Intel® 5000 Series Chipset Memory Controller Hub (MCH)

Specification Update

September 2009

Reference Number: 313069-009

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. 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.

Intel products are not intended for use in medical, life saving, life sustaining, critical control or safety systems, or in nuclear facility applications.

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 Intel® 5000 Series Chipset Memory Controller Hub (MCH) 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 ordering number and are referenced in this document, or other Intel literature may be obtained by calling 1-800-548-4725 or by visiting Intel's website at http://developer.intel.com/products/index.htm.

Contents

Revision History

Revision	Description	Date
001	Initial release of Specification Update.	May 2006
002	Added Compliance Issues section and issue “PCIe compliance test failure: CBDMA Interrupt Line Register” and “PCIe compliance test failure: MSI Address Register”. Updated Software Guidance for MSI handling section. Added Specification Clarification: SMBus and PCIe interoperability time out recommendation. Added erratum “RID=CRID is sticky across warm reset“	June 2006
003	Added errata 28 through 40	December 2006
004	Added G1 stepping and errata to summary table. Added 5000P, 5000V, and 5000X to Section 3.6. added 5000X to Section 3.7	January 2007
005	Added Spec clarifications #2-#4 to section 3.4, Compliance Issues #1 and #2 to section 3.5, Documentation Changes #1 and #2 to section 3.6 and section 8 Fixed affected steppings for errata #23	June 2007
006	Added errata 41, and 42 Added Document change 3	October 2007
007	Added errata 43	May 2008
008	Added errata 44	February 2009
009	Added errata 45, 46	September 2009

Preface

This document is an update to the Intel® 5000 Series Chipset specifications contained in the Affected Documents/Related Documents table below. This document is a compilation of sightings, device and documentation errata, specification clarifications and changes. It is intended for hardware system manufacturers and software developers of applications, operating systems, or tools.

Information types defined in Nomenclature are consolidated into the specification update and are no longer published in other documents.

This document may also contain information that was not previously published.

Affected Documents/Related Documents

Document Title	Reference Number
Intel® 5000P/5000V/5000Z Chipset Memory Controller Hub (MCH) Datasheet	313071
Intel® 5000X Chipset Memory Controller Hub (MCH) Datasheet	313070
Intel® 5000P/5000V/5000Z/5000X Chipset Memory Controller Hub (MCH) Thermal Mechanical Design Guide	313067
Interrupt Swizzling Solution for the Intel® 5000 Chipset Series based Platforms - Application Note	314337

Nomenclature

Errata are design defects or errors. These may cause the Intel 5000 Series chipset MCH 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.

Specification Changes are modifications to the current published specifications. These changes will be incorporated in any new release of the specification.

Specification Clarifications describe a specification in greater detail or further highlight a specification's impact to a complex design situation. These clarifications will be incorporated in any new release of the specification.

Documentation Changes include typos, errors, or omissions from the current published specifications. These will be incorporated in any new release of the specification.

Note: Errata remain in the specification update throughout the product's lifecycle, or until a particular stepping is no longer commercially available. Under these circumstances, errata removed from the specification update are archived and available upon request. Specification changes, specification clarifications and documentation changes are removed from the specification update when the appropriate changes are made to the appropriate product specification or user documentation (datasheets, manuals, and so forth).

Summary Table of Changes

The following table indicates the sightings, errata, specification changes, specification clarifications, or documentation changes which apply to the Intel® 5000 Chipset Series. Intel may fix some of the errata in a future stepping of the component, and account for the other outstanding issues through documentation or specification changes as noted. This table uses the following notations:

Codes Used in Summary Table

X: Erratum exists in the stepping indicated. Specification Change or Clarification that applies to this stepping.

B: BIOS Update corrects this issue.

(No mark)/(Blank box): This erratum is fixed in listed stepping or specification change does not apply to listed stepping.

Plan fix: This erratum may be fixed in a future stepping of the component.

Fixed: This erratum has been previously fixed.

No Fix: There are no plans to fix this erratum.

UI: Currently under investigation, final status not determined. This is in terms of when the issue will get fixed or if it will get fixed.

Shaded: A shaded row indicates this erratum is either new or modified from the previous version of the document

Errata

Table 1. Errata (Sheet 1 of 3)

Number	Steppings				Status	Errata
	B2	G0[1]	B3	G1		
1	X	X	X	X	No Fix	PCI Express* Auto Link Negotiation Occasionally Fails to Correctly Detect Link Width
2	X	X	X	X	No Fix	MCH B1Err Errors Logged Incorrectly
3	X	X	X	X	No Fix	PCI Express* IBIST on x8/x16 Port Will Not Stop Testing of Entire Port if an Error is Detected on Any Lane Other Than Lowest 4 Lanes of the Port
4	X	X	X	X	No Fix	MCH Thermal Sensor Reporting Incorrect Values
5	X	X	X	X	No Fix	Crystal Beach Channel Completion Address Logged Twice in FERR_CHANCMP Register
6	X	X	X	X	No Fix	PCI Express* Receiver Error Logged Upon Putting Link in Disable State
7	X	X	X	X	No Fix	PCI Express Hot-Plug ABP Bit Set While Attention Button is Pressed
8	X	X	X	X	No Fix	Leakage from 1.5V VCC to 1.2V VTT
9	X	X	X	X	No Fix	Surprise Link Down Error Reported When PCI Express Slot Power is Removed During Hot-plug Event
10	X	X	X	X	No Fix	System Hang with Large Number of Transaction Retries
11	X	X	X	X	No Fix	INTL[7:2] Registers Are Not Implemented as Read/Write

Table 1. Errata (Sheet 2 of 3)

Number	Steppings				Status	Errata
	B2	G0[1]	B3	G1		
12	X	X	X	X	No Fix	FBD Alert Packet Check Before M21 Error Incorrectly Checks Wrong Channel
13	X	X	X	X	No Fix	SMI Escalation via ERR[2:0]# Pins May Result in IERR#
14	X	X	X	X	No Fix	Read From Remote Branch in Memory Mirrored Mode May Result in Data Corruption
15	X	X	X	X	No Fix	CPU May Record Signal Glitches When MCH is Being Reset
16	X	X	X	X	No Fix	Coalesce Mode Cannot Be Used with Max Payload Size of 256B
17	X	X	X	X	No Fix	Outstanding Write Transaction in Mirrored Mode Will Cause System Hang
18	X	X			Fixed	PCIe* Link Width Degradation During Reset Tests
19	X	X			Fixed	PCIe Surprise Link Down or Link Degrade During L1 Exit
20	X	X			Fixed	L1 entry Hangs When x4 Links Have Downshifted to x2
21	X	X			Fixed	MCH PCI Express L0s issue
22	X	X	X	X	No Fix	MCH may log F2Err during shutdown special cycle initiated due to FSB timeout
23	X	X	X		Fixed	System hang with multiple retries and locks
24	X	X	X		Fixed	Patrol Scrub with CRC error issue
25	X	X	X	X	No Fix	Header of malformed TLPs on a x16 port not always logged
26	X	X	X	X	No Fix	Illegal addresses within the 40-bit address space in the channel completion address register does not generate cmp_addr_err
27	X	X	X	X	No Fix	RID=CRID is sticky across warm reset
28	X	X	X	X	No Fix	SLD could causes spurious completions
29	X	X	X	X	No Fix	IBIST does not capture failed lanes properly
30	X	X	X	X	No Fix	IBIST RX logic does not stop when IBISTR is reset
31	X	X	X	X	No Fix	FBD Channel Index is incorrect when logging some error conditions
32	X	X	X	X	No Fix	PCIE x16 link goes to recovery with a x1 card and L0s
33	X	X	X	X	No Fix	SLD on PCIE port during P2P Posted requests can causes ESI to hang
34	X	X	X	X	No Fix	PEXGCTRL.PME_TO_ACK may not be set when a Turn Off Acknowledge TLP have been Received from All Ports
35	X	X	X	X	No Fix	Masked Completer Abort Status Errors may be Reported in the UNCERRSTS[0] Register
36	X	X	X	X	No Fix	SMBus 2.0 Specification TLOW:SEXT may be exceeded on SMBus 0 when the North Bridge is clocked with a 266 MHz BUSCLK
37	X	X	X	X	No Fix	The First Uncorrectable Fatal bit of the Root Error Status Register may be Incorrectly Set when a Second Uncorrectable Fatal Error is Received
38	X	X	X	X	No Fix	Bit [3], INTxST, of the PCISTS Register may be Cleared when Bit [10], INTxDisable, of the PCICMD Register is Set
39	X	X	X	X	No Fix	The DMA Engines Next Channel Error Register is not Updated When Subsequent Errors are Detected
40	X	X	X	X	No Fix	Store to Write-Through (WT) Memory Data May be Seen in Wrong Order by Two Subsequent Loads When Snoop Filter is Enabled
41	X	X	X	X	No Fix	PCI-Express transaction IO ordering queue overflow
42	X	X	X	X	No Fix	System failures during spare copy
43	X	X	X	X	No Fix	Read transactions may be delayed

