembly (PBA) number used for Intel manufactured Network Interface Cards (NICs) is stored in the EEPROM.

Note that through the course of hardware ECOs, the suffix field is incremented. The purpose of this information is to enable customer support (or any user) to identify the revision level of a product.

Network driver software should not rely on this field to identify the product or its capabilities.

Current PBA numbers have exceeded the length that can be stored as hex values in these two words. For these PBA numbers the high word is a flag (0xFAFA) indicating that the PBA is stored in a separate PBA block. The low word is a pointer to a PBA block.

PBA Number	Word 0x15	Word 0x16
G23456-003	FAFA	Pointer to PBA Block

The PBA block is pointed to by word 0x16.

Word Offset	Description	Reserved
0x0	Length in words of the PBA block (default 0x6).	
0x1 ... 0x5	PBA number stored in hexadecimal ASCII values.	

The PBA block contains the complete PBA number including the dash and the first digit of the 3-digit suffix. For example:

PBA Number	Word Offset 0	Word Offset 1	Word Offset 2	Word Offset 3	Word Offset 4	Word Offset 5
G23456-003	0006	4732	3334	3536	2D30	3033

Older PBA numbers starting with (A,B,C,D,E) are stored directly in words 0x15 and 0x16. The dash itself is not stored nor is the first digit of the 3-digit suffix, as it is always 0b for relevant products.

PBA Number	Byte 1	Byte 2	Byte 3	Byte 4
123456-003	12	34	56	03

2. Updates to PXE/iSCSI EEPROM Words

Change:

Words 0x30 and 0x34 (bits 2:0) are now defined as follows:

Bit(s)	Value	Port Status	CLP (Combo) Executes	iSCSI Boot Option ROM CTRL-D Menu	FCoE Boot Option ROM CTRL-D Menu
2:0	0	PXE	PXE	Displays port as PXE. Allows changing to Boot Disabled, iSCSI Primary or Secondary.	Displays port as PXE. Allows changing to Boot Disabled, FCoE enabled.
	1	Boot Disabled	NONE	Displays port as Disabled. Allows changing to iSCSI Primary/Secondary.	Displays port as Disabled. Allows changing to FCoE enabled.
	2	iSCSI Primary	iSCSI	Displays port as iSCSI Primary. Allows changing to Boot Disabled, iSCSI Secondary.	Displays port as iSCSI. Allows changing to Boot Disabled, FCoE enabled.
	3	iSCSI Secondary	iSCSI	Displays port as iSCSI Secondary. Allows changing to Boot Disabled, iSCSI Primary.	Displays port as iSCSI Allows changing to Boot Disabled, FCoE enabled.
	4	FCoE	FCOE	Displays port as FCoE. Allows changing port to Boot Disabled, iSCSI Primary or Secondary.	Displays port as FCoE Allows changing to Boot Disabled.
	7:5	Reserved	Same as disabled.	Same as disabled.	Same as disabled.
4:3	Same a before.				
5	Bit 5: formerly used to indicate iSCSI enable / disable, is no longer valid and is not checked by software.				
15:7	Same a before.				

3. NC-SI Pull-Down Resistor Value Change

Change:

Previous version 1.2 documentation suggested a 100 Ω pull-down resistor value for NC-SI. It was determined that the 100 Ω value was to strong and prevented Input High Voltage (V_{IH}) from reaching 2.0V minimum. To correct, replace all 100 Ω NC-SI pull-down resistors with 10 K Ω pull-down resistors.

4. RXMTRL.UDPT Initial Value

Change:

If the Time Sync (IEEE 1588) feature is used, the RXMTRL.UDPT field should be initialized to 0x13F.

This is fixed in ixgbe v3.11.20.

5. The Flow Director FDIRErr(0) Bit In The Rx Descriptor Is Valid Only If The FLM Bit Is Set

Change:

The FDIRErr(0) bit in Rx descriptor (length error) is valid only if the FLM bit is set (a packet matches a flow director filter) in the Extended Status of the Advanced Receive Descriptor.

2.3 Documentation Updates

1. X540-AT2 and X540-BT2 Current Consumption

Note: This information now appears in the *Intel® Ethernet Controller X540 Datasheet*, revision 2.4.

