Subject: [Q] ide cdrom in native mode leads to irq storm?
Posted by vaverin on Tue, 24 Oct 2006 07:37:13 GMT
View Forum Message <> Reply to Message

there is node with Intel 7520-based motherboard (MSI-9136), IDE cdrom (hda) and SATA disc and 2.6.19-rc3 linux kernel.

When I set IDE controller into the native mode, I get irq storm on the node and this interrupt is disabled. If this interrupt is shared, the other subsystems are stop working too.

When I switch the IDE controller into legacy mode, all works correctly.

I've tried to use noapic, acpi=off, pci=routeirq, irqpoll options but it does not help.

This issue is reproduced on the old kernels (2.6.15-1.2054_FC5smp and latest RHEL4 kernel) too.

Is it probably a known issue and is there any work-around?

thank you,
Vasily Averin

bootlogs, /proc/interrupts and lspci are below:

```
Linux version 2.6.19-rc3 (vvs@dhcp0-157) (gcc version 3.3.5 20050117
(prerelease) (SUSE Linux)) #1 SMP Tue Oct 24 11:02:23 MSD 2006
...
Uniform Multi-Platform E-IDE driver Revision: 7.00alpha2
ide: Assuming 33MHz system bus speed for PIO modes; override with idebus=xx
ICH5: IDE controller at PCI slot 0000:00:1f.1
ACPI: PCI Interrupt 0000:00:1f.1[A] -> GSI 18 (level, low) -> IRQ 17
ICH5: chipset revision 2
ICH5: 100% native mode on irq 17
    ide0: BM-DMA at 0x1460-0x1467, BIOS settings: hda:DMA, hdb:pio
    ide1: BM-DMA at 0x1468-0x146f, BIOS settings: hdc:pio, hdd:pio
Probing IDE interface ide0...
hda: ATAPI-CD ROM-DRIVE-52MAX, ATAPI CD/DVD-ROM drive
ide0 at 0x1490-0x1497,0x1486 on irq 17
Probing IDE interface ide1...
Probing IDE interface ide1...
...
libata version 2.00 loaded.
ata_piix 0000:00:1f.2: version 2.00ac6
ata_piix 0000:00:1f.2: MAP [ P1 -- P0 -- ]
ACPI: PCI Interrupt 0000:00:1f.2[A] -> GSI 18 (level, low) -> IRQ 17
PCI: Setting latency timer of device 0000:00:1f.2 to 64
```

```
ata1: SATA max UDMA/133 cmd 0x1F0 ctl 0x3F6 bmdma 0x1470 irq 14
ata2: SATA max UDMA/133 cmd 0x170 ctl 0x376 bmdma 0x1478 irq 15
scsi0 : ata_piix
ata1.00: ATA-7, max UDMA/133, 156301488 sectors: LBA48 NCQ (depth 0/32)
ata1.00: ata1: dev 0 multi count 16
ata1.00: configured for UDMA/133
scsi1 : ata_piix
ATA: abnormal status 0x7F on port 0x177
scsi 0:0:0:0: Direct-Access     ATA      ST380811AS       3.AA PQ: 0 ANSI: 5
SCSI device sda: 156301488 512-byte hdwr sectors (80026 MB)
sda: Write Protect is off
sda: Mode Sense: 00 3a 00 00
SCSI device sda: drive cache: write back
SCSI device sda: 156301488 512-byte hdwr sectors (80026 MB)
sda: Write Protect is off
sda: Mode Sense: 00 3a 00 00
SCSI device sda: drive cache: write back
 sda: sda1 sda2 sda3 sda4 < sda5 >
