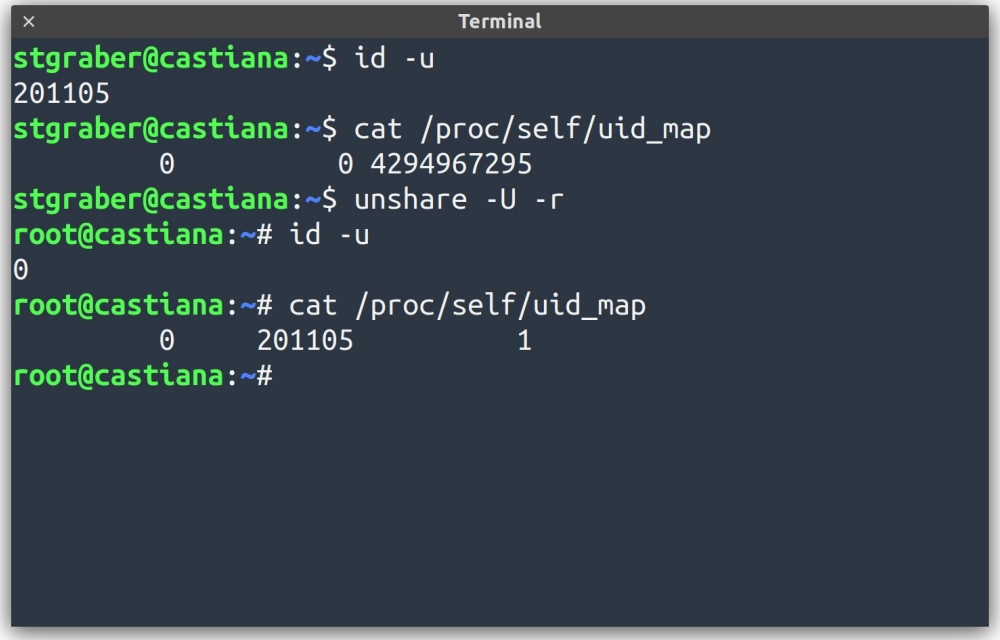
Isolated user namespaces

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

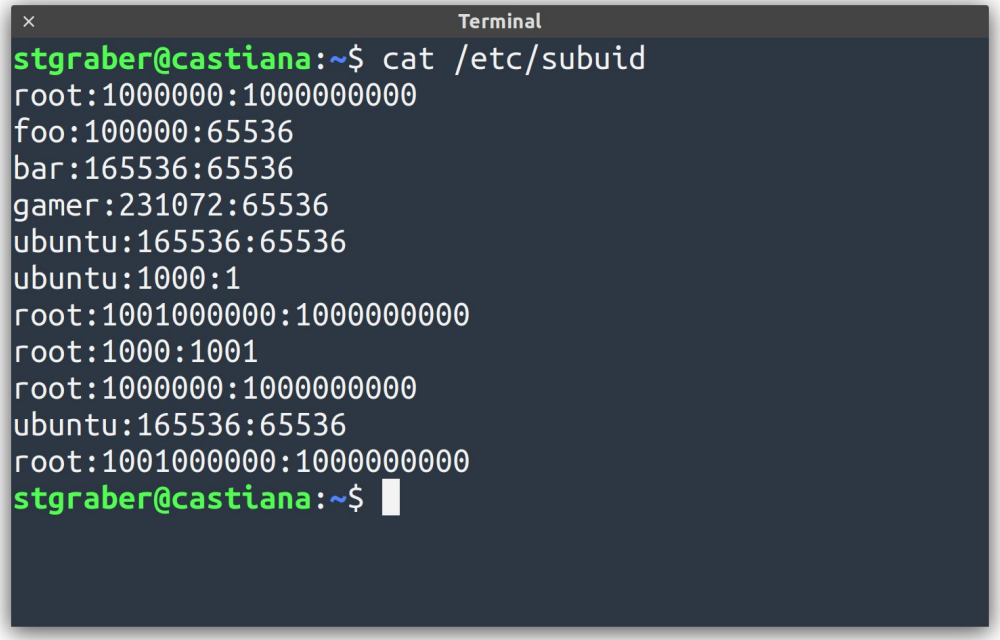
Aleksandr Mikhalitsyn
Software engineer, Canonical
aleksandr.mikhalitsyn@canonical.com

What are user namespaces anyway?



Demo

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 k{u,g}id_t 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?

Terminal

```
int main(int argc, char **argv)
{
        if (unshare(CLONE_NEWUSER)) {
                exit(1);
        }

        int fd_isol = open("/proc/self/isolated_uns", O_WRONLY);
        if (fd_isol < 0) {
                exit(1);
        }

        if (write(fd_isol, "yes", 3) < 0) {
                exit(1);
        }
        close(fd_isol);

        if (setuid(0) || setgid(0)) {
                exit(1);
        }

        execvp(argv[1], &argv[1]);

        return 0;
}
                                                    34,1          Bot
```

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

Demo

Isn’t that a huge change?

```
fs/9p/v9fs.h                            |    4
fs/crypto/keyring.c                     |    2
fs/mnt_idmapping.c                      |  113 ++++++++++++++++++-
fs/proc/array.c                         |    8 +
fs/proc/base.c                          |   53 +++++++++
fs/smb/server/smbacl.c                  |    4
fs/verity/signature.c                   |    2
include/linux/mnt_idmapping.h           |   60 ++++++++--
include/linux/skbuff.h                  |    2
include/linux/uidgid.h                  |  110 ++++++++++++++++---
include/linux/uidgid_types.h            |   16 ++
include/linux/user_namespace.h          |    4
kernel/auditsc.c                        |    6 -
kernel/ucount.c                         |    2
kernel/user.c                           |    1
kernel/user_namespace.c                 |  268 ++++++++++++++++++++++++++++++++++++++++++++---
net/core/fib_rules.c                    |    5
net/wireless/reg.c                      |    2
security/integrity/digsig.c             |    4
security/integrity/ima/ima_mok.c        |    2
security/integrity/ima/ima_policy.c |   16 +-
security/keys/persistent.c              |    2
security/safesetid/lsm.c                |    6 +
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/

[2] Simplified user namespace allocation https://lpc.events/event/11/contributions/982/

[3] Linux kernel patches: https://github.com/mihalicyn/linux/commits/isolated_userns

[4] LXC patch: https://github.com/mihalicyn/lxc/commits/isolated_userns

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