2. NC-SI_CRS_DV Pull-up/Pull-down Requirement

Pin NC-SI_CRS_DV (ball H1) described in Table 2-6 requires a 10 KΩ pull-down resistor instead of a 10 KΩ pull-up resistor. This change will be reflected in the next release of the *Intel® Ethernet Controller X540 Datasheet*.

2.4 Errata

1. Flow Director: Length Error Bit Not Updated on a Remove Operation

Problem:

In order to avoid high latency, the length of the Flow Director (FD) filters linked list is limited. The length limit is programmable (FDIRCTRL.Max-Length field). If a linked list exceeds this limit, a length error is reported in the FDIRErr.length field in the Rx descriptor.

This erratum exists because once a filter is assigned to have the length-error attribute, it stays with this attribute even if an error condition doesn't exist anymore (such as a previous filter was removed from the list).

Implication:

When the FD table is programmed with many filters while dynamic filter removal is used, the driver might get an indication for over length lists (FDIRErr.length) even though the linked lists are not too long. This indication could be used by the software driver to remove filters from the table. Note that the current software driver does not use the dynamic filter removal option.

Workaround:

Software - Reset Flow Director (FD) tables when max-length indication is observed, or hold image of all the FD table and update the FD table (holding the image is less recommended).

The FD table is the hardware internal memory structure. Clearing this table means that the packet buffer memory of FD is cleared and linked to the empty link-list and head/tail CSRs are initialized. All other CSR are re-configured by software.(see Datasheet section 7.1.2.7).

Status:

B-Step; NoFix

2. Flow Director: Filter Might Lose the Length-Error Attribute in Perfect-Match Mode

Problem:

In order to avoid high latency, the length of the Flow Director (FD) filters linked list is limited. The length limit is programmable (FDIRCTRL.Max-Length field). If a linked list exceeds this limit, a length error is reported in the FDIRErr.length field in the Rx descriptor.

In some rare cases a filter that has the length-error attribute might change the attribute to No-Length-Error. As a result, the FD table includes long lists, which are not reported to software. Once a packet matches these filters it causes a slightly higher latency in the device.

Implication:

There is no expected impact. In the cases where this indication is important, we expect other filters to indicate length-error.

FD tables are reset, which lowers the probability of reaching this case.There is also no impact to packet counters.

Workaround:

None.

Status:

B-Step; NoFix

3. Flow Director: L4 Packet Type Gives Wrong Indication

Problem:

The MSB of the L4 Packet Type (L4TYPE) field in the Flow Director Filters Command Register (FDIRMC[6]) might give a wrong value during read access.

The flow director filters operate with the correct parameters.

Implication:

No impact on functionality. Software should ignore the read result of this bit.

Workaround:

None. Make sure that in a read to verify successful write, this bit is ignored.

Status:

B-Step; NoFix

4. Replaced by Erratum #37.

5. No Length Error on VLAN Packets With Bad Type/Length Field

Problem:

The X540 does not assert length error for VLAN packets that have a bad *Type/Length* field in the MAC header.

Implication:

There is no impact on system level performance. The packets are posted to the host as with any other packets.

Workaround:

None.

Status:

B-Step; NoFix

6. GPRC and GORCL/H Also Count Missed Packets

Problem:

GPRC (Good Packets Received Count) and GORCL/H (Good Octets Received Count) count missed packets and missed packets bytes.

Implication:

None.

Workaround:

Statistics are available indirectly for these registers. This workaround is included in Intel drivers.

- For GPRC — Subtract MPC (Missed Packet Count) from GPRC. Alternatively, use QPRC.
- For GORCL/H — use QBRCL/H (Quad Bytes Received).

Status:

B-Step; NoFix

7. FCoE: In Order to Read DMA-Rx FCoE Context, CSRs Need to Add a Dummy Write

Problem:

There is a need to add a dummy write before the read of an FCoE context CSRs (FCDMARW) to avoid context corruption.

Implication:

No Impact.

Workaround:

Write FCDMARW twice while having the required FCoE read index valid and 0b in the RE and WE bits.

Note: No workaround in current Intel drivers.

