Module: Sandboxing

seccomp

Yan Shoshitaishvili
Arizona State University

More Complex: Syscall Filtering

Modern sandboxes *heavily* restrict permitted system calls through the use a kernel-level sandboxing mechanism: seccomp.

seccomp allows developers to write complex rules to:

- allow certain system calls
- disallow certain system calls
- filter allowed and disallowed system calls based on argument variables

seccomp rules are inherited by children!

These rules can be quite complex (see http://man7.org/linux/man-pages/man3/seccomp_rule_add.3.html).

How does seccomp work?

seccomp uses the kernel functionality eBPF.

extended **B**erkeley **P**acket **F**ilters are programs that run in an *in-kernel*, *"provably-safe"* virtual machine.

Can instrument kernel functionality in very general ways.

- originally used to filter network traffic (iptables)
- i.e., used to implement system-wide syscall tracing: https://github.com/iovisor/bcc

Used with `seccomp()` to apply syscall filters to processes.