Subject: forcedeth hangs on loading
Posted by vaverin on Fri, 30 Nov 2007 07:52:29 GMT
View Forum Message <> Reply to Message

Hi Ayaz,

While testing new OpenVZ kernel based on RHEL5.1 kernel 2.6.18-53.el5 we have observed that node loads forcedeth driver too long:

Nov 28 14:04:19 ts49 kjournald starting. Commit interval 5 seconds
Nov 28 14:04:19 ts49 EXT3-fs: mounted filesystem with ordered data mode.
...
Nov 28 19:34:47 ts49 modprobe   D ffff8101ff772120   0 4101   4098
          (NOTLB)
Nov 28 19:34:47 ts49 ffff8101fe7e1c28 0000000000000086 0000000000000246
00000010000200d0
Nov 28 19:34:47 ts49 ffff8101ff772120 ffff8103ff93c930 000012229b7b3bd7
00002bcc55b0dda7
Nov 28 19:34:47 ts49 ffff8101ff772328 ffff8101ff93a000 ffffffff802de500
ffff8103ff93c930
Nov 28 19:34:47 ts49 Call Trace:
Nov 28 19:34:47 ts49 [<ffffffff80062cc2>] schedule_timeout+0x8a/0xad
Nov 28 19:34:47 ts49 [<ffffffff80091bf9>] process_timeout+0x0/0x5
Nov 28 19:34:47 ts49 [<ffffffff800925c0>] msleep+0x21/0x2c
Nov 28 19:34:47 ts49 [<ffffffff8819fc8f>] :forcedeth:nv_probe+0x873/0xd4d
Nov 28 19:34:47 ts49 [<ffffffff801427f8>] pci_device_probe+0x100/0x180
Nov 28 19:34:47 ts49 [<ffffffff8019ff2c>] driver_probe_device+0x52/0xaa
Nov 28 19:34:47 ts49 [<ffffffff801a005b>] __driver_attach+0x65/0xb6
Nov 28 19:34:47 ts49 [<ffffffff8019fff6>] __driver_attach+0x0/0xb6
Nov 28 19:34:47 ts49 [<ffffffff8019f96d>] bus_for_each_dev+0x43/0x6e
Nov 28 19:34:47 ts49 [<ffffffff8019f5b3>] bus_add_driver+0x7e/0x130
Nov 28 19:34:47 ts49 [<ffffffff801429dc>] __pci_register_driver+0x57/0x7e
Nov 28 19:34:47 ts49 [<ffffffff800aa815>] sys_init_module+0x16aa/0x185f
Nov 28 19:34:47 ts49 [<ffffffff8005e166>] system_call+0x7e/0x83
...
Nov 28 21:02:02 ts49 eth0: forcedeth.c: subsystem: 010f1:2912 bound to 0000:00:08.0
Nov 28 21:02:02 ts49 ACPI: PCI Interrupt Link [LMA2] enabled at IRQ 18
Nov 28 21:02:02 ts49 GSI 22 sharing vector 0x3A and IRQ 22
Nov 28 21:02:02 ts49 ACPI: PCI Interrupt 0000:00:09.0[A] -> Link [LMA2] -> GSI
18 (level, high) -> IRQ 58
Nov 28 21:02:02 ts49 PCI: Setting latency timer of device 0000:00:09.0 to 64
Nov 28 21:02:02 ts49
Nov 28 21:02:02 ts49 forcedeth: using HIGHDMA
Nov 28 21:02:02 ts49 eth1: forcedeth.c: subsystem: 010f1:2912 bound to 0000:00:09.0

nv_probe() was called under semaphore and blocks lots of other modprobe's from loading other drivers.

linux-2.6-net-forcedeth-update-to-driver-version-0-60.patch patch from RHEL5.1 added the following piece of code to nv_probe():

```
+                       for (i = 0; i < 5000; i++) {
+                               msleep(1);
+                               if (nv_mgmt_acquire_sema(dev)) {
<skip>
+                               }
+                       }
```

obviously, this loops 5000 times and calls nv_mgmt_acquire_sema() inside, which in the worst case does msleep(500) 10 times.
So this loop can last 5000*10*0.5sec = 25000sec = 6.94 hours
This is exactly what we face in the bug: boot hanged at 14:04:19, continued at 21:02:02, i.e. it took ~25063 seconds.

As a workaround we have changed number of loops from 5000 to 5.