Status:

B-Step; NoFix

8. In 100 Mb/s Link Mode, CSR Access to DMA-Rx Might Reach an Internal Timeout

Problem:

In 100 Mb/s link mode, internal clocks are slower, and access of an internal register can lead to timeout.

Implication:

An unknown value is returned on the PCI Express* (PCIe*) interface.

Workaround:

Software — in 100 Mb/s link mode programmers need to disable aggregation in DMA-Rx (set RDRXCTL.AGGDIS=1b) and to extend the PCIe timeout extension to 32 µs (set PCIEMISC.TO_extension to 011b).

When aggregation is disabled, expect an impact on performance for packets below 128 bytes in length.

Note: Programmers should not increase the timeout extension beyond 32 µs to avoid PCIe system issues.

Status:

B-Step; NoFix

9. MACSec: When PN=0b, a Packet is Not Dropped

Problem:

According to the MACSec specification, frames with PN=0 (packet number) in the sectag should be counted as bad tags/packets. The X540 will consider these packets as late packets and they will be incorrectly identified as a late packets instead of a bad tag/ packets. So they are dropped, but for the wrong reason (late packet instead of bad tagged).

Implication:

MACSec Rx statistic counters might report inaccurate values.

Workaround:

None.

Note: No workaround in current Intel software device drivers.

Status:

B-Step; NoFix

10. MACSec Statistics: LSECRXUC, LSECRXNUSA and LSECRXUNSA Statistics Counters Not implemented According to Specification

Problem:

InPktsUnchecked (LSECRXUC) statistic is not provided (LSECRXUC does not count correctly).

InPktsNotUsingSA (LSECRXNUSA) and InPktsUnusedSA (LSECRXUNSA) should be defined per SA. In this implementation, these are captured by a single counter.

Implication:

Statistics defined in the MACSec standard cannot be provided.

Workaround:

None.

Note: No workaround in current Intel drivers. Once MACSec is included in Intel drivers, this workaround will be applied.

Status:

B-Step; NoFix

11. Cause of an Interrupt Might Never be Cleared

Problem:

If the cause of an interrupt is set by the Extended Interrupt Cause Set (EICS) register writing just before the interrupt line is set, then it might not be cleared. This means that there might be a deadlock that prevents the interrupt line from rising.

This erratum only occurs when all three modes referenced are used at the same time: non-PBA mode, Auto Clear (of the cause), No Auto Mask.

PBA is Pending Bit Array mode. During this mode the device is able to capture additional interrupts during the interval between initial interrupt and driver access to the device.

Implication:

The X540 stops issuing interrupts.

Workaround:

When operating using the above configurations, software should manually clear the cause by writing a 1b to the specific bit in the relevant EICR/EICR1/EICR2/VTEICR0-63 register (after the interrupt occurs and the EICS was written). This workaround is included in Intel drivers.

Status:

B-Step; NoFix

12. The X540 Doesn't Meet the Timing Requirements for PAUSE Operation in 1 GbE Speed

Problem:

In 1 GbE speed, the X540 responds to a received pause frame after a longer time than defined in the IEEE 802.3 specification.

Implication:

Specification conformance. The response gap is small.

Workaround:

None.

Status:

B-Step; NoFix

13. The X540 Doesn't Meet the Timing Requirements for PAUSE Operation in 100 Mb/s

Problem:

In 100 Mb/s speed, the X540 responds to a received pause frame after a longer time than defined in the IEEE 802.3 specification.

Implication:

Specification conformance. No system impact with low traffic.

Workaround:

None.

Status:

B-Step; NoFix

14. NC-SI Additional Multicast Packets Might Be Forwarded to the MC

Problem:

If the MC enables multicast filtering for IPv6 neighbor advertisement and/or IPv6 router advertisement, additional multicast packets are forwarded to the MC. The additional packets forwarded are:

1. Packets with the ICMPv6 header's message type: 135, 137.
2. IPv6 neighbor advertisement.
3. IPv6 router advertisement.

Implication:

Additional packets might be forwarded to the MC.

Workaround:

The MC should filter the different multicast packets.

