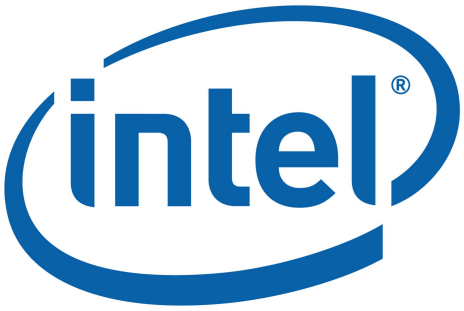


Intel® Ethernet Controller X710-TM4/ AT2

Specification Update

Ethernet Networking Division (ND)

October 2019

Revision 1.0
615119-002

Revision History

Revision	Date	Comments
1.0[1]	October 30, 2019	**Errata added or updated:** • 53. PCIe Replay Timer Can Occasionally be Lower Than PCIe Spec Requirements (Added)
0.1	August 29, 2019	Initial release (Intel Public).

1. There were no versions of this document released between Revision 0.1 and Revision 1.0.

1. Introduction

This document applies to the Intel® Ethernet Controller X710-TM4/AT2 (X710-TM4/AT2).

This document is an update to a published specification, the *Intel® Ethernet Controller X710-TM4/AT2 Datasheet*. It is intended for use by system manufacturers and software developers. All product documents are subject to frequent revision and new order numbers might apply. New documents might be added. Be sure you have the latest information before finalizing your design.

References to PCIe Express* (PCIe*) in this document refer to PCIe v3.0 (2.5GT/s, 5GT/s, and 8GT/s).

For more information on supported features, see the *Intel® Ethernet Controller X710-TM4/AT2 Feature Support Matrix*. This document is updated periodically. Please ensure that you have the latest version.

1.1 Product Code and Device Identification

Product Codes: X710-TM4 and X710-AT2.

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

Table 1-1 Markings

Device	Stepping	Top Marking	Q-Specification[1]	Description
X710-TM4	B1	EZX710TM4	QURS	Intel® Ethernet Controller X710 for 10 GBASE_T, 10 GbE SFP+ and 10 GbE backplane
X710-AT2	B1	EZX710AT2	QUZ8	Intel® Ethernet Controller X710 for 10GBASE-T

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

Table 1-2 Device IDs

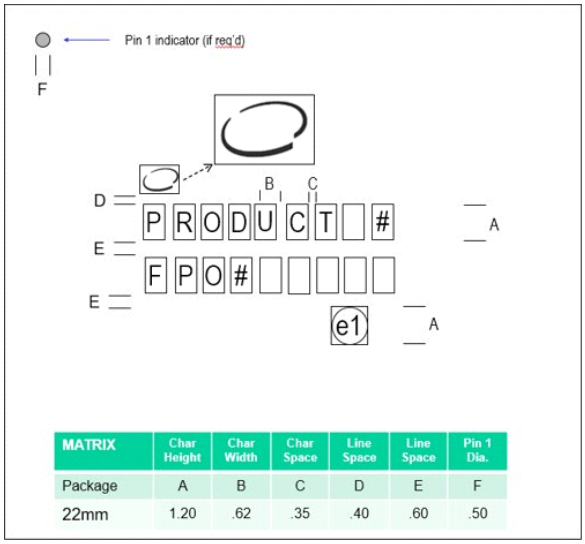
Branding String	Interface Type	Device ID	Vendor ID	Revision ID
				B1
Intel® Ethernet Controller X710 for 10 GbE SFP+	SFI	0x104E	0x8086	0x01
Intel® Ethernet Controller X710 for 10 GbE backplane	KR/SFI	0x104F	0x8086	0x01
Intel® Ethernet Controller X710 for 10GBASE-T	10GBASE-T	0x15FF	0x8086	0x01

Table 1-3 MM Numbers

Product	S-Specification	Tray MM#	Tape and Reel MM#
X710-TM4	SLMM6	976476	
X710-TM4	SLMM7		976477
X710-AT2	SLMR5	980538	
X710-AT2	SLMR6		980539

1.2 Marking Diagrams

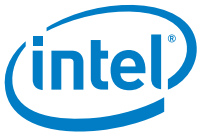
Figure 1-1 Example Component with Identifying Marks



MATRIX	Char Height	Char Width	Char Space	Line Space	Line Space	Pin 1 Dia.
Package	A	B	C	D	E	F
22mm	1.20	.62	.35	.40	.60	.50

Figure 1-2 Marking Diagram

- LINE1:Swirl Logo.
- LINE2: Product Number. For example, EZX710TM4.
- LINE3: Lot Trace Code. For example, T641AS01.
- LINE4: Pb Free Symbol.



1.3 Nomenclature Used in This Document

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

Table 1-4 Nomenclature

Name	Description
A0, B0, etc.	Stepping to which the status applies.
Device	Where applicable, indicates the specific device to which an errata applies. Possible values are: • X710 — 10 GbE device
Doc	Document change or update that will be implemented.
Documentation Changes	Typos, errors, or omissions from the current published specifications. These changes will be incorporated in the next release of the specifications.
Errata	Design defects or errors. Errata might 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.
Eval	Plans to fix this erratum are under evaluation.
Fix Planned	This erratum is intended to be fixed in a future stepping of the component.
Fix Planned in NVM	This erratum is intended to be fixed in a future NVM version.
Fixed	This erratum has been fixed.
Fixed in NVM	This erratum has been fixed in NVM X.XX.
NoFix	There are no plans to fix this erratum.
Software Clarifications	Applies to Intel drivers, EEPROM loads, etc.
Specification Changes	Modifications to the current published specifications. These changes will be incorporated in the next release of the specifications.
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.

