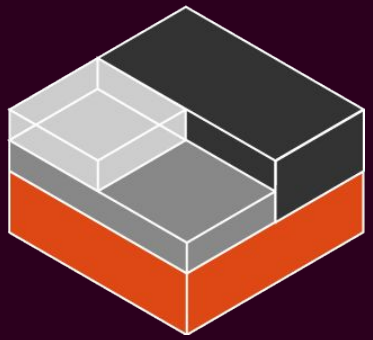
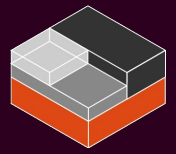
fs mounts in userns

OSDN
Kiev



Christian Brauner
Kernel Engineer and LXD Maintainer, Canonical Ltd.
christian@brauner.io
@brau_ner
https://brauner.github.io/

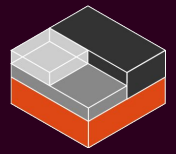
Namespaces



Abbreviations used in this talk:

- ns := namespace
- userns := user namespace
- fs := filesystem

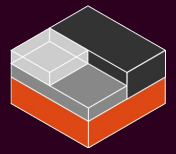
Namespaces



mount, PID, UTS, IPC, cgroup, network, user

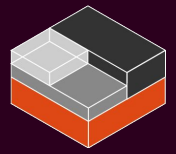
(time, device, ima)

Namespaces



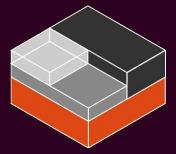
no real privilege separation for most ns
→ introduce ns for privilege separation

User namespace
requirements



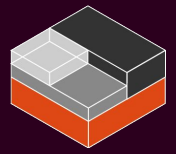
- separate host ids from userns ids
- userns root id privileged over userns
  https://asciinema.org/a/197170
- nesting should be possible
- userns root id not privileged over any resources it does not own
- unprivileged user should be able to safely create a userns

User namespace



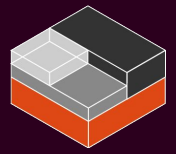
- bijective mapping between host ids and userns ids
- isomorphism to retain permission model

User namespace



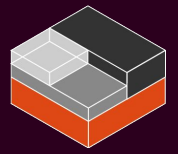
```
https://asciinema.org/a/lphpBinvqFxDCnYWsed77g4PD

lxc-usernsexec -m b:0:1000:1 -m b:1:1001:1 -m
b:2:1002:1 -m b:3:1003:1 -- bash

cat /proc/self/uid_map

lxc-usernsexec -m b:0:1000:4 -- bash

cat /proc/self/uid_map
```

User namespace



- capabilities
- owning user namespace
- resources

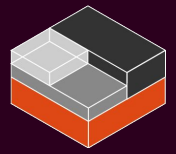
Denying fs mounts in userns



some problems

- device files
- sid bits
- fcaps
- unmapped ids

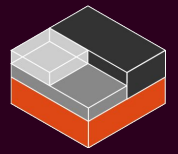
Enabling fs mounts in userns



changing vfs infrastructure

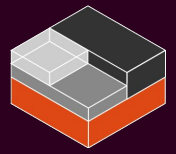
- `s_user_ns`
- `s_iflags & SB_I_NODEV`
- `VFS_CAP_REVISION_3`
  `https://asciinema.org/a/195209`

Allowing fs mounts in userns(?)



vfs robustness != fs robustness

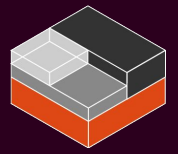
Allowing fs mounts in userns(?)



FUSE

- filesystem in userspace
- kernel module + libfuse library
- userspace can write its own fs
- any fs code now runs in userspace

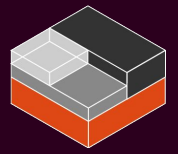
Allowing fs mounts in userns(?)



ext4, overlayfs, XFS, btrfs

- seccomp
- new mount API
- lklfuse

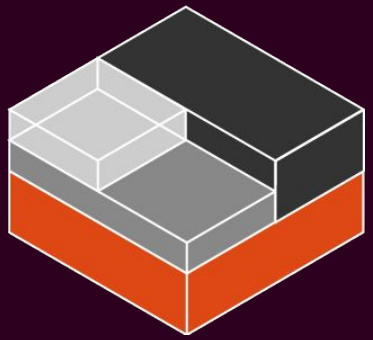
the people in the background



- Eric Biederman
- Serge Hallyn
- Seth Forshee
- Stéphane Graber

fs mounts in userns

Container Camp
London, UK



Christian Brauner
Software Engineer, Canonical Ltd.
christian@brauner.io
@brau_ner
https://brauner.github.io/