Subject: [RFC][PATCH 2/3] Signal semantics for /sbin/init
Posted by Sukadev Bhattiprolu on Thu, 30 Aug 2007 06:20:29 GMT
View Forum Message <> Reply to Message

(This is Oleg's patch with my pid ns additions. Compiled and unit tested
on 2.6.23-rc3-mm1 with other patches in this set)

Currently, /sbin/init is protected from unhandled signals by the
"current == child_reaper(current)" check in get_signal_to_deliver().
This is not enough, we have multiple problems:

- this doesn't work for multi-threaded inits, and we can't
  fix this by simply making this check group-wide.

- /sbin/init and kernel threads are not protected from
  handle_stop_signal(). Minor problem, but not good and
  allows to "steal" SIGCONT or change ->signal->flags.

- /sbin/init is not protected from __group_complete_signal(),
  sig_fatal() can set SIGNAL_GROUP_EXIT and block exec(), kill
  sub-threads, set ->group_stop_count, etc.

Also, with support for multiple pid namespaces, we need an ability to
actually kill the sub-namespace's init from the parent namespace. In
this case it is not possible (without painful and intrusive changes)
to make the "should we honor this signal" decision on the receiver's
side.

Hopefully this patch (adds 43 bytes to kernel/signal.o) can solve
these problems.

Notes:

- Blocked signals are never ignored, so init still can receive
  a pending blocked signal after sigprocmask(SIG_UNBLOCK).
  Easy to fix, but probably we can ignore this issue.

- this patch allows us to simplify de_thread() playing games
  with pid_ns->child_reaper.

(Side note: the current behaviour of things like force_sig_info_fault()
 is not very good, init should not ignore these signals and go to the
 endless loop. Exit + panic is imho better, easy to chamge)

Oleg.

---
 kernel/signal.c | 60 ++--------------

```
 1 file changed, 46 insertions(+), 14 deletions(-)

Index: 2.6.23-rc3-mm1/kernel/signal.c
===================================================================
--- 2.6.23-rc3-mm1.orig/kernel/signal.c 2007-08-29 22:53:20.000000000 -0700
+++ 2.6.23-rc3-mm1/kernel/signal.c 2007-08-29 23:10:16.000000000 -0700
@@ -26,6 +26,7 @@
 #include <linux/freezer.h>
 #include <linux/pid_namespace.h>
 #include <linux/nsproxy.h>
+#include <linux/hardirq.h>

 #include <asm/param.h>
 #include <asm/uaccess.h>
@@ -39,11 +40,42 @@

 static struct kmem_cache *sigqueue_cachep;

+static int sig_init_ignore(struct task_struct *tsk)
+{

-static int sig_ignored(struct task_struct *t, int sig)
+ // Currently this check is a bit racy with exec(),
+ // we can _simplify_ de_thread and close the race.
+ if (likely(!is_container_init(tsk->group_leader)))
+  return 0;
+
+ /*
+  * If signal is from an ancestor pid namespace, do not
+  * ignore the signal.
+  */
+ if (task_ancestor_pid_ns(current, tsk))
+  return 0;
+
+ if (in_interrupt())
+  return 0;
+
+ return 1;
+}
+
+static int sig_task_ignore(struct task_struct *tsk, int sig)
 {
- void __user * handler;
+ void __user * handler = tsk->sighand->action[sig-1].sa.sa_handler;
+
+ if (handler == SIG_IGN)
+  return 1;
```

```
+ if (handler != SIG_DFL)
+  return 0;
+
+ return sig_kernel_ignore(sig) || sig_init_ignore(tsk);
+}
+
+static int sig_ignored(struct task_struct *t, int sig)
+{
  /*
   * Tracers always want to know about signals..
   */
@@ -58,10 +90,7 @@ static int sig_ignored(struct task_struc
  if (sigismember(&t->blocked, sig))
   return 0;

- /* Is it explicitly or implicitly ignored? */
- handler = t->sighand->action[sig-1].sa.sa_handler;
- return   handler == SIG_IGN ||
-  (handler == SIG_DFL && sig_kernel_ignore(sig));
+ return sig_task_ignore(t, sig);
 }

 /*
@@ -568,6 +597,9 @@ static void handle_stop_signal(int sig,
    */
   return;

+ if (sig_init_ignore(p))
+  return;
+
  if (sig_kernel_stop(sig)) {
   /*
    * This is a stop signal.  Remove SIGCONT from all queues.
@@ -1863,12 +1895,6 @@ relock:
   if (sig_kernel_ignore(signr)) /* Default is nothing. */
    continue;

-  /*
-   * Global init gets no signals it doesn't want.
-   */
-  if (is_global_init(current))
-   continue;
-
   if (sig_kernel_stop(signr)) {
    /*
     * The default action is to stop all threads in
@@ -2320,6 +2346,13 @@ int do_sigaction(int sig, struct k_sigac
  k = &current->sighand->action[sig-1];
```

```
  spin_lock_irq(&current->sighand->siglock);
+
+ if (current->signal->flags & SIGNAL_GROUP_EXIT) {
+  spin_unlock_irq(&current->sighand->siglock);
+  /* The return value doesn't matter, SIGKILL is pending */
+  return -EINTR;
+ }
+
  if (oact)
   *oact = *k;

@@ -2338,8 +2371,7 @@ int do_sigaction(int sig, struct k_sigac
    *   (for example, SIGCHLD), shall cause the pending signal to
    *   be discarded, whether or not it is blocked"
    */
-  if (act->sa.sa_handler == SIG_IGN ||
-      (act->sa.sa_handler == SIG_DFL && sig_kernel_ignore(sig))) {
+  if (sig_task_ignore(current, sig)) {
    struct task_struct *t = current;
    sigemptyset(&mask);
    sigaddset(&mask, sig);
```