sd 0:0:0:0: Attached scsi disk sda
...
irq 17: nobody cared (try booting with the "irqpoll" option)
 [<c0145eea>] __report_bad_irq+0x2a/0xa0
 [<c014602f>] note_interrupt+0xaf/0xe0
 [<c0146888>] handle_fasteoi_irq+0xc8/0xe0
 [<c01059f9>] do_IRQ+0x69/0xd0
 [<c0103ace>] common_interrupt+0x1a/0x20
 =======================
handlers:
[<c02b30c0>] (ide_intr+0x0/0x170)
Disabling IRQ #17
...
hda: lost interrupt
ide-cd: cmd 0x3 timed out
hda: lost interrupt
ide-cd: cmd 0x3 timed out
...
hda: lost interrupt
ide-cd: cmd 0x1e timed out
hda: lost interrupt


# cat /proc/interrupts
           CPU0       CPU1       CPU2       CPU3
  0:      15923      15011      22615      15936   IO-APIC-edge      timer
  1:          0          0          0          8   IO-APIC-edge      i8042
  6:          3          0          0          1   IO-APIC-edge      floppy
  8:          0          0          0          1   IO-APIC-edge      rtc
  9:          0          0          0          0   IO-APIC-fasteoi   acpi
```

```
 12:        99         0         0         6   IO-APIC-edge      i8042
 14:      3432        69       295        19   IO-APIC-edge      libata
 15:         0         0         0         0   IO-APIC-edge      libata
 17:     99999         0         0         1   IO-APIC-fasteoi   ide0
 18:         0         0         0         0   IO-APIC-fasteoi   uhci_hcd:usb2
 19:      8429         0         0         1   IO-APIC-fasteoi   eth0
 21:         0         0         0         0   IO-APIC-fasteoi   uhci_hcd:usb1
 22:         0         0         0         0   IO-APIC-fasteoi   ehci_hcd:usb3
NMI:         0         0         0         0
LOC:     69336     69336     69338     69329
ERR:         0
MIS:         0

# lspci -vn
...
00:1f.0 0601: 8086:25a1 (rev 02)
        Flags: bus master, medium devsel, latency 0

00:1f.1 0101: 8086:25a2 (rev 02) (prog-if 8f)
        Subsystem: 8086:24d0
        Flags: bus master, medium devsel, latency 0, IRQ 17
        I/O ports at 1490 [size=8]
        I/O ports at 1484 [size=4]
        I/O ports at 1488 [size=8]
        I/O ports at 1480 [size=4]
        I/O ports at 1460 [size=16]
        Memory at d0001800 (32-bit, non-prefetchable) [size=1K]

00:1f.2 0101: 8086:25a3 (rev 02) (prog-if 8a)
        Flags: bus master, 66MHz, medium devsel, latency 0, IRQ 17
        I/O ports at <unassigned>
        I/O ports at <unassigned>
        I/O ports at <unassigned>
        I/O ports at <unassigned>
        I/O ports at 1470 [size=16]

00:1f.3 0c05: 8086:25a4 (rev 02)
        Subsystem: 8086:24d0
        Flags: medium devsel, IRQ 16
        I/O ports at 1440 [size=32]
```

