

Re: [PATCH] fix cpufreq_stats attrs removal

Re: [PATCH] fix cpufreq_stats attrs removal

Source: <http://linux.derkeiler.com/Mailing-Lists/Kernel/2007-03/msg11390.html>

- *From:* Alexey Dobriyan <adobriyan@xxxxx>
 - *Date:* Fri, 30 Mar 2007 12:09:03 +0400
-

On Thu, Mar 22, 2007 at 06:02:01PM +0100, Mattia Dongili wrote:

On Wed, Mar 21, 2007 at 08:10:42PM -0800, Andrew Morton wrote:

I ain't picky, but as a short-term thing it'd be kinda nice if it didn't oops the kernel.

There are other symptoms to this same bug:

1. unload p4-clockmod: /sys/.../cpu0/cpufreq is removed all together
2. load p4-clockmod: /sys/.../cpu0/cpufreq appears but no 'stats' subdir (yes, cpufreq_stats is loaded)
3. rmmmod cpufreq_stats: Oops!

Call Trace:

```
[<c0183f5b>] remove_dir+0x33/0xc4
[<c0184fca>] remove_files+0x1a/0x28
[<c018503b>] sysfs_remove_group+0x63/0x71
[<f898c38d>] cpufreq_stat_cpu_callback+0x51/0x8a [cpufreq_stats]
[<f898c477>] cpufreq_stats_exit+0x47/0x4b [cpufreq_stats]
[<c012f145>] sys_delete_module+0x190/0x1b7
[<c0140073>] do_wp_page+0x231/0x3e7
[<c0102e17>] syscall_call+0x7/0xb
```

The problem is cpufreq_stats doesn't know when a cpufreq driver is removed and doesn't cleanup. I guess this affects any setup with cpufreq_stats.

The attached patch seems to solve both symptoms and yes... it's quite invasive as it introduce one more cpufreq policy notification (REMOVED).

BTW: the patch is against .21-rc4-mm1 but applies with some fuzz to 2.6.20 too

Also, it doesn't work.

/sys/*/cpufreq/stats dir stays if you cd into it _before_ rmmmod.

Re: [PATCH] fix cpufreq_stats attrs removal

```
# modprobe p4-clockmod
$ /sys/devices/system/cpu/cpu0/cpufreq/stats
# rmmod p4-clockmod
$ cat time_in_state
```

```
p4-clockmod: P4/Xeon(TM) CPU On-Demand Clock Modulation available
BUG: unable to handle kernel paging request at virtual address 6b6b6b6f
printing eip:
c01955e1
*pde = 00000000
Oops: 0000 [#1]
last sysfs file: devices/system/cpu/cpu0/cpufreq/stats/time_in_state
Modules linked in: speedstep_lib ohci_hcd af_packet e1000 ehci_hcd uhci_hcd usbcore
CPU: 0
EIP: 0060:[<c01955e1>] Not tainted VLI
EFLAGS: 00010202 (2.6.21-rc5-mm1 #2)
EIP is at sysfs_open_file+0xb6/0x26f
eax: 6b6b6b6b ebx: c03b3cb0 ecx: 00000000 edx: f732772c
esi: 00000000 edi: c0681400 ebp: f36fad10 esp: f3595ee0
ds: 007b es: 007b fs: 00d8 gs: 0033 ss: 0068
Process cat (pid: 7545, ti=f3594000 task=f3773510 task.ti=f3594000)
Stack: 00001000 f376f26c f732772c f376f26c f36fad10 f376f26c f3595f38 c019552b
c015ced8 c18d3190 f3746a18 f36fad10 00008000 f3595f38 fffff9c c015d052
f36fad10 00000000 00000000 c015d094 00000000 f3595f38 f3746a18 c18d3190
Call Trace:
[<c019552b>] sysfs_open_file+0x0/0x26f
[<c015ced8>] __dentry_open+0xa4/0x191
[<c015d052>] nameidata_to_filp+0x31/0x3a
[<c015d094>] do_filp_open+0x39/0x40
[<c015ce23>] get_unused_fd+0xa1/0xb2
[<c02db41f>] _spin_unlock+0x14/0x1c
[<c015ce23>] get_unused_fd+0xa1/0xb2
[<c015d0d5>] do_sys_open+0x3a/0x6d
[<c015d143>] sys_open+0x1c/0x20
[<c0103c96>] sysenter_past_esp+0x5f/0x99
=====
INFO: lockdep is turned off.
Code: 0f 85 24 01 00 00 8b 43 04 85 c0 74 0f 83 38 02 0f 84 c4 01 00 00 ff 80 80 01 00 00 8b 54 24 08 8b 42
28 85 c0 0f 84 62 01 00 00 <8b> 40 04 85 c0 0f 84 57 01 00 00 8b 40 04 89 44 24 0c 8b 74 24
EIP: [<c01955e1>] sysfs_open_file+0xb6/0x26f SS:ESP 0068:f3595ee0
Slab corruption: start=f73276e0, len=256
Redzone: 0x5a2cf071/0x5a2cf071.
Last user: [<c0185810>](load_elf_binary+0xa79/0x1a19)
060: 6b 6b 6b 6b 6c 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b
Prev obj: start=f73275d4, len=256
Redzone: 0x5a2cf071/0x5a2cf071.
Last user: [<c01864fb>](load_elf_binary+0x1764/0x1a19)
000: 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b
010: 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b
Next obj: start=f73277ec, len=256
Redzone: 0x5a2cf071/0x5a2cf071.
```

