LXCFS: Not just for LXC anymore

Serge Hallyn

LXC project

August 24, 2016

About me

- 2003 - bsdjail
- 2005 - containers
- 2010 - lxc
- 2013 - unprivileged containers
  - User namespaces
  - Network
  - Cgroup manipulation
- 2014 - lxcfs

Cgmanager

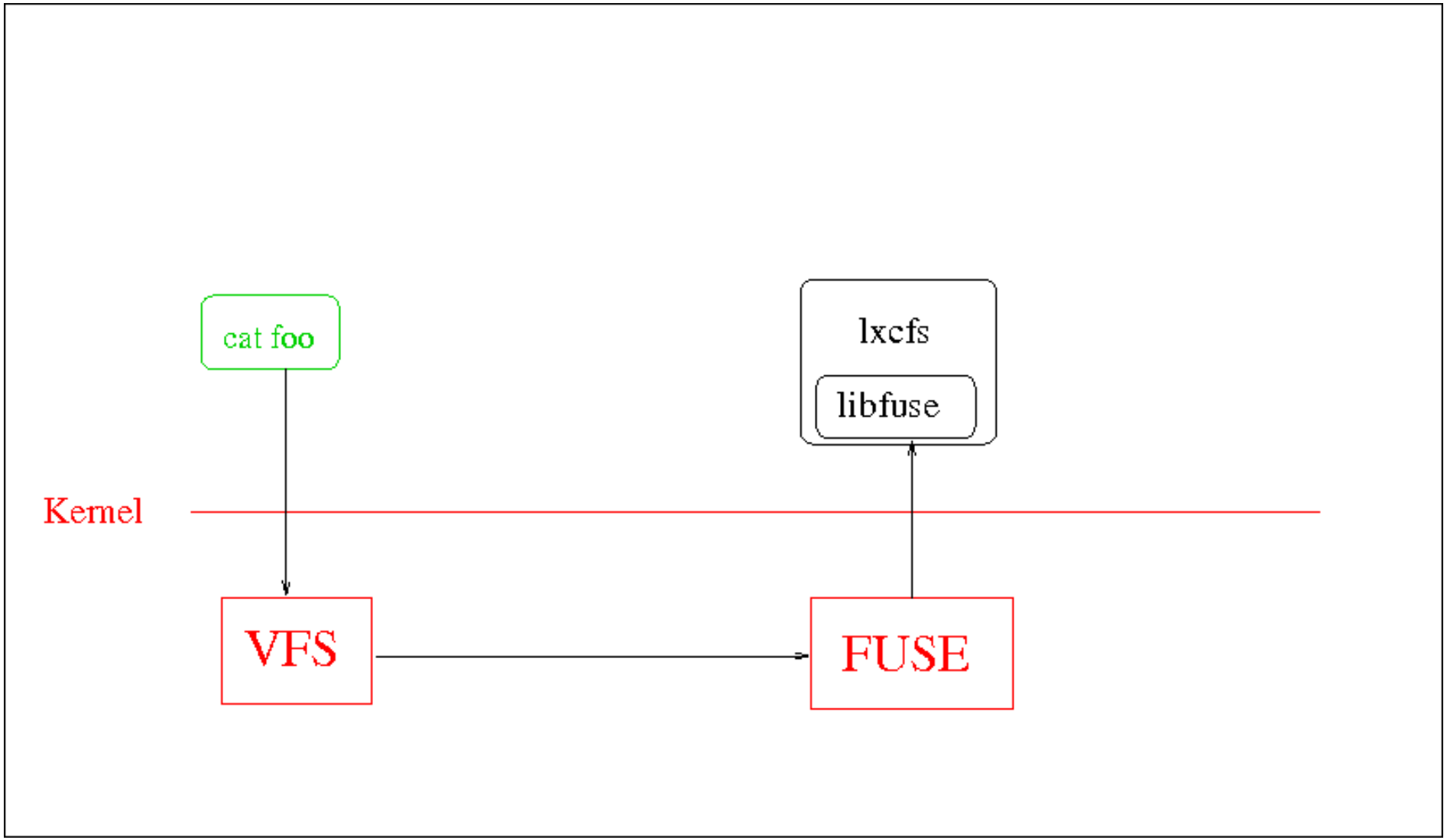
- Central cgroup manager
- Goals
  - Simplify container manager code
    (Ignore questions of mountpoints)
  - Delegate cgroups to users
  - Safely support unprivileged, nested containers
- DBus interface
  - Simplify integration
  - Built on libnih

