APP SEPARATION NO NEED CONTAINERS

VON

DOCKER, LXC, LXD, SYSTEMD-NSPAWN, CHROOT, OVERLAYFS UND FIREJAIL

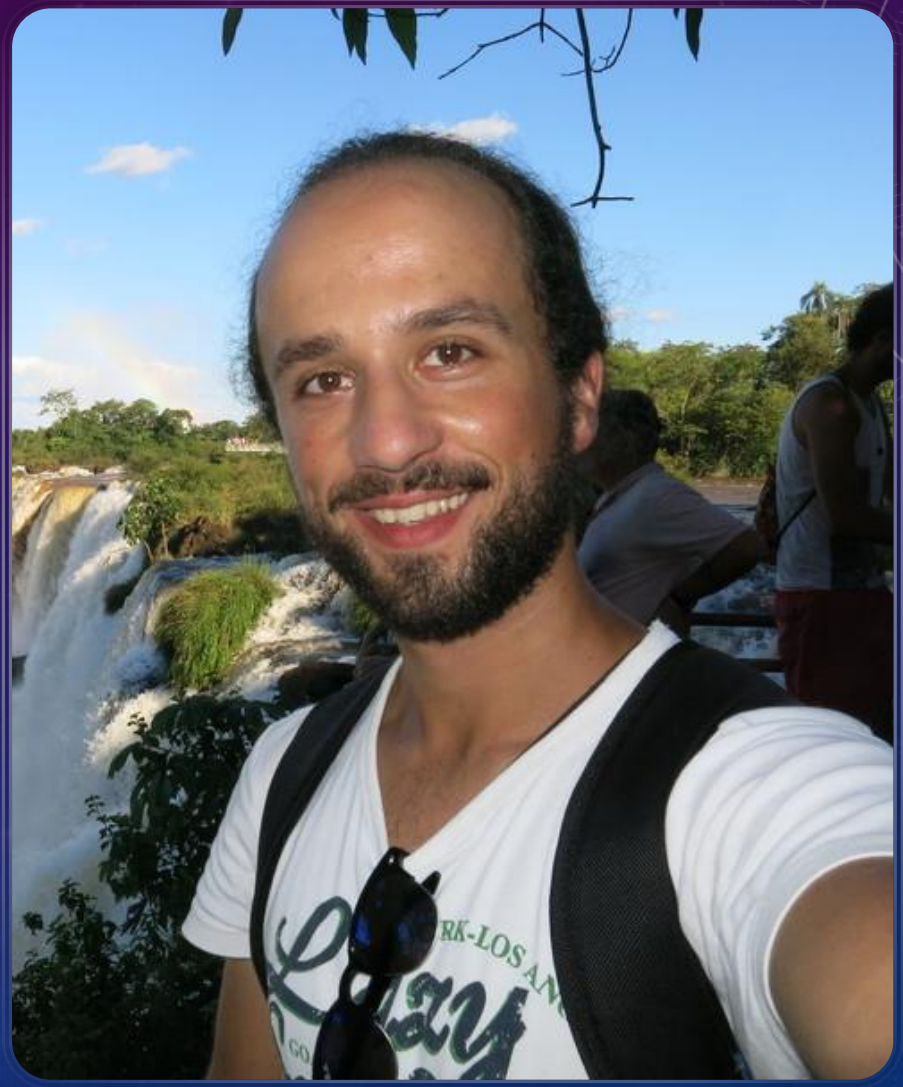
Felix Bauer (felix@ai4me.de)

SICHERHEITSHINWEIS

- Bitte die Rückenlehne gerade stellen und den Anweisungen des Sicherheitspersonals folge leisten