Status:

B-Step; NoFix

15. SMBus: Unread Packets Received On One Port Might Cause Loss of Ability To Receive on Other Port

Problem:

The X540's two ports share an internal memory. When packets are received by one of the ports and not read by the MC, they are stored in the shared memory. When this memory fills up, no more packets can be received from either ports.

Implication:

Loss of packets. The MC should be aware of the previous behavior.

Workaround:

1. Make use of a SMBus alert timeout mechanism.
2. Momentarily disable receives by the other port.

Status:

B-Step; NoFix

16. NC-SI: Packet Loss When the MC Sends Packets to Both Ports and One Port Has Link Down

Problem:

NCSI Rx (MC-to-LAN) FIFO is shared between both ports. When one of the LAN port’s Tx buffer is congested because of link failure or flow control, the NCSI Rx FIFO gets congested and as a result the packets for the second port also get dropped and are not sent to the LAN.

Implication:

Loss of packets. The MC should be aware of the problem.

Workaround:

The MC should monitor the link status and stop sending packets to a specific port if the link is down.

Status:

B-Step; NoFix

17. FCoE: Exhausted Receive Context is not Invalidated if Last Buffer Size is Equal to User Buffer Size

Problem:

If the last buffer of an FCoE context doesn’t have sufficient room for the FC payload, the context is considered exhausted and must be invalidated by hardware.

The FCoE context is not invalidated as required under the following scenarios:

- FCoE last buffer size (FCDMARW.LASTSIZE) equals the exact user buffer size (FCBUFF.BUFFSIZE).
- FCoE DDP last payload byte in a mid packet written to the last byte of the last allocated buffer (the packet fills in the exact buffer value).
- Extra FCoE packet(s) are received in the problematic context.

Implication:

- Invalid host memory access.
- Hardware does not invalidate FCoE context when exhausted and does not assert error status to software.

Workaround:

FCoE context last buffer must be smaller than the context buffer size.

If it’s necessary to configure a last buffer to equal buffer size, the following flow should be used:

- Allocate the extra user-buffer in the context list. Set it in the context buffer list and then increment FCBUFF.BUFFCNT to reflect a possible usage of an additional buffer.
- Set FCDMARW.LASTSIZE = 0x1.
- If flow ends and the extra buffer is used, the flow is invalid and exhausted.

If FCDMARW.LASTSIZE = FCBUFF.BUFFSIZE, the number of used DDP buffers is limited to 255. The FCBUFF.BUFFCNT value should be programmed for less than 256.

Note: The workaround is included in ixgbe v3.2.10 and in Intel’s Windows* drivers, starting with Release 16.4 version 2.9.66.0.

Status:

B-Step; NoFix

18. Gen1 Tx Compliance Pattern Test Mode (Wrong Disparity)

Problem:

Tx compliance is a test mode in which the X540 transmits a continuous sequence of symbols for electrical characterization purposes. In X540 A0, Gen1, this pattern is not compliant with the specification. The transmitted symbol sequence is COM-, D21.5, COM-, D10.2 (instead of COM-, D21.5, COM+, D10.2). Note that in Gen2, the test mode works correctly.

Implication:

PCIe specification compliance issue.

Workaround:

Use JTAG flow to configure the PCIe core to transmit the right pattern.

Status:

B-Step; NoFix

19. LEDs Cannot Be Configured To Blink in LED_ON Mode

Problem:

When the LEDx_Mode field of a specific LED is set to 1110b in the LEDCTL register (0x00200), the respective LED is in LED_ON mode. This LED should be always asserted when the mode is set to LED_ON. The LED should also blink based on the LEDx_BLINK setting; however, due to a device limitation, the LED does not blink regardless of the LEDx_BLINK value.

Implication:

LEDs cannot be configured to blink in LED_ON mode.

Workaround:

The software driver should switch between LED_ON and LED_OFF mode to make the LED blink.

Status:

B-Step; NoFix

20. NVM Missing or Blank

Problem:

Due to analog defaults (overridden with NVM load), the X540 might not reach PCIe link with certain link partners if an NVM is missing or blank.

Implication:

The X540 might fail to show up on the PCIe bus when an NVM is missing or blank.

