Subject: Re: [PATCH] allow a task to join a pid namespace
Posted by ebiederm on Tue, 05 Jun 2012 17:18:52 GMT
View Forum Message <> Reply to Message

Oleg Nesterov <oleg@redhat.com> writes:

> On 06/04, Glauber Costa wrote:
>>
>> Currently, it is possible for a process to join existing
>> net, uts and ipc namespaces. This patch allows a process to join an
>> existing pid namespace as well.
>
> I can't understand this patch... but probably I missed something,
> I never really understood setns.

The idea with setns is akin to callusermodehelper in the kernel.

>From outside a container we want to allow an appropriately privileged
user to create a process inside the container.

We run into all kinds of interesting gotchas with entering the
pid namespace:
 - Disjoint process trees.
 - Ensuring all processes are gone when we exit a pid namespace.
 - Not letting an empty pid namespace accept more processes.

We really only have two possbilities here.
- Allocate a new struct pid that is a superset of our current struct
  pid but having additional processes ids inside a new pid namespace.

  Along with all of the appropriate sanity checks to make that safe.

- Just modify the pid namespace the child processes of setns will use.

I lean towards the second option as that seems to have the best semantic
match to practical applications, and fewer kernel races to contend with,
but I might be persuadable.

However we do this we need to fix the bugs in pid namespace cleanup,
and deal with the issues that disjoint process trees bring to waiting
for all processes in a pid namespace to exit.

Ugh. Getting the waking up of zap_pid_ns_processes right and handling
the reaping of zombines in the cases of disjoint process trees is going
to be interesting.

Eric