Table 1. Errata (Sheet 3 of 3)

Number	Steppings				Status	Errata
	B2	G0[1]	B3	G1		
44	X	X	X	X	No Fix	CRC errors logged by AMB during C2 or S1 transition
45	X	X	X	X	No Fix	PCI express TLP packets not flagged “Malformed” under certain conditions.
46	X	X	X	X	No Fix	Error Source Identification (ID) is not properly reporting the Requestor ID when the uncorrectable (Non-fatal/fatal) error is detected in the PCI express Root Port

Specification Changes

Number	SPECIFICATION CHANGES
N/A	None

Specification Clarifications

Number	SPECIFICATION CLARIFICATIONS
1	Software Guidance for MSI handling
2	SMBus and PCIe interoperability time out recommendation
3	Intel® 5000 Series Chipsets Interrupt Swizzling Recommendation
4	Issues with some PCI Express adapters during Link Training and L1 exit

Compliance Issues

Number	Compliance Issues
1	PCIe compliance test failure: CBDMA Interrupt Line Register
2	PCIe compliance test failure: MSI Address Register

Documentation Changes

Number	DOCUMENTATION CHANGES
1	3.8.8.36 PEXGCTRL - PCI Express Global Control Register
2	3.17.12.2 PCI_Express Device Number Assignment and Header Log
3	PEX[7:2,0]CDTHROTTLE: PCI Express Throttle Control

Component Identification via Programming Interface

The Intel 5000 chipset series can be identified by the following register contents:

MCH Version	Stepping	Vendor ID[1]	Device ID[2]	Revision Number[3]
5000P	B-2	8086h	25D8h	92h
5000V	B-2	8086h	25D4h	92h
5000X	B-2	8086h	25C0h	12h
5000P	B-3	8086h	25D8h	93h
5000V	B-3	8086h	25D4h	93h
5000X	B-3	8086h	25C0h	13h
5000Z	B-3	8086h	25D0h	93h
5000X	G-0	8086h	25C0h	30h
5000P	G-1	8086h	25D8h	B1h
5000V	G-1	8086h	25D4h	B1h
5000X	G-1	8086h	25C0h	31h

Notes:

1. The Vendor ID corresponds to bits 15:0 of the Vendor ID Register located at offset 00 - 01h in the PCI function 0 configuration space.
2. The Device ID corresponds to bits 15:0 of the Device ID Register located at offset 02 - 03h in the PCI function 0 configuration space.
3. The Revision Number corresponds to bits 7:0 of the Revision ID Register located at offset 08h in the PCI function 0 configuration space.

Component Marking Information

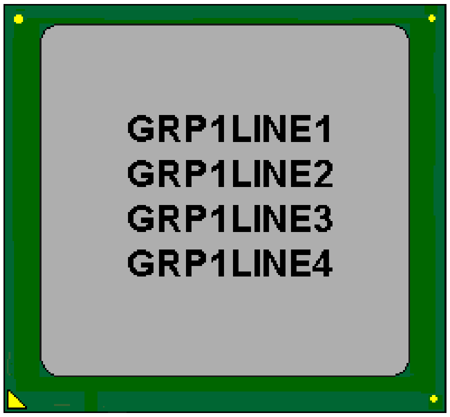
The Intel 5000 chipset series can be identified by the following component markings:

MCH	Stepping	S-Spec	Top Marking
5000X	B-2	SL96T	NQ5000X
5000P	B-2	SL96Z	NQ5000P
5000V	B-2	SL96X	NQ5000V
5000Z	B-2	SL96V	NQ5000Z
5000X Pb-free	B-2	SL96U	QG5000X
5000P Pb-free	B-2	SL972	QG5000P
5000V Pb-free	B-2	SL96Y	QG5000V
5000Z Pb-free	B-2	SL96W	QG5000Z
5000X	B-3	SL9LV	NQ5000X3
5000P	B-3	SL9LT	NQ5000P3
5000V	B-3	SL9LR	NQ5000V3
5000Z	B-3	SL9LP	NQ5000Z3
5000X Pb-free	B-3	SL9LW	QG5000X3
5000P Pb-free	B-3	SL9LU	QG5000P3
5000V Pb-free	B-3	SL9LS	QG5000V3
5000X Pb-free	G-0	SL9LX	QG5000XG
5000X	G-1	SL9TG	NQ5000X

MCH	Stepping	S-Spec	Top Marking
5000P	G-1	SL9TN	NQ5000P
5000V	G-1	SL9TJ	NQ5000V
5000Z	G-1	SL9TL	NQ5000Z
5000X Pb-free	G-1	SL9TH	QG5000X
5000P Pb-free	G-1	SL9TP	QG5000P
5000V Pb-free	G-1	SL9TK	QG5000V
5000Z Pb-free	G-1	SL9TM	QG5000Z

The Intel 5000 chipset series stepping can be identified by the following component markings:

Figure 1. Top-Side Marking Example



Errata

1. PCI Express* Auto Link Negotiation Occasionally Fails to Correctly Detect Link Width

Problem: The MCH occasionally fails to correctly detect the width of a hot-plugged PCI Express* card. This is restricted to the case where cards are "hot-plugged" with system power on.

Implication: The PCI Express auto link negotiation feature can not be relied on to correctly detect the link width of a hot-plugged PCI Express card during a hot add operation.

Workaround: Use the PEWIDTH[3:0] bits to set the width of the plugged in card. These bits are defined in the following table:

PEWIDTH[3:0]	Port0 (ESI)	Port2	Port3	Port4	Port5	Port6	Port7
0000	x4	x4	x4	x4	x4	x4	x4
0001	x4	x4	x4	x4	x4	x8	
0010	x4	x4	x4	x8		x4	x4
0011	x4	x4	x4	x8		x8	
0100	x4	x4	x4	x16			
others	Reserved						
1000	x4	x8		x4	x4	x4	x4
1001	x4	x8		x4	x4	x8	
1010	x4	x8		x8		x4	x4
1011	x4	x8		x8		x8	
1100	x4	x8		x16			
others	Reserved						
1111	All port widths determined by link negotiation.						

Status: No Fix

2. MCH B1Err Errors Logged Incorrectly

Problem: B1Err logging does not operate correctly. In the event that an internal data manager parity error occurs B1Err may not be logged. In addition, B1Err may get logged if poisoned data is passed to the internal data manager from the FSB and PCI Express interfaces.

Implication: B1Err error reporting is unreliable.

Workaround: B1Err error logging should be disabled. The lack of B1Err error logging does not cause internal data manager errors to go undetected. An internal data manager parity error will be flagged at the destination interface.

Any parity error in the data manager will cause one of the following:

- A data parity error on the FSB if the destination is FSB
- A poisoned data pattern will be written to memory if the destination is system memory
- EP bit will be set on outbound PCI Express packets if the destination is one of the PCI Express ports

Status: No Fix.