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. SFP+ Cable EEPROM Overwrite on Power Down	N/A
2. PCIe Re-timers Might Cause Replay Timer Timeout Correctable Errors	N/A
3. I^2C Minimum Time Between Transactions	N/A
4. Qualified Module Bit	N/A
5. L2 Padding and L4 Checksum Offloads	N/A
6. Malicious Driver Detection MAX_BUFF Event	N/A
7. Intel® Ethernet Controller X710-TM4/AT2 Throughput Limit	N/A
8. Small Packets Performance Degrade when Using Private VLAN	N/A
9. The X710-TM4/AT2 Packet Drop Rate is Limited to 27 MPPS	N/A
10. Expansion ROM is Exposed in Blank Flash Programming Mode	N/A

Table 2-2 Summary of Specification Changes

Specification Change	Status
1. Ingress Mirroring Cannot be Changed on the Fly	N/A
2. RSS Field Selection is Globally Defined	N/A
3. NC-SI Get Controller Packet Statistics Command Limitations	N/A
4. SMBus Minimum Packet Size	N/A
5. Support of the Admin Queue Command “Set Loopback modes command (opcode:0x0618)”	N/A
6. Logging of PCIe Correctable Receiver Error	N/A
7. PRTPM_SAL and PRTPM_SAH are Re-loaded from NVM on PCIe Reset	N/A
8. VEB Statistics Disable NVM Bit	N/A
9. Unicast Hash Filtering Removal	N/A
10. Input Reference Clock Rise/Fall Times	N/A
11. Set Local LLDP MIB when DCBX Agent is Disabled or Stopped	N/A
12. Teredo UDP Tunneling Offload Support	N/A
13. GLQF_PCNT Counters	N/A

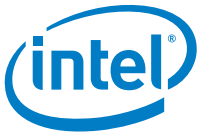


Table 2-2 Summary of Specification Changes [continued]

Specification Change	Status
14. Flash CS Negation Time	N/A
15. Parsing of MPLS Headers	N/A

Table 2-3 Summary of Documentation Updates

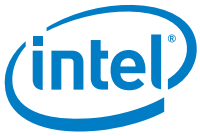
Documentation Update	Status
None	N/A

Table 2-4 Summary of Errata; Errata Include Steppings

Erratum	Status
1. TX Performance Degradation for Small Cloud Packets	B1=Yes; NoFix
2. PCIe Subsystem ID Incorrectly Reported for All PCI Functions Except Function 0	B1=Yes; NoFix
3. Illegal Byte Error Statistical Counter Inaccuracy	B1=Yes; NoFix
4. Receive Performance Degradation with Specific Cloud Header	B1=Yes; NoFix
5. MCTP Discovery Error when Replacing Active PF	B1=Yes; NoFix
6. RX Queue Disable is Reported Done Before It is Disabled	B1=Yes; NoFix
7. TX Descriptor Might be Read Twice	B1=Yes; NoFix
8. Immediate Interrupts are Delayed in Very Loaded System	B1=Yes; NoFix
9. ECRC Bits are Not RO when ECRC is Disabled	B1=Yes; NoFix
10. NC-SI I/Os Output Rise Slew Rate is Higher Than Specification	B1=Yes; NoFix
11. TC Strict Priority Does Not Work as Expected	B1=Yes; NoFix
12. Management-only Packets Cannot be Ignored for Wake-Up	B1=Yes; NoFix
13. Common Clock Configuration Bit Specification Compliance	B1=Yes; NoFix
14. Low Latency TC Might be Momentarily Starved	B1=Yes; NoFix
15. Round Robin (RR) Bandwidth Distribution is Traffic Dependent	B1=Yes; NoFix
16. L2 Tag Stored in the Wrong RX Descriptor Field	B1=Yes; NoFix
17. Internal VLAN is Not Reflected in RX Descriptor	B1=Yes; NoFix
18. Transmit Queue Group with Single Queue Enabled Performance	B1=Yes; NoFix
19. A Switching Table Might Reduce Small Packets Performance	B1=Yes; NoFix
20. Set Binding Command is Not Functional for IPv4	B1=Yes; NoFix
21. Cloud Traffic Over VEB is Transmitted to LAN	B1=Yes; NoFix
22. VLAN Prune is Not Functional	B1=Yes; NoFix
23. INTENA_MSK Setting Might Clear Interrupt	B1=Yes; NoFix
24. Manageability Checksum Filtering of IPv6 Packets	B1=Yes; NoFix
25. Link Remains Up During Power Saving State	B1=Yes; NoFix

Table 2-4 Summary of Errata; Errata Include Steppings [continued]

Erratum	Status
26. PRTDCB_RUP2TC and PRTDCB_TC2PFC are Not Writable	B1=Yes; NoFix
27. AER Header Log Might be Invalid	B1=Yes; NoFix
28. A CfgWr to a VF TLP with Error Might Generate an Error Message with Wrong VF Number	B1=Yes; NoFix
29. No LAN-to-BMC Pass-through Traffic in Dr State	B1=Yes; NoFix
30. MNG Packets are Dropped while a Function-Level Reset to PF 0 is in Progress	B1=Yes; NoFix
31. DCBx Resume of a Port Affects Other Ports	B1=Yes; NoFix
32. A Global SDP Might be Affected by a Specific Port Power State	B1=Yes; NoFix
33. Legacy SMBus: Failure to De-assert Alert Signal when Not Using ARA Cycle	B1=Yes; NoFix
34. Get Link Status AQ Command Might Return Incorrect Status	B1=Yes; NoFix
35. A Function-level Reset Might Affect Other Functions	B1=Yes; NoFix
36. Rx Packet Drops Even with Priority Flow Control	B1=Yes; NoFix
37. DCBx Configuration Might Change After LLDP Stops	B1=Yes; NoFix
38. PCIe Interrupt Status Bit	B1=Yes; NoFix
39. Glitch on SDP Outputs During GLOBR	B1=Yes; NoFix
40. Function-Level Reset Fails to Complete	B1=Yes; NoFix
41. Incorrect Flexible Payload Extraction from Flow Director Filter to Receive Descriptor	B1=Yes; NoFix
42. Aux Power Detected Bit Not Implemented	B1=Yes; NoFix
43. SGMII Receiver Sensitivity	B1=Yes; NoFix
44. IEEE 802.3 Clause 73 AN Does Not Support Parallel Detection	B1=Yes; NoFix
45. IEEE 802.3 Clause 73 AN Echoed Nonce Field is Zero	B1=Yes; NoFix
46. KR Transmitter Output Waveform Violations	B1=Yes; NoFix
47. 10GBASE-KR wait_timer Value Smaller Than Specification	B1=Yes; NoFix
48. Receive Queue Disable Can Get Stuck	B1=Yes; NoFix
49. Set DCB Parameters AQC (Opcode 0x303) Might Return EINVAL Even when It Succeeds	B1=Yes; NoFix
50. Receive IP Packets in a Low-Latency Traffic Class Are Not Fully Processed	B1=Yes; NoFix
51. Activity LED Might Blink Regardless if Link is Up or Down for a Port	B1=Yes; NoFix
52. EMP Reset After Using Intel QCU Tool	B1=Yes; NoFix
53. PCIe Replay Timer Can Occasionally be Lower Than PCIe Spec Requirements	B1=Yes; NoFix



2.1 Specification Clarifications

1. SFP+ Cable EEPROM Overwrite on Power Down

After PCIe Reset, the normal X710-TM4/AT2 operation might include I^2C transactions to the SFP+ cable EEPROM. Under certain timing conditions, these transactions might coincide with power ramping down. This could lead to an unintentional I^2C write command, causing the cable EEPROM contents to be overwritten on cables that do not have write protection, and making them inoperable.

