Linux Containers

Process Management

Namespaces

Chroot

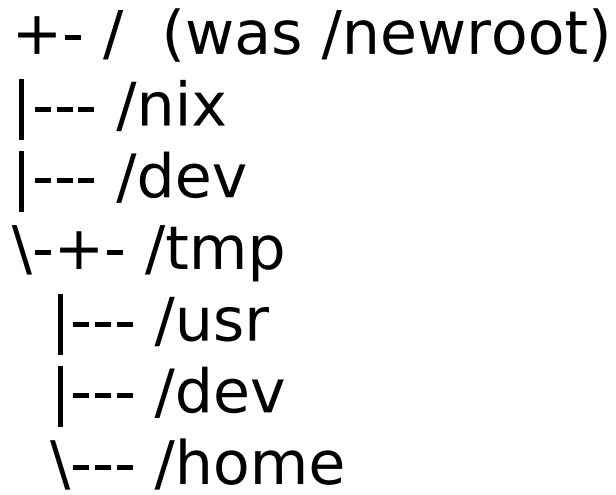
```
# ls -la /proc/self/root
/
# chroot /newroot sleep 1000 & pid=$!
[4] 29988
# ls -la /proc/$pid/root
/newroot
```

Pivot Root


+- /
|--- /usr
|--- /dev
|--- /home
\-+- /newroot
|--- /nix
|--- /dev
\--- /tmp

```
# pivot_root /newroot /newroot/tmp
# umount /tmp
```

Warning: modifies current mount namespace


+- / (was /newroot)
|--- /nix
|--- /dev
\-+- /tmp
  |--- /usr
  |--- /dev
  \--- /home

mount --bind /dev /newroot/dev

Avoiding Mount Namespace Change

- openat()
- linkat()
- renameat()

Mount Namespace

```
# unshare --mount /bin/bash
# mount --make-rprivate /
# mount -t tmpfs tmpfs /tmp
# ls -la /tmp
total 16
drwxrwxrwt 2 root root  40 Jan 19 22:32 .
drwxr-xr-x 1 root root 142 Dec 13 16:14 ..
#
```

Network Namespace

```
# ip netns create isolated
# ip netns exec isolated ip addr
1: lo: <LOOPBACK> mtu 65536 qdisc noop state DC
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00
```

```
# ip netns exec isolated ip link set dev lo up
# ip netns exec isolated ip addr
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group default
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
       valid_lft forever preferred_lft forever
```

ılt

```
# ip netns exec isolated wget google.com
--2015-01-19 22:49:23--  http://google.com/
Resolving google.com (google.com)... 173.194.113.194, ...
Connecting to google.com (google.com)|173.194.113.194|:80...
failed: Network is unreachable.
```

How Wget Resolves IP?

```
# ip netns exec isolated \
  strace -f -s 100 wget google.com
```

write(2, "Resolving google.com (google.com)... ", 37Resolving google.com (google.com)...) =
socket(PF_LOCAL, SOCK_STREAM|SOCK_CLOEXEC|SOCK_NONBLOCK, 0) = 3
connect(3, {sa_family=AF_LOCAL, sun_path="/var/run/nscd/socket"}, 110) = 0
sendto(3, "\2\0\0\0\r\0\0\0\6\0\0\0hosts\0", 18, MSG_NOSIGNAL, NULL, 0) = 18
poll([{fd=3, events=POLLIN|POLLERR|POLLHUP}], 1, 5000) = 1
recvmsg(3, {msg_name(0)=NULL, msg_iov(2)=[...]...}) = 14

) = 37

```
# ip link add ext type veth peer name int
# ip link set int netns isolated
# ip addr add dev ext 192.168.17.1/24
# ip netns exec \
  ip addr add dev int 192.168.17.2/24
```

```
# unshare --net bash
# echo $$
12356
...
# touch /run/netns/isolated
# mount --bind /proc/12356/ns/net \
    /run/netns/isolated
```

```
# nsenter --net=/run/netns/isolated /bin/bash
```

The setns() system call

```
# unshare --uts
# hostname something
```

- ~~resolv.conf~~
- ~~hosts~~

unshare --ipc

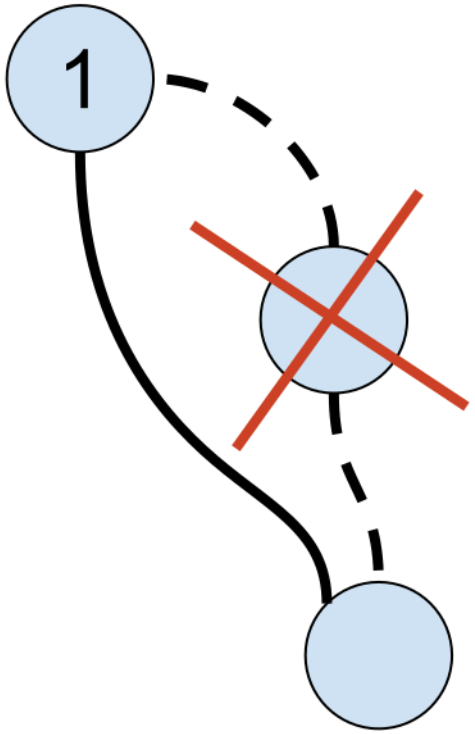
```
# unshare --pid --fork sh -c 'echo $$'
1
```

Pid Namespace

- /proc filesystem
- security: ps, kill, strace

Pid 1

- KILL
- reparenting
- Term -> Ignore


1

Not-a-Pid-1

- KILL -> prctl(PR_SET_PDEATHSIG)
- reparenting -> prctl(PR_SET_CHILD_SUBREAPER)
- Term -> Ignore (sigaction)

unshare --user

User Namespaces

- namespaces for unprivileged users
- mapping of uids/gids

docker --net=host
docker --pid=host

Competitors

- systemd-nspawn/machinectl
- runc
- rkt

Rust

- https://crates.io/crates/unshare

- nix
- pbuilder
- network shaping
- packet loss
- vagga