3. PCI Express* IBIST on x8/x16 Port Will Not Stop Testing of Entire Port if an Error is Detected on Any Lane Other Than Lowest 4 Lanes of the Port

Problem: During PCIe IBIST "stop on error" testing, if an error is detected on a x4 port then the test will stop for the entire port and the error is logged. Likewise, if an error is detected on the lower 4 lanes of a x8 port, the test will stop for the entire port and the error is logged. If an error is detected on the upper 4 lanes of a x8 port, the error is logged, the test stops running on the upper 4 lanes, but the test will continue on the lower 4 lanes of the port.
Likewise, if an error is detected on the lowest 4 lanes of a x16 port, the test will stop for the entire port and the error is logged. If an error is detected on any of the upper 12 lanes of a x16 port, the error is logged, the test stops running on the 4-lane group that the error occurred on, but the test will continue on the other 12 lanes of the port.

Implication: IBIST test will continue running in spite of errors if errors occur anywhere other than lowest 4 lanes of a x8/x16 port.

Workaround: If "stop on error" functionality is desired, changes will have to be made to the IBIST scripts to enable this functionality by polling the global status registers. If an error occurred in only the upper lanes, halt IBIST (if in loop continuous mode) as the IBIST engine for that set of 4 lanes has stopped (if stop on error is enabled). If an error occurred in the lowest 4 lanes, all of the IBIST engines will be forced to stop (exit from LoopBack). Check the loop count error status register to determine what set of 4 lanes failed first. One thing to note is that the upper lanes will likely record an error in all lanes whereas the lower 4 may or may not depending on the nature of the real error. Whether the ports are in x4, x8 or x16 mode, the DIO0IBSTAT and DIO0IBERR registers need to be checked to determine which set of 4 lanes had an error and which set of 4 lanes did not.

Status: No Fix

4. MCH Thermal Sensor Reporting Incorrect Values

Problem: The values reported in the MCH Thermal Sensor Output Register (THRMSR_OP) do not match the actual temperature of the MCH.

Implication: Thermal sensor cannot be used for thermal management activities.

Workaround: None

Status: No Fix

5. Crystal Beach Channel Completion Address Logged Twice in FERR_CHANCMP Register

Problem: A 32-bit channel completion address is logged in both the high order 32 bits and the low order 32 bits of the 64-bit FERR_CHANCMP register in Crystal Beach.

Implication: Upper 32 bits of the 64-bit completion address are not logged.

Workaround: The correct upper 32 bits of the FERR_CHANCMP register can be obtained by getting the channel number logged from the FERR_CHANSTS.FERR_DMA Channel Number field (Device 8, Function 0, Offset 84h, bits 4:3) and reading the CHANCMP (Device 8, Function 1, Offsets: 218h, 198h, 118h, 98h) register for that particular channel. This should cover all usage models, unless software reprograms the CHANCMP register after the error occurred and before this reading occurs.

Status: No Fix

6. PCI Express* Receiver Error Logged Upon Putting Link in Disable State

Problem: When the link is told to go to Disable it is possible for the receive side to incorrectly log an error.

Implication: Receiver error incorrectly logged.

Workaround: To enter link disable, set receiver error mask bit in EMASK_COR_PEX and then set the link disable bit. To exit link disable, clear link disable bit and then clear the receiver error mask bit in EMASK_COR_PEX.

Status: No Fix

7. PCI Express Hot-Plug ABP Bit Set While Attention Button is Pressed

Problem: The Attention Button Pressed (ABP) bit (bit 0) in the PEXSLOTSTS register (Device 0, 2-7, Function 0, Offset 86h) is asserted while the hotplug attention button is pressed. Software is supposed to clear this bit after the field has been read and processed but this clear operation could happen before the button is released, causing it to be set again.

Implication: Multiple attention button presses could be registered and processed during a single button press.

Workaround: Hot-plug drivers in Operating Systems capable of supporting native PCI Express hot-plug will not be supported. BIOS should use the _OSC method to maintain control of hot-plug events thereby only allowing ACPI hot-plug support. BIOS should implement a wait of 200 ms after the interrupt is received if using attention button. If using attention jumper, implement a wait of 2200 ms. Please contact your Intel representative to get the latest revision of the *Intel® 5000 Chipset BIOS Specification Update* for more details on the workaround.

Status: No Fix

8. Leakage from 1.5V VCC to 1.2V VTT

Problem: The latest revision of the *Intel® Xeon® Processor-Based Servers – Platform Design Guide (PDG)* specifies that the MCH 1.5 V VCC power rail must ramp ahead of the MCH 1.2 V VTT power rail. During this power-up Series, a leakage path exists within the MCH from the 1.5 V VCC pins to the 1.2 V VTT pins. This leakage path only exists while the MCH is powering up.

Implication: Higher than expected current seen on MCH 1.2V VTT power rail while the 1.5V VCC power rail is ramping. This leakage path does not impact the reliability or functionality of the MCH.

Workaround: None

Status: No Fix

9. Surprise Link Down Error Reported When PCI Express Slot Power is Removed During Hot-plug Event

Problem: When power to a PCI Express slot is removed during a hot remove event, a surprise link down error is incorrectly logged.

Implication: Surprise link down error is incorrectly logged.

Workaround: Hot-plug drivers in operating systems capable of supporting native PCI Express hot-plug will not be supported.

Perform the following:
When turning OFF power to slot:
Set bit 5 on EMASK_UNCOR_PEX register
Set bit 10 on PEXSLOTCTRL register

When turning ON power to slot:
Clear bit 5 on EMASK_UNCOR_PEX register
Clear bit 10 on PEXSLOTCTRL register

BIOS should use the _OSC method to maintain control of hot-plug events, thereby only allowing ACPI hot-plug support. Please contact your Intel representative to get the latest revision of the *Intel® 5000 Chipset BIOS Specification Update* for more details on the workaround.

Status: No Fix

10. System Hang with Large Number of Transaction Retries

Problem: Platforms can experience a system hang. This hang is characterized by a large number of transaction retries and repeated code fetches, as well as conflicting writes to the same address. No commercially available software has been observed to cause this condition in Intel's validation environment.

Implication: Possible system hang.

Workaround: Revision 0.70 of the Memory Reference Code (MRC) and later revisions include a BIOS workaround that resolves all known instances of this erratum. Please contact your Intel representative to get the latest revision of the *Intel® 5000 Chipset BIOS Specification Update* for more details on the workaround. Intel recommends inclusion of this workaround for all Intel 5000 Series chipset MCH steppings.

Status: No Fix

11. INTL[7:2] Registers Are Not Implemented as Read/Write

Problem: The INTL[7:2] registers are implemented as Read Only (RO) when they should be implemented as Read/Write (RW) as specified in the *PCI Local Bus Specification*.

Implication: PCI compliance tests may report errors.

Workaround: None

Status: No Fix. When making a WHQL submission for "PCI Compliance Test", see WHDC Errata ID number 1522. See also Compliance Issue 597049 in this document.

12. FBD Alert Packet Check Before M21 Error Incorrectly Checks Wrong Channel

Problem: After a southbound CRC error, the AMB drives alert packets. The MCH is supposed to check that the incoming packet is not an alert packet before signalling M21 (FBD Northbound CRC Error on FBD Sync Status). Instead, the MCH incorrectly performs this check on the other channel in the branch. The result is that when an alert packet is received, it can be incorrectly logged as a CRC error on a sync packet, or that a sync packet parity error might not be logged due to an alert packet being received on the other channel at the same time.

Implication: Two possible implications:

1. M21 error logged incorrectly upon receiving alert packet.
2. M21 error not logged when there is a sync parity error on one channel and alert packets received on the other channel.

Workaround: For Implication #1, both M13 (indicating SB CRC error) and M21 will be set - M13 as FERR and M21 as NERR. In this case, ignore M21 as M13 is the correct error in this situation. For Implication #2, M13 gets logged as FERR but M21 is not logged. In this case, once M13 is resolved, the M21 error will be logged as FERR if it happens again. M21 is a parity error on a sync packet. Sync packets contain thermal trip information and are sent at a regular interval (40-42 frames). If a sync packet has a parity error and arrives at the same time as an alert packet is received on the other channel, the sync packet will get lost and M21 is not logged. However, the thermal trip info will be sent in a later sync packet.