NVM v7.1 provides updated timing of the I^2C transactions to avoid unintentional modification of the cable EEPROM contents.

2. PCIe Re-timers Might Cause Replay Timer Timeout Correctable Errors

The addition of PCIe re-timers add to the total channel latency. According to PCI-SIG ECN extension devices, latency is defined as "the time from when the last bit of a Symbol is received at the input pins of one Pseudo Port to when the equivalent bit is transmitted on the output pins of the other Pseudo Port". The ECN allows for a maximum of 64 symbol x latency per PCIe re-timer for 8 GT/s speed.

The PCIe ACK/NACK round trip delay is incremented according to the number of re-timers used in Tx/Rx lanes. The extra delay added by a re-timer might cause the X710-TM4/AT2 Replay_Timer to expire, causing replay timer timeout correctable errors. X710-TM4/AT2 design does not take into consideration the extension devices ECN.

If a design must include re-timers, and if Replay_Timer timeout correctable errors are seen, please contact your Intel representative for support.

3. I^2C Minimum Time Between Transactions

The SFF-8431 Specification requires that the minimum time between STOP and START on an I^2C bus (Tbuf) should be at least 20 μs. The time measured in the X710-TM4/AT2 is less than required by the specification. No functional implication should be expected.

4. Qualified Module Bit

According to the *Intel® Ethernet Controller X710-TM4/AT2 Datasheet*, the Qualified Module bit in Admin Queue Get link status response (Byte 3, bit 7) is cleared when the module is not found in pre-configured list of qualified modules. In addition, this bit can be cleared in case that there is a contradiction between the module and device configuration. For example, NVM of BASE-T with external optical module.

5. L2 Padding and L4 Checksum Offloads

When using UDP/TCP checksum offloading on Tx packets (L4T in the Tx descriptor is 01b or 11b), any L2 padding at the end of the packet must be all zeros.

When using SCTP CRC offloading on Tx packets (L4T in the Tx descriptor is 10b), L2 padding should not be used.

6. Malicious Driver Detection MAX_BUFF Event

When the first Tx descriptor of a TSO packet contains both header and payload, it is counted twice in the malicious detection of MAX_BUFFS. Therefore, an MDD event is reported if the first segment is spread over eight descriptors, while it should only cause an MDD event if there are more than eight descriptors.

This can result in spurious Malicious Driver Detection events.

Software drivers must limit the first segment of a TSO packet to seven descriptors instead of eight. This restriction has been implemented in Intel drivers Release 21.3.

7. Intel® Ethernet Controller X710-TM4/AT2 Throughput Limit

Small Packet throughput limit:

For packets below 160 bytes there is a hardware packet processing limit for the entire device of ~37 Mpps. This results in limited throughput for:

- The Intel® Ethernet Controller X710-TM4 (4x10 GbE mode) when using 3 or 4 port 10 GbE operation.

Note: The Intel® Ethernet Controller X710-TM4 has an expected total throughput for the entire device of 40 Gb/s in each direction.

8. Small Packets Performance Degrade when Using Private VLAN

When using private VLAN, the device uses a VLAN pruning filter that slows down performance for small packets. For example, the total max MPPS (Million Packets Per Second) achievable drops from ~73 MPPS to ~39 MPPS.

Note: Depending on software driver capabilities, if only a single VLAN is applied, the driver might choose to use a port VLAN instead of a private VLAN and thus avoid the use of the VLAN pruning filter and associated performance penalty.

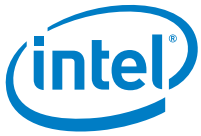
9. The X710-TM4/AT2 Packet Drop Rate is Limited to 27 MPPS

The packet drop rate was limited to 27 MPPS instead of 37 MPPS to improve device functionality robustness.

Maximum drop packet rate is now 27 MPPS. In case of enabling a queue without handling its descriptors, there might be a case of massive packets drop (i.e. broadcast) which will effect the overall traffic bandwidth. This case is forbidden to keep the device operation at full bandwidth.

10. Expansion ROM is Exposed in Blank Flash Programming Mode

In blank flash programming mode, the expansion ROM will be exposed, but might point to invalid pre boot driver code.



2.2 Specification Changes

1. Ingress Mirroring Cannot be Changed on the Fly

Changing of Ingress Mirroring configuration during traffic might cause a bad configuration.

2. RSS Field Selection is Globally Defined

RSS field selection is done globally and cannot be configured differently per PF or VF.

- Functions that require the Hash (RSS) filters on IPv4 packets should set all IPv4 PCTYPEs in the PFQF_HENA / VFQF_HENA (PCTYPEs 31, 33…36).
- Functions that require the Hash filters on IPv6 packets should set all IPv6 PCTYPEs in the PFQF_HENA / VFQF_HENA (PCTYPEs 41, 43…46).
- Functions that require the Hash filters on FCoE packets should set all FCoE PCTYPEs in the PFQF_HENA / VFQF_HENA (PCTYPEs 48…50).

3. NC-SI Get Controller Packet Statistics Command Limitations

Counter 0 of NC-SI "Get Controller Packet Statistics" command returns the value of "Valid Bytes Received" instead of "Total Bytes Received".

4. SMBus Minimum Packet Size

The minimum Ethernet packet size transmitted by BMC over SMBus supported by the X710-TM4/AT2 is 17 bytes.

5. Support of the Admin Queue Command "Set Loopback modes command (opcode:0x0618)"

The Set Loopback mode command (opcode:0x0618) is supported. For full details see the *Intel® Ethernet Controller X710-TM4/AT2 Datasheet*, Revision 2.0 or later.

