Namespaces and Capabilities

Overview and Recent Developments

Linux Security Summit Europe
Edinburgh, Scotland

Christian Brauner
christian@brauner.io
christian.brauner@ubuntu.com
@brau_ner
https://brauner.github.io/

Capabilities

- splitting root into units of privilege

- **libcap(3)**
  - **libcap 2.26** released on 10 September 2018 with full ambient capability and namespaced filesystem capability support
  - libcap has moved to https://git.kernel.org/pub/scm/libs/libcap/libcap.git/ and Andrew Morgan is back maintaining it

Namespaces

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

(time, device*, ima**)

* technically not a real namespace
** will likely not be a namespace but tied to a namespace

User namespace

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
- 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

- capabilities
- owning user namespace
- resources

4.10

- **(user,mnt)** infrastructure to enable unprivileged mount
  - 412ac77a9d3ec015524dacea905471d66480b7ac
- **(user)** added a user_ns owner to mm_struct so you can have sensible ptrace permissions checks across user namespace boundaries on e.g. exec
  - bfedb589252c01fa505ac9f6f2a3d5d68d707ef4
- **(net)** SIOCGSKNS: add an ioctl to get a socket network namespace
  - c62cce2caee558e18aa05c01c2fd3b40f07174f2

4.11

- **(user)** limit inotify instances inside a user namespace
  - 1cce1eea0aff51201753fcaca421df825b0813b6
- **(user)** addition of namespace ioctl()s to query hierarchy and properties of namespaces
  - d95fa3c76a66b6d76b1e109ea505c55e66360f3c (NS_GET_OWNER_UID)
  - e5ff5ce6e20ee22511398bb31fb912466cf82a36 (NS_GET_PARENT, NS_GET_USERNS)
- **(user,mnt)** infrastructure to enable unprivileged mounts
  - 93faccbbfa958a9668d3ab4e30f38dd205cee8d8

4.12

- **(pidns)** expose pidns_for_children in /proc/<pid>/ns/
  - eaa0d190bfe1ed891b814a52712dcd852554cb08

- **(pidns)** add pid namespace support to fuse takes care to translate pids when the userspace process servicing fuse requests is running in a pid namespace
  - 0b6e9ea041e6c932f5b3a86fae2d60cbcfad4dd2

- **(all)** include namespace info in perf output via PERF_RECORD_NAMESPACES
  - e422267322cd319e2695a535e47c5b1feeac45eb

4.13

- **(mnt)** improve umount performance dramatically (0.06s vs 60s with overlapping mount propagation trees)
  - 296990deb389c7da21c78030376ba244dc1badf5
- **(cgroup)** add "nsdelegate" to allow cgroup delegation by considering cgroup namespaces delegation boundaries
  - 5136f6365ce3eace5a926e10f16ed2a233db5ba9

4.14

- **(user)** introduce namespaced file capabilities
  - 8db6c34f1dbc8e06aa016a9b829b06902c3e1340

4.15

- **(user)** bump limit of allowed user namespace mappings from 5 to 340
  - aa4bf44dc851c6bdd4f7b61b5f2c56c84dfe2ff0
  - 6397fac4915ab3002dc15aae751455da1a852f25
  - 11a8b9270e16e36d5fb607ba4b60db2958b7c625
  - 3edf652fa16562fb57a5a4b996ba72e2d7cdc38b
  - d5e7b3c5f51fc6d34e12b6d87bfd30ab277c4625
  - ece66133979b211324cc6aff9285889b425243d2
  - 3fda0e737e906ce73220b20c27e7f792d0aac6a8

4.16

- **(net)** query peer network namespaces (RTM_NEWLINK, RTM_DELLINK, RTM_SETLINK)
  - 7c4f63ba8243024929855530188814559822 41d6
  - c310bfcb6e1be993629c5747accf8e1c65fbb255
  - b61ad68a9fe85d29d5363eb36860164a049723cf
  - 5bb8ed075428b71492734af66230aa0c07fcc515
  - 7973bfd8758d05c85ee32052a3d7d5d0549e91b4
  - 4ff66cae7f10b65b028dc3bdaaad9cc2989ef6ae