Subject: Re: [Q] ide cdrom in native mode leads to irq storm?
Posted by vaverin on Tue, 24 Oct 2006 07:53:00 GMT

View Forum Message <> Reply to Message

Vasily Averin wrote:
> there is node with Intel 7520-based motherboard (MSI-9136), IDE cdrom (hda) and

> SATA disc and 2.6.19-rc3 linux kernel.
>
> When I set IDE controller into the native mode, I get irq storm on the node and
> this interrupt is disabled. If this interrupt is shared, the other subsystems
> are stop working too.
>
> When I switch the IDE controller into legacy mode, all works correctly.
>
> I've tried to use noapic, acpi=off, pci=routeirq, irqpoll options but it does
> not help.

When I use irqpoll option I get the following oops in create_empty_buffers():
it is not expected that alloc_page_buffers(page, blocksize, 1) can return NULL,
but it does it because of requested blocksize is more than PAGE_SIZE.

Unfortunately I have not any ideas how to fix this issue correctly.

thank you,
Vasily Averin

BUG: unable to handle kernel NULL pointer dereference at virtual address 00000000
printing eip:
c0191790
*pde = 37b31001
Oops: 0002 [#1]
SMP
Modules linked in: thermal processor fan button battery asus_acpi ac lp
parport_pc parport floppy ehci_hcd uhci_hcd sg e1000 i2c_i801 i2c_core ide_cd
cdrom shpchp usbcore
CPU: 0
EIP: 0060:[<c0191790>] Not tainted VLI
EFLAGS: 00010296 (2.6.19-rc3 #1)
EIP is at create_empty_buffers+0x30/0xb0
eax: 00000000 ebx: c16e1360 ecx: c16e1360 edx: 00000000
esi: 00000000 edi: 00000000 ebp: f7a720ac esp: f7f3bc5c
ds: 007b es: 007b ss: 0068
Process lvm.static (pid: 2249, ti=f7f3a000 task=f7bce550 task.ti=f7f3a000)
Stack: c16e1360 00010000 00000001 00010000 00000000 f7a72150 c0192491 c16e1360
00010000 00000000 00000011 f7f3bcb8 c01059fe 00000000 00000000 00000001
00000440 00010000 00000003 c16e0740 f7a72150 00000004 c0103ace 00000000
Call Trace:
[<c0192491>] block_read_full_page+0x251/0x3a0
[<c01059fe>] do_IRQ+0x6e/0xd0
[<c0103ace>] common_interrupt+0x1a/0x20
[<c0147cac>] add_to_page_cache+0x9c/0xc0
[<c014f515>] read_pages+0x45/0x100
[<c0195a10>] blkdev_get_block+0x0/0x80
[<c014d035>] __alloc_pages+0x55/0x320

[<c014f73d>] __do_page_cache_readahead+0x16d/0x180
[<c014f8b9>] blockable_page_cache_readahead+0x59/0xd0
[<c014fb3e>] page_cache_readahead+0x13e/0x1f0
[<c0148980>] do_generic_mapping_read+0x4c0/0x600
[<c0148de4>] generic_file_aio_read+0x214/0x250
[<c0148ac0>] file_read_actor+0x0/0x110
[<c016bbee>] do_sync_read+0xde/0x130
[<c0136e60>] autoremove_wake_function+0x0/0x60
[<f8846d08>] usb_hcd_irq+0x28/0x70 [usbcore]
[<c0145e48>] misrouted_irq+0xd8/0x150
[<c0146016>] note_interrupt+0x96/0xe0
[<c016bcfe>] vfs_read+0xbe/0x1a0
[<c016c101>] sys_read+0x51/0x80
[<c0103147>] syscall_call+0x7/0xb
=======================
Code: 00 00 53 83 ec 0c 8b 5c 24 1c 89 74 24 08 8b 44 24 20 8b 7c 24 24 89 1c 24
89 44 24 04 e8 69 f4 ff ff 89 c6 89 c2 90 8d 74 26 00 <09> 3a 89 d0 8b 52 04 85
d2 75 f5 89 70 04 8b 43 10 83 c0 44 e8
EIP: [<c0191790>] create_empty_buffers+0x30/0xb0 SS:ESP 0068:f7f3bc5c
<3>irq 17: nobody cared (try booting with the "irqpoll" option)
[<c0145eea>] __report_bad_irq+0x2a/0xa0
[<c014602f>] note_interrupt+0xaf/0xe0
[<c0146888>] handle_fasteoi_irq+0xc8/0xe0
[<c01059f9>] do_IRQ+0x69/0xd0
[<c0103ace>] common_interrupt+0x1a/0x20
[<c0101082>] mwait_idle_with_hints+0x32/0x40
[<c01010a8>] mwait_idle+0x18/0x30
[<c0100ef3>] cpu_idle+0x73/0x90
[<c0552a5a>] start_kernel+0x1ca/0x220
[<c0552370>] unknown_bootoption+0x0/0x1e0
=======================
handlers:
[<c02b30c0>] (ide_intr+0x0/0x170)
Disabling IRQ #17

Subject: Re: [Q] ide cdrom in native mode leads to irq storm?
Posted by vaverin on Fri, 27 Oct 2006 13:17:34 GMT
View Forum Message <> Reply to Message

Vasily Averin wrote:
> Vasily Averin wrote:
>> there is node with Intel 7520-based motherboard (MSI-9136), IDE cdrom (hda) and
>> SATA disc and 2.6.19-rc3 linux kernel.
>>
>> When I set IDE controller into the native mode, I get irq storm on the node and
>> this interrupt is disabled. If this interrupt is shared, the other subsystems
>> are stop working too.

>>
>> When I switch the IDE controller into legacy mode, all works correctly.

