Subject: Re: [PATCH 1/2] signals: kill(-1) should only signal processes in the same namespace
Posted by Daniel Hokka Zakrisso on Thu, 17 Jul 2008 18:39:14 GMT
View Forum Message <> Reply to Message

Eric W. Biederman wrote:
> "Daniel Hokka Zakrisson" <daniel@hozac.com> writes:
>
>> Pavel Emelyanov wrote:
>>> Daniel Hokka Zakrisson wrote:
>>>> While moving Linux-VServer to using pid namespaces, I noticed that
>>>> kill(-1) from inside a pid namespace is currently signalling every
>>>> process in the entire system, including processes that are otherwise
>>>> unreachable from the current process.
>>>
>>> This is not a "news" actually, buy anyway - thanks :)
>>
>> And yet nobody's fixed it... Kind of a critical thing, if you actually
>> want to use them, since most distribution's rc-scripts do a kill(-1,
>> SIGTERM), followed by kill(-1, SIGKILL) when halting (which, needless to
>> say, would be very bad).
>>
>>>> This patch fixes it by making sure that only processes which are in
>>>> the same pid namespace as current get signalled.
>>>
>>> This is to be done, indeed, but I do not like the proposed
>>> implementation,
>>> since you have to walk all the tasks in the system (under
>>> tasklist_lock,
>>> by the way) to search for a couple of interesting ones. Better look at
>>> how
>>> zap_pid_ns_processes works (by the way - I saw some patch doing so some
>>> time ago).
>>
>> The way zap_pid_ns_processes does it is worse, since it signals every
>> thread in the namespace rather than every thread group. So either we
>> walk
>> the global tasklist, or we create a per-namespace one. Is that what we
>> want?
>
> Can you please introduce kill_pidns_info and have both
> kill_something_info and zap_pid_ns_processes call this common
> function?

Looks like you've already done that. :-) (Referring to Sukadev's email.)
Is there any reason we don't just merge that patch?

> We want to walk the set of all pids in a pid namespace. /proc does

> this and it is the recommended idiom. If walking all of the pids in a
> pid namespace is not fast enough we can accelerate that.
>
> You are correct signalling every thread in a namespace is worse, in
> fact it is semantically incorrect. zap_pid_ns_processes gets away
> with it because it is sending SIGKILL.   Therefore kill_pidns_info
> should skip sending a signal to every task that is not the
> thread_group_leader.
>
> We need to hold the tasklist_lock to prevent new processes from
> joining the list of all processes. Otherwise we could run the code
> under the rcu_read_lock.
>
> Eric

--
Daniel Hokka Zakrisson

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