Subject: Re: [RFC][PATCH 2/3] Signal semantics for /sbin/init
Posted by Oleg Nesterov on Thu, 30 Aug 2007 07:00:32 GMT
View Forum Message <> Reply to Message

On 08/29, sukadev@us.ibm.com wrote:
>
> --- 2.6.23-rc3-mm1.orig/kernel/signal.c 2007-08-29 22:53:20.000000000 -0700
> +++ 2.6.23-rc3-mm1/kernel/signal.c 2007-08-29 23:10:16.000000000 -0700
> @@ -26,6 +26,7 @@
> #include <linux/freezer.h>
> #include <linux/pid_namespace.h>
> #include <linux/nsproxy.h>
> +#include <linux/hardirq.h>
>
> #include <asm/param.h>
> #include <asm/uaccess.h>
> @@ -39,11 +40,42 @@
>
> static struct kmem_cache *sigqueue_cachep;
>

> +static int sig_init_ignore(struct task_struct *tsk)
> +{
>
> -static int sig_ignored(struct task_struct *t, int sig)
> + // Currently this check is a bit racy with exec(),
> + // we can _simplify_ de_thread and close the race.
> + if (likely(!is_container_init(tsk->group_leader)))
> + return 0;
> +
> + /*
> + * If signal is from an ancestor pid namespace, do not
> + * ignore the signal.
> + */
> + if (task_ancestor_pid_ns(current, tsk))
> + return 0;
> +

This patch was intended as a fix for the current behaviour, and a preparation
for the pid_ns requirements. If possible, it should go ahead of all other
pid_ns changes, imho.

In any case, we shouldn't mix all this in one patch.

If there are no objections to this approach, I'll re-send the patch on weekend.

> @@ -2320,6 +2346,13 @@ int do_sigaction(int sig, struct k_sigac
>   k = &current->sighand->action[sig-1];
>
>   spin_lock_irq(&current->sighand->siglock);
> +
> + if (current->signal->flags & SIGNAL_GROUP_EXIT) {
> + spin_unlock_irq(&current->sighand->siglock);
> + /* The return value doesn't matter, SIGKILL is pending */
> + return -EINTR;
> + }

This is not needed. Please look at

 http://marc.info/?l=linux-kernel&m=118763989718143

I specially removed this chunk in a separate patch, no need to re-introduce.

Oleg.

Subject: Re: [RFC][PATCH 2/3] Signal semantics for /sbin/init
Posted by Sukadev Bhattiprolu on Thu, 30 Aug 2007 07:11:44 GMT
View Forum Message <> Reply to Message

Oleg Nesterov [oleg@tv-sign.ru] wrote:
| On 08/29, sukadev@us.ibm.com wrote:
| >
| > --- 2.6.23-rc3-mm1.orig/kernel/signal.c 2007-08-29 22:53:20.000000000 -0700
| > +++ 2.6.23-rc3-mm1/kernel/signal.c 2007-08-29 23:10:16.000000000 -0700
| > @@ -26,6 +26,7 @@
| > #include <linux/freezer.h>
| > #include <linux/pid_namespace.h>
| > #include <linux/nsproxy.h>
| > +#include <linux/hardirq.h>
| >
| > #include <asm/param.h>
| > #include <asm/uaccess.h>
| > @@ -39,11 +40,42 @@
| >
| > static struct kmem_cache *sigqueue_cachep;
| >
| > +static int sig_init_ignore(struct task_struct *tsk)
| > +{
| >
| > -static int sig_ignored(struct task_struct *t, int sig)
| > + // Currently this check is a bit racy with exec(),
| > + // we can _simplify_ de_thread and close the race.
| > + if (likely(!is_container_init(tsk->group_leader)))
| > + return 0;
| > +
| > + /*
| > + * If signal is from an ancestor pid namespace, do not
| > + * ignore the signal.
| > + */
| > + if (task_ancestor_pid_ns(current, tsk))
| > + return 0;
| > +
|
| This patch was intended as a fix for the current behaviour, and a preparation
| for the pid_ns requirements. If possible, it should go ahead of all other
| pid_ns changes, imho.
|
| In any case, we shouldn't mix all this in one patch.
|
| If there are no objections to this approach, I'll re-send the patch on weekend.

Yes I think your patch looks good. I will apply the other two patches on top
of this.

|
| > @@ -2320,6 +2346,13 @@ int do_sigaction(int sig, struct k_sigac
| >   k = &current->sighand->action[sig-1];
| >
| >   spin_lock_irq(&current->sighand->siglock);
| > +
| > + if (current->signal->flags & SIGNAL_GROUP_EXIT) {
| > + spin_unlock_irq(&current->sighand->siglock);
| > + /* The return value doesn't matter, SIGKILL is pending */
| > + return -EINTR;
| > + }
|
| This is not needed. Please look at
|
| http://marc.info/?l=linux-kernel&m=118763989718143
|
| I specially removed this chunk in a separate patch, no need to re-introduce.

Sorry I missed that. I will remove it.

Thanks,

Suka

_____
Containers mailing list
Containers@lists.linux-foundation.org
https://lists.linux-foundation.org/mailman/listinfo/containers