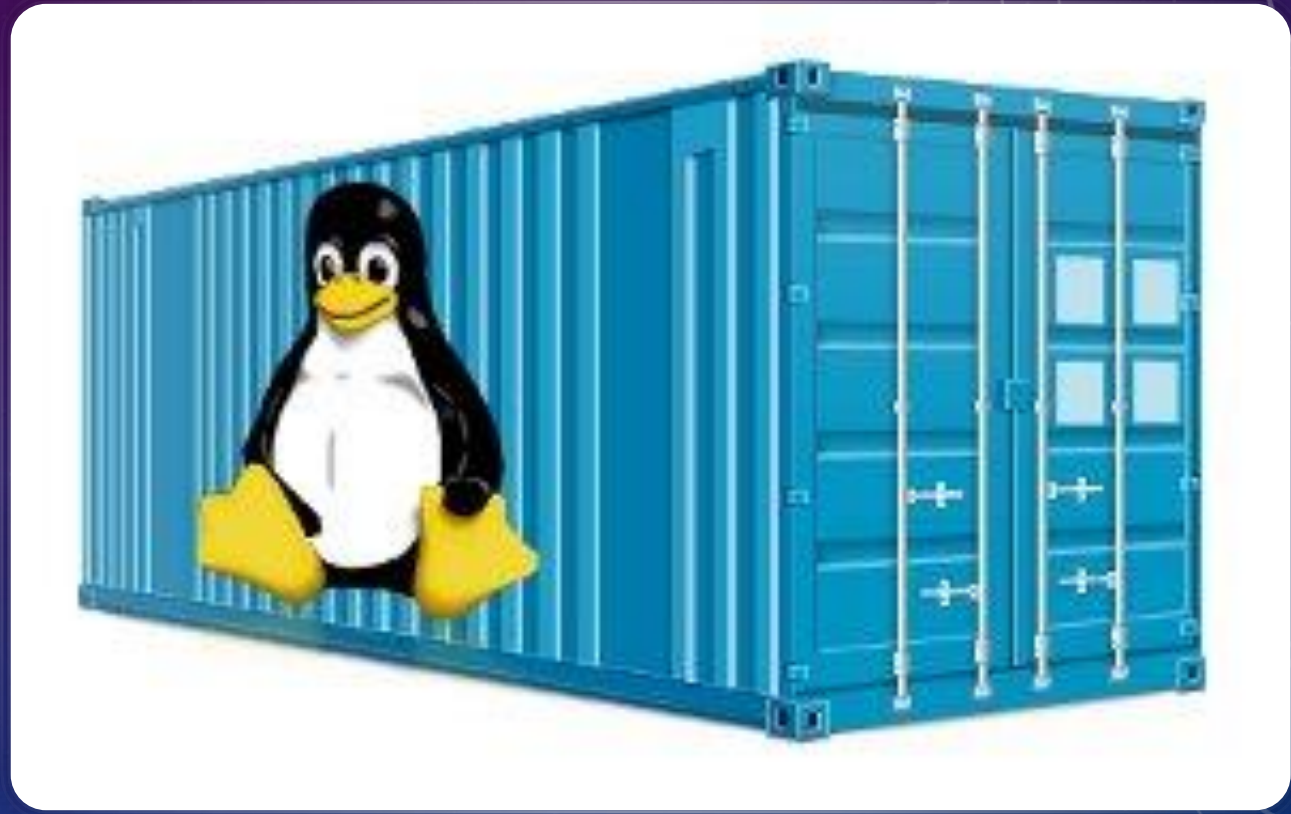
FELIX BAUER

- IT Security Consultant und Engineer bei science + computing ag an Atos Company
- Ethical Hacker
- Fablab Neckarälbler