Workaround:

Pre-program the NVM prior to powering on the X540.

Status:

B-Step; NoFix

21. Management Component Transport Protocol (MCTP) in D3 State

Problem:

MCTP Get Inventory command not functional during D3.

Implication:

MCTP Get Inventory command not functional during D3.

Workaround:

None.

Status:

B-Step; Fixed

22. LED Does Not Blink In Invert Mode

Problem:

LEDx_IVRT bit in LEDCTL register (offset 0x00200) is ignored if the respective LEDx_BLINK bit is set. This issue is relevant only if LEDx_MODE is programmed to one of the modes where LEDx_BLINK is used (MAC_ACTIVITY, FILTER_ACTIVITY, LINK_UP, LINK_1G, and LINK_10G).

Implication:

LED stays lit during idle time.

Workaround:

If LEDx_IVRT must be set together with a blink effect, use LINK_ACTIVITY mode instead of the modes using LEDx_BLINK (MAC_ACTIVITY, FILTER_ACTIVITY, LINK_UP, LINK_1G, and LINK_10G).

Status:

B-Step; NoFix

23. In Certain Configurations, LPLU (at S5) Can Link at 1 GbE

Problem:

In certain configurations, LPLU (at S5) can link at 1 GbE instead of 100 Mb/s if the following conditions take place:

1. Setting LPLU disable 10 GbE and LPLU disable 1 GbE in the NVM.

2. Linking at 1 GbE at S0.

3. Going down to S5.

4. Link is 1 GbE instead of 100 Mb/s.

Implication:

LPLU links at 1 GbE instead of 100 Mb/s.

Workaround:

1. Set LPLU to 10 GbE disable in the NVM (links at 100 Mb/s if the link partner is capable).
2. Set both LPLU disable 10 GbE and LPLU disable 1 GbE in the NVM. Make sure your S0 link is 10 GbE and when going down to S5 it ends with a 100 Mb/s link.

Status:

B-Step; Fixed

Note: NVM starting with v4.3 and later will incorporate this fix.

24. External POR Assertion

Problem:

When the BYPASS_POR signal is asserted high (Power On Reset bypass), the X540 might not established link due to PHY reset limitations.

Implication:

When set to 1b, BYPASS_POR disables the internal POR circuit and uses the LAN_PWR_GOOD pin as a POR indication. Note that this might not work due to PHY reset limitations.

Workaround:

Setting BYPASS_POR to 0b (instead of 1b), maintains the functionality of using the LAN_PWR_GOOD pin as a POR indication.

Status:

B-Step; NoFix

25. The Allow Link Down (ALD) Feature Doesn't Work While Using Function Swap

Problem:

The ALD feature doesn't work when function swap is applied.

When one function in D3, a second in D0, and when ALD is applied to the D3 function.

Implication:

The ALD feature doesn't work while using port swap.

Workaround:

None.

Status:

B-Step; NoFix

26. PCIe Gen2 TX Common Return Loss

Problem:

Spec com: < -6dB @ 50 MHz-2.5 GHz; Some lanes showed failures @ 2.5 GHz - the violation is 0.5-2.45 db above specification limit.

Implication:

No impact on customers.

Workaround:

None.

Status:

B-Step; NoFix

27. Double Image Policy Flow Is Not Applicable to PHY Image Module

Problem:

NVM double image policy flow is used to protect the update of big Flash modules, but it is not applicable to the PHY image module.

Implication:

The PHY image module update is not protected by the double image policy.

Workaround:

None

Status:

B-Step; NoFix

28. Flow Director Filters Configuration Issue

Problem:

Before an X540 receive path enable, the default value of both RXCTRL.RXEN and SECRXCTL.RX_DIS is zero. If the flow director filters are configured in this state, the receive data buffer might not be configured correctly.

Implication:

Receive hang.

Workaround:

If RXCTRL.RXEN is clear, set SECRXCTL.RX_DIS and wait for a SECRXSTAT.SECRX_RDY indication before configuring the flow director filters.

This workaround is implemented in the Intel ixgbe driver 3.8.21.

