Isolated user namespaces

Stéphane Graber, Owner at Zabbly
https://stgraber.org / stgraber@stgraber.org
@stgraber / @stgraber@hackyderm.io

Aleksandr Mikhalitsyn
Software engineer, Canonical
aleksandr.mikhalitsyn@canonical.com
What are user namespaces anyway?
So what’s wrong exactly?
What can we do about that?

Get more of them!
A lot more of them!
About 4.2 billion times as many!
What can we do about that?

Make \texttt{\textit{k\{u,g\}id\_t \textit{uint64}}}
But isn’t that breaking everything?

No!

- Extended uid/gid type is in-kernel only
- Userspace remains 32bit
- Persistent data remains 32bit
So, how does that work?
What about filesystems?

- Anything that can be mounted from within the namespace is fine and will work (tmpfs, fuse, …)

- Anything else you’ll have issues reading and writing to unless you’re using VFS idmap
Isn’t that a huge change?

23 files changed, 611 insertions(+), 81 deletions(-)
What’s next?

- Post the RFC patchset
- VFS idmap corner cases
- cgroupfs & cgroup namespace handling
- SCM_CREDS and some other boundary crossings
- Nested containers
  - Isolated userns in isolated userns
  - Regular userns in isolated userns
Questions?

Stéphane Graber, Owner at Zabbly
https://stgraber.org / stgraber@stgraber.org
@stgraber / @stgraber@hackyderm.io

Aleksandr Mikhalitsyn
Software engineer, Canonical
aleksandr.mikhalitsyn@canonical.com
Links

[1] Isolated dynamic user namespaces https://lpc.events/event/7/contributions/836/


[5] cgroupfs and cgroup namespace:
https://github.com/torvalds/linux/blob/3ca112b71f35dd5d99fc4571a56b5fc6f0c15814/kernel/cgroup/cgroup.c#L2169