6. Logging of PCIe Correctable Receiver Error

The optional error logging of correctable receiver error is disabled in the X710-TM4/AT2, which is allowed as described in:

> *PCI Express® Base Specification, Revision 3.0, November 10, 2010, Section 7.10.5, Correctable Error Status Register Footnote 101*

PHY layer receiver error detection and recovery mechanisms are operational such that there is no functional implication to the device or system operation. Please note that both Correctable Error Status Register[0] and Correctable Error Mask Register[0] are implemented such that the X710-TM4/AT2 is a PCI-SIG compliant device.

7. PRTPM_SAL and PRTPM_SAH are Re-loaded from NVM on PCIe Reset

PRTPM_SAL and PRTPM_SAH registers are re-loaded from the NVM on PCIe reset. Therefore, only the station address values stored in the NVM can be used for Wake-On-LAN.

8. VEB Statistics Disable NVM Bit

In X710-TM4/AT2 NVM 7.1:

EMP Settings Module Header Section - Word Offset #3 - Bit 0 - VEB Statistics Disable.

Description:

When set to 0b - VEB statistics are enabled.
When set to 1b - VEB statistics are disabled (default).

The statistics counters disabled by this bit are:

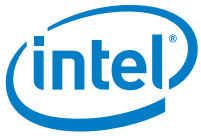
GLPRT_RUPP[0]
GLSW_GOTCH/L
GLVEBVL_GOTC_[n]
GLVEB_TCBCH/L[n]
GLVEB_TCPCH/L[n]
GLSW_UPTCH/L[n]
GLSW_MPTCH/L[n]
GLSW_BPTCH/L[n]
GLSW_GORCH/L[n]
GLVEB_VLBCH/L[n]
GLVEB_RCBCH/L[n]
GLVEB_RCPCH[n]
GLSW_UPRCH/L[n]
GLSW_MPRCH/L[n]
GLSW_BPRCH/L[n]
GLSW_RUPP_[n]
GLVEB_VLUPCH/L[n]
GLVEB_VLMPCH/L[n]
GLVEB_VLBPCH/L[n]

This bit can be used to disable VEB statistics and improve 64-byte packet performance in SR-IOV or any other configuration that has more than one VSI connected to a port.

Note: The EVB Protocols Enable bit originally mapped to bit 0 of the same NVM word is now mapped to bit 1.

Starting from NVM 7.1, the NVM bit is no longer used. VEB statistics are enabled/disabled on a per-VEB basis using a new flag in the Add VEB command. See the *Intel® Ethernet Controller X710-TM4/AT2 Datasheet* for details. Intel drivers that accompany this NVM release disable the VEB statistics on all VEBs by default.

Software drivers before Release 24.0 (i40e 1.4.25, ixl 1.4.26, VMware ESX i40e v1.4.26) should not be used with NVM 7.1 and above.



9. Unicast Hash Filtering Removal

Unicast Hash filtering is removed from the X710-TM4/AT2 switching elements options and should not be used. Unicast MAC Addresses can be filtered by Perfect filtering (up to 2K entries) or Promiscuous filtering.

10. Input Reference Clock Rise/Fall Times

The minimum input reference clock rise and fall times (Tr/Tf) listed in table *Input Reference Clock Electrical Characteristics* in the *Intel® Ethernet Controller X710-TM4/AT2 Datasheet* changes from 300 ps to 50 ps. This is a relaxation of the electrical specification for external oscillators.

Note: The routing of the input clock between the external oscillator and the X710-TM4/AT2 balls must be routed as a differential pair. Target 100 Ω differential impedance.

11. Set Local LLDP MIB when DCBX Agent is Disabled or Stopped

According to the *Intel® Ethernet Controller X710-TM4/AT2 Datasheet*, if the Set Local LLDP MIB command (0x0A08) is received while the FW DCBX agent is disabled or stopped, the MIB is parsed by firmware and used to configure the local DCB settings of the port, with no DCB TLV exchange with the peer performed by firmware.

Starting from NVM 7.1, if the Set Local LLDP MIB command is received while the DCBx specific agent is stopped, the command returns an EPERM error. If the command is received while the LLDP agent is stopped, it sets the local MIB without exchanging LLDP with peer, and returns SUCCESS.

12. Teredo UDP Tunneling Offload Support

Starting with NVM 7.1, Intel removed support for UDP Teredo tunneling offload.

Note: This feature was not supported in Intel Drivers.

13. GLQF_PCNT Counters

GLQF_PCNT counters do not wrap around.

Software should periodically clear these counters by writing any value. Note that the counter might miss a few events during the clearing process.

Starting in i40e v 2.2.x the driver keeps a count in software. However, there is a possibility of missing a few counts. Other Intel drivers do not use the counter.

14. Flash CS Negation Time

The X710-TM4/AT2's minimum value of the FLSH_CE_N High Time (tCS) specification was incorrectly reported in the *Intel® Ethernet Controller X710-TM4/AT2 Datasheet* as 25 ns. The actual minimum value of t_{CS} is 80 ns, except for the case of consecutive Read Status Register commands when the minimum value is 70 ns.

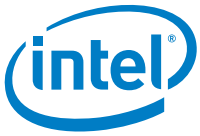
This specification is consistent with the requirements of the supported flash devices listed in Section 14 of the *Intel® Ethernet Controller X710-TM4/AT2 Datasheet*, except for the Micron/Numonyx M25PX64, which requires a minimum CS negation time (t_{CSH}) of 80 ns. No failures have been reported due to this specification mismatch.

15. Parsing of MPLS Headers

Starting from NVM image 7.1, the X710-TM4/AT2 identifies and skips up to 2 MPLS labels as described in Section *MPLS Header(s)* of the *Intel® Ethernet Controller X710-TM4/AT2 Datasheet*.

2.3 Documentation Updates

None.



2.4 Errata

1. TX Performance Degradation for Small Cloud Packets

Problem:

Happening for GRE+IPv6+TCP without payload. Degradation is expected to give 33 Gb/s instead of 34 Gb/s.

Implication:

This is seen if GRE+IPV6+TCP Packet is transmitted with no payload. This not typical packet format, and is not expected is most use cases.

Workaround:

None.

Status: B1=Yes; NoFix

2. PCIe Subsystem ID Incorrectly Reported for All PCI Functions Except Function 0

Problem:

All PCIe functions except Function 0 report a Subsystem ID of 0x0000 in the configuration space (including related Virtual Functions) regardless of the value programmed in the NVM.