4.17

- **(user,mnt)** make unprivileged fuse mounts work with ima
  - dbf107b2a7f36fa635b40e0b554514f599c75b33
  - c9582eb0ff7d2b560be60eafab29183882cdc82b
  - 8cb08329b0809453722bc12aa912be34355bcb66
  - 73f03c2b4b527346778c711c2734dbff3442b139
  - 57b56ac6fecb05c3192586e4892572dd13d972de
- **(user,mnt)** devpts: resolve devpts bind-mounts
  - a319b01d9095da6f6c54bd20c1f1300762506255
- **(user,net)** uevent injection (device namespaces)
  - 94e5e3087a67c765be98592b36d8d187566478d5
  - 692ec06d7c92af8ca841a6367648b9b3045344fd

4.18

- **(user,mnt)** finalize infrastructure to enable unprivileged mounts (aka getting away with regressing userspace)
  - 593d1ce854dff93b3c9066e897192eb676b09c46
  - 55956b59df336f6738da916dbb520b6e37df9fbd
  - 0031181c49ca94b14b11f08e447f40c6ebc842a4
  - bc6155d1326092f4c29fe05a32b614249620d88e
  - b1d749c5c34112fab5902c43b2a37a0ba1e5f0f1
  - f3f1a18330ac1b717cd7a32adff38d965f365aa2

4.18

- **(user,mnt)** enable unprivileged fuse mounts
  - e45b2546e23c2d10f8585063a15c745a7603fac9
  - 4ad769f3c346ec3d458e255548dec26ca5284cf6
- **(user,net)** uevent namespacing (device namespaces)
  - 26045a7b14bc7a5455e411d820110f66557d6589
  - a3498436b3a0f8ec289e6847e1de40b4123e1639

Current Patchsets

- **(time)** introduce time namespaces
  - https://lkml.org/lkml/2018/9/19/950
- **(mnt)** AT_{BENEATH, NO_PROCLINKS, NO_SYMLINKS, THIS_ROOT, XDEV}
  - https://lists.linuxfoundation.org/pipermail/containers/2018-October/039525.html
- **(net)** query peer network namespaces (RTM_GETADDR)
  - https://lists.linuxfoundation.org/pipermail/containers/2018-September/039351.html
- **(mnt)** new mount api
  -

Future Patchsets

- **(mnt)** recursive read-only bind mounts for old and new mount api
  - (new mount api) https://lkml.org/lkml/2018/9/24/1096
  - (old mount api) https://github.com/brauner/linux/commits/2018-09-05/ro (MS_REC_RDONLY)
- **(mnt)** make umount{2}() reversible (tucked mounts)
  - https://github.com/brauner/linux/tree/2018-10-07/tucked_mounts
- **(mnt)** handle mount propagation in statfs() syscall
  - https://lkml.org/lkml/2018/5/25/397
- **(user)** introduce new ns ioctl()s NS_{IS_INIT,ACCESS}
  - https://github.com/brauner/linux/tree/2018-10-01/ns_is_init

Future Patchsets

- **(lsm)** lsm namespacing/stacking
  - http://namei.org/presentations/selinux_namespacing_lca2018.pdf
- **(seccomp)** seccomp trap from userspace
  - https://lists.linuxfoundation.org/pipermail/containers/2018-September/039419.html

New things to argue about

- **(capabilities)** split CAP_SYS_ADMIN
  - likely a major task, very hard to convince the right people, likely requires a whole new kernel config option (e.g. CONFIG_CAPABILITY_V2=y)

Namespaces and Capabilities

Overview and Recent Developments

Linux Security Summit Europe
Edinburgh, Scotland

Christian Brauner
christian@brauner.io
christian.brauner@ubuntu.com
@brau_ner
https://brauner.github.io/