Subject: Re: [RFC] [PATCH 0/4] uid_ns: introduction
Posted by serue on Wed, 08 Nov 2006 21:54:49 GMT

View Forum Message <> Reply to Message

Quoting Herbert Poetzl (herbert@13thfloor.at):
> On Wed, Nov 08, 2006 at 01:34:09PM -0700, Eric W. Biederman wrote:
> > Trond Myklebust <trond.myklebust@fys.uio.no> writes:
> >
> > > On Wed, 2006-11-08 at 01:52 +0100, Herbert Poetzl wrote:
> > >> On Mon, Nov 06, 2006 at 10:18:14PM -0600, Serge E. Hallyn wrote:
> > >> > Cedric has previously sent out a patchset
> > >> > (http://lists.osdl.org/pipermail/containers/2006-August/000078.html)
> > >> > impplementing the very basics of a user namespace. It ignores
> > >> > filesystem access checks, so that uid 502 in one namespace could
> > >> > access files belonging to uid 502 in another namespace, if the
> > >> > containers were so set up.
> > >> >
> > >> > This isn't necessarily bad, since proper container setup should
> > >> > prevent problems. However there has been concern, so here is a
> > >> > patchset which takes one course in addressing the concern.
> > >> >
> > >> > It adds a user namespace pointer to every superblock, and to
> > >> > enhances fsuid equivalence checks with a (inode->i_sb->s_uid_ns ==
> > >> > current->nsproxy->uid_ns) comparison.
> > >>
> > >> I don't consider that a good idea as it means that a filesystem
> > >> (or to be precise, a superblock) can only belong to one specific
> > >> namespace, which is not very useful for shared setups
> > >>
> > >> Linux-VServer provides a mechanism to do per inode (and per
> > >> nfs mount) tagging for similar 'security' and more important
> > >> for disk space accounting and limiting, which permits to have
> > >> different disk limits, quota and access on a shared partition
> > >>
> > >> i.e. I do not like it
> > >
> > > Indeed. I discussed this with Eric at the kernel summit this summer and
> > > explained my reservations. As far as I'm concerned, tagging superblocks
> > > with a container label is an unacceptable hack since it completely
> > > breaks NFS caching semantics.

So from your pov the same objection would apply to tagging vfsmounts,
or not?

What is the scenario where the caching is broken? It can't be multiple
clients accessing the same NFS export from the same NFS service container,
since that would just be an erroneous setup, right?

> > As I recall there are two basic issues.
> >
> > Putting the default on the mount structure instead of the superblock
> > for filesystems that are not uid namespaces aware sounded reasonable,
> > and allowed certain classes of sharing between namespaces where they
> > agreed on a subset of the uids (especially for read-only data).
>
> yes, that is especially interesting for --bind mounts
> when you 'know' that you will dedicate a certain
> sub-tree to one context/guest

Ok, so you wouldn't object to a patch which tagged vfsmounts?

I guess a NULL vfsmnt->user_ns pointer would mean ignore user_ns and
only apply uid checks (useful for ro bind mount of /usr into multiple
containers).

That of course wouldn't preclude also tagging inodes in later patches.

If you do object, then I can jump straight to tagging inodes with a
container, though that seems more likely to interfere conceptually
with any filesystems which are uid namespace aware.

thanks,
-serge
_____
Containers mailing list
Containers@lists.osdl.org
https://lists.osdl.org/mailman/listinfo/containers