Contribution ID: **266** Type: **not specified**

Seccomp Syscall Interception

Tuesday, September 10, 2019 3:45 PM (15 minutes)

Recently the kernel landed seccomp support for SECCOMP_RET_USER_NOTIF which enables a process (watchee) to retrieve a fd for its seccomp filter. This fd can then be handed to another (usually more privileged) process (watcher). The watcher will then be able to receive seccomp messages about the syscalls having been performed by the watchee.

We have integrated this feature into userspace and currently make heavy use of this to intercept mknod() syscalls in user namespaces aka in containers.
If the mknod() syscall matches a device in a pre-determined whitelist the privileged watcher will perform the mknod syscall in lieu of the unprivileged watchee and report back to the watchee on the success or failure of its attempt. If the syscall does not match a device in a whitelist we simply report an error.

This talk is going to show how this works and what limitations we run into and what future improvements we plan on doing in the kernel.

I agree to abide by the anti-harassment policy

Yes

Primary author: Mr BRAUNER, Christian

Presenter: Mr BRAUNER, Christian

Session Classification: Containers and Checkpoint/Restore MC