Subject: Re: [PATCH 0/9] namespaces: Introduction
Posted by ebiederm on Mon, 22 May 2006 17:23:38 GMT
View Forum Message <> Reply to Message

Andrew Morton <akpm@osdl.org> writes:

> Andrey Savochkin <saw@sw.ru> wrote:
>>
>> I have a practical proposal.
>> We can start with presenting and merging the most interesting part, network
>> containers. We discuss details, possible approaches, and related subsystems,
>> until networking is finished to its utmost detail.
>> This will create an example of virtualization of a non-trivial subsystem,
>> and we will have to agree on basic principles of virtualization of related
>> subsystems like proc.
>>
>> Virtualization of networking presents a lot of challenges and decision-making
>> points with respect to user-visible interfaces: proc, sysctl, netlink events
>> (and netlink sockets themselves), and so on. This code will also become
>> immediately useful as an improvement over chroot.
>> I am sure that when we come to a mutually acceptable solution with respect to
>> networking, virtualization of all other subsystems can be implemented and
>> merged without many questions.
>>
>> What do people think about this plan?
>
> It sounds like that feature might be the
> most-likely-to-cause-maintainer-revolt one, in which case yes, it is
> absolutely definitely the one to start with.

It sounds like a case of: That first step is a doozy.
We should be able to resolve proc and sysctl issues with just
the uts namespace. So I don't think we necessarily have to take
everything at once.

Beyond that the real sticky issue and what leads to most of the peculiar
cases is the one thing not addressed by doing the network namespace.
How do we keep someone inside a namespace from accessing files in /proc
and other places that they should not be able to access.

It occurred to me that most of the permission checking issues trivially
go away if you make the permission checks test for equality of
the tuple (uid namespace, uid). At which point a lot of the reasons
people have previously put forth for completely reorganizing proc and
sysfs go away, because users not in their uid namespace will only be
able to access world readable and world writable files. Anything else
will simply be inaccessible.

So I think we need to have a serious look at the uid/gid namespace.

This is a bit of a pain because this brings us face to face with the uid mapping problem we have avoided for years on network filesystems, and makes it a problem on local filesystems as well.

Getting both the uid/gid namespace and the network namespace should get the bulk of the infrastructure problems solved.

I am even happy to do the network namespace first on the understanding that permission checking problems caused by different users with the same uid should be ignored until we have handled the uid/gid namespace.

Eric