Implication:

No functional impact to the device or drivers. However, this might impact the branding of the device if the Subsystem ID is used to select the device branding string.

Workaround:

None.

Status: B1=Yes; NoFix

3. Illegal Byte Error Statistical Counter Inaccuracy

Problem:

Short packets with bad symbols that arrive back-to-back might not be counted by GLPRT_ILLERRC.

Implication:

GLPRT_ILLERRC is inaccurate.

Workaround:

None.

Status: B1=Yes; NoFix

4. Receive Performance Degradation with Specific Cloud Header

Problem:

A small performance degradation is expected when receiving back-to-back GRE+IPv6+TCP cloud frames with 128-byte Header and almost no payload.

Implication:

Expected 33 Gb/s instead of 34 Gb/s.

Workaround:

None.

Status: B1=Yes; NoFix

5. MCTP Discovery Error when Replacing Active PF

Problem:

MCTP Discovery might respond with a wrong PF ID when BMC is replacing the active PF. Expected to be a rare scenario on specific machines.

Implication:

PF replacement might not work for MCTP.

Workaround:

None.

Status: B1=Yes; NoFix

6. RX Queue Disable is Reported Done Before It is Disabled

Problem:

RX Queue disable is reported done before it is disabled.

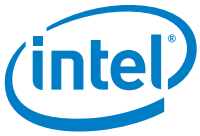
Implication:

An RX Hang could result if the software re-enables the queue too early.

Workaround:

An RX Queue should be reused only after a minimum delay of 50 ms. This workaround is implemented in Intel Software Release 24.0.

Status: B1=Yes; NoFix



7. TX Descriptor Might be Read Twice

Problem:

A TX Descriptor might be read more than once in corner case conditions.

Implication:

Negligible.

Workaround:

None.

Status: B1=Yes; NoFix

8. Immediate Interrupts are Delayed in Very Loaded System

Problem:

In a case where there are ten or more active queues in the system, and some of the queues are assigned with immediate interrupts, the interrupt delay might exceed the value specified in the Datasheet (”ITR and immediate interrupts jitter” table).

Implication:

Low performance impact

Workaround:

None.

Status: B1=Yes; NoFix

9. ECRC Bits are Not RO when ECRC is Disabled

Problem:

ECRC bits in PCIe AER registers are writable even when ECRC is disabled.

Implication:

Specification compliance issue.

Workaround:

None.

Status: B1=Yes; NoFix

10. NC-SI I/Os Output Rise Slew Rate is Higher Than Specification

Problem:

NC-SI I/Os output rise time might be as low as 400 ps, while the NC-SI Specification requires a minimum of 500 ps.

Implication:

Specification compliance issue.

Workaround:

Place a 30 Ω resistor in serial to the pad.

Status: B1=Yes; NoFix

11. TC Strict Priority Does Not Work as Expected

Problem:

An UP might not get exclusive priority if PCIe bandwidth is insufficient (although gets higher priority).

Implication:

RX TC strict priority limitation.

Workaround:

None.

Status: B1=Yes; NoFix

12. Management-only Packets Cannot be Ignored for Wake-Up

Problem:

Due to a “NoTCO” wake-up capability malfunction, a wake event might be issued for packets that are expected to be routed to the BMC exclusively.

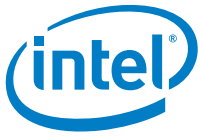
Implication:

Management-only packets cannot be ignored for wake-up.

Workaround:

None.

Status: B1=Yes; NoFix



13. Common Clock Configuration Bit Specification Compliance

Problem:

Common clock configuration bit should be writable for all PFs, but it is not always writable for a PF > 0.

Implication:

Specification compliance issue.

Workaround:

None.

Status: B1=Yes; NoFix

14. Low Latency TC Might be Momentarily Starved

Problem:

Low Latency TC might be momentarily starved under TPB Non-Strict Priority (RR) policy when both Bulk and Low Latency traffic are pending.

Implication:

Low Latency TC impact.

Workaround:

None.

Status: B1=Yes; NoFix

15. Round Robin (RR) Bandwidth Distribution is Traffic Dependent

Problem:

Under RR RX Policy, RX bandwidth might be distributed unevenly among ports and TCs if PCIe bandwidth is smaller than incoming traffic, or traffic is a stream of small packets (smaller than 128 bytes).

Implication:

Uneven traffic distribution under RR.

Workaround:

Use Strict Priority policy instead of Round Robin.

Status: B1=Yes; NoFix

16. L2 Tag Stored in the Wrong RX Descriptor Field

Problem:

If two L2 tags (for example: VLAN and S-TAG) are programmed to be extracted to the receive descriptor, and the receive packet includes only a single L2 tag, the extracted L2 tag is always posted in the L2TAG1 field if L2TSEL is set to 1b, or to L2TAG2 if L2TSEL is set to 0b.

Implication:

In the following cases there are no implications:

1. If the receive data flow always includes two L2 tags.
2. If the receive data might include packets with a single L2 tag, but are always the same tag type (first or second).

If the receive data flow that might include packets with only one L2 tag (which can be either the first or the second tag), software cannot identify which of the two enabled L2 tags were extracted to the receive descriptor.

Workaround:

If the receive data flow includes packets with only one L2 tag, and software is not able to identify if it is the first or the second tag, it should not enable more than a single L2 tag to be extracted to the receive descriptor.

Status: B1=Yes; NoFix

17. Internal VLAN is Not Reflected in RX Descriptor

Problem:

When SHOWIV field is set in the receive queue context, the internal VLAN is stripped, but it is not inserted in the RX descriptor as expected.

Implication:

Internal VLAN is not reflected in RX descriptor.

Workaround:

None.

Status: B1=Yes; NoFix

18. Transmit Queue Group with Single Queue Enabled Performance

Problem:

A transmit queue Group with single Queue enabled might have performance limitations when scheduling consecutive packets.

Implication:

TX Performance issue.

Workaround:

None.

Status: B1=Yes; NoFix

19. A Switching Table Might Reduce Small Packets Performance

Problem:

If the switching table is relatively full, it might reduce performance with a continuous stream of packets smaller than 160 bytes. A data stream that includes a mix of small and big packets should not experience any degradation.

Implication:

Small packets performance impact.

Workaround:

Avoid switching table fullness.