I have reproduced the same issue on the another node:

ASUSTeK P5GD1-VM,
Intel 915G chipset,
ICH6 IDE controller,
IDE dvdrom: SONY DVD-ROM DDU1615 (hda),
sata disk: WDC WD1600JS-00M

when I switch IDE controller to the native mode, I see "Disabling IRQ" message,
then kernel generates an oops in create_empty_buffers(), like I've reported earlier.

Could somebody please help me to troubleshoot this issue? I've seen this issue
on the customer nodes and would like to know how I can work-around this issue
without any changes inside motherboard BIOS.

thank you,
Vasily Averin

Subject: Re: [Q] ide cdrom in native mode leads to irq storm?
Posted by Alan Cox on Fri, 27 Oct 2006 14:19:39 GMT
View Forum Message <> Reply to Message

Ar Gwe, 2006-10-27 am 17:17 +0400, ysgrifennodd Vasily Averin:
> Could somebody please help me to troubleshoot this issue? I've seen this issue
> on the customer nodes and would like to know how I can work-around this issue
> without any changes inside motherboard BIOS.

If its an IRQ routing triggered problem you probably can't, at least not
the IDE error. The oops wants debugging further because it shouldn't
have oopsed on that error merely given up.

Subject: [Q] PCI Express and ide (native) leads to irq storm?
Posted by vaverin on Tue, 14 Nov 2006 09:08:45 GMT
View Forum Message <> Reply to Message

Alan Cox wrote:
> Ar Gwe, 2006-10-27 am 17:17 +0400, ysgrifennodd Vasily Averin:
>> Could somebody please help me to troubleshoot this issue? I've seen this issue
>> on the customer nodes and would like to know how I can work-around this issue
>> without any changes inside motherboard BIOS.
>
> If its an IRQ routing triggered problem you probably can't, at least not

> the IDE error. The oops wants debugging further because it shouldn't
> have oopsed on that error merely given up.

Alan,
I've reproduced this issue on linux 2.6.19-rc5 kernel.

As far as I see if IDE controller is switched into native mode it shares irq
together with one of PCI Express Ports. It seems for me the last device is
guilty in this issue, becuase of it shares IDE irq on all the checked nodes.
and I do not know the ways to change their irq number or disable this device at all.

I means the following devices:

on Intel 915G-based nodes
0000:00:1c.2 Class 0604: 8086:2664 (rev 03)
0000:00:1c.2 PCI bridge: Intel Corporation 82801FB/FBM/FR/FW/FRW (ICH6 Family)
PCI Express Port 3 (rev 03)

on Intel E7520 node:
00:04.0 0604: 8086:3597 (rev 0a)
00:05.0 0604: 8086:3598 (rev 0a)
00:04.0 PCI bridge: Intel Corporation E7525/E7520 PCI Express Port B (rev 0a)
00:05.0 PCI bridge: Intel Corporation E7520 PCI Express Port B1 (rev 0a)

I've checked Intel chipset spec updates but do not found any related issues.

Please see http://bugzilla.kernel.org/show_bug.cgi?id=7518 for details

thank you,
Vasily Averin

Subject: Re: [Q] PCI Express and ide (native) leads to irq storm?
Posted by Tejun Heo on Wed, 15 Nov 2006 10:46:52 GMT
View Forum Message <> Reply to Message

Vasily Averin wrote:
> Alan Cox wrote:
>> Ar Gwe, 2006-10-27 am 17:17 +0400, ysgrifennodd Vasily Averin:
>>> Could somebody please help me to troubleshoot this issue? I've seen this issue
>>> on the customer nodes and would like to know how I can work-around this issue
>>> without any changes inside motherboard BIOS.
>> If its an IRQ routing triggered problem you probably can't, at least not
>> the IDE error. The oops wants debugging further because it shouldn't
>> have oopsed on that error merely given up.
>
> Alan,
> I've reproduced this issue on linux 2.6.19-rc5 kernel.

