Subject: Re: [RFC] Virtualization steps
Posted by Herbert Poetzl on Fri, 24 Mar 2006 21:19:17 GMT
View Forum Message <> Reply to Message

On Fri, Mar 24, 2006 at 08:19:59PM +0300, Kirill Korotaev wrote:
> Eric, Herbert,
>
> I think it is quite clear, that without some agreement on all these
> virtualization issues, we won't be able to commit anything good to
> mainstream. My idea is to gather our efforts to get consensus on most
> clean parts of code first and commit them one by one.
>
> The proposal is quite simple. We have 4 parties in this conversation
> (maybe more?): IBM guys, OpenVZ, VServer and Eric Biederman. We
> discuss the areas which should be considered step by step. Send
> patches for each area, discuss, come to some agreement and all 4
> parties Sign-Off the patch. After that it goes to Andrew/Linus.
> Worth trying?

sounds good to me, as long as we do not consider
the patches 'final' atm .. because I think we should
try to test them with _all_ currently existing solutions
first ... we do not need to bother Andrew with stuff
which doesn't work for the existing and future 'users'.

so IMHO, we should make a kernel branch (Eric or Sam
are probably willing to maintain that), which we keep
in-sync with mainline (not necessarily git, but at
least snapshot wise), where we put all the patches
we agree on, and each party should then adjust the
existing solution to this kernel, so we get some deep
testing in the process, and everybody can see if it
'works' for him or not ...

things where we agree that it 'just works' for everyone
can always be handed upstream, and would probably make
perfect patches for Andrew ...

> So far, (correct me if I'm wrong) we concluded that some people don't
> want containers as a whole, but want some subsystem namespaces. I
> suppose for people who care about containers only it doesn't matter, so
> we can proceed with namespaces, yeah?

yes, the emphasis here should be on lightweight and
modular, so that those folks interested in full featured
containers can just 'assemble' the pieces, while those
desiring service/space isolation pick their subsystems
one by one ...

> So the most easy namespaces to discuss I see:
> - utsname

yes, that's definitely one we can start with, as it seems that we already have _very_ similar implementations

> - sys IPC

this is something which is also related to limits and should get special attention with resource sharing, isolation and control in mind

> - network virtualization

here I see many issues, as for example Linux-VServer does not necessarily aim for full virtualization, when simple and performant isolation is sufficient.

don't get me wrong, we are _not_ against network virtualization per se, but we isolation is just so much simpler to administrate and often much more performant, so that it is very interesting for service separation as well as security applications

just consider the 'typical' service isolation aspect where you want to have two apaches, separated on two IPs, but communicating with a single sql database

> - netfilter virtualization

same as for network virtualization, but not really an issue if it can be 'disabled'

of course, the ideal solution would be some kind of hybrid, where you can have virtual interfaces as well as isolated IPs, side-by-side ...

> all these were discussed already somehow and looks like there is no
> fundamental differencies in our approaches (at least OpenVZ and Eric,
> for sure).
>
> Right now, I suggest to concentrate on first 2 namespaces - utsname
> and sysvipc. They are small enough and easy. Lets consider them
> without sysctl/proc issues, as those can be resolved later. I sent the
> patches for these 2 namespaces to all of you. I really hope for some
> _good_ critics, so we could work it out quickly.

will look into them soon ...

best,
Herbert

> Thanks,
> Kirill