Status: No Fix.

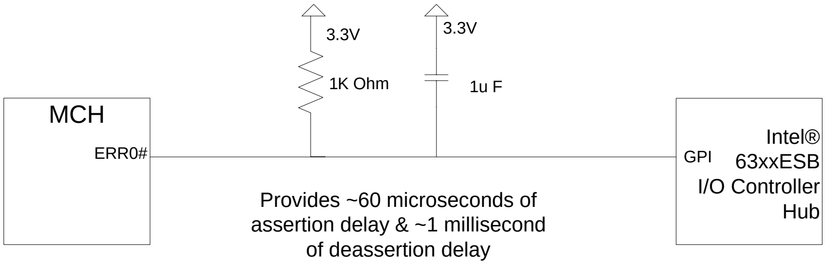
13. SMI Escalation via ERR[2:0]# Pins May Result in IERR#

Problem: Any condition routed to MCH ERR[2:0]# output pins for SMI# escalation that would result in data poisoning on FSB will result in a CPU IERR# assertion due to race condition. The race condition exists between poisoned data presented to the FSB by the MCH and the uncorrectable error escalated to SMI# via ERR[2:0]# pins. If the CPU observes an SMI# assertion before the poisoned data is presented to the FSB, the CPU asserts MCERR# followed by IERR#. The CPU expects the SMI# assertion to occur at least 10 BCLKs after the associated MCERR# assertion. Conditions that result in data poisoning on FSB are uncorrectable memory ECC errors or poisoned TLPs received from any PCI Express port.

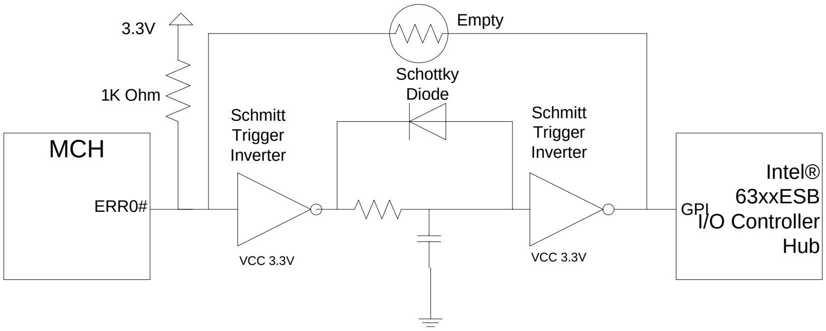
Implication: CPU may assert IERR# and cause a system hang before the error may be logged in SMM.

Workaround: The following workarounds have been identified:

1. Log errors upon reboot
   - System components (MCH, CPU, Southbridge) preserve errors across warm resets via sticky error registers
   - Requires a mechanism to reset system upon IERR# (server management controller, depress system reset button, and so forth)
2. Delay SMI# assertion by inserting a 60µs-150µs delay on the ERR0# pin
   - Two recommended circuit solutions (other solutions may be possible):



Provides ~60 microseconds of assertion delay & ~1 millisecond of deassertion delay



RC filter network can be designed for 60 us – 150 us assertion delay. Diode minimizes deassertion delay.

   - Route errors that result in FSB data poisoning to the ERR0# pin (IO4, M2, M4-13, M15)
3. Disable FSB data parity checking on CPU
   - Prevents CPU from asserting MCERR# on poisoned data

Status: No Fix

14. Read From Remote Branch in Memory Mirrored Mode May Result in Data Corruption

Problem: In memory mirrored mode a read that incurs an error (CRC or Uncorrectable ECC Error) on the first attempt on one branch will be reissued on the remote branch. If the FBD link does not train successfully in the remote branch after the fast link reset the MCH will return 0s to the requestor's read request that originally failed on the first branch. Subsequent reads to the remote branch will also return 0s to the requestor. Note that a concurrent fast link reset is also issued to the remote branch when an error is incurred on the original branch.

Implication: Data corruption may occur if the MCH returns 0s to the requestor in the event the remote branch did not recover after the fast link reset.

Workaround: To prevent data corruption disable Fast Link Reset Timeout (B0:D16h:F1 Offset 19Ch bit 16) when in memory mirrored operation.

Status: No Fix

15. CPU May Record Signal Glitches When MCH is Being Reset

Problem: When the MCH is reset via RSTIN# the CPU may record any one or more of the following errors: address strobe glitch (MSR IA32_MCi_STATUS bit 23), data strobe glitch (MSR IA32_MCi_STATUS bit 22), P/N data strobes out of sync (MSR IA32_MCi_STATUS bit 21), PIC & FSB data parity (MSR IA32_MCi_STATUS bit 19), RSP parity (MSR IA32_MCi_STATUS bit 18), or FSB address parity (MSR IA32_MCi_STATUS bit 16). This can happen when the MCH asserts CPURST# just after the MCH drives an FSB transaction. This may happen because RSTIN# and CPURST# maintain an asynchronous relationship with each other.

Workaround: None.

Status: No Fix

16. Coalesce Mode Cannot Be Used with Max Payload Size of 256B

Problem: When Max Payload Size is set to 256B (PEXDEVCTRL[7:2,0] register, Device 7-2,0, Function 0, Offset 74h, bits 7:5 set to 001), and read completion coalescing is enabled (PEXCTRL[7:2,0] register, Device 7-2,0, Function 0, Offset 48h, bit 10 set to 1), the system may hang. If MPS of 256B is used, coalescing must be disabled. If coalescing is desired, MPS must be set to 128B.

Implication: A system hang may occur.

Workaround: If MPS of 256B is desired, use the following settings:
Set Max Payload Size to 256B (PEXDEVCTRL[7:2,0] register, Device 7-2,0, Function 0, Offset 74h, bits 7:5 set to 001)
Disable coalescing (PEXCTRL[7:2,0] register, Device 7-2,0, Function 0, Offset 48h, bit 10 set to 0)

If coalescing is desired, use the following settings:
Enable coalescing (PEXCTRL[7:2,0] register, Device 7-2,0, Function 0, Offset 48h, bit 10 set to 1)
Set Max Payload Size to 128B (PEXDEVCTRL[7:2,0] register, Device 7-2,0, Function 0, Offset 74h, bits 7:5 set to 000)

In all cases, the following coalesce settings should be used:
Set COALESCE_MODE to 00 (PEXCTRL[7:2,0] register, Device 7-2,0, Function 0, Offset 48h, bits 25:24)
Use Max_rdcmp_lmt_EN default setting of 0 (PEXCTRL[7:2,0] register, Device 7-2,0, Function 0, Offset 48h, bit 12)
Use COALESCE_FORCE default setting of 0 (PEXCTRL[7:2,0] register, Device 7-2,0, Function 0, Offset 48h, bit 11)

Status: No Fix

17. Outstanding Write Transaction in Mirrored Mode Will Cause System Hang

Problem: If a write transaction that failed once cannot be replayed the second time on the local branch as a result of an unsuccessful fast link reset or AMB overtemp condition, the system may hang because the MCH does not drop the request as it should.

Implication: A system hang will occur. As a result of this erratum, all RAS situations which involve a fast reset failure will not be handled. A possible list of such failing scenarios is given below. Note that this list is not exhaustive.

- Hard open contact(s) on arbitrary FB-DIMM connector pins
- Arbitrary stuck bit/lane in the FBD channel
- AMB input clock failure
- AMB fails to train after a Fast Reset
- AMB overtemp shutdown
- General AMB failure (not due to overtemp)

Workaround: None

Status: No Fix

18. PCIe* Link Width Degradation During Reset Tests

Problem: PCIe* Links may downshift to a smaller width during warm resets due to elastic buffer read/write pointer misalignment. Warm resets are characterized by toggling of the RSTIN# input signal to the MCH without toggling the PWRGD input.

Implication: Performance impact on ports which are expected to normally train at non-degraded link width.

