

Completely unprivileged containers

Serge Hallyn

LXC project

November 4, 2016

- Do not exist
 - ▶ Cgroups
 - ▶ Namespaces
 - ▶ LSMs
- Tools written to hide this
 - ▶ `docker run --rm -it ubuntu bash`
 - ▶ `lxc launch ubuntu:xenial x1`
- These tools require root or privileged group
 - ▶ Probably worth it for convenience, but
 - ▶ Not inherently required

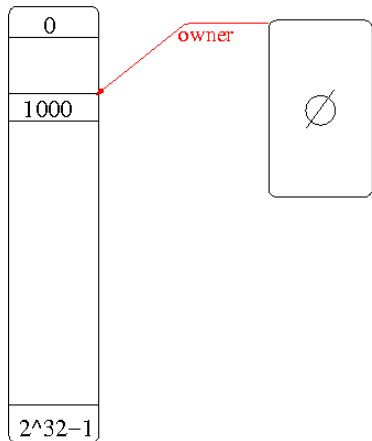
Using namespaces by hand

```
sudo lxc-unshare -s "MOUNT|PID" -M -- bash
```

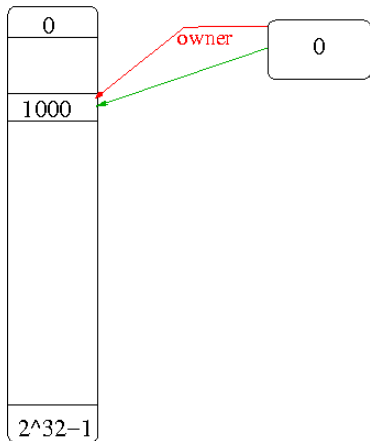
Uid namespace

- Map userspace uids (`uid_t`) to kernel kuid
- Namespace id 0 is privileged over namespace's resources and may unshare all namespaces
- By default, uid map is $\{0 : 2^{32} - 1\} \rightarrow \{0 : 2^{32} - 1\}$
- Any user may unshare
- New uid namespace has no mapping (\emptyset)
- Unprivileged user may map own uid to any namespace id
- Setuid-root programs delegate subuids
 - ▶ `newuidmap` using `/etc/subuid`
 - ▶ `newgidmap` using `/etc/subgid`

UID Namespaces



(a) Empty UID Namespace



(b) Unprivileged UID Namespace

UID Namespaces

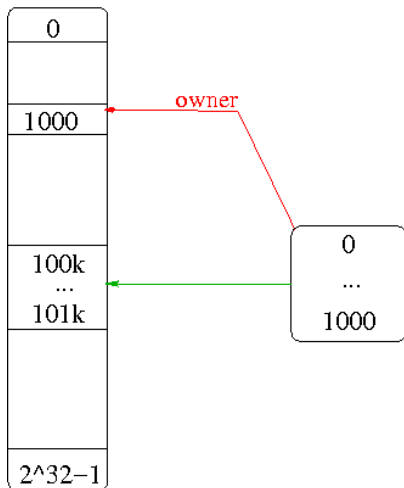


Figure: Delegated UID Namespace

Putting it together

Let's make a rootfs

```
mkdir rootfs.dir
lxc-usernsexec -m b:0:1000:1 -m b:1:100000:1 -- chown 1:1 rootfs.dir
lxc-usernsexec -m b:0:100000:65536 -- tar -C rootfs.dir -Jxf rootfs.tar.xz
lxc-usernsexec -- lxc-unshare -s "MOUNT|PID" sh
    touch rootfs.dir/dev/null
    mount --bind /dev/null rootfs.dir/dev/null
    chroot rootfs.dir sh
    mount -t proc proc /proc
    adduser ubuntu
lxc-usernsexec -- rm -rf rootfs.dir/*
```

Networking

- User namespace can unshare network namespace which it then owns
- User ns cannot “hook into” host namespace
- Solution: delegate bridges
 - ▶ Be careful! nics can spoof each other
 - ▶ `/etc/lxc/lxc-usernet`: user veth bridge number
- `lxc-user-nic`
 - ▶ Creates veth pair
 - ▶ Inserts one into container
 - ▶ Hooks other into specified host bridge (if permitted)

Summary

- Creating your own containers requires:

- ▶ Delegated subuids

and `newuidmap` and `newgidmap`

- ▶ Delegated bridge

and `lxc-user-nic`

- ▶ Delegated cgroup

`pam` - not as crucial

```
echo "session optional pam_cgfs.so -c freezer,memory,name=systemd" >>  
/etc/pam.d/common-services
```

Using LXC

```
lxc-create -t download -n x1 -- \
    -d ubuntu -r xenial -a amd64
lxc-start -n x1
lxc-attach -n x1
lxc-stop -n x1
lxc-destroy -n x1
```

- Container exists under `$HOME/.local/share/lxc/x1`

Questions/Comments?

- <http://linuxcontainers.org>
- `lxc-{users,devel}@lists.linuxcontainers.org`
- `serge@hallyn.com`