Status: B1=Yes; NoFix

20. Set Binding Command is Not Functional for IPv4

Problem:

Set Binding command is not functional for IPv4.

Implication:

No manageability traffic after command.

Workaround:

None.

Status: B1=Yes; NoFix

21. Cloud Traffic Over VEB is Transmitted to LAN

Problem:

Cloud traffic over VEB is transmitted to LAN.

Implication:

Cloud traffic over VEB is transmitted to LAN.

Workaround:

None.

Status: B1=Yes; NoFix

22. VLAN Prune is Not Functional

Problem:

Default action for VLAN Prune table is not set after "Add VLAN AQ" command.

Implication:

VLAN Prune is not functional.

Workaround:

Additional VLAN Prune configuration should be done by software.

Status: B1=Yes; NoFix

23. INTENA_MSK Setting Might Clear Interrupt

Problem:

A write access to a xxINT_DYN_CTLx CSR with INTENA_MSK bit (bit 31) set to 0 clears the corresponding interrupt bit in the PBA array.

Implication:

There is a possibility of an Interrupt missing. However, current Intel software implementation has this bit set to 1 except when enabling or disabling interrupts.

Workaround:

INTENA_MSK should be set in all CSR write accesses other than INTENA bit change.

Status: B1=Yes; NoFix

24. Manageability Checksum Filtering of IPv6 Packets

Problem:

The IPv6 checksum calculation could be incorrect for received packets that contain either a Routing (Type 2) Extension Header or a Destination Options Extension Header that includes a Home Address option.

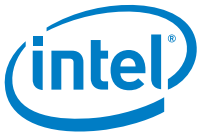
Implication:

If the manageability filtering is configured to drop packets with checksum errors, IPv6 manageability packets with the extension headers described above could be incorrectly dropped.

Workaround:

Do not enable checksum filtering for manageability if the IPv6 Extension Headers described above are used on manageability traffic.

For SMBus: The Enable Xsum Filtering to MNG bit should be 0b in the Update Management Receive Filter Parameters command and in the Set Common Filters Receive Control Bytes command if these commands are used.



For NCSI: Do not use the Enable Checksum Offloading Command (Intel OEM Command 0x23).

Status: B1=Yes; NoFix

25. Link Remains Up During Power Saving State

Problem:

Intel X710/XL710-based devices might maintain link, regardless of system power state, as long as power is provided to the device.

Implication:

Link remains up during power saving state.

Workaround

None

Status: B1=Yes; NoFix

26. PRTDCB_RUP2TC and PRTDCB_TC2PFC are Not Writable

Problem:

PRTDCB_RUP2TC (0x1C09A0) and PRTDCB_TC2PFC (0x001C0980) CSRs cannot be written directly by software when CSR protection is enabled.

Implication:

Programming this CSR is required if the software is configuring DCB on the device.

Workaround:

For PRTDCB_RUP2TC: Write to PRTDCB_RUP2TC as usual, then use a Direct Admin command with the following values to complete the write transaction.

For PRTDCB_TC2PFC: Write to PRTDCB_TC2PFC as usual, then use a Direct Admin command with the following values to complete the write transaction.

Field	Byte	Value PRTDCB_RUPTC	Value PRTDCB_TC2PFC
Flags	0-1	0x0	0x0
Opcode	2-	0xFF04	0xFF04
Data Length	4-5	0x0	0x0
Return Value/VFID	6-7	0x0	0x0
Cookie	8-15	Arbitrary value defined by software	Arbitrary value defined by software
Param 0	16,--19	0	0
Param 1	20-23	(0x000AC440 + 0x4 * PRT)	(0x000AC200 +0x4 * PRT)
Data Address High	24-27	0	0
Data Address Low	28-31	<CSR Write Data>	<CSR Write Data>

Status: B1=Yes; NoFix

27. AER Header Log Might be Invalid

Problem:

If more than two uncorrectable function-specific errors are reported to VFs connected to the same PF, the Advanced Error Reporting (AER) Header Log (PCIe Configuration Registers offset 0x11C... 0x128) might be invalid.

This occurs only in case that one or more of the two errors have been cleared by the host, and a 3rd one arrives later for a VF connected to the same PF. In this case, the header log of this last error might be corrupted.

Implication:

Error source debug limitation. Uncommon systems suffering from multiple uncorrectable errors might have invalid AER Header Log.

Workaround:

PCIe trace data collected by a protocol analyzer can alternatively be used to recognize the TLP that is causing the error.

Status: B1=Yes; NoFix

28. A CfgWr to a VF TLP with Error Might Generate an Error Message with Wrong VF Number

Problem:

When a CfgWr TLP that is poisoned or has a parity error is received by the X710-TM4/AT2, an error message with the wrong VF number might be generated. Note that the status is correctly reported in the respective VF Status registers.

Implication:

PCIe error message with wrong Requester ID.

Workaround:

When the OS gets an error message and the status registers bits are cleared, it should poll the other VFs' status registers.

Status: B1=Yes; NoFix

29. No LAN-to-BMC Pass-through Traffic in Dr State

Problem:

While in Dr state and pass-through is enabled, the X710-TM4/AT2 should keep pass-through functionality active. However, LAN-to-BMC traffic is not functional in Dr state.

Implication:

Cannot maintain manageability pass-through traffic while the system is in Soft Off G2/S5 state.

Workaround:

An NVM workaround is available in NVM 7.1.

Status: B1=Yes; NoFix

30. MNG Packets are Dropped while a Function-Level Reset to PF 0 is in Progress

Problem:

When Function-Level Reset (FLR) is applied to PF 0, it also resets the LAN-to-BMC pass-through flow.

Implication:

LAN-to-BMC pass-through traffic stops while FLR is applied to PF 0.

Workaround:

None.

Status: B1=Yes; NoFix

31. DCBx Resume of a Port Affects Other Ports

Problem:

When DCBx resumes a port's traffic, done after port draining is performed, traffic might also be resumed for other ports.

Implication:

A port might be unintentionally resumed.

Workaround:

None.

Status: B1=Yes; NoFix

32. A Global SDP Might be Affected by a Specific Port Power State

Problem:

When a GPIO is defined as a global SDP and its behavior is unrelated to any specific port, the GLGEN_GPIO_CTL.*PRT_NUM_NA* bit should be set, and the SDP value should be controlled by the GLGEN_GPIO_SET register regardless of port state.