Workaround: BIOS is suggested to perform the following workaround as described in pseudocode during BIOS initialization. Please contact your Intel representative to get the latest revision of the *Intel® 5000 Chipset BIOS Specification Update* for more details on the workaround.

Note: This workaround additionally provides a foundation for workarounds identified for erratum 5015621

Workaround Pseudocode

```
Counter 2 (Cnt_2) should be sticky
;read LnkNeg, LnkCap, DnLnkCap and DnLnkNeg
;Cnt_1 = 0
;while ((LnkNeg != LnkCap) && (DnLnkCap != DnLnkNeg) && (DllAct == 1) && (Cnt_2 < 3)){
    ;while (((LnkNeg != LnkCap) && (DnLnkCap != DnLnkNeg) && (DllAct == 1) && (Cnt_1 < 3)) || TO_flag == 1) {
        ;TO_flag = 0
        ;Cnt_1++
        ;set bit 4 in register 0x7C //Drive link to disable
        ;wait 3mS
        ;set bit 23 in register 0x31C //Reset LTSSM
        ;wait 10mS
        ;clr bit 0 in register 0x110 //Clear receiver errors
        ;clr bit 4 in register 0x7C //Clear the link disable bit
        ;clr bit 23 in register 0x31C //Release LTSSM
        ;wait for bit DllAct == 1 or timeout of 120mS. If timeout then TO_flag = 1
        ;read LnkNeg, LnkCap, DnLnkCap and DnLnkNeg
    ;}
    ;read rcverr (bit 0 offset 0x110)
    ;if ((LnkNeg != LnkCap) && (DnLnkCap != DnLnkNeg)) && (rcverr == 1)){
        ;Cnt_2++
        ;0xCF9 = 0x6 // Warm reset
    ;}
;}
;read LnkNeg for each port and save for later use in erratum 501621's workaround routine.
; Note: Upon hot-plug events the stored LnkNeg width will need to be updated because the links would have retrained.
```

Status: Plan Fix

19. PCIe Surprise Link Down or Link Degrade During L1 Exit

Problem: The MCH may clock in noise at the receivers during electrical idle. This condition may result in 2 conditions.

1. Links may experience a surprise link down condition with an IO19Err assertion upon L1 exit. The link will retrain.
2. Links may alternatively downshift link width upon L1 exit with an IO12Err (Receiver Error) assertion. In this case IO19Err will **not** be asserted.

This issue does not affect x1 link configurations.

Implication: Condition #1 will result in an assertion of UNCERRSTS[7:2].IO19Err. Operating systems compliant to the PCI Power Management Specification Rev. 1.1 will restore the endpoint context upon L1 exit. Wake on LAN with the “interesting” packet method may be impacted if the “interesting” packet is stored in volatile state on the adapter. The surprise link down event in this “interesting” packet scenario will reset the endpoint which may cause the network adapter to be unresponsive to “interesting” packets. The “interesting” packet state may be reset and lost upon a surprise link down event. The OS will not be able to restore volatile space on the adapter because it is unaware of its existence.

Condition #2 will result in a link width downshift with an assertion of CORERRSTS[7:2].IO12Err. A performance impact may be experienced with a downshifted link. For example, a x8 link may downshift to x4, x2 or x1.

To workaround condition #1, Intel suggests that BIOS suppress the escalation of UNCERRSTS[7:2].IO19Err. This is performed by setting UNCERRMSK[7:2].IO19Msk. This software workaround allows standard operating systems to survive the issue. Current operating systems restore the state of endpoints upon exiting the L1 state per PCI Power Management Specification Rev. 1.1. To minimize susceptibility to the Wake on LAN “interesting” packet usage model the “magic” packet method may be employed or x1 WOL adapters be employed.

To reduce susceptibility to condition #2 Intel suggests the following SMI-based workaround to up-configure the link width to full width before the link returns to the L0 state. Once the link has returned to the L0 state operating systems compliant to PCI Power Management Rev. 1.1 will restore the endpoint context appropriately. Pseudocode is described here.

Condition 2 Pseudocode

Assumptions:

1. Final un-degraded link widths are stored in a table during POST = saved link width
2. Correctable errors should be unmasked and routed to a error pin Err[1] or Err[2]
3. Err[1] or Err[2] routed to a GPI capable of generating SMI. On GPI-generated SMI (appropriate GPE bit set):
4. All Correctable receiver errors in the presence of a degraded link are a result of the errata

```
Check for receive error bit in CORERRSTS for
( all ports with rcv err bit set )

{
CheckLaneWidth:

        If current lane width  == saved link width,  Go To SBR_Done
        If SBR_Count >= 3, Go To SBR_Done
        Else, Issue Secondary Bus Reset (SBR) on the effected port
        (1.5 ms between assert and de-assert)
        wait 100ms (required per 6.6. PCI Express Reset – Rules)
        SBR_Count++
        Jmp  CheckLaneWidth
SBR_Done:

        Re-enable error-reporting and re-arm the GPI SMI generation
        Claim SMI as serviced and resume from SMM mode back to OS
}
```

Please contact your Intel representative to get the latest revision of the *Intel® 5000 Chipset BIOS Specification Update* for more details on the workaround.

Status: Plan Fix

20. L1 entry Hangs When x4 Links Have Downshifted to x2

Problem: If a PCI Express link downshifts from x4 width to x2 and subsequently enters L1, a system hang may occur. Only x4 strapped ports are susceptible to this issue. Should a x4 to x2 downshift occur, there is a 50% chance an L1 entry will cause a system hang. This condition occurs because of arbitration issues between the link & PHY layer which cause the L1 handshake to stall on a x4 strapped link that has downshifted to x2. All other strapped widths are not affected (x16, x8, x1).

Implication: A system hang may occur.

Workaround: A runtime hardware monitor workaround is available. It will detect x4 to x2 downshifts and force the link width to x1 to prevent x2 negotiation and the potential system hang. This workaround is only available for **only one** x4 strapped port on the MCH. Once this is accomplished the workaround for erratum 501621 can be used to up-configure the link width to full width. Operating systems compliant to PCI Power Management Rev. 1.1 will restore the endpoint context appropriately upon L1 exit. Please contact your Intel representative to get the latest revision of the *Intel® 5000 Chipset BIOS Specification Update* for more details on the workaround.

Status: Plan Fix

21. MCH PCI Express L0s issue

Problem: A downstream PCI Express device on exit from L0s state may be exposed to a Surprise Link Down condition.

Implication: Downstream PCI Express device may reset due to the surprise link down condition. Intel 5000 Chipset Series-based platforms will not support any device entering the L0s state.

Workaround: BIOS should not enable L0s.

Status: Plan Fix.

22. MCH may log F2Err during shutdown special cycle initiated due to FSB timeout

Problem: BNB will flag a FSB F2Err when a CPU issues a Shutdown Special Cycle due to FSB timeout with the deferred response enable (DEN#) not asserted. The CPU cycle is valid per FSB protocol.

Implication: An invalid FSB F2Err error is logged. This issue does not affect the Dual-Core Intel® Xeon® Processor 5000/5063 Series.

Workaround: There are 2 possible workarounds
1. BIOS may disable error logging for FSB F2Err by setting bit[1] of EMASK_FSB[1:0] registers (Device:Function:Offset = 0n16:0:0x492,0x192).
2. Upon reboot error management software can check the CPU machine check registers for a FSB timeout status. If the FSB timeout status is asserted and the MCH is flagging F2Err then the F2Err assertion can be ignored.

Status: No Fix.

23. System hang with multiple retries and locks

Problem: Intel 5000 Chipset Series based platforms can experience a system hang. This hang is characterized by one BREQ agent performing a large number of retried reads and writes from separate cores, while another BREQ agent on the same bus is performing bus locks. Intel has not observed this erratum with any commercially available software.

Implication: A system hang may occur. The B2, B3, and G0 MCH steppings are only compatible with the Intel Xeon Processor 5000/5063 Series and Intel Xeon 5100 processors.

Workaround: None

Status: Plan Fix.

24. Patrol Scrub with CRC error issue

Problem: If a Patrol Scrub encounters a CRC error it will write back poisoned data in to memory.

