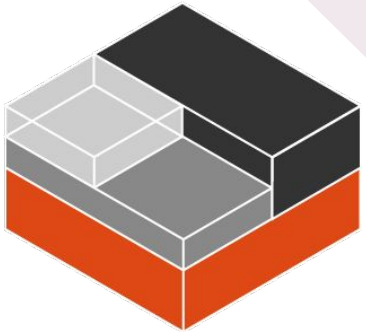
Supervising and emulating syscalls



Christian Brauner
LXD maintainer & kernel engineer

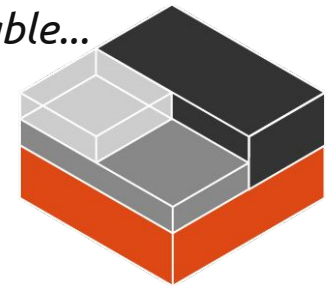
@brau_ner
https://brauner.io
christian.brauner@ubuntu.com

CANONICAL ubuntu

seccomp <3 syscalls

- **Syscalls**
  Allow userspace to communicate with the kernel (a fancy request handler).

- **Seccomp**
  Allows to intercept system calls and then denies or allows them.
  The kernel never blocks in seccomp.
  Seccomp runs before the syscall number is looked up in the syscall table...

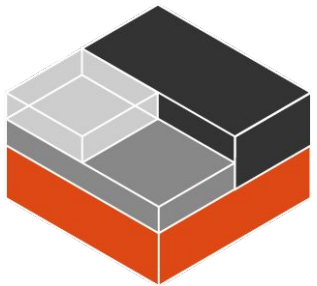


seccomp: notify

- **Allows running less privileged containers**
  Unprivileged containers can be granted very specific privileges.

- **Seccomp asks userspace for return value and errno**
  Execution does NOT continue in the kernel, userspace must do the work.

- **Initial support landed in 5.0**
  Userspace requires un-released libseccomp.

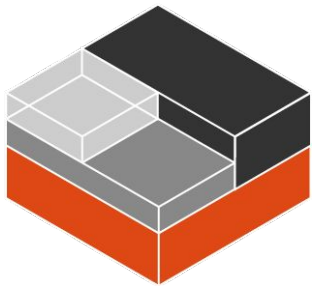


seccomp: resume syscalls

- **Builds on top of existing notify target**
  Effectively a new type of return value from userspace.

- **Allows for complex userspace filtering**
  For cases where the kernel cannot filter on some arguments.

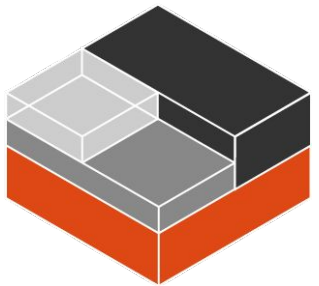
- **No raised privileges**
  Execution continues in the kernel with original privileges.



seccomp + pidfd_getfd()

- **Inventing new syscalls**
  Seccomp runs before the syscall number is verified to be valid...
  Retrieving file descriptors from target.

- **File descriptor retrieval**
  Retrieve a file descriptor from a target process and perform actions on it.



Questions ?

We have LXD stickers,
come get them in front!

Christian Brauner
LXD maintainer & kernel engineer

@brau_ner
https://brauner.io
christian.brauner@ubuntu.com