However, the *PRT_NUM_NA* bit does not take effect, and SDP output is tri-stated or driven high (depending on GLGEN_GPIO_CTL.*OUT_CTL*) when the port specified at the *PRT_NUM* field is in power-down state.

Implication:

A global SDP unrelated to any specific port is disabled according to a port power state.

Workaround:

This should be taken in account in the board design. In some cases, it might just be a matter of inverting the polarity.

Status: B1=Yes; NoFix

33. Legacy SMBus: Failure to De-assert Alert Signal when Not Using ARA Cycle

Problem:

In legacy SMBus mode, the MC might get an indication of outstanding events through the SMBALRT_N line. The MC should then do an ARA cycle to get the indicating function. It can instead read the status of all functions. If the MC fails to do this, and reads only a single function status, the SMBALRT_N line will never de-assert, even if the timeout expires.

Implication:

SMBALRT_N is not de-asserted.

Workaround:

Poll status of all functions.

Status: B1=Yes; NoFix

34. Get Link Status AQ Command Might Return Incorrect Status

Problem:

If there is an I^2C access error when executing the Get Link Status AQ command, the X710-TM4/AT2 might falsely provide a link down response.

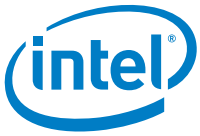
Implication:

A transient error in accessing the external module via I^2C causes the software device driver to report a link flap to the system.

Workaround:

If a Get Link Status response shows a link de-assertion, the Get Link Status command should be repeated.

Status: B1=Yes; NoFix



35. A Function-level Reset Might Affect Other Functions

Problem:

When a function-level reset is applied (PFR, VFR or VMR), under rare conditions it might affect the Tx of a different function.

Implication:

Tx hang.

Workaround:

To prevent the failure, ensure that all queues belonging to the entity to be reset are disabled before initiating the reset. If this cannot be ensured, the failure could occur and the software device driver should use a CORER to recover from a Tx hang that cannot be cleared by a function-level reset.

Status: B1=Yes; NoFix

36. Rx Packet Drops Even with Priority Flow Control

Problem:

When using flow control, the expectation is for no Rx packet drops caused by a Receive Packet Buffer overflow. In the situation where Priority Flow Control (PFC) is enabled on some traffic classes, but not on all enabled traffic classes, there is a possibility for the Receive Packet Buffer to fill up and drop packets belonging to any traffic class.

Implication:

PFC is not completely effective in preventing Receive Packet Buffer overflows under small-packet stress conditions.

Workaround:

None.

Status: B1=Yes; NoFix

37. DCBx Configuration Might Change After LLDP Stops

Problem:

After a Stop LLDP Agent AQ command, the LLDP agent should be stopped but DCBx configuration should stay unchanged. However, if a CORER or GLOBR is asserted when the LLDP agent is stopped, the configuration might be changed.

Implication:

DCBx configuration unstable when LLDP agent stops.

Workaround:

Prior to a Stop LLDP Agent AQ command, software should read the MIB (Get LLDP MIB AQ command). After LLDP stops, software should write a previous MIB (Set Local LLDP MIB AQ command).

Status: B1=Yes; NoFix

38. PCIe Interrupt Status Bit

Problem:

The *Interrupt Status* bit in the Status register of the PCIe configuration space is not implemented and is not set as described in the PCIe specification.

Implication:

When using shared legacy PCI interrupts, software might use this bit to determine if the X710-TM4/AT2 has a pending interrupt. Since the bit is not implemented, the software might not handle the interrupt, resulting in a continuous interrupt assertion.

There is no implication when using MSI or MSI-X.

Workaround:

The *Interrupt Status* bit should not be used. Avoid using shared legacy PCI interrupts.

Status: B1=Yes; NoFix

39. Glitch on SDP Outputs During GLOBR

Problem:

GPIO pins that are defined as SDP outputs (*PIN_FUNC* is 000b and *PIN_DIR* is 1b in GLGEN_GPIO_CTL) can have a high-to-low glitch during GLOBR if *OUT_CTL* is 0b.

The same applies when the port specified in GLGEN_GPIO_CTL.*PRT_NUM* is enabled/disabled.

Implication:

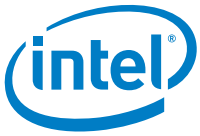
The implication depends on the use of the SDP. For example, an SDP used as a QSFP+ reset signal might cause the module to malfunction due to a short reset assertion.

Workaround:

One of the following:

- If the SDP is supposed to be high during GLOBR, set *OUT_CTL* to 1b.
- For a general-purpose 2-state SDP output (*PHY_PIN_NAME* is 0x3F), set *PIN_FUNC* to 001b (LED) and use the *LED_MODE* field (0000b or 1111b) to control the output value.

Status: B1=Yes; NoFix



40. Function-Level Reset Fails to Complete

Problem:

In rare cases, the hardware activity of a function-level reset (PFR, VFR, or VMR) might fail to complete.

Implication:

PFR: Software times out while waiting for the PFR to complete. The firmware gets stuck and the firmware watchdog timer expires, triggering an EMPR.

VFR/VMR: Software times out while waiting for the reset to complete.

Workaround:

PFR: Software should re-initialize the device after the EMPR.

VFR/VMR: After a timeout waiting for the reset to complete, software should retry the reset by clearing and then setting the reset trigger bit (GLGEN_VFRTRIG.*VFSWR* for VFR or VSIGEN_RTRIG.*VMSWR* for VMR) and then restarting the polling for reset completion. After three retry attempts, abort with an error.

Status: B1=Yes; NoFix

41. Incorrect Flexible Payload Extraction from Flow Director Filter to Receive Descriptor

Problem:

When programming a Flow Director filter, if *FD_STATUS* is 10b, the FLEXOFF value provided in the programming descriptor is used incorrectly, and the wrong bytes are extracted to the receive descriptor.

Implication:

Incorrect descriptor content.

Workaround:

To get four bytes starting from offset N of the flexible payload to the receive descriptor, the value N-2 should be used for FLEXOFF. Byte offsets 0 and 1 cannot be extracted with *FD_STATUS* of 01b.

Status: B1=Yes; NoFix

42. Aux Power Detected Bit Not Implemented

Problem:

The *Aux Power Detected* bit in the Device Status register of the PCIe Configuration Space is not implemented. The bit is always 0b.