>
> As far as I see if IDE controller is switched into native mode it shares irq
> together with one of PCI Express Ports. It seems for me the last device is
> guilty in this issue, becuase of it shares IDE irq on all the checked nodes.
> and I do not know the ways to change their irq number or disable this device at all.
>
> I means the following devices:
>
> on Intel 915G-based nodes
> 0000:00:1c.2 Class 0604: 8086:2664 (rev 03)
> 0000:00:1c.2 PCI bridge: Intel Corporation 82801FB/FBM/FR/FW/FRW (ICH6 Family)
> PCI Express Port 3 (rev 03)
>
> on Intel E7520 node:
> 00:04.0 0604: 8086:3597 (rev 0a)
> 00:05.0 0604: 8086:3598 (rev 0a)
> 00:04.0 PCI bridge: Intel Corporation E7525/E7520 PCI Express Port B (rev 0a)
> 00:05.0 PCI bridge: Intel Corporation E7520 PCI Express Port B1 (rev 0a)
>
> I've checked Intel chipset spec updates but do not found any related issues.
>
> Please see http://bugzilla.kernel.org/show_bug.cgi?id=7518 for details

Okay, I tracked this one down. It's pretty interesting.

In short, some piix controllers including ICH7, when put into enhanced mode (PCI native mode), uses BMDMA Interrupt bit as interrupt pending/clear bit for *all* commands. ie. Reading STATUS does NOT clear IRQ even for PIO commands. 1 should be written to BMDMA Interrupt bit to clear IRQ. That's what's causing IRQ storm. IDE driver does what it's supposed to do but IRQ is just stuck at low active.

Fortunately, libata is immune to the problem because it does ap->ops->irq_clear(ap) in ata_host_intr() regardless of command type in flight. So, not loading IDE piix and using libata to drive all piix ports solves the problem.

I guess this behavior is unique to some piixs in enhanced mode considering wide use of IDE driver. Fixing this in IDE driver is pain in the ass because IRQ handler is scattered all over the place. I'm thinking about adding big warning message saying "IRQ storm can occur and you better switch to libata if that happens". But if anyone else is up for the job of fixing IDE, please don't hesitate.

Thanks.

--
tejun

Subject: Re: [Q] PCI Express and ide (native) leads to irq storm?
Posted by Jeff Garzik on Wed, 15 Nov 2006 18:49:21 GMT
View Forum Message <> Reply to Message

Tejun Heo wrote:
> In short, some piix controllers including ICH7, when put into enhanced
> mode (PCI native mode), uses BMDMA Interrupt bit as interrupt
> pending/clear bit for *all* commands. ie. Reading STATUS does NOT clear

Yep. I thought I had mentioned this, ages ago.

> Fortunately, libata is immune to the problem because it does
> ap->ops->irq_clear(ap) in ata_host_intr() regardless of command type in
> flight. So, not loading IDE piix and using libata to drive all piix
> ports solves the problem.

Yep, that's intentional :)

Jeff

Subject: Re: [Q] PCI Express and ide (native) leads to irq storm?
Posted by vaverin on Thu, 16 Nov 2006 08:45:04 GMT
View Forum Message <> Reply to Message

Tejun Heo wrote:
> Vasily Averin wrote:
>> Alan Cox wrote:
>>> Ar Gwe, 2006-10-27 am 17:17 +0400, ysgrifennodd Vasily Averin:
>>>> Could somebody please help me to troubleshoot this issue? I've seen this issue
>>>> on the customer nodes and would like to know how I can work-around this issue
>>>> without any changes inside motherboard BIOS.
>>> If its an IRQ routing triggered problem you probably can't, at least not
>>> the IDE error. The oops wants debugging further because it shouldn't
>>> have oopsed on that error merely given up.
>> Alan,
>> I've reproduced this issue on linux 2.6.19-rc5 kernel.
>>
>> As far as I see if IDE controller is switched into native mode it shares irq
>> together with one of PCI Express Ports. It seems for me the last device is
>> guilty in this issue, becuase of it shares IDE irq on all the checked nodes.
>> and I do not know the ways to change their irq number or disable this device at all.
>>
>> I means the following devices:
>>
>> on Intel 915G-based nodes
>> 0000:00:1c.2 Class 0604: 8086:2664 (rev 03)