Cgmanager

- Central cgroup manager
- Goals
  - Simplify container manager code
    (Ignore questions of mountpoints)
  - Delegate cgroups to users
  - Safely support unprivileged, nested containers
- DBus interface
  - Simplify integration
  - Built on libnih
- Systemd in containers
  - Systemd wants to believe it owns cgroups
  - Requires cgroupfs interface

Enter lxcfs

FUSE filesystem



Enter lxcfs

- Cgroupfs virtualization
  - Over cgmanager DBus interface
    ... Over cgroupfs virtual fs
    ... over cgroup kernel feature
  - Worked - but performance cost became high
    ... Especially with systemd

Enter lxcfs

- Cgroupfs virtualization
  - Over cgmanager DBus interface
    ... Over cgroupfs virtual fs
    ... over cgroup kernel feature
  - Worked - but performance cost became high
    ... Especially with systemd
  - Drop cgmanager, use own native cgroupfs mountsk

Enter lxcfs

- Cgroupfs virtualization
  - Over cgmanager DBus interface
    . . . Over cgroupfs virtual fs
    . . . over cgroup kernel feature
  - Worked - but performance cost became high
    . . . Especially with systemd
  - Drop cgmanager, use own native cgroupfs mountsk
  - Finally obsolete - cgroup namespaces