Re: [PATCH] fix cpufreq_stats attrs removal

Last user: [<c01864fb>](load_elf_binary+0x1764/0x1a19)

000: 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b

010: 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b 6b

```
--- linux-2.6.20/drivers/cpufreq/cpufreq.c 2007-03-22 17:00:38.000000000 +0100
+++ linux-2.6.20.dirty/drivers/cpufreq/cpufreq.c 2007-03-22 16:51:09.000000000 +0100
@@ -989,6 +989,10 @@ static int __cpufreq_remove_dev (struct
```

```
unlock_policy_rwlock_write(cpu);
```

```
+ /* notify of policy cancellation */
+ blocking_notifier_call_chain(&cpufreq_policy_notifier_list,
+ CPUFREQ_REMOVE, data);
+
kobject_unregister(&data->kobj);
```

```
kobject_put(&data->kobj);
```

```
diff -rup linux-2.6.20/drivers/cpufreq/cpufreq_stats.c
```

```
linux-2.6.20.dirty/drivers/cpufreq/cpufreq_stats.c
```

```
--- linux-2.6.20/drivers/cpufreq/cpufreq_stats.c 2007-03-22 17:00:38.000000000 +0100
```

```
+++ linux-2.6.20.dirty/drivers/cpufreq/cpufreq_stats.c 2007-03-22 17:06:24.000000000
+0100
```

```
@@ -257,18 +257,23 @@ static int
```

```
cpufreq_stat_notifier_policy (struct notifier_block *nb, unsigned long val,
void *data)
```

```
{
- int ret;
+ int ret = 0;
struct cpufreq_policy *policy = data;
struct cpufreq_frequency_table *table;
unsigned int cpu = policy->cpu;
- if (val != CPUFREQ_NOTIFY)
- return 0;
- table = cpufreq_frequency_get_table(cpu);
- if (!table)
- return 0;
- if ((ret = cpufreq_stats_create_table(policy, table)))
- return ret;
- return 0;
+ switch (val) {
+ case CPUFREQ_NOTIFY:
+ table = cpufreq_frequency_get_table(cpu);
+ if (!table)
+ break;
+ ret = cpufreq_stats_create_table(policy, table);
+ break;
+
+ case CPUFREQ_REMOVE:
+ cpufreq_stats_free_table(cpu);
+ break;
```

Re: [PATCH] fix cpufreq_stats attrs removal

```
+ }  
+ return ret;  
}
```

```
static int  
@@ -371,8 +376,7 @@ __exit cpufreq_stats_exit(void)  
CPUFREQ_TRANSITION_NOTIFIER);  
unregister_hotcpu_notifier(&cpufreq_stat_cpu_notifier);  
for_each_online_cpu(cpu) {  
- cpufreq_stat_cpu_callback(&cpufreq_stat_cpu_notifier,  
- CPU_DEAD, (void *) (long)cpu);  
+ cpufreq_stats_free_table(cpu);  
}  
}
```

```
--- linux-2.6.20/include/linux/cpufreq.h 2007-03-22 17:00:47.000000000 +0100  
+++ linux-2.6.20.dirty/include/linux/cpufreq.h 2007-03-22 16:10:37.000000000 +0100  
@@ -96,6 +96,7 @@ struct cpufreq_policy {  
#define CPUFREQ_ADJUST (0)  
#define CPUFREQ_INCOMPATIBLE (1)  
#define CPUFREQ_NOTIFY (2)  
+#define CPUFREQ_REMOVE (3)  
  
#define CPUFREQ_SHARED_TYPE_NONE (0) /* None */  
#define CPUFREQ_SHARED_TYPE_HW (1) /* HW does needed coordination */
```

—

To unsubscribe from this list: send the line "unsubscribe linux-kernel" in
the body of a message to majordomo@xxxxxxxxxxxxxxxxxxx

More majordomo info at <http://vger.kernel.org/majordomo-info.html>

Please read the FAQ at <http://www.tux.org/lkml/>