>> 0000:00:1c.2 PCI bridge: Intel Corporation 82801FB/FBM/FR/FW/FRW (ICH6 Family)
>> PCI Express Port 3 (rev 03)
>>
>> on Intel E7520 node:
>> 00:04.0 0604: 8086:3597 (rev 0a)
>> 00:05.0 0604: 8086:3598 (rev 0a)
>> 00:04.0 PCI bridge: Intel Corporation E7525/E7520 PCI Express Port B (rev 0a)
>> 00:05.0 PCI bridge: Intel Corporation E7520 PCI Express Port B1 (rev 0a)
>>
>> I've checked Intel chipset spec updates but do not found any related issues.
>>
>> Please see http://bugzilla.kernel.org/show_bug.cgi?id=7518 for details
>
> Okay, I tracked this one down. It's pretty interesting.
>
> In short, some piix controllers including ICH7, when put into enhanced
> mode (PCI native mode), uses BMDMA Interrupt bit as interrupt
> pending/clear bit for *all* commands. ie. Reading STATUS does NOT clear
> IRQ even for PIO commands. 1 should be written to BMDMA Interrupt bit
> to clear IRQ. That's what's causing IRQ storm. IDE driver does what
> it's supposed to do but IRQ is just stuck at low active.
>
> Fortunately, libata is immune to the problem because it does
> ap->ops->irq_clear(ap) in ata_host_intr() regardless of command type in
> flight. So, not loading IDE piix and using libata to drive all piix
> ports solves the problem.

I've disabled IDE support in the config and recompiled the kernel.
It seems you are right, problem go away, new kernel was booted without any
problems and works well.

> I guess this behavior is unique to some piixs in enhanced mode
> considering wide use of IDE driver. Fixing this in IDE driver is pain
> in the ass because IRQ handler is scattered all over the place. I'm
> thinking about adding big warning message saying "IRQ storm can occur
> and you better switch to libata if that happens". But if anyone else is
> up for the job of fixing IDE, please don't hesitate.

I'm very happy that we have found the cause of this issue, however it seems for
me you do not understand fully its severity for linux end-users.

At the present moment this issue is present in all vendor kernels, and they
cannot be installed on the huge number of end-user nodes. Moreover, end-user
nodes can have installed old Linux distribution where initscripts do not loads
all the detected modules at the boot-time. Linux may be installed and the
following situation is possible: kernel was booted and works well until some
user will going to access the CDROM.

>From end-users point of view this issue looks mystic and very dump: is the linux
stable? is it ready for desktop? $%^&#! It crashes when I accessing the CDROM! :(

As a linux support engeneer I've seen this issue several times on the user-nodes
and it was very hard to understand what's happened and how to prevent this issue
in the future. First question is resolved now but from support point of view it
is very important to find some workaround against this issue on existing
distributions. Right now I see only one way: if this issue is detected on the
user node, we can add something like "ide=disable" into kernel commandline.

Probably the better solution exists?

thank you,
Vasily Averin

Subject: Re: [Q] workaround for ide (native) leads to irq storm?
Posted by vaverin on Fri, 17 Nov 2006 12:54:43 GMT
View Forum Message <> Reply to Message

Vasily Averin wrote:
> Tejun Heo wrote:
>> Vasily Averin wrote:
>>> I've reproduced this issue on linux 2.6.19-rc5 kernel.
>>>
>>> Please see http://bugzilla.kernel.org/show_bug.cgi?id=7518 for details
>>
>> Fortunately, libata is immune to the problem because it does
>> ap->ops->irq_clear(ap) in ata_host_intr() regardless of command type in
>> flight. So, not loading IDE piix and using libata to drive all piix
>> ports solves the problem.
>
> I've disabled IDE support in the config and recompiled the kernel.
> It seems you are right, problem go away, new kernel was booted without any
> problems and works well.
>
> As a linux support engeneer I've seen this issue several times on the user-nodes
> and it was very hard to understand what's happened and how to prevent this issue
> in the future. First question is resolved now but from support point of view it
> is very important to find some workaround against this issue on existing
> distributions. Right now I see only one way: if this issue is detected on the
> user node, we can add something like "ide=disable" into kernel commandline.

I've tried to find the some work around for this issue. "hda=noprobe" helps, CD
is not detected on the node and all other devices on the node works well...

However if I have additional device who uses the same irq the issue returns
back. When I enable USB support on my testnode, one of USB controllers requests

the same IRQ line. And IRQ storm occurs again when I load uhci_hcd driver on the node. It is very strange for me: we do not have any IDE devices in this case.

I would note, that I've seen the same behaviour when I detach the CDROM manually.

I've updated the bug.

Thank you,
Vasily Averin