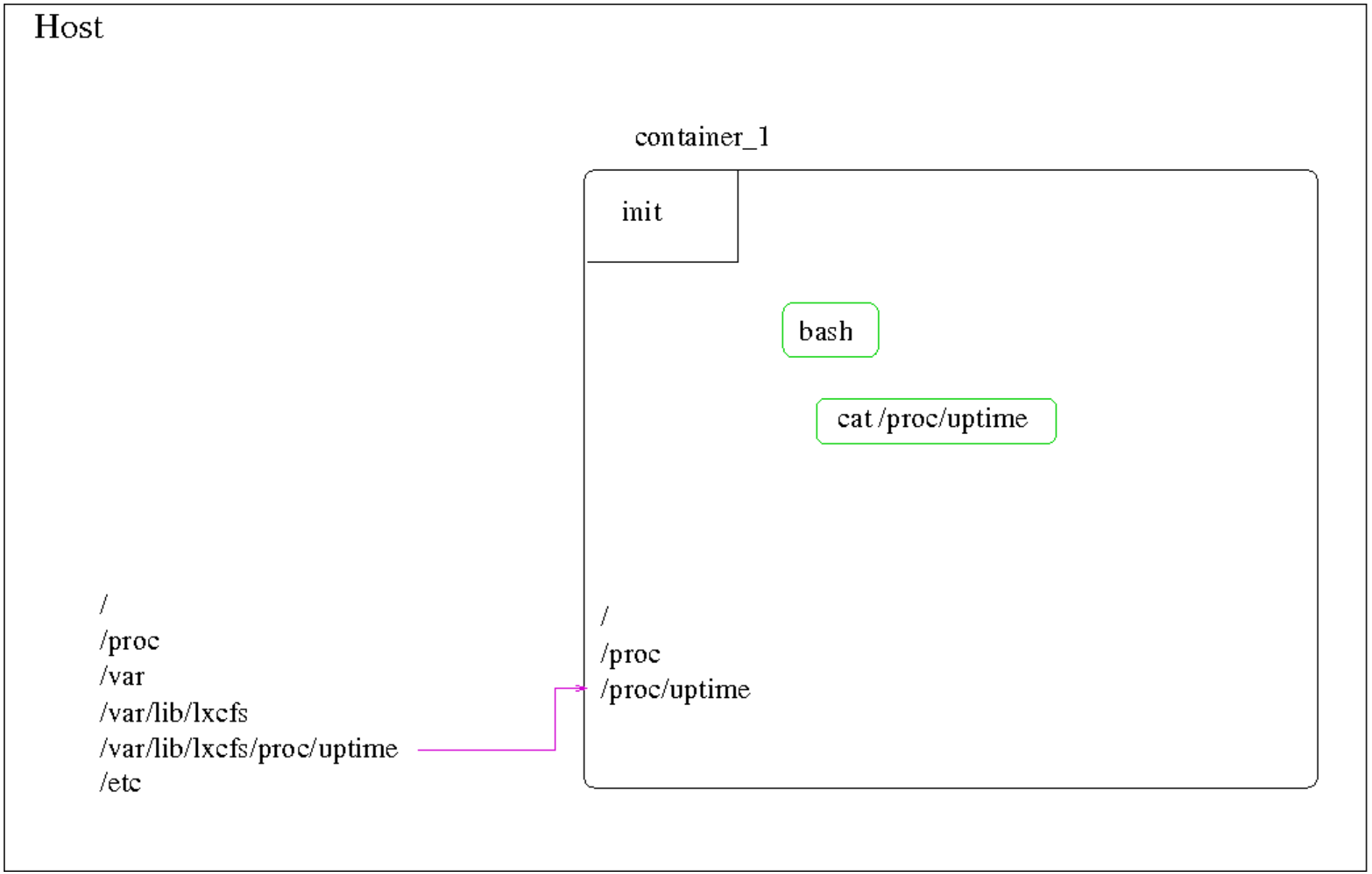
Enter lxcfs

- Procfs virtualization
  - Host resources >> container's
  - Some userspace tries to be civil
  - `top`, `ps` ... look at `/proc`
  - Show only available memory, cpus
  - Show actual container uptime

Proc virtualization options

- Kernel /proc virtualization
  - Deemed unacceptable for years
  - Muddled by lack of "container" in kernel
  - Might be acceptable as a **new** procfs
- Teach userspace
- Libresource
  - Some community interest
  - Did not gain traction
  - High bar to usefulness - need to
    - design a useful API
    - adapt existing tools (procps, top, etc)
- Fuse

Basic design



LXCFS

Supported files:

- cpuinfo
- meminfo
- stat
- uptime
- diskstats
- swaps

Other FUSE based proc virtualization

- original lxc approach (dlezcano)
  - since 2008
  - no longer compiles
  - never really tested/supported
- libvirt
  - since 2012
  - not a standalone project
  - only supports meminfo
- cgroupfs
  - written in go
  - standalone, works in docker
  - supports cpuinfo, diskstats, meminfo, stat

Complications

- Maintaining lxcfs container mounts across security upgrades
  - Do not restart on upgrades
  - Dlopen private library on each operation
  - Reload under lock after SIGUSR1 handler
- Private cgroup mounts
  - Don't confuse docker, libvirt with our mounts
  - Don't pin any host mounts in our namespace
  - Slightly different solution from cgmanager:
    - Open sparse namespace
    - Keep open fd for each mounted controller dir.

Requested features

- support `/sys/devices/system/cpu`
- loadavg
  - Problem - seems to require polling and tracking data
  - Could benefit from new kernel support

Bugs

- ram reported incorrect
- swapfree incorrect
- ps ux returns no btime in `/proc/stat`

How to use

- Run lxcfs: `lxcfs /var/lib/lxcfs`
- Mount lxcfs into containers at container startup
  - lxd - automatic
  - lxc - `/usr/share/lxc/config/common.conf.d/00-lxcfs.conf`
  - docker - map files with -v (requires very recent patch)

```
docker run \
        -v /var/lib/lxcfs/proc/cpuinfo:/proc/cpuinfo \
        -v /var/lib/lxcfs/proc/diskstats:/proc/diskstats \
        -v /var/lib/lxcfs/proc/meminfo:/proc/meminfo \
        -v /var/lib/lxcfs/proc/stat:/proc/stat \
        -v /var/lib/lxcfs/proc/swaps:/proc/swaps \
        -v /var/lib/lxcfs/proc/uptime:/proc/uptime \
        -it ubuntu bash
```

Questions/Comments?

- http://linuxcontainers.org
- http://github.com/lxc/lxcfs
- lxc-{users,devel}@lists.linuxcontainers.org
- serge@hallyn.com