Container SW Workshop

Attendance

CERN/HEP: Lukas Heinrich, Jakob Blomer, Clemens Lange, Ricardo Rocha, Alessandra Forti (Vidyo)

Google: Abdelfettah SGHIOUAR

Uni Michigan: Bob Killen, Jeffrey Sica

NTT: Akihiro Suda, Kohei Tokunaga

IBM: Phil Estes, Nuri Twebti

CSCS: Kean Mariotti, Filipe Cruz

Jessie Frazelle, Christian Kniep

Main Topics

Container Software Distribution

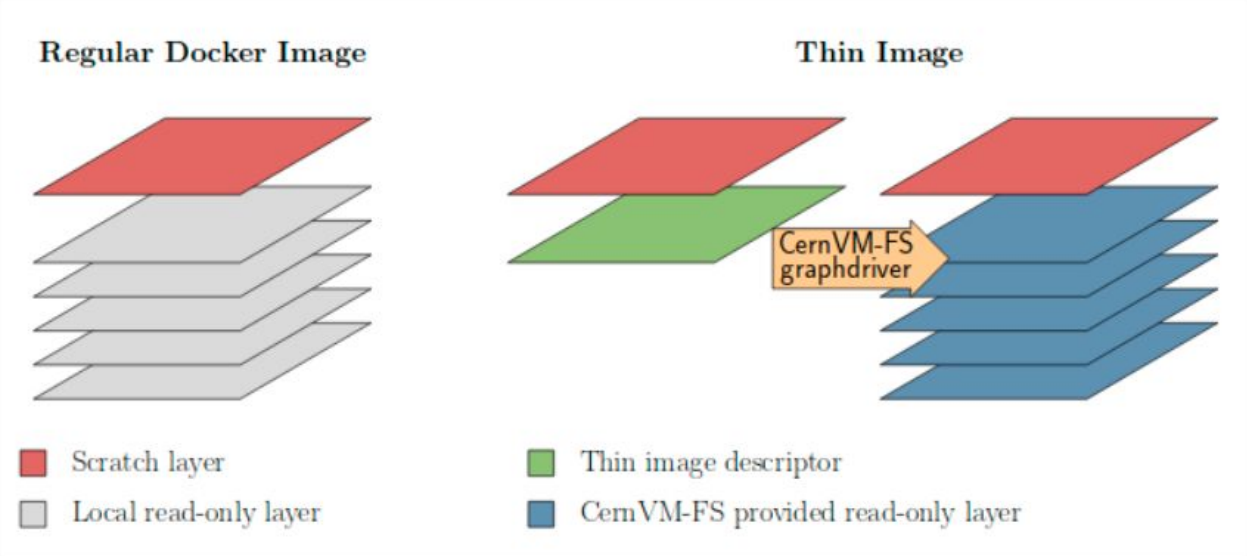
Rootless Containers

Container SW Distribution

Presented by Jakob Blomer

Follow up on the initial implementation of the Docker CVMFS Graph Driver

Custom Graph Driver / Snapshotter Approach



Ideally: native support for (some) unpacked layers on a read-only file sytem

Container SW Distribution

Presented by Jakob Blomer

Follow up on the initial implementation of the Docker CVMFS Graph Driver

Goal is to move the logic into a containerd snapshotter

- As containerd looks like the best common denominator
- Docker, Kubernetes with cri-containerd

https://github.com/containerd/containerd/issues/2943

Similar request from Google's CRFS (Container Registry Filesystem) and Filegrain (IPFS backed storage)

The state of rootless

Long and detailed presentation from Akihiro, check it out here

Rootless Containers

- Run containers, runtimes, and orchestrators as a non-root user

- Don't confuse with:
  - `usermod -aG docker penguin`
  - `docker run --user`
  - `dockerd --userns-remap`

What Rootless Containers cannot

- If a container was broke out, the attacker still might be able to
  - Mine cryptocurrencies
  - Springboard-attack to other hosts

- Not effective for kernel / VM/ HW vulns
  - But we could use gVisor together for mitigating some of them

