

[PATCH 0/3][BUGFIX] configs: Fix deadlock rmdir() vs rename()

Source: <http://linux.derkeiler.com/Mailing-Lists/Kernel/2008-06/msg05062.html>

- *From:* Louis Rilling <Louis.Rilling@xxxxxxxxxxx>
 - *Date:* Thu, 12 Jun 2008 15:31:26 +0200
-

Hi,

This patchset fixes the deadlock described below. I did not exactly implement what I described earlier, since I found this more elegant solution meanwhile. Please see the third patch header for a detailed explanation of the fix.

The following procedure can trigger a deadlock in configs (see <http://www.ussg.iu.edu/hypermail/linux/kernel/0806.1/0380.html> for a patch that makes it easier to trigger):

```
# mkdir /config/cluster/foo
# cd /config/cluster/foo
# mv heartbeat/dead_threshold node/bar
```

and in another shell, right after having launched test_deadlock:

```
# rmdir /config/cluster/foo
```

First, lockdep warns as usual (see below), and after two minutes (standard task deadlock parameters), we get the dead lock alerts:

<log>

```
=====
[ INFO: possible recursive locking detected ]
2.6.26-rc5 #13
=====
```

```
rmdir/3997 is trying to acquire lock:
(&sb->s_type->i_mutex_key#11){---}, at: [<ffffffff802d2131>] configs_detach_prep+0x58/0xaa
```

```
but task is already holding lock:
(&sb->s_type->i_mutex_key#11){---}, at: [<ffffffff80296070>] vfs_rmdir+0x49/0xac
```

other info that might help us debug this:

2 locks held by rmdir/3997:

```
#0: (&sb->s_type->i_mutex_key#3/1){---}, at: [<ffffffff80297c77>] do_rmdir+0x82/0x108
#1: (&sb->s_type->i_mutex_key#11){---}, at: [<ffffffff80296070>] vfs_rmdir+0x49/0xac
```

[PATCH 0/3][BUGFIX] configs: Fix deadlock rmdir() vs rename()

stack backtrace:

Pid: 3997, comm: rmdir Not tainted 2.6.26-rc5 #13

Call Trace:

```
[<ffffffff8024aa65>] __lock_acquire+0x8d2/0xc78
[<ffffffff802495ec>] find_usage_backwards+0x9d/0xbe
[<ffffffff802d2131>] configs_detach_prep+0x58/0xaa
[<ffffffff8024b1de>] lock_acquire+0x51/0x6c
[<ffffffff802d2131>] configs_detach_prep+0x58/0xaa
[<ffffffff80247dad>] debug_mutex_lock_common+0x16/0x23
[<ffffffff805d63a4>] mutex_lock_nested+0xcd/0x23b
[<ffffffff802d2131>] configs_detach_prep+0x58/0xaa
[<ffffffff802d327b>] configs_rmdir+0xb8/0x1c3
[<ffffffff80296092>] vfs_rmdir+0x6b/0xac
[<ffffffff80297cac>] do_rmdir+0xb7/0x108
[<ffffffff80249d1e>] trace_hardirqs_on+0xef/0x113
[<ffffffff805d74c4>] trace_hardirqs_on_thunk+0x35/0x3a
[<ffffffff8020b0cb>] system_call_after_swaps+0x7b/0x80
```

INFO: task test_deadlock:3996 blocked for more than 120 seconds.

"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.

test_deadlock D 0000000000000001 0 3996 3980

ffff81007cc93d78 0000000000000046 ffff81007cc93d40 ffffffff808ed280

fffffff808ed280 ffff81007cc93d28 ffffffff808ed280 ffffffff808ed280

fffffff808ed280 ffffffff808ea120 ffffffff808ed280 ffff81007cdcaa10

Call Trace:

```
[<ffffffff802955e3>] lock_rename+0x11e/0x126
[<ffffffff805d641e>] mutex_lock_nested+0x147/0x23b
[<ffffffff802955e3>] lock_rename+0x11e/0x126
[<ffffffff80297838>] sys_renameat+0xd7/0x21c
[<ffffffff805d74c4>] trace_hardirqs_on_thunk+0x35/0x3a
[<ffffffff80249d1e>] trace_hardirqs_on+0xef/0x113
[<ffffffff805d74c4>] trace_hardirqs_on_thunk+0x35/0x3a
[<ffffffff8020b0cb>] system_call_after_swaps+0x7b/0x80
```

INFO: lockdep is turned off.

INFO: task rmdir:3997 blocked for more than 120 seconds.

"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.

rmdir D 0000000000000000 0 3997 3986

ffff81007cdb9dd8 0000000000000046 0000000000000000 ffffffff808ed280

fffffff808ed280 ffff81007cdb9d88 ffffffff808ed280 ffffffff808ed280

fffffff808ed280 ffffffff808ea120 ffffffff808ed280 ffff81007cde0a50

Call Trace:

```
[<ffffffff802d2131>] configs_detach_prep+0x58/0xaa
[<ffffffff805d641e>] mutex_lock_nested+0x147/0x23b
[<ffffffff802d2131>] configs_detach_prep+0x58/0xaa
[<ffffffff802d327b>] configs_rmdir+0xb8/0x1c3
[<ffffffff80296092>] vfs_rmdir+0x6b/0xac
[<ffffffff80297cac>] do_rmdir+0xb7/0x108
[<ffffffff80249d1e>] trace_hardirqs_on+0xef/0x113
[<ffffffff805d74c4>] trace_hardirqs_on_thunk+0x35/0x3a
```

[PATCH 0/3][BUGFIX] configs: Fix deadlock rmdir() vs rename()

[<ffffff8020b0cb>] system_call_after_swaps+0x7b/0x80

INFO: lockdep is turned off.

</log>

The issue here is that the VFS locks the `i_mutex` of the source and target directories of the rename in source -> target order (because none is ascendent of the other one), while `configs_detach_prep()` takes them in default group order (or reverse order, I'm not sure), following the order specified by the groups' creator.

Louis

--

Dr Louis Rilling Kerlabs

Skype: louis.rilling Batiment Germanium

Phone: (+33|0) 6 80 89 08 23 80 avenue des Buttes de Coesmes

<http://www.kerlabs.com/> 35700 Rennes

--

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

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

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