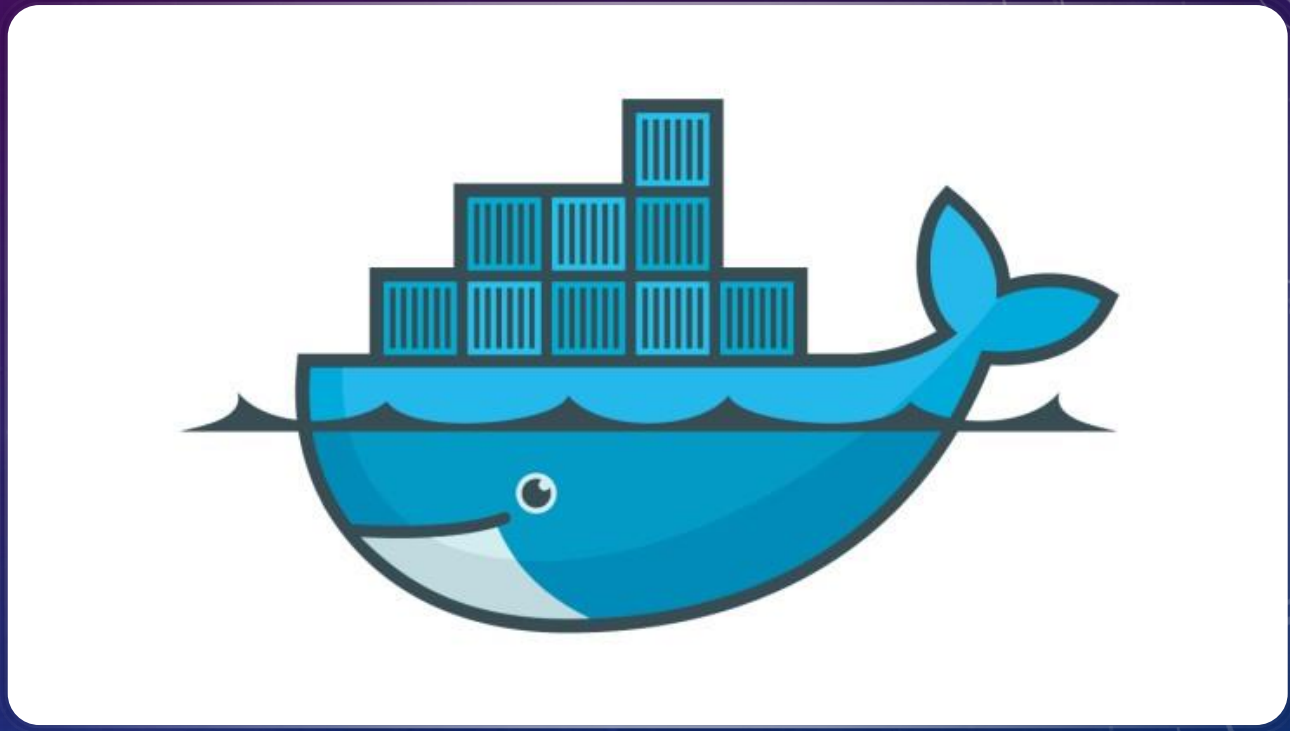
WAS IST EIN CONTAINER?

- VM ohne virtuelle Hardware
- Separierungsmechanismus
- Sicherheitsmechanismus
- `pacman -S` auf Ubuntu
- Anleitung zum unsauberen Arbeiten
- Komplexität reduzieren
- Black Box Computing



WAS IST DOCKER?