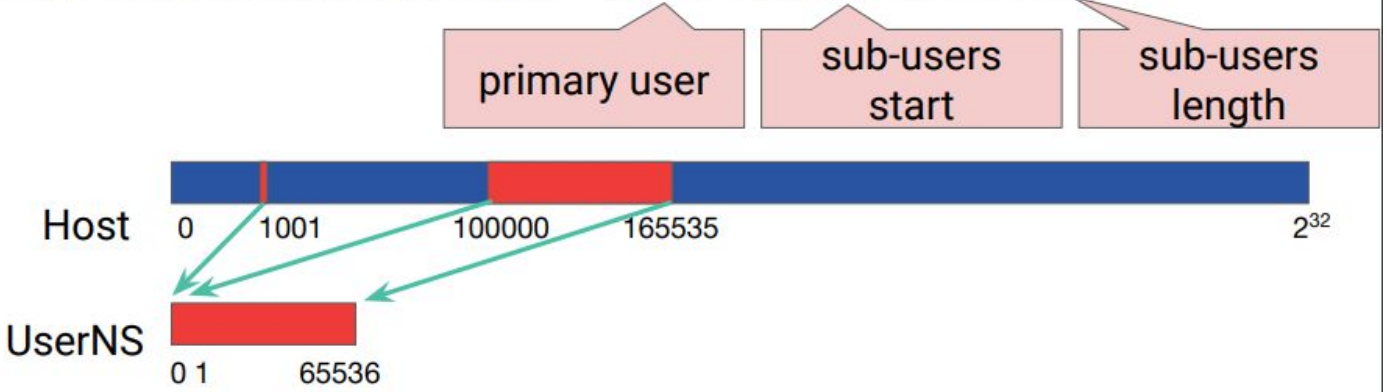
User Namespaces

User Namespaces

```
$ docker-rootless run -v /:/host -it alpine
/ # ls -ln /host/dev/sda
brw-rw---- 1 65534    65534   8,    0 May  1 12:00
/host/dev/sda
/ # cat /host/dev/sda
cat: can't open '/host/dev/sda': Permission denied
```

Sub-users (and sub-groups)

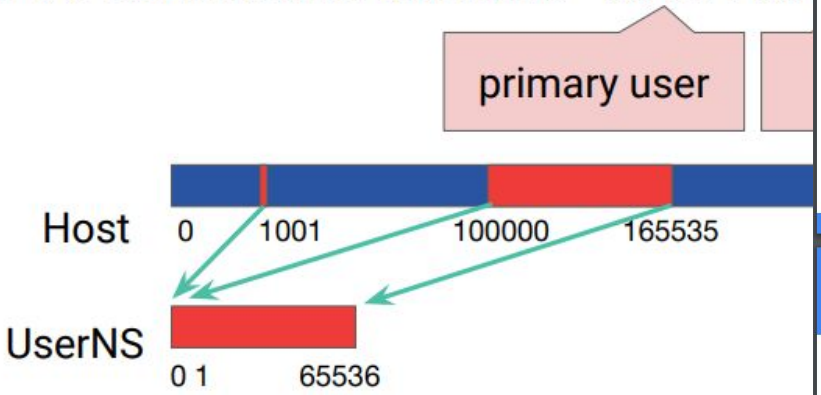
- If `/etc/subuid` contains "`1001:100000:65536`"



- Having 65536 sub-users should be enough for most containers

Sub-users (and sub-groups

- If `/etc/subuid` contains "`1001:10`



- Having 65536 sub-users should be en containers

Sub-users (and sub-groups)

- Sub-users are configured via SUID binaries `/usr/bin/{newuidmap, newgidmap}`
- SETUID binary can be dangerous; `newuidmap` & `newgidmap` had two CVEs so far:
  - CVE-2016-6252 (CVSS v3: 7.8): integer overflow issue
  - CVE-2018-7169 (CVSS v3: 5.3): supplementary GID issue

Sub-users (and sub-groups)

- Also hard to maintain sub-users
  - LDAP / AD
  - Nesting user namespaces might need huge number of sub-users

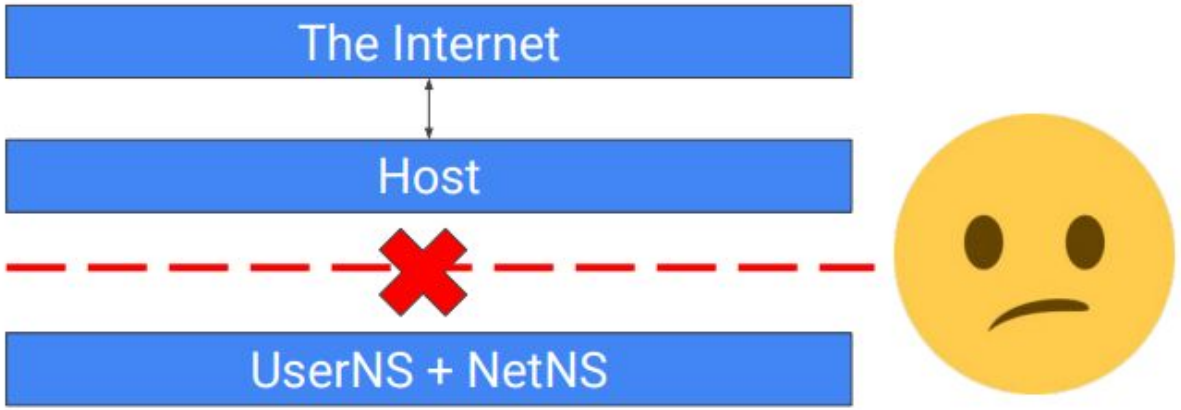
Sub-users (and sub-groups)

