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