Status:

B-Step; NoFix

29. PCIe Compliance Pattern is Not Transmitted When Connected to a x4/x2/x1 Slot

Problem:

If the PCIe compliance pattern is activated by setting the *Enter Compliance* bit in the Link Control 2 register, the X540 is able to transmit the compliance pattern only if it is connected to a x8 slot. If it is connected to a x4, x2 or x1 slot, the unconnected lanes falsely cause a premature exit from the compliance state and the pattern is not transmitted.

If a passive test load is applied on all lanes, the X540 goes to a compliance state and transmits the pattern accordingly, regardless of the internal lane width configuration.

Implication:

A PCIe compliance pattern cannot be transmitted if the X540 is connected to a x4 or narrower PCIe slot.

Workaround:

The X540 is still able to transmit the compliance pattern when connected to a x4/x2/x1 slot if entering the Polling.Compliance state due to detecting eight consecutive TS1 Ordered Sets in Polling.Active with the *Compliance Receive* bit (bit 4 of Symbol 5) asserted.

Status:

B-Step; NoFix

30. PF’s MSI TLP Might Contain the Wrong Requester ID when a VF Uses MSI-X

Problem:

When using IOV, if a PF uses MSI interrupts and one or more VFs use MSI-X interrupts, some of the MSI TLPs for the PF might contain the wrong Requester ID.

Implication:

There could be missing interrupts on the PF since the incorrect Requester ID could result in the virtualization mechanism misrouting or dropping TLPs.

Workaround:

If any VFs use MSI-X, all PFs should also use MSI-X.

Status:

B-Step; NoFix

31. PCIe Rx Termination During Power Up

Problem:

According to the PCIe Specification, the receiver is required to present a high-impedance termination any time adequate power is not provided to the receiver until the device is out of reset. The X540 does not present high-impedance termination for a period of a few hundred microseconds right after power on, and then restores the high-impedance termination right after that for at least 100 milliseconds until it's out of reset.

Typically, this issue is unseen if the root complex exits reset at the same time or after the X540. In systems where the root complex exits reset before the X540 is powered up, this issue might cause a false detection of the X540 Rx by the root complex, and its LTSSM moves to the polling state. However, since the X540 is under reset for at least 100 milliseconds, the root complex LTSSM should time out and it won't affect the PCIe link connection.

Implication:

Momentary detection of the X540 receiver when it is still under reset. However, for a fully compliant root complex there is no implication.

Workaround:

Synchronizing the reset of the X540 and an upstream device avoids any false Rx detection.

Status:

B-Step; NoFix

32. EICR Bit 23 Can Be Read As Set

Problem:

Implication:

Unexpected software device driver behavior.

Workaround:

The software device driver should ignore this bit.

Status:

B-Step; NoFix

33. NC-SI: Get NC-SI Pass-through Statistics Response Might Contain Incorrect Packet Counts

Problem:

The X540 maintains packet counters that are used in the Get NC-SI Pass-through Statistics Response. These counters are halted during PCIe reset.

Implication:

If a PCIe reset has occurred since the previous Get NC-SI Pass-through Statistics Response, the packet count values could be lower than the actual packet counts.

Workaround:

The packet counts in the Get NC-SI Pass-through Statistics Response can be used for debug purposes, but they should not be used for maintaining reliable statistics.

Status:

B-Step; NoFix

34. IPv4 Checksum Error Might Be Reported For Multicast Frames Over 12 KB

Problem:

IPE (IPv4 Checksum Error) might be rarely set in the Rx descriptor of multicast frames over 12 KB even though their checksum is valid.

Implication:

An IPE (IPv4 Checksum Error) error can incorrectly be reported by the X540.

Workaround:

To avoid the erratum condition, limit the size of jumbo frames to less than or equal to 12 KB.

If using jumbo frames over 12 KB, software should re-calculate the IPV4 Header Checksum if RDESC.IPE is set.

The Intel Windows* and Linux* drivers limit the size of jumbo frames to less than or equal to 9 KB and are not exposed to this erratum.

Status:

B-Step; NoFix

35. Flow Director: Collision Indication Can Be Cleared By Adding A New Filter

Problem:

A Flow Director collision indication of the last Signature filter can be unintentionally cleared by adding a subsequent Signature filter.