Implication: A M12Err or M9Err followed by M4 Uncorrectable ECC error (if M4 is enabled) will be flagged. Poisoned data will be written to the memory location and a machine check will happen if the corrupted memory location is read.

Workaround: Turn off Patrol Scrub. Demand Scrub can be enabled in Non-Mirrored mode. Note: demand scrub is not supported in Mirrored mode. Turn off Patrol Scrub by setting Bus 0, Device 16, Function 1, Offset 40h, bit 7 to 0.

Status: Plan Fix.

25. Header of malformed TLPs on a x16 port not always logged

Problem: Intel 5000 Chipset Series based platforms may not log header information for an illegal length malformed TLP on a x16 port. This happens when the STP and END occur in the same symbol time as follows: H H H E/IL. Where H means Header Dword and E/IL is END symbol with illegal length. Symbol time means same clock cycle from the transaction layers point of view. For example a x16 port will see 4 DWords of information in a single cycle.

Implication: HDRLOG registers may not contain the header of malformed TLP.

Workaround: None.

Status: No Fix.

26. Illegal addresses within the 40-bit address space in the channel completion address register does not generate cmp_addr_err

Problem: The Intel 5000 Chipset DMA engine will not flag a cmp_addr_err (DMA12) when illegal addresses within the 40-bit address space are programmed in the channel completion address register.

Implication: cmp_addr_err (DMA12) bit will not get set when illegal addresses within the 40-bit address space are programmed in the channel completion address register. The error is flagged correctly when an address greater than 40 bits is programmed in the channel completion address register. However, in all cases where an illegal address is programmed, the illegal address will be caught and the transaction will be master aborted

Workaround: None.

Status: No Fix.

27. RID=CRID is sticky across warm reset

Problem: The Revision Identification Register (RID) is keeping the value of Compatibility Revision ID (CRID) across a warm reset (RID=CRID) if RID is set to CRID prior to the warm reset. After a warm reset, RID should have the value of Stepping Revision ID (SRID) (RID=SRID).

Implication: When RID is set to show CRID's value and then a warm reset occurs RID will not be set to show SRID's value. RID will continue to show CRID's value. After a warm reset, it is expected that RID will show SRID's value.

Workaround: Prior to setting RID=CRID BIOS/ Software can store a copy of RID=SRID in a Sticky Scratch Pad Register, SPADS3[7:0] for example. After a warm reset BIOS/ Software can get the value of SRID from the Sticky Scratch Pad Register.

Status: No Fix.

28. SLD could causes spurious completions

Problem: When a link goes down due to surprise link down (SLD), Intel 5000 chipset may not drain all transaction layer packets (TLPs) before allowing the link to come back up.

Implication: In cases, where Intel 5000 chipset has not fully drained all downstream completions, a downstream device coming out of reset will see spurious completions.

Workaround: After an SLD, it is recommended that software clear the downstream device and root port of all errors.

Status: No Fix.

29. IBIST does not capture failed lanes properly

Problem: The register that holds lane statistics is not sticky.

Implication: This means that only those lanes that detect an error last will be remembered. For example, if lanes 1 & 2 failed in IBIST cycle N and lanes 3 & 4 failed in cycle N+M, the error status register will only reflect lanes 3 & 4 as failed and will not show lanes 1 & 2.

Workaround: None

Status: No Fix.

30. IBIST RX logic does not stop when IBISTR is reset

Problem: When IBIST is started in continuous mode and the IBISTR bit is cleared, the IBIST rx logic does not stop.

Implication: IBIST rx logic will continue comparing whatever data shows up at the input, even after the IBISTR bit is cleared.

Workaround: Clear the continuous mode bit before or along with the start bit.

Status: No Fix.

31. FBD Channel Index is incorrect when logging some error conditions

Problem: In the FB-DIMM First Fatal Errors (FERR_FAT_FBD) register and FB-DIMM First Non-Fatal Errors (FERR_NF_FBD) register the channel index (FBDChan_Indx) is not correctly reporting which lane is experiencing an error.

Implication: FBDChan_Indx may not report the correct channel in which the error occurred.

Workaround: None.

Status: No Fix.

32. PCIE x16 link goes to recovery with a x1 card and L0s

Problem: If a PCIe x16 link coming off of the Intel 5000 chipset MCH degrades to a x1 link and the L0s state is enabled the system may hang.

Implication: This means that if a PCIe card plugged into the Intel 5000 chipset MCH's x16 port degrades to a x1 link and the L0s state is enable the system may hang under heavy load conditions. A x1 PCIe card plugged into the Intel 5000 chipset MCH's x16 port with the L0s state enabled may also hang.

Workaround: BIOS should not enable L0s.

Status: No Fix.

33. SLD on PCIE port during P2P Posted requests can causes ESI to hang

Problem: Under specific conditions a Surprise Link Down (SLD) could cause the ESI port to hang. For this to happen a peer to peer (P2P) write, with the source connected to the south bridge and destination connected to the north bridge must be in progress. If an SLD occurs on the target link during certain phases of the MCH processing the request the ESI port may hang.

Implication: This bug causes the ESI port to hang which will hang the system.

Workaround: None.

Status: No Fix.

34. PEXGCTRL.PME_TO_ACK may not be set when a Turn Off Acknowledge TLP have been Received from All Ports

Problem: When commanded to send PME Turn Off messages by setting Bit[1], PME_TURN_OFF, of the PEXGCTRL Register (Device 19, Function 0, Offset 17Ch) and a PME_TO_ACK TLP is received from a port before another port has sent its PME Turn Off message then Bit[0], PME_TO_ACK, of the PEXGCTRL Register (Device 19, Function 0, Offset 17Ch) may not be set when all the acknowledges have been received.

Implication: When a system is transitioning from S0 to S3, S4 or S5 a system hang may occur if the PCI Express Global Control Registers Completion Timeout value has been set too low.

Workaround: Set Bits[31:18], Completion Timeout, of the PEXGCTRL Register (Device 19, Function 0, Offset 17Ch) to a value greater than 10 mS.

Status: For the steppings affected, see the Summary Table of Changes.

35. Masked Completer Abort Status Errors may be Reported in the UNCERRSTS[0] Register

Problem: When Bit [15], IO7MSK, of the UNCERRMSK[0] register (Device 0, Function 0, Offset 108h) is set to mask Completer Abort status errors, uncorrectable Completer Abort status errors may still be logged in Bit [15], IO7ERR, of the UNCERRSTS[0] register (Device 0, Function 0, Offset 104h) register.

Implication: A masked Completer Abort status error may be reported.

Workaround: None Identified.

Status: For the steppings affected, see the Summary Table of Changes.

36. SMBus 2.0 Specification TLOW:SEXT may be exceeded on SMBus 0 when the North Bridge is clocked with a 266 MHz BUSCLK

Problem: The SMBus TLOW:SEXT specification of 25 mS MAX, which applies only to Slave ports, may be exceeded on SMBus 0 by devices with a BUSCLK of 266 MHz.

Implication: A TLOW of ~31 mS may occur on SMBus 0. A Master is allowed to abort the transaction in progress to any slave that violates the TLOW:SEXT specification.

Workaround: None Identified.

Status: For the steppings affected, see the Summary Table of Changes.

37. The First Uncorrectable Fatal bit of the Root Error Status Register may be Incorrectly Set when a Second Uncorrectable Fatal Error is Received

Problem: Bit [4], FRST_UNCOR_FATAL, of the RPERRSTS Register (Device: 0, 2-3 and 4-7, Function 0, Offset 130h) may be incorrectly set if an Uncorrectable Fatal Error is received and Bit [2], ERR_FAT_NOFAT_RCVD, of the RPERRSTS Register (Device 0, 2-3 and 4-7, Function 0, Offset 130h) is set.

Implication: An Uncorrectable Fatal Error which was not the first error received may be incorrectly indicated as the type of error that was first received.

Workaround: None Identified.

Status: For the steppings affected, see the Summary Table of Changes.