- Alternative way: Single-mapping mode
- Does not require `newuidmap/newgidmap`
- Ptrace and/or Seccomp can be used for intercepting syscalls to emulate sub-users
  - `user.rootlesscontainers` xattr can be used for chown emulation

Network Namespaces

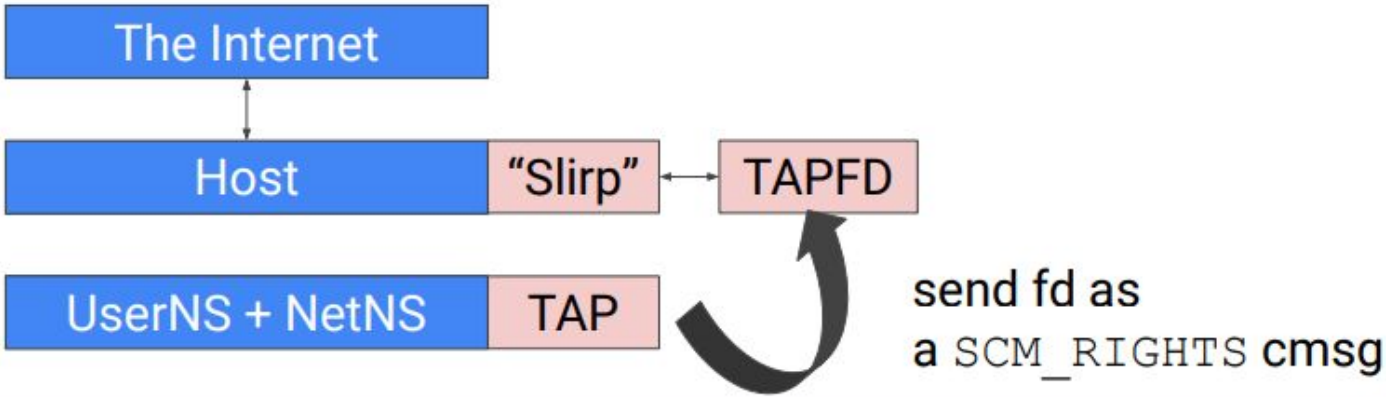
Network Namespaces

- But an unprivileged user cannot set up `veth` pairs across the host and namespaces, i.e. No internet connection



Network Namespaces

- `lxc-user-nic` SUID binary allows unprivileged users to create veth, but we are not huge fun of SUID binaries
- Our approach: use completely unprivileged usermode network ("Slirp") with a TAP device



Network Namespaces

Benchmark of several “Slirp” implementations:

	MTU=1500	MTU=4000	MTU=16384	MTU=65520
vde_plug	763 Mbps	Unsupported	Unsupported	Unsupported
VPNKit	514 Mbps	526 Mbps	540 Mbps	Unsupported
slirp4netns	1.07 Gbps	2.78 Gbps	4.55 Gbps	9.21 Gbps
cf. rootful veth	52.1 Gbps	45.4 Gbps	43.6 Gbps	51.5 Gbps

- slirp4netns (our own implementation based on QEMU Slirp) is the fastest because it avoids copying packets across the namespaces

Benchmark: iperf3 (netns -> host), measured on Travis CI. See rootless-containers/rootlesskit#12

Snapshotting / FS

Snapshotting

- OverlayFS is currently unavailable in UserNS (except on Ubuntu kernel)
- FUSE-OverlayFS can be used instead with kernel 4.18+
- XFS reflink can be also used to deduplicate files (but slow)

containerd plans

containerd plans

Implement FUSE-OverlayFS snapshotter plugin

Add support for cgroup2 (waiting for OCI spec to support it first)

Support running containerd in gVisor (mitigate potential kernel vulns)

Follow Up

Keep following the remote snapshotter PR

Pity Lantao from Google could not make it

https://github.com/containerd/containerd/issues/2943

(contact Brad from Google to see what they are doing)

Follow Akihiro's work on rootless for the whole stack

runc, docker, kubernetes, …

Seems we will just get it at some point

Akihiro Suda @_AkihiroSuda_ · 9h

Rio is going to switch from Knative Build w/ buildkitd to Tekton w/ buildkit-daemonless

Update tekton by StrongMonkey · Pull Request #31…

github.com

1 4 21

Lukas Heinrich @lukasheinrich_ · 6h

we briefly talked about this before but would it be possible to do the same thing to the containerd daemon ? iirc buildkit-daemonless just starts and tears down the buildkit daemon with the build.

1

Akihiro Suda

@_AkihiroSuda_

Following

Replying to @lukasheinrich_

probably it will be possible in the future

1:18 PM - 3 Jul 2019

1 Like