Implication:

Flow Director collision indication is missing.

Workaround:

None.

Status:

B-Step; NoFix

36. RXMEMWRAP Register Content Is Inaccurate

Problem:

RXMEMWRAP register (0x03190) content is inaccurate:

- RX Buffer Wrap Around Counter values could be inaccurate.
- RX Buffer Empty bits are not reliable in the presence of FCoE or TCP-no-payload packets.

Implication:

Incorrect status read.

Workaround:

Use the RXUSED register for an indication as to whether or not the Rx buffer is empty.

Status:

B-Step; NoFix

37. Flow Director Statistics Inaccuracy

Problem:

- FDIRMATCH (0x0EE58) should count the number of packets that matched any flow director filter.
- FDIRMISS (0x0EE5C) should count the number of packets that missed matching any flow director filter.
- FDIFSTAT.FADD (0x0EE54, bits 7:0) should count the number of failed added filters due to no space in the filter table.

These counters might be incremented by two instead of one.

Implication:

The counters can't be used for exact statistics. Counters should be used as an approximate indication on miss/match/failed addition of filters.

Workaround:

None.

Status:

B-Step; NoFix

3. Software Clarifications

Table 3-1. Summary or Software Clarifications

Software Clarifications	Status
1. While in TCP Segmentation Offload, Each Buffer is Limited to 64 KB	N/A
2. RSC Performance Tradeoff	N/A
3. Serial Interfaces Programmed By Bit Banging	N/A
4. Identity Network Adapter Port By Blinking LED	N/A
5. PF/VF Drivers Should Configure Registers That Are Not Reset By VFLR	N/A

1. While in TCP Segmentation Offload, Each Buffer is Limited to 64 KB

Problem Description:

The X540 supports 256 KB TCP packets; however, each buffer is limited to 64 KB since the data length field in the descriptor is only 16 bits. This restriction can complicate things for the driver if the operating system passes down a scatter/gather element greater than 64 KB in length. This issue can be avoided by limiting the offload size to 64 KB.

Investigation has concluded that the increase in data transfer size does not provide any noticeable improvements in LAN performance. As a result, Intel network software drivers limit the data transfer in all drivers to 64 KB.

Please note that Linux operating systems only support 64 KB data transfers.

2. RSC Performance Tradeoff

Problem Description:

The RSC feature is used to merge receive frames into the same descriptor structure with a shared header, improving receiving packet performance.

It should be noted that if small Rx data buffers are used (2 KB), RSC may involve a high rate of partial cache line PCIe transactions, which have a performance penalty from a memory access perspective.

In overloaded systems (more than 2 x 10 Gb/s LAN ports traffic load) the use of RSC may adversely affect Rx data throughput. Therefore, there is a performance tradeoff regarding the usage of the RSC feature.

To improve throughput in overloaded systems, the user can use large receive data buffers (larger than 2 KB or may opt to turn of RSC.

3. Serial Interfaces Programmed By Bit Banging

Problem Description:

When bit banging on a serial interface (such as SPI, I^{2}C, or MDIO), it is often necessary to perform consecutive register writes with a minimum delay between them. However, simply inserting a software delay between the writes can be unreliable due to hardware delays on the CPU and PCIe interfaces. The delay at the final hardware interface might be less than intended if the first write is delayed by hardware more than the section write. To prevent such problems, a register read should be inserted between the first register write and the software delay. For example: write, read, software delay, write.

4. Identity Network Adapter Port By Blinking LED

Problem Description:

Intel device drivers and supported tools include a feature that provides network adapter port identification by blinking LED2. This feature assumes that LED2 is connected as the Link/Activity LED as recommended in the reference schematics.

5. PF/VF Drivers Should Configure Registers That Are Not Reset By VFLR

Problem Description:

The following registers are not reset by VFLR and need to be configured by PF or VF in case of a change to a new configuration (such as VF OS transition):

VFRDH/T, VFTDH/T, VFPSRTYPE, VFSRRCTL, VFRXDCTL, VFTXDCTL, VFTDWBAL/H, VFDCA_RXCTRL, and VFDCA_TXCTRL.

NOTE: This page intentionally left blank.