38. Bit [3], INTxST, of the PCISTS Register may be Cleared when Bit [10], INTxDisable, of the PCICMD Register is Set

Problem: When Bit [10], INTxDisable, of the PCICMD register (Device 0, 2-3 and 4-8, Function 0, Offset 04h) is set to disable interrupts then Bit [3], INTxST, of the PCISTS register (Device 0, 2-3 and 4-8, Function 0, Offset 06h) register may be cleared.

Implication: An interrupt status that is pending may be incorrectly cleared.

Workaround: None identified.

Status: For the steppings affected, see the Summary Table of Changes.

39. The DMA Engines Next Channel Error Register is not Updated When Subsequent Errors are Detected

Problem: The NERR_CHANERR register (Device: 8, Function: 0, Offset: BCH) is not updated when additional error are detected, only the second error is logged.

Implication: The latest error status is not captured.

Workaround: None Identified.

Status: For the steppings affected, see the Summary Table of Changes.

40. Store to Write-Through (WT) Memory Data May be Seen in Wrong Order by Two Subsequent Loads When Snoop Filter is Enabled

Problem: If the data of a store transaction to WT memory is used by two subsequent loads of one thread and another thread performs a store to the same address, and the Intel® 5000X chipset filters out the snoop request triggered by the second store, then the first load may get the data from external memory or the L2 cache as written by another core while the second load gets the data straight from the WT store transaction.

Implication: Software that uses WT memory with shared data may violate proper store ordering. Intel has not observed this erratum with any commercially available software.

Workaround: After the write operation to shared data area of WT memory type, software may use the SFENCE instruction before accessing this data.

Status: For the steppings affected, see the Summary Table of Changes.

41. PCI-Express transaction IO ordering queue overflow

Problem: Under some corner case scenarios when the system is stressed with heavy IO traffic a hang condition could occur as the transactions to/from PCIE port are blocked due to IO order queue overflow

Implication: The overflow condition will eventually cause the system to deadlock because many outstanding transactions are unable to make forward progress. System level implication is a system hang with IERR or MCERR or PCIe fatal error assertion NMI, if enabled.

Workaround: Please refer to the latest Intel® 5000 Series MCH BIOS Specification update.

Status: No-Fix.

42. System failures during spare copy

Problem: In some rare cases a memory transaction may not complete when patrol scrub is enabled and a DIMM spare copy is started.

Implication: A system hang may occur

Workaround: None.

Status: No fix.

43. Read transactions may be delayed

Problem: Under a certain sequence of read and write transactions issued from processors or bus mastering I/O devices, the read transaction may be delayed.

Implication: A read transaction may be delayed. Intel has not observed this behavior with any commercially available software.

Workaround: None.

Status: For the steppings affected, see the Summary Table of Changes.

44. CRC errors logged by AMB during C2 or S1 transition

Problem: AMB occasionally logs CRC errors while the SB lanes are transitioning into Electrical Idle for C2 or S1 power state (MCH FBD fails to set IgnoreERR in the SYNC packet to the AMB preceding Electrical Idle entry).

Implication: Results in bogus errors being set in the AMB error log.

Workaround: The following AMB register fields must be cleared upon C2 or S1 exit: Function 1, Offset 90h and 94h, bit 0.

Status: No Fix

45. PCI express TLP packets not flagged “Malformed” under certain conditions.

Problem: PCI express Transaction Layer Protocol (TLP) packets that are of exactly 256B in length will not be flagged with Malformed TLP when the Maximum Payload Size (MPS) is set to 128B in the PEXDEVCTRL register. All packet sizes greater than 128B except for exactly 256B are correctly reported as a Malformed TLP in the error reporting registers. This does not affect systems with the MPS set to 256B, in this case all packets greater than 256B are correctly reported as Malformed TLP in the error reporting registers.

Implication: When the MPS is set to 128B, if the PCI express end point incorrectly transmits a packet that is exactly 256B the Intel® 5100 MCH will process the packet and will not report a malformed TLP error. **Note**: There is an errata 16 on MPS setting to 256B, please refer to item 16 for details.

Workaround: None

Status: No Fix

46. Error Source Identification (ID) is not properly reporting the Requestor ID when the uncorrectable (Non-fatal/fatal) error is detected in the PCI express Root Port

Problem: The event collector for uncorrectable error source ID in the Root Complex of the PCI express ports reported in RPERRSID[7:2,0] register under bits[31:16] ERR_FAT_NOFAT_SID field is not capturing the Requestor ID of the source when a Fatal or Non Fatal error is received by the Root Port.

Implication: The value reported in the RPERRSID[7:2,0][ERR_FAT_NOFAT_SID] does not represent the source of the uncorrectable (Non-fatal/fatal) error detected by the root port.

Workaround: Do not use RPERRSID[7:2,0][ERR_FAT_NOFAT_SID] information when a uncorrectable error is detected by the PCI express Root Port.

Status: No Fix

Specification Changes

This document revision contains no Specification Changes.

Specification Clarifications

1. Software Guidance for MSI handling

There are two conditions under which the MCH expects software to handle Message Signaled Interrupts (MSI) appropriately. The first is if one or more interrupt status bits are set to ‘1’ and a new bit gets set to ‘1’. The MCH will send an MSI for the new bit. This may cause extraneous Interrupt Service Routine (ISR) calls. The second condition occurs when one or more interrupt status bits are set to ‘1’ and software clears some (but not all) bits. The MCH will not send an MSI for the remaining uncleared bits. This may cause a lost interrupts.

Interrupt service routines (ISR) should record all events in the status registers that they process and clear all events detected. There should be no lingering status upon ISR exit. It is software’s responsibility to handle MSI’s.

If an ISR does not follow the requirement to read the Interrupt Control register (INTRCTRL), and the ISR is called twice back-to-back such that it reads Attention Status register (ATTNSTATUS) both times, then a 2nd MSI could overwrites the first one and hang the system.

2. SMBus and PCIe interoperability time out recommendation

The System Management Bus (SMBus) Specification Version 2.0, Section 3.1.1 SMBus common AC specifications states the timeout value to be 25 ms and the PCI EXPRESS BASE SPECIFICATION, REV. 1.0a section 2.8 Completion Timeout Mechanism specifies PCIe global timeout of 50 ms. Due to an implementation problem with the Intel 5000 Chipset SMBus logic, once a timeout occurs, SMBus hangs until a hard-reset. Any SMBus transaction through the Intel 5000 Chipset MCH targeting any device outside of the Intel 5000 Chipset MCH taking longer than 25 ms to complete will cause the SMBus to hang.

Intel recommends that the PCI Express Global Control Register (PEXGCTRL) Completion Timeout be programmed to less than the SMBus time out. Currently Intel sets the timeout to approximately 20 ms (Completion timeout register value = 744h) in BIOS. Please refer to the Intel® 5000P/5000V/5000Z Chipset Memory Controller Hub (MCH) External Design Specification (EDS) Rev 2.0 (document order# 21099) for detailed information on PEXGCTRL register. For an Intel 5000 Chipset based system to encounter this error (SMBus hang) two conditions must be met. One, the Intel 5000 Chipset based system would need to have a PEXGCTRL Completion Timeout value greater than 25 ms. Two, an SMBus master, such as a BMC, on the Intel 5000 Chipset based system would have to read data from a PCI or PCIe device through the MCH and if this operation takes more than 25 ms, then the SMBus will hang. It is important to note that an SMBus master reading from the Intel 5000 Chipset MCH (Bus 0) will not encounter the SMBus hang. Systems that do not access any SMBUS devices outside of the Intel 5000 Chipset MCH through the Intel 5000 Chipset MCH can safely set the timeout value to 50 ms to conform to the PCI EXPRESS BASE SPECIFICATION. There is no planned fix for this.