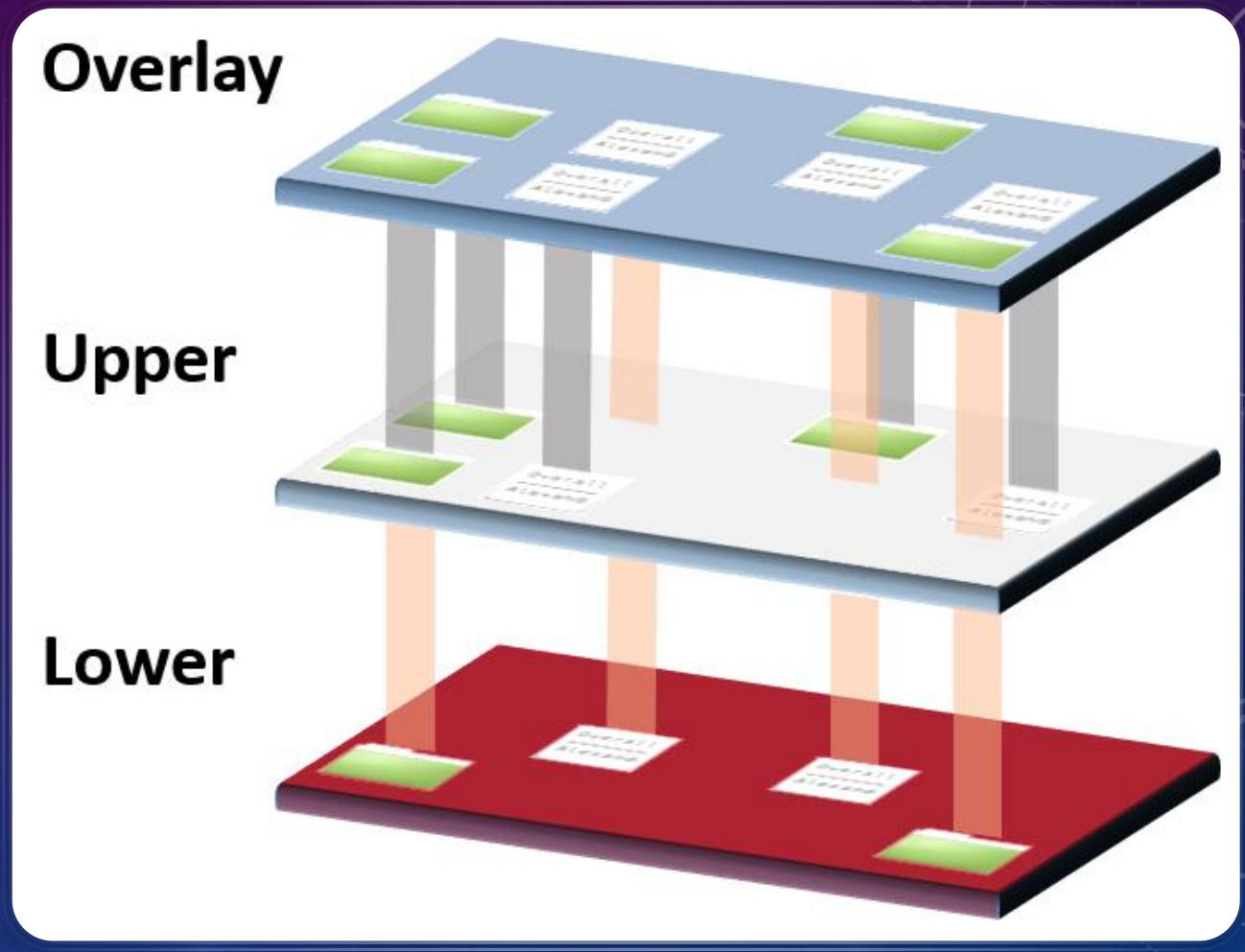
- Container Engine
- Image Repository
- Layer



```
# mount -t overlay overlay -o lowerdir=/lower,upperdir=/upper,workdir=/work /merged
```

UND WAS IST EIN OVERLAYFS?

- Dateisystem (Union Filesystem)
- Verzeichnisse werden übereinander gelegt
- Lower, Upper, Workdir
- Lower ist read-only
- Upper ist schreibbar



EIN PROZESS WIRD EINGESPERRT

- Ein Prozess kann nicht auf Verzeichnisse außerhalb zugreifen
- Dateisystemseparierung
- Das kann `CHROOT` auch

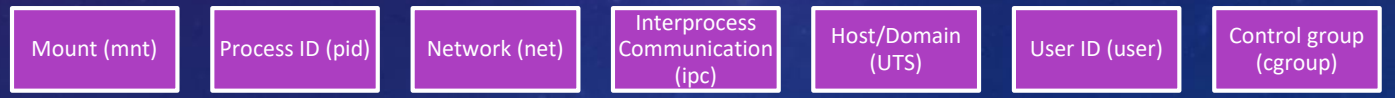
CHANGE ROOT (CHROOT)

- Prozess und Kindprozesse können nicht auf Verzeichnisse außerhalb zugreifen
- `Root` im `chroot` kann leicht ausbrechen
- Nicht verschachtelbar