https://bugzilla.redhat.com/show_bug.cgi?id=405521

```
[root@ts49 ~]# lspci -n
00:00.0 0500: 10de:0369 (rev a2)
00:01.0 0601: 10de:0364 (rev a3)
00:01.1 0c05: 10de:0368 (rev a3)
00:02.0 0c03: 10de:036c (rev a1)
00:02.1 0c03: 10de:036d (rev a2)
00:04.0 0101: 10de:036e (rev a1)
00:05.0 0101: 10de:037f (rev a3)
00:05.1 0101: 10de:037f (rev a3)
00:05.2 0101: 10de:037f (rev a3)
00:06.0 0604: 10de:0370 (rev a2)
00:08.0 0680: 10de:0373 (rev a3)
00:09.0 0680: 10de:0373 (rev a3)
00:0a.0 0604: 10de:0376 (rev a3)
00:0d.0 0604: 10de:0378 (rev a3)
00:0f.0 0604: 10de:0377 (rev a3)
00:18.0 0600: 1022:1100
00:18.1 0600: 1022:1101
00:18.2 0600: 1022:1102
00:18.3 0600: 1022:1103
00:19.0 0600: 1022:1100
00:19.1 0600: 1022:1101
00:19.2 0600: 1022:1102
00:19.3 0600: 1022:1103
01:04.0 0300: 1002:515e (rev 02)

00:08.0 0680: 10de:0373 (rev a3)
```

```
        Subsystem: 10f1:2912
        Control: I/O+ Mem+ BusMaster+ SpecCycle- MemWINV- VGASnoop- ParErr-
Stepping- SERR- FastB2B-
        Status: Cap+ 66MHz+ UDF- FastB2B+ ParErr- DEVSEL=fast >TAbort- <TAbort-
<MAbort- >SERR- <PERR-
        Latency: 0 (250ns min, 5000ns max)
        Interrupt: pin A routed to IRQ 66
        Region 0: Memory at d0009000 (32-bit, non-prefetchable) [size=4K]
        Region 1: I/O ports at 3808 [size=8]
        Region 2: Memory at d0005800 (32-bit, non-prefetchable) [size=256]
        Region 3: Memory at d0005400 (32-bit, non-prefetchable) [size=16]
        Capabilities: [44] Power Management version 2
                Flags: PMEClk- DSI- D1+ D2+ AuxCurrent=0mA
PME(D0+,D1+,D2+,D3hot+,D3cold+)
                Status: D0 PME-Enable+ DSel=0 DScale=0 PME-
        Capabilities: [70] MSI-X: Enable- Mask- TabSize=8
                Vector table: BAR=2 offset=00000000
                PBA: BAR=3 offset=00000000
        Capabilities: [50] Message Signalled Interrupts: 64bit+ Queue=0/3 Enable+
                Address: 00000000fee00000  Data: 4042
        Capabilities: [6c] HyperTransport: MSI Mapping
00: de 10 73 03 07 00 b0 00 a3 00 80 06 00 00 00 00
10: 00 90 00 d0 09 38 00 00 00 58 00 d0 00 54 00 d0
20: 00 00 00 00 00 00 00 00 00 00 00 00 f1 10 12 29
30: 00 00 00 00 44 00 00 00 00 00 00 00 0a 01 01 14
40: f1 10 12 29 01 70 02 fe 00 01 00 00 0c 00 00 00
50: 05 6c 87 01 00 00 e0 fe 00 00 00 00 42 40 00 00
60: 00 00 00 00 00 00 00 00 ff 00 00 00 08 00 03 a8
70: 11 50 07 00 02 00 00 00 03 00 00 00 00 00 00 00
80: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
90: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
a0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
b0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
c0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
d0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
e0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
f0: 00 00 00 00 11 00 00 00 42 01 00 00 00 00 00 00

00:09.0 0680: 10de:0373 (rev a3)
        Subsystem: 10f1:2912
        Control: I/O+ Mem+ BusMaster+ SpecCycle- MemWINV- VGASnoop- ParErr-
Stepping- SERR- FastB2B-
        Status: Cap+ 66MHz+ UDF- FastB2B+ ParErr- DEVSEL=fast >TAbort- <TAbort-
<MAbort- >SERR- <PERR-
        Latency: 0 (250ns min, 5000ns max)
        Interrupt: pin A routed to IRQ 58
        Region 0: Memory at d000b000 (32-bit, non-prefetchable) [size=4K]
        Region 1: I/O ports at 3810 [size=8]
```

```
        Region 2: Memory at d000a000 (32-bit, non-prefetchable) [size=256]
        Region 3: Memory at d0005c00 (32-bit, non-prefetchable) [size=16]
        Capabilities: [44] Power Management version 2
                Flags: PMEClk- DSI- D1+ D2+ AuxCurrent=0mA
PME(D0+,D1+,D2+,D3hot+,D3cold+)
                Status: D0 PME-Enable- DSel=0 DScale=0 PME-
        Capabilities: [70] MSI-X: Enable- Mask- TabSize=8
                Vector table: BAR=2 offset=00000000
                PBA: BAR=3 offset=00000000
        Capabilities: [50] Message Signalled Interrupts: 64bit+ Queue=0/3 Enable-
                Address: 0000000000000000  Data: 0000
        Capabilities: [6c] HyperTransport: MSI Mapping
00: de 10 73 03 07 00 b0 00 a3 00 80 06 00 00 00 00
10: 00 b0 00 d0 11 38 00 00 00 a0 00 d0 00 5c 00 d0
20: 00 00 00 00 00 00 00 00 00 00 00 00 f1 10 12 29
30: 00 00 00 00 44 00 00 00 00 00 00 00 0b 01 01 14
40: f1 10 12 29 01 70 02 fe 00 00 00 00 01 00 00 00
50: 05 6c 86 01 00 00 00 00 00 00 00 00 00 00 00 00
60: 00 00 00 00 00 00 00 00 ff 00 00 00 08 00 03 a8
70: 11 50 07 00 02 00 00 00 03 00 00 00 00 00 00 00
80: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
90: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
a0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
b0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
c0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
d0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
e0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
f0: 00 00 00 00 11 00 00 00 42 01 00 00 00 00 00 00
```

Thank you,
Vasily Averin