Subject: [patch -mm 4/4] add child reaper to pid_namespace
Posted by Cedric Le Goater on Fri, 08 Sep 2006 17:00:08 GMT
View Forum Message <> Reply to Message

Add per-pid-space child-reaper. This is needed so processes are reaped within
the same pid space and do not spill over to the parent pid space. Its also
needed so containers preserve existing semantic that pid == 1 would reap
orphaned children.

This is based on Eric Biederman's patch: http://lkml.org/lkml/2006/2/6/285

Signed-off-by: Sukadev Bhattiprolu <sukadev@us.ibm.com>
Signed-off-by: Cedric Le Goater <clg@fr.ibm.com>
Cc: Eric Biederman <ebiederm@xmission.com>
Cc: Dave Hansen <haveblue@us.ibm.com>
Cc: Serge Hallyn <serue@us.ibm.com>

```
 fs/exec.c                  |    5 +++--
 include/linux/pid.h        |    5 +++--
 include/linux/pid_namespace.h |    9 +++++++----
 include/linux/sched.h      |    1 -
 init/main.c                |    5 ++---
 kernel/exit.c              |   26 ++++++++++++++++++++++++++--------
 kernel/pid.c               |    3 ++-
 kernel/signal.c            |   10 ++++++++--
 8 files changed, 41 insertions(+), 23 deletions(-)

Index: 2.6.18-rc6-mm1/init/main.c
===================================================================
--- 2.6.18-rc6-mm1.orig/init/main.c
+++ 2.6.18-rc6-mm1/init/main.c
@@ -49,6 +49,7 @@
 #include <linux/buffer_head.h>
 #include <linux/debug_locks.h>
 #include <linux/lockdep.h>
+#include <linux/pid_namespace.h>

 #include <asm/io.h>
 #include <asm/bugs.h>
@@ -623,8 +624,6 @@ static int __init initcall_debug_setup(c
 }
 __setup("initcall_debug", initcall_debug_setup);

-struct task_struct *child_reaper = &init_task;
-
 extern initcall_t __initcall_start[], __initcall_end[];

 static void __init do_initcalls(void)
```

```
@@ -724,7 +723,7 @@ static int init(void * unused)
   * assumptions about where in the task array this
   * can be found.
   */
- child_reaper = current;
+ init_pid_ns.child_reaper = current;

  smp_prepare_cpus(max_cpus);

Index: 2.6.18-rc6-mm1/fs/exec.c
===================================================================
--- 2.6.18-rc6-mm1.orig/fs/exec.c
+++ 2.6.18-rc6-mm1/fs/exec.c
@@ -38,6 +38,7 @@
 #include <linux/binfmts.h>
 #include <linux/swap.h>
 #include <linux/utsname.h>
+#include <linux/pid_namespace.h>
 #include <linux/module.h>
 #include <linux/namei.h>
 #include <linux/proc_fs.h>
@@ -620,8 +621,8 @@ static int de_thread(struct task_struct
   * Reparenting needs write_lock on tasklist_lock,
   * so it is safe to do it under read_lock.
   */
- if (unlikely(tsk->group_leader == child_reaper))
-  child_reaper = tsk;
+ if (unlikely(tsk->group_leader == tsk->nsproxy->pid_ns->child_reaper))
+  tsk->nsproxy->pid_ns->child_reaper = tsk;

  zap_other_threads(tsk);
  read_unlock(&tasklist_lock);
Index: 2.6.18-rc6-mm1/include/linux/pid.h
===================================================================
--- 2.6.18-rc6-mm1.orig/include/linux/pid.h
+++ 2.6.18-rc6-mm1/include/linux/pid.h
@@ -35,8 +35,9 @@ enum pid_type
  *
  * Holding a reference to struct pid solves both of these problems.
  * It is small so holding a reference does not consume a lot of
- * resources, and since a new struct pid is allocated when the numeric
- * pid value is reused we don't mistakenly refer to new processes.
+ * resources, and since a new struct pid is allocated when the numeric pid
+ * value is reused (when pids wrap around) we don't mistakenly refer to new
+ * processes.
  */

 struct pid
```

```
Index: 2.6.18-rc6-mm1/include/linux/sched.h
===================================================================
--- 2.6.18-rc6-mm1.orig/include/linux/sched.h
+++ 2.6.18-rc6-mm1/include/linux/sched.h
@@ -1380,7 +1380,6 @@ extern NORET_TYPE void do_group_exit(int
 extern void daemonize(const char *, ...);
 extern int allow_signal(int);
 extern int disallow_signal(int);
-extern struct task_struct *child_reaper;

 extern int do_execve(char *, char __user * __user *, char __user * __user *, struct pt_regs *);
 extern long do_fork(unsigned long, unsigned long, struct pt_regs *, unsigned long, int __user *,
int __user *);
Index: 2.6.18-rc6-mm1/kernel/exit.c
===================================================================
--- 2.6.18-rc6-mm1.orig/kernel/exit.c
+++ 2.6.18-rc6-mm1/kernel/exit.c
@@ -22,6 +22,7 @@
 #include <linux/file.h>
 #include <linux/binfmts.h>
 #include <linux/nsproxy.h>
+#include <linux/pid_namespace.h>
 #include <linux/ptrace.h>