- https://asciinema.org/a/4Cr7nrhyNS9KMr2dfkkdX114q

NAMESPACES UND CGROUPS

- Resourcen Verwaltung im Linux Kernel
- „Views“ für Prozesse und Prozessgruppen



BEISPIEL NETWORK NAMESPACE

- Virtualisieren den Network Stack
- Enthält am Anfang nur ein loopback Interface
- Netzwerkkarten können nur jeweils in einem Netzwerknamespace sein
- Jeder hat seine eigenen IP-Adressen, Routing Tabelle, Sockets, Verbindungen, Firewall …

DOCKER

- Fast wie eine VM
- https://asciinema.org/a/roCDxNFp45QfYofIIYGz5rM4x

FIREJAIL

- Einzelne Anwendungen kapseln
- https://asciinema.org/a/443QdfHIbxbrokg2xRA7dJJFZ

```
include /etc/firejail/wget.local
include /etc/firejail/globals.local

blacklist /tmp/.X11-unix

include /etc/firejail/disable-common.inc
include /etc/firejail/disable-passwdmgr.inc
include /etc/firejail/disable-programs.inc

include /etc/firejail/whitelist-var-common.inc

caps.drop all
netfilter
no3d
nodvd
nogroups
nonewprivs
noroot
nosound
notv
novideo
protocol unix,inet,inet6
#seccomp
#seccomp.drop exec,execve
# strace -qcf ./service 2>&1 | cut -c 52- | tr '\n'
',
seccomp.keep
read,write,open,close,stat,lseek,mmap,mprotect,brk,rt_
sigaction,rt_sigprocmask,ioctl,readv,select,getpid,soc
ket,connect,sendto,bind,setsockopt,getsockopt,clone,ex
ecve,uname,fcntl,getuid,getgid,geteuid,getegid,prctl,a
rch_prctl,futex,set_tid_address,eventfd2,select,ioctl,
sendto,open,read,close,stat,lseek,mmap,mprotect,brk,rt
_sigaction,rt_sigprocmask,readv,getpid,socket,connect,
bind,setsockopt,getsockopt,clone,execve,uname,fcntl,ge
tuid,getgid,geteuid,getegid,prctl,arch_prctl,set_tid_a
ddress,eventfd2
shell none

private-dev

noexec ${HOME}
noexec /tmp
```

SYSTEMD-NSPAWN

- Verzeichnis booten
- https://asciinema.org/a/rnWyXkaQN8YhS2XNww0xVpArt

```
#!/usr/bin/env bash
#
# Felix Bauer
# cc by sa 4.0
#

name=$(pwgen)
root=/

# test run as root
# test $name must not exist

tmpupper=$(mktemp -p . -d --suffix=$name.tmpupper)
chown $SUDO_USER $tmpupper
chmod 755 $tmpupper
tmpwork=$(mktemp -p . -d --suffix=$name.tmpwork)
chown $SUDO_USER $tmpwork
chmod 755 $tmpwork

mkdir $name
chown $SUDO_USER $name

mount -t overlay overlay -o \
lowerdir=$root,upperdir=$tmpupper,workdir=$tmpwork $name

if [ ! $? -eq 0 ]
then
  echo "mount didn't work"
  exit 1
else
  echo $tmpupper > $name.info
  echo $tmpwork >> $name.info

  # no root in container test
  #systemd-nspawn -D $name --setenv=DISPLAY=:0 \
--setenv=XAUTHORITY=~/.Xauthority \
--bind-ro=$XAUTHORITY:/XAUTHORITY --bind=/tmp/.X11-unix \
-u $SUDO_USER

  # with root
  systemd-nspawn -D $name --bind=/tmp/.X11-unix # -n"
fi

machinectl stop $name
umount $name
rm -rf $(cat $name.info) $name $name.info
```

FAZIT

- Es gibt verschiedene Mechanismen um „least privilege“ umzusetzen
- Wrapper und Anwendungen mit nativer Untersützung

- `setuid, chroot, namespaces, cgroups, seccomp filter …`

VIELEN DANK

- felix@ai4me.de