On the way to safe containers

Linux Security Summit 2016
Toronto, Canada

Stéphane Graber
LXD project leader, Canonical Ltd.

stgraber@ubuntu.com @stgraber
https://www.stgraber.org

Tycho Andersen
Software engineer, Canonical Ltd.

tycho.andersen@canonical.com
https://tycho.ws
LXD: the container lighter-visor

What it **IS**

➔ Simple
*Clean command line interface, simple REST API and clear terminology.*

➔ Fast
*No virtualization overhead so as fast as bare metal.*

➔ Secure
*Safe by default. Combines all available kernel security features.*

➔ Scalable
*From a single container on a developer’s laptop to thousands of containers per host in a datacenter.*
LXD: the container lighter-visor

What it IS

nova-lxd

command line tool

your own client/script?

LXD REST API

LXD

LXC

Linux kernel

Host A

LXD

LXC

Linux kernel

Host B

LXD

LXC

Linux kernel

Host C

LXD

LXC

Linux kernel

Host D

LXD

LXC

Linux kernel

Host ...

Host ...
LXD: the container lighter-visor

What it **ISN’T**

➔ Another virtualization technology

*LXD tries to offer as similar a user experience as that of a virtual machine but it doesn’t itself virtualize anything, you always get access to the real hardware and the real native performance.*

➔ A fork of LXC

*LXD uses LXC’s API to manage the containers behind the scene.*

➔ Another application container manager

*LXD only cares about full system containers and doesn’t care about what runs inside the container.*
LXD: the container lighter-visor

Security

- Namespaces
- LSMs
- Capabilities
- CGroups
LXD: the container lighter-visor

Resource limits

- CPU
- Memory
- Disk
- Network
- Kernel resources
Safe containers on Linux

Shared kernel resources

➔ Inotify handles
➔ Network tables
➔ PTS devices
➔ Ulimits
Container checkpoint/restore

Some oddities

sysctl writing is strange:

➔ netns sysctls change values for namespace that open()s

➔ IPC/UTS namespace changes values for namespace that write()s

◆ But nobody can open() them besides real root
  ● But you can set these values with sethostname() anyway
Container checkpoint/restore

More TODOs

→ PTRACE_O_SUSPEND_SECCOMP for LSMs? Capabilities?
  ◆ May need to mount("proc") worst case
  ◆ create/connect unix socket
  ◆ “This feature gives me the creeps”

→ Checkpoint of nested namespaces (docker, vsftpd)
Demo time!
LXD: the container hypervisor

Let’s recap

➔ Unprivileged containers are safe by design
➔ LSMs and other kernel features can be used as an additional safety net
➔ It is still much easier to DoS the kernel than we’d like
➔ Lots of requests for additional unprivileged interfaces, some are reasonable, some not so much
➔ Checkpoint/restore is hard
Questions?

LXD stickers are available at the front!

Stéphane Graber
LXD project leader, Canonical Ltd.
stgraber@ubuntu.com @stgraber
https://www.stgraber.org

Tycho Andersen
Software engineer, Canonical Ltd.
tycho.andersen@canonical.com
https://tycho.ws

https://linuxcontainers.org/lxd
https://github.com/lxc/lxd