Subject: [PATCH 2/20] Move exit_task_namespaces()
Posted by Pavel Emelianov on Tue, 07 Aug 2007 09:29:38 GMT
View Forum Message <> Reply to Message

Makve task release its namespaces after it has reparented all his children
to child_reaper, but before it notifies its parent about its death.

The reason to release namespaces after reparenting is that when task exits
it may send a signal to its parent (SIGCHLD), but if the parent has already
exited its namespaces there will be no way to decide what pid to dever to
him - parent can be from different namespace.

The reason to release namespace before notifying the parent it that when
task sends a SIGCHLD to parent it can call wait() on this taks and release
it. But releasing the mnt namespace implies dropping of all the mounts in
the mnt namespace and NFS expects the task to have valid sighand pointer.

Thanks to Oleg for pointing out some races that can apear and helping with
patches and fixes.

Signed-off-by: Pavel Emelyanov <xemul@openvz.org>
Cc: Oleg Nesterov <oleg@tv-sign.ru>

---

 exit.c | 2 +-
 1 files changed, 1 insertion(+), 1 deletion(-)

```
--- ./kernel/exit.c.ve2 2007-07-27 12:45:46.000000000 +0400
+++ ./kernel/exit.c 2007-07-27 12:46:08.000000000 +0400
@@ -796,6 +796,7 @@ static void exit_notify(struct task_stru
   * jobs, send them a SIGHUP and then a SIGCONT.  (POSIX 3.2.2.2)
   */
  forget_original_parent(tsk);
+ exit_task_namespaces(tsk);

  write_lock_irq(&tasklist_lock);
  /*
@@ -1003,7 +1004,6 @@ fastcall NORET_TYPE void do_exit(long co

  tsk->exit_code = code;
  proc_exit_connector(tsk);
- exit_task_namespaces(tsk);
  exit_notify(tsk);
 #ifdef CONFIG_NUMA
  mpol_free(tsk->mempolicy);
```