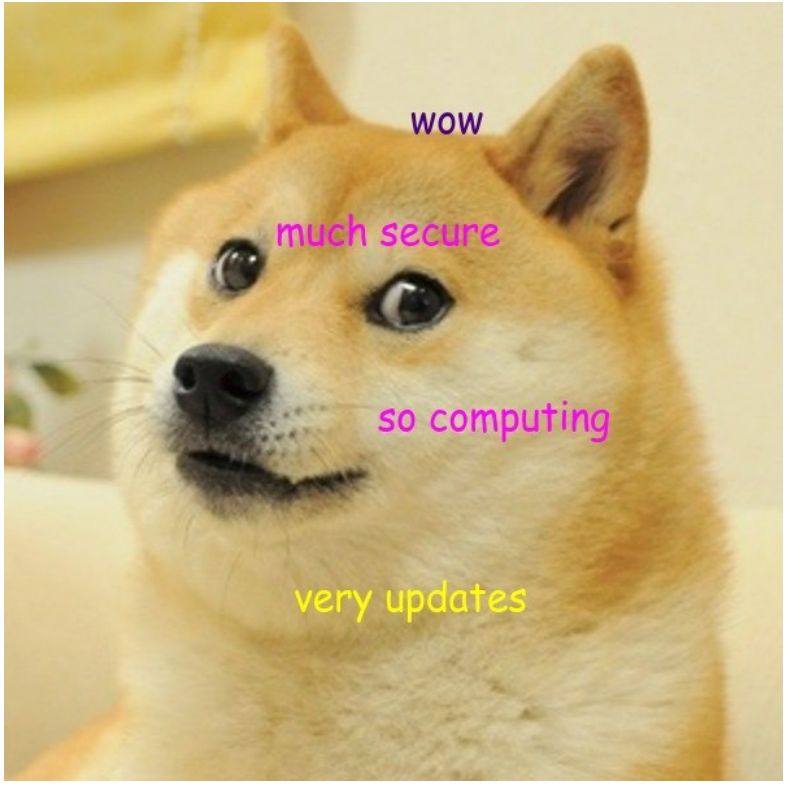
seccomp update

v4.9 – v4.14



https://outflux.net/slides/2017/lss/seccomp.pdf

Linux Security Summit, Los Angeles 2017
Kees Cook <keescook@chromium.org>
(pronounced “Case”)

What is seccomp?

- Programmatic kernel attack surface reduction
- Used by:
  - Chrome
  - Android (minijail)
  - vsftpd
  - OpenSSH
  - Systemd (“SystemCallFilter=...”)
  - LXC (blacklisting?)
  - … and you too! (easiest via libseccomp)

Architecture support

- x86: v3.5
- s390: v3.6
- arm: v3.8
- mips: v3.15
- arm64: v3.19, AKASHI Takahiro
- powerpc: v4.3, Michael Ellerman
- tile: v4.3, Chris Metcalf
- um: v4.5, Mickaël Salaün
- parisc: v4.5, Helge Deller
- No new architecture support this year!

coredumps

- Since v4.11, from Mike Frysinger
    - (with a tweak from Kees Cook)
- Uncaught SIGSYS (i.e. RET_KILL) now generates a coredump
- Now matches existing docs for SIGSYS!
  - See “man 7 signal” *cough*

Dynamic logging

- For v4.14, from Tyler Hicks
- New sysctls (audit is unchanged):
  - kernel.seccomp.actions_avail (read-only list of all seccomp actions)
  - kernel.seccomp.actions_logged (writable list of seccomp actions)
    - Never log any RET_ALLOW
    - Optionally log all RET_KILL
- New filter flag: SECCOMP_FILTER_FLAG_LOG
  - If this filter is hit, log the result (modulo actions_logged)
- New action: SECCOMP_RET_LOG
  - Like RET_ALLOW, but log it (modulo actions_logged)
- New op: SECCOMP_GET_ACTION_AVAIL (programmatic version of actions_avail)

Process killing

- For v4.14, from Kees Cook
- SECCOMP_RET_KILL now an alias for SECCOMP_RET_KILL_THREAD since that's what it does
- SECCOMP_RET_PROCESS added to kill the entire thread group
  - SECCOMP_RET_* ordered by value, and KILL_THREAD was 0 …
  - Steal reserved high bit to make KILL_PROCESS be -1!
  - Needs updated SECCOMP_RET_ACTION mask, available as SECCOMP_RET_ACTION_FULL

Regression tests

- tools/testing/selftests/seccomp/seccomp_bpf.c
- v4.13:
  - Make test harness generally available: Mickaël Salaün
  - Fix threading vs Bionic: Paul Lawrence
- v4.14:
  - Action availability, filter flags, RET_LOG: Tyler Hicks
  - Sane ptrace, RET_KILL_PROCESS: Kees Cook
- Tile arch support testing remains missing

Wanted: deep argument inspection

- seccomp must not access userspace memory
  - check would race with syscall usage
  - double-read would result in poor performance
- Possible solutions
  - Landlock
  - flag an LSM to perform checks at LSM hook time
  - cached argument copying requires teaching syscall infrastructure about the cache

Questions?

https://outflux.net/slides/2017/lss/seccomp.pdf

@kees_cook

keescook@chromium.org

keescook@google.com

kees@outflux.net