

Re: [discuss] Re: [Perfctr-devel] Re: Enabling RDPMC in user space by default

Source: <http://linux.derkeiler.com/Mailing-Lists/Kernel/2005-12/msg00046.html>

- *From:* Andi Kleen <ak@xxxxxxx>
 - *Date:* Wed, 30 Nov 2005 17:23:15 +0100
-

On Wed, Nov 30, 2005 at 08:01:59AM -0800, Stephane Eranian wrote:

> Andi,

>

> On Tue, Nov 29, 2005 at 11:51:55PM +0100, Andi Kleen wrote:

>> On Tue, Nov 29, 2005 at 02:19:15PM -0800, Stephane Eranian wrote:

>>>

>>> On AMD you only have 4 counters. That's not a lot for some measurements.

>>

>> Disabling the NMI watchdog for that is out of question. It's a important

>> debugging device and without it kernel bug reports are much worse.

>> It increased the quality of x86-64 bug reports over the years

>> considerably and I'm unwilling to give that up.

>>

>

> So if I understand correctly, the kernel would program a counter

It always did that on x86-64. On i386 it's an option, but off by default because it breaks on some broken BIOS.

[imho it should be actually made default there too with the bad systems just blacklisted.]

> to count elapsed cycles while executing a ring 0 and ring 3. The watchdog

> works by polling on the counter and after a certain delta is reached it

> triggers an NMI interrupt which, in turn, causes a kernel crash and the

> (bug) report. Is that the correct behavior?

The watchdog is driven by the performance counter (this means it has varying frequency, but that's not a big issue for the watchdog)

It underflows every second in the fastest case or very slowly (if the machine is idle). Every time it underflows it checks if the per CPU timer has been ticking, and if it hasn't for some time it triggers an oops.

In theory other sources could be used, but there aren't any generic ones. At some point the local APIC timer was used, but that was disabled because it ticked at HZ and caused too much overhead.

If the motherboard has a usable watchdog timer in the chipset i wouldn't be completely opposed to using that too, but the problem is that many chipsets don't have them or only broken and it's fairly important that this works reliably to get better bug reports.

>>> but it does only implement 47bits. At a high clock rate, this can wrap
>>> around fairly rapidly. It all depends on what is the intended usage model.
>>
>> TSC also doesn't count cycles in many circumstances (different frequency
>> depending on P states or not synchronized over CPUs, even running
>> at completely different frequencies etc.)
>>
>
> Well the big difference here is that once the counter reaches 2^{47} , it goes
> back to zero, i.e., it wraps around silently by default. If you are polling
> it, you will suddenly see a huge delta and think that a long period of time
> has elapsed when in fact it is just one cycle. The TSC may not count all cycles
> but the user sees the counter has continuously increasing.

The obvious solution would be to set an underflow interrupt at 2^{46} or so and then reset the counters. For that you would need to count down though.

>
> Also are you sure that the PERFCTR0/PERFSEL0 are not affected when going
> into lower power state? I know by experience that one IA-64, for instance,
> the counters are seriously affected.

They stop ticking in idle. Yes, that's ok if you just want to measure cycles because there are no cycles in idle.

It's not ok for timing (wall clock time) purposes, but it's also not intended for that. If you want time use gettimeofday

They will also clock slower if the CPU is in a P state (runs with lower frequency), but for measurements that's also wanted and expected I believe. e.g with RDTSC on Intel right now if you are in a lower P state you will get wrong results.

Basically it's a good cycle timer for instruction measurements and nothing more.

Not ticking in idle actually helps with that because it makes it totally clear to everybody that it's not a wall clock :)

> As Ray mentioned, it all depends on what the user/sysdamin is after.
> Some people maybe okay with disbaling NMI in favor of more counters.
> Obviously others people are not.

I cannot stop them from hacking the kernel, but I don't think

Re: [discuss] Re: [Perfctr-devel] Re: Enabling RDPMC in user space by default

I will make it easy for them to do this in a stock kernel
(or at least not until they provide an reliable alternative watchdog
time source)

>> This means there is one alternative – some of the newer chipsets
>> have external watchdogs that could be also used (using the ACPI WDOG
>> table). If someone writes a nice NMI driver for these then on system
>> with working WDOG it could replace the perfctr based timeout and free
>> the perfctr. That would need some code to allocate and deallocate
>> perfctrs though.
>>
> I discussed the idea of an abitration layer to access performance
> counters with David Gibson a coule of months ago. I do believe this is
> an important mechanism to have and I would like to restart the discussion
> on this topic. You are providing us another reason why this could
> be useful.

There have been many tries of that over the years. The attempt
from IBM from a few years back they did for dprobes didn't look that bad
actually. There were a few others.

At least for the NMI watchdog it is not 100% needed – the
other code can just be changed to leave perfctr 0 alone.

–Andi

–

To unsubscribe from this list: send the line "unsubscribe linux-kernel" in
the body of a message to majordomo@xxxxxxxxxxxxxxxxx
More majordomo info at <http://vger.kernel.org/majordomo-info.html>
Please read the FAQ at <http://www.tux.org/lkml/>

• ***Follow-Ups:***

- ◆ ***Re: [discuss] Re: [Perfctr-devel] Re: Enabling RDPMC in user space by default***
◇ From: Stephane Eranian

• ***References:***

- ◆ ***Re: [discuss] Re: [Perfctr-devel] Re: Enabling RDPMC in user space by default***
◇ From: Stephane Eranian

- Prev by Date: ***Re: [NET] Remove ARM dependency for dm9000 driver***
- Next by Date: ***Re: [PATCH 0/9] x86-64 put current in r10***
- Previous by thread: ***Re: [discuss] Re: [Perfctr-devel] Re: Enabling RDPMC in user space by default***
- Next by thread: ***Re: [discuss] Re: [Perfctr-devel] Re: Enabling RDPMC in user space by default***
- Index(es):
 - ◆ ***Date***
 - ◆ ***Thread***