 #include <linux/profile.h>
 #include <linux/mount.h>
@@ -48,7 +49,6 @@
 #include <asm/mmu_context.h>

 extern void sem_exit (void);
-extern struct task_struct *child_reaper;

 static void exit_mm(struct task_struct * tsk);

@@ -259,7 +259,8 @@ static int has_stopped_jobs(int pgrp)
 }

 /**
- * reparent_to_init - Reparent the calling kernel thread to the init task.
+ * reparent_to_init - Reparent the calling kernel thread to the init task
+ * of the pid space that the thread belongs to.
  *
  * If a kernel thread is launched as a result of a system call, or if
  * it ever exits, it should generally reparent itself to init so that
@@ -277,8 +278,8 @@ static void reparent_to_init(void)
  ptrace_unlink(current);
  /* Reparent to init */
  remove_parent(current);
- current->parent = child_reaper;
```

```
- current->real_parent = child_reaper;
+ current->parent = current->nsproxy->pid_ns->child_reaper;
+ current->real_parent = current->nsproxy->pid_ns->child_reaper;
  add_parent(current);

  /* Set the exit signal to SIGCHLD so we signal init on exit */
@@ -662,7 +663,8 @@ reparent_thread(struct task_struct *p, s
  * When we die, we re-parent all our children.
  * Try to give them to another thread in our thread
  * group, and if no such member exists, give it to
- * the global child reaper process (ie "init")
+ * the child reaper process (ie "init") in our pid
+ * space.
  */
 static void
 forget_original_parent(struct task_struct *father, struct list_head *to_release)
@@ -673,7 +675,10 @@ forget_original_parent(struct task_struc
  do {
   reaper = next_thread(reaper);
   if (reaper == father) {
-   reaper = child_reaper;
+   /*
+    * FIXME: which reaper to use ?
+    */
+   reaper = init_pid_ns.child_reaper;
    break;
   }
  } while (reaper->exit_state);
@@ -861,8 +866,13 @@ fastcall NORET_TYPE void do_exit(long co
   panic("Aiee, killing interrupt handler!");
  if (unlikely(!tsk->pid))
   panic("Attempted to kill the idle task!");
- if (unlikely(tsk == child_reaper))
-  panic("Attempted to kill init!");
+ if (unlikely(tsk == tsk->nsproxy->pid_ns->child_reaper)) {
+  if (tsk->nsproxy->pid_ns != &init_pid_ns)
+   tsk->nsproxy->pid_ns->child_reaper = init_pid_ns.child_reaper;
+  else
+   panic("Attempted to kill init!");
+ }
+

  if (unlikely(current->ptrace & PT_TRACE_EXIT)) {
   current->ptrace_message = code;
Index: 2.6.18-rc6-mm1/kernel/signal.c
===================================================================
--- 2.6.18-rc6-mm1.orig/kernel/signal.c
+++ 2.6.18-rc6-mm1/kernel/signal.c
```

```
@@ -24,6 +24,8 @@
 #include <linux/ptrace.h>
 #include <linux/signal.h>
 #include <linux/capability.h>
+#include <linux/pid_namespace.h>
+#include <linux/nsproxy.h>
 #include <asm/param.h>
 #include <asm/uaccess.h>
 #include <asm/unistd.h>
@@ -1994,8 +1996,12 @@ relock:
   if (sig_kernel_ignore(signr)) /* Default is nothing. */
    continue;

-  /* Init gets no signals it doesn't want.  */
-  if (current == child_reaper)
+  /*
+   * Init of a pid space gets no signals it doesn't want from
+   * within that pid space. It can of course get signals from
+   * its parent pid space.
+   */
+  if (current == current->nsproxy->pid_ns->child_reaper)
    continue;

   if (sig_kernel_stop(signr)) {
Index: 2.6.18-rc6-mm1/include/linux/pid_namespace.h
===================================================================
--- 2.6.18-rc6-mm1.orig/include/linux/pid_namespace.h
+++ 2.6.18-rc6-mm1/include/linux/pid_namespace.h
@@ -8,15 +8,16 @@
 #include <linux/nsproxy.h>

 struct pidmap {
-       atomic_t nr_free;
-       void *page;
+ atomic_t nr_free;
+ void *page;
 };

 #define PIDMAP_ENTRIES         ((PID_MAX_LIMIT + 8*PAGE_SIZE - 1)/PAGE_SIZE/8)

 struct pid_namespace {
-       struct pidmap pidmap[PIDMAP_ENTRIES];
-       int last_pid;
+ struct pidmap pidmap[PIDMAP_ENTRIES];
+ int last_pid;
+ struct task_struct * child_reaper;
 };
```

```
 extern struct pid_namespace init_pid_ns;
Index: 2.6.18-rc6-mm1/kernel/pid.c
===================================================================
--- 2.6.18-rc6-mm1.orig/kernel/pid.c
+++ 2.6.18-rc6-mm1/kernel/pid.c
@@ -62,7 +62,8 @@ struct pid_namespace init_pid_ns = {
  .pidmap = {
   [ 0 ... PIDMAP_ENTRIES-1] = { ATOMIC_INIT(BITS_PER_PAGE), NULL }
  },
- .last_pid = 0
+ .last_pid = 0,
+ .child_reaper = &init_task
 };

 /*

--
```

_____

Containers mailing list
Containers@lists.osdl.org
https://lists.osdl.org/mailman/listinfo/containers