Subject: Re: [PATCH] allow a task to join a pid namespace
Posted by Daniel Lezcano on Tue, 05 Jun 2012 09:30:08 GMT
View Forum Message <> Reply to Message

On 06/04/2012 06:51 PM, Oleg Nesterov wrote:
> 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.

Hi Oleg,

let me clarify why is needed setns. In the world of container, setns allows to administrate the container from outside. One good example is to shutdown the container. The users setup their hosts with the init's services to startup the containers when the system starts, but they have no way to invoke 'shutdown' from inside the container when the system goes down except doing some trick with the signals. The setns syscall with the pid namespace support will allow to do that.

Also a complete setns support will allow to write some administrative tools to have a global view of the different separated resources running in several containers.

For example, if you are the administrator of the host and you have hundred of containers running on it, you can use setns to run netstat within each container and build a view of the different network stack. The same applies for 'ps' or 'top'.

Without setns, things are much more complicated and in some cases, impossible. For instance, you can run a daemon inside the container, send command to it and redirect its output to the fifo  but that increase the number of processes and has some limitations. Also that means the command you want to run is present in the container's FS.

The setns syscall is highly needed for the VRF, where a single process can handle thousand of network namespaces and switch from a network namespace to another network namespace with one syscall. The usage of the file descriptors pins the namespace and prevent it from being destroyed when switching from one namespace to another.

In other words, +1 for pid ns support with setns :)

-- Daniel