Subject: Re: The issues for agreeing on a virtualization/namespaces implementation.
Posted by ebiederm on Wed, 08 Feb 2006 02:49:08 GMT

View Forum Message <> Reply to Message

Alexey Kuznetsov <kuznet@ms2.inr.ac.ru> writes:

> Hello!
>
>> >2) What is the syscall interface to create these namespaces?
>> >   - Do we add clone flags?
>> >   (Plan 9 style)
>>
>> Like that approach .. flexible .. particular when one has well specified
>> namespaces.
>>
>> >   - Do we add a syscall (similar to setsid) per namespace?
>> >   (Traditional unix style)?
>>
>> Where does that approach end .. what's wrong with doing it at clone() time ?
>
> That most of those namespaces need a special setup rather than a plain copy?
>
> F.e. what are you going to do with NETWORK namespace? The only valid thing
> to do is to prepare a new context and to configure its content (addresses,
> routing tables, iptables...) later. So that, in this case it is natural
> to inherit the context through clone() and to create new context
> with a separate syscall.

With a NETWORK namespace what I implemented was that you get a empty
namespace with a loopback interface.

But setting up the namespace from the inside is clearly the sane thing
todo.

> Seems, only PID space needs to be setup at clone time. All the rest of
> suggested namespaces are more convenient to change with separate syscalls.

Actually I think I can setup a PID space in a syscall as well.
It is a little odd that your session, and process group change but since
I was keeping 2 pids on the PID space leader I could easily do it.
The fact that getpid() would start returning 1 might be confusing to a some
processes though so clone seems to be the natural time to do it.

> I would suggest to combine both approaches. Those namespaces, which can be
> naturally copied while clone() (f.e. the best example is already existing
> CLONE_NEWNS) deserve a clone() flag. The rest are preserved through clone()
> and forked and configured later.

Sounds reasonable. We make the decision on a case by case base whatever make sense for the patch and the namespace.

The only real advantage to clone is you can create a bunch of namespaces all in one shot. Of course that makes sys_clone a little slower.

Eric