Implication:

PCIe specification compliance, but this issue is not detected by the existing compliance testing.

It is not expected that any software uses this bit. If it is being used, the workaround should be implemented by the software.

Workaround:

Use the most-significant bit of PMCR.*PME_Support* instead.

Status: B1=Yes; NoFix

43. SGMII Receiver Sensitivity

Problem:

The SGMII specification requires a maximum receiver sensitivity of 100 mV peak-to-peak. The The X710-TM4/AT2 receiver sensitivity can be as high as 190 mV peak-to-peak.

Implication:

No expected implication, since the input signal voltage would normally be high enough.

Workaround:

Ensure that the input signal is strong enough when using an SGMII connection.

Status: B1=Yes; NoFix

44. IEEE 802.3 Clause 73 AN Does Not Support Parallel Detection

Problem:

When using Clause 73 auto-negotiation, parallel detection is not supported.

Implication:

Inability to link with legacy devices that do not have Clause 73 AN enabled.

Workaround:

None.

Status: B1=Yes; NoFix

45. IEEE 802.3 Clause 73 AN Echoed Nonce Field is Zero

Problem:

During the IEEE 802.3 Clause 73 auto-negotiation process, the last message page has the Echoed Nonce field set to 00000b even if the ACK bit is 1b.

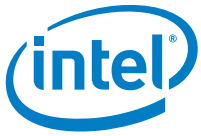
Implication:

With certain link partners this might cause auto-negotiation failures.

Workaround:

None.

Status: B1=Yes; NoFix



46. KR Transmitter Output Waveform Violations

Problem:

The KR transmitter does not meet the IEEE 802.3 Clause 72.7.1.11 transmitter output waveform requirements for R_{pre} when both C(1) and C(-1) are disabled and c(0) is maximum.

Implication:

Conformance issue. Not expected to impact functionality.

Workaround:

None.

Status: B1=Yes; NoFix

47. 10GBASE-KR wait_timer Value Smaller Than Specification

Problem:

The 10GBASE-KR wait_timer is defined by IEEE 802.3 to have a value between 100 and 300 training frames. The actual value is 75 training frames.

Implication:

Potential training failure with some link partners.

Workaround:

None.

Status: B1=Yes; NoFix

48. Receive Queue Disable Can Get Stuck

Problem:

If there are no descriptors available for a receive queue that belongs to a no-drop TC and the queue is disabled at the same time that a packet arrives for the queue, the queue disable can get stuck.

Implication:

Head-of-line blocking continues despite an attempt to disable the queue.

Workaround:

To avoid this situation, the driver should try to ensure that there are always Rx descriptors available, especially when disabling an Rx queue.

If a head-of-line blocking situation does occur, it is handled as usual when the PFCTIMER expires.

Status: B1=Yes; NoFix

49. Set DCB Parameters AQC (Opcode 0x303) Might Return EINVAL Even when It Succeeds

Problem:

The Set DCB Parameters AQ command (opcode 0x303) might return EINVAL even when it succeeds.

Implication:

Software driver does not know whether the command succeeded.

Workaround:

Ignore the return code.

Status: B1=Yes; NoFix

50. Receive IP Packets in a Low-Latency Traffic Class Are Not Fully Processed

Problem:

Receive packets that contain an IP header and belong to a low-latency traffic class (as defined by PRTDCB_RETSC.*LLTC*) are not fully processed by the X710-TM4/AT2. The following processing is not performed on these packets:

- Validating the IP checksum.
- Validating the L4 checksum.
- Stripping/extracting the VLAN from a tunneled packet.

Implication:

Performing this data processing in software results in lower overall performance of the product.

Workaround:

When using DCBx, ETS should be enabled for all active TCs.

Status: B1=Yes; NoFix

51. Activity LED Might Blink Regardless if Link is Up or Down for a Port

Problem:

X710-TM4/AT2 Activity LEDs toggle as a result of BMC/HOST transmit packets regardless of the port link state. Activity LEDs are MAC_ACT or FILTER_ACT (set by the field *LED_MODE* - 1101 or 1110 respectively).

Implication:

The Activity LED might be blinking even if link is down.

Workaround:

BMC/HOST should transmit packets only when link is up.

Status: B1=Yes; NoFix

52. EMP Reset After Using Intel QCU Tool

Problem:

EMP reset occurs after changing the device configuration using Intel QCU tools. For example, changing from 2x40 to 4x10.

Implication:

Device hang and later EMP reset that removes the manageability configuration.

Workaround:

None.

Status: B1=Yes; NoFix

53. PCIe Replay Timer Can Occasionally be Lower Than PCIe Spec Requirements

Problem:

PCIe Replay Timer can occasionally be lower than PCIe spec requirements

Implication:

PCIe Replay Timer Timeouts and Replay_Num rollover correctable errors. Due to this error, reduced PCIe performance is possible.

These errors can only be observed on platforms with downstream port ACK latencies that are beyond maximum PCIe spec limits.

Workaround:

None.

Status: B1=Yes; NoFix

3. Software Clarifications

Table 3-1 Summary of Software Clarifications

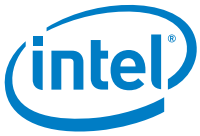
Software Clarification	Status
1. X710-TM4/AT2 Option ROM Should Not be Integrated in the BIOS	N/A
2. VXLAN Guidance for VMware vSphere	N/A

1. X710-TM4/AT2 Option ROM Should Not be Integrated in the BIOS

Previous generations of Intel networking controllers allowed the Option ROM to be stored in the flash attached to the device, or in the BIOS flash. The X710-TM4/AT2 requires the Option ROM, if one is used, to be stored in the flash attached to the X710-TM4/AT2. This is done to maintain alignment of the pre-boot code with the internal X710-TM4/AT2 firmware when upgrades are necessary.

2. VXLAN Guidance for VMware vSphere

For VXLAN traffic in production VMware vSphere environments with the X710-TM4/AT2, use the 1.3.38 ESXi driver or later. For the latest driver version currently available, please reference the VMware Compatibility Guide.



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.

This document (and any related software) is Intel copyrighted material, and your use is governed by the express license under which it is provided to you. Unless the license provides otherwise, you may not use, modify, copy, publish, distribute, disclose or transmit this document (and related materials) without Intel's prior written permission. This document (and related materials) is provided as is, with no express or implied warranties, other than those that are expressly stated in the license.

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.