For further clarification an example scenario of a case in which a change to the recommended time out is required. In this scenario a video card is unable to respond within the recommended time out. Furthermore, the default Windows driver maps the video frame buffer in MMIO space as Uncacheable Speculative Write Combining (USWC) memory. The CPU may speculatively execute a load (read) to the frame buffer region. The video controller may be unable to respond to the speculative read and could issue a retry. The system will continue to reissue the load to the video controller which may respond back with a retry. After 20 mS, the Intel 5000 Chipset timer expires and generates an NMI followed by a Windows blue screen. To work around this problem set the PEXGCTRL Completion Timeout to max value of 3FFFh (approximately 220 ms). The PEXGCTRL timeout should be set to 744h for any SMBus transactions through the Intel 5000 Chipset targeting a PCI Express device outside of the 5000 series chipset. Once the SMBus transactions are completed, the PEXGCTRL timeout should be set back to the max value of 3FFFh. Ensure that the system is silent on the PCI Express side before changing the timer value (3FFFh to 744h and vice versa).

3. Intel® 5000 Series Chipsets Interrupt Swizzling Recommendation

The Intel® 5000 Series Chipset MCH has interrupt swizzling logic to rebalance and distribute inbound PCIe interrupts for performance and load balancing considerations. BNB.INTxSWZCTRL[7:2,0].INTxSWZ (D[7:2,0]:F0:Offset_4Fh[1:0]) provides software/BIOS the ability to swizzle the PCIe interrupt (INTx) from each PCI Express Port and remap them to a different interrupt pin.

INTA (as depicted in Figure-2) is usually overloaded since it is reserved per the PCI/PCI Express spec for PCI/PCI Express devices (if the device uses interrupts), so is of particular concern. Figure-1 depicts interrupt swizzling where INTx is distributed. Figure-1 and Figure-2 depict an example of interrupt swizzling and should ONLY be used for illustration purposes, as system board interrupt routing is very platform specific.

For optimal system performance, it is recommended that system BIOS utilizes this register to achieve a more balanced PCIe interrupt distribution. Note that the PCI interrupt routing table in the system BIOS needs to be modified to match the particular swizzling scheme selected.

For additional information please refer to document # 314337 *Interrupt Swizzling Solution for the Intel® 5000 Chipset Series based Platforms - Application Note* available on www.intel.com.

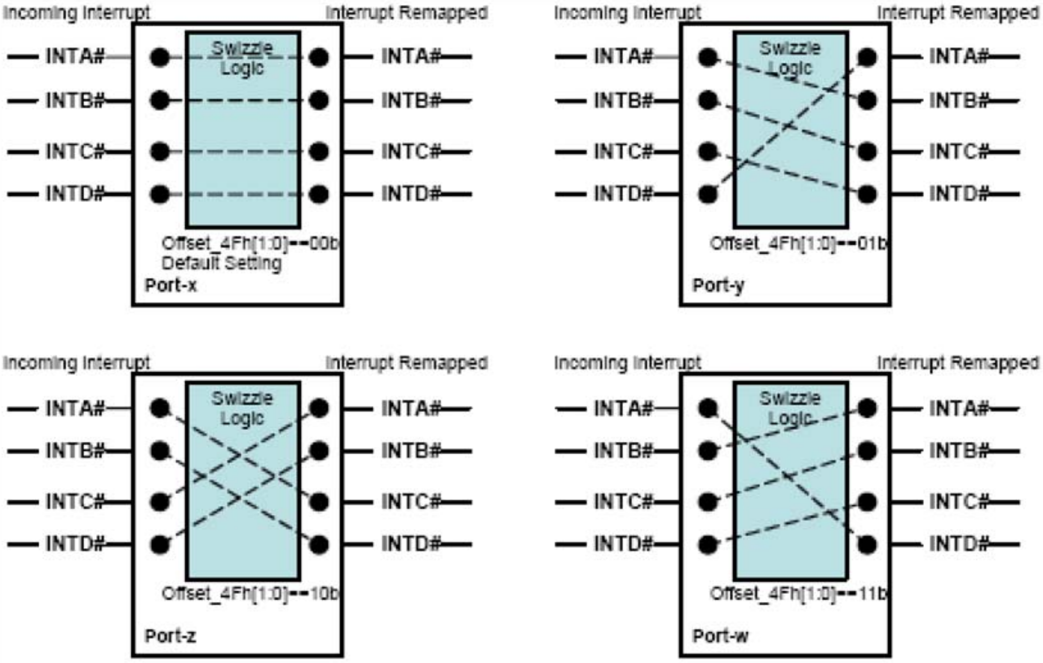


Figure-1: Interrupt Swizzling

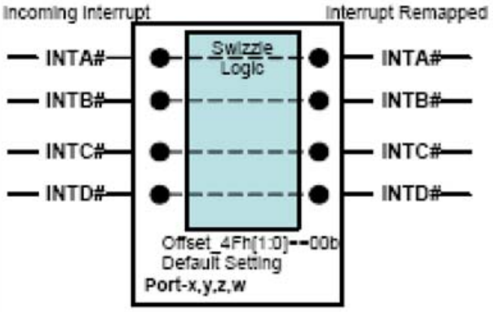


Figure-2: No Interrupt Swizzling

4. Issues with some PCI Express adapters during Link Training and L1 exit

Intel® 5000 Series Chipset MCH PCI Express devices rely on a Skip Ordered Set to perform link deskew during a system or link reset, and during an exit from an L1 state. This is expected to be received during the final stages of training, Config.Complete and Recovery.RcvrCfg respectively.

Some adapters have been observed to stop sending Skip-Order Sets in Config.Idle or Recovery.Idle states and this may cause issues with link training or errors to be logged during the training process even-though the MCH is operating within spec.

Compliance Issues

1. PCIe compliance test failure: CBDMA Interrupt Line Register

Problem: The PCIe compliance test TD_1_13 Interrupt Pin - Interrupt Line Register Test reports a RW failure for the interrupt Line Register bits[7:0] in Intel 5000 Chipset devices. According to the PCIe Specification (PCI EXPRESS BASE SPECIFICATION, REV. 1.1, Section 7.5.1.5), the Interrupt Line Register bits[7:0] should be RW. The INTL register is implemented as Read/Write Once (RWO) in Intel 5000 Chipset devices. The register becomes RO after BIOS initializes it, violating the PCIe specification. This is a minor violation because the Crystal Beach device does not have interrupt lines and INTL is not used by the Intel 5000 Chipset.

Implication: NA

Workaround: NA

Status: No Fix. Microsoft has decided not to make this a requirement for any OS WHQL. See also Errata 501247 in this document.

2. PCIe compliance test failure: MSI Address Register

Problem: The PCIe compliance test TD_1_06 MSI Capability Structure Test reports a RO failure for the MSIAR bits[31:20] in Intel 5000 Chipset devices. According to the PCI Specification (Conventional PCI, REV. 3.0, Section 6.8.1.4. Message Address for MSI), the Message Address Register bits[31:2] should be RW. The MSIAR in the Intel 5000 Chipset is a root port and is fixed to Intel specific IO_APIC range of 0xFEE*h to route the MSI ensuring proper functionality of the MSI architecture.

Implication: NA

Workaround: NA

Status: No Fix.

Documentation Changes

1. PEXGCTRL - PCI Express Global Control Register

Device: 19
Function: 0
Offset: 17Ch
Version: Intel 5000P chipset, Intel 5000V chipset, Intel 5000Z chipset

Bit	Attr	Default	Description
1	RWST	0	**PME_TURN_OFF: Send PME Turn Off Message** When set, the Intel 5000 Chipset MCH will issue a PME Turn Off Message to all enabled PCI Express ports excluding the ESI port. The Intel 5000 Chipset MCH will clear this bit once the Message is sent. • NOTE: In the Intel 5000 Chipset MCH implementation, an end device that is D3 PM state and the Link being in L2 will not respond to any transaction to the device until it is woken up by the WAKE# signal in the platform. Under these conditions, if software sets the PME_Turn_OFF (bit 1) of this register, the Intel 5000 Chipset MCH will not send the message until the Link is brought back into L0. i.e. PME_TURN_OFF bit will remain set until the message is dispatched. Furthermore, a surprise link Down error is logged. † Expected Usage: Software should not set this bit if the link is already in L2 prior.

§