



Fakultät Informatik

Softwaretechnik und Medieninformatik

Bachelorthesis

Evaluierung verschiedener Container-Technologien

Corvin Schapöhler

751301

Semester 2018

Firma: NovaTec GmbH

Betreuer: Dipl.-Ing. Matthias Haeussler

Erstprüfer: Prof. Dr.-Ing. Dipl.-Inform. Kai Warendorf

Zweitprüfer: Prof. Dr. Dipl.-Inform. Dominik Schoop

Ehrenwörtliche Erklärung

Hiermit versichere ich, Corvin Schapöhler, dass ich die vorliegende Bachelorarbeit mit dem Titel „Evaluierung verschiedener Container-Technologien“ selbständig und ohne fremde Hilfe verfasst und keine anderen als die angegebene Literatur und Hilfsmittel verwendet habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen wurden, sind in jedem Fall unter Angabe der Quelle kenntlich gemacht. Die Arbeit ist noch nicht in gleicher oder anderer Form als Prüfungsleistung vorgelegt worden.

Stuttgart, 28. März 2018

Ort, Datum

Corvin Schapöhler

Kurzfassung

Stichwörter: *Container, Docker, Cloud Native, OCI, Linux, rkt, Evaluation*

Abstract

Keywords: *Container, Docker, Cloud Native, OCI, Linux, rkt, Evaluation*

Inhaltsverzeichnis

Ehrenwörtliche Erklärung	I
Kurzfassung	II
Abstract	II
1 Einleitung	1
1.1 Motivation	1
1.2 Aufbau der Arbeit	1
2 Grundlagen	3
2.1 Standards	4
2.2 Funktionsweise	6
2.3 Eigene Implementierung	11
Abbildungsverzeichnis	A
Tabellenverzeichnis	B
Listings	C
Glossar	D
Akronyme	E
Literatur	F

1 Einleitung

1.1 Motivation

Die Welt wird immer stärker vernetzt. Durch den Drang, Anwendungen für viele Nutzer zugänglich zu machen besteht der Bedarf an Cloud-Diensten wie Amazon Web Services. Eine dabei immer wieder auftretende Schwierigkeit ist es, die Skalierbarkeit der Services zu gewährleisten. Selbst wenn viele Nutzer gleichzeitig auf einen Service zugreifen, darf dieser nicht unter der Last zusammenbrechen.

Bis vor einigen Jahren wurde diese Skalierbarkeit durch Virtuelle Maschinen (VMs) gewährleistet. Doch neben großem Konfigurationsaufwand haben VMs auch einen großen Footprint und sind für viele Anwendungen zu ineffizient. Eine Lösung für dieses Problem stellen Container dar.

Diese Arbeit gibt einen Einblick in das Thema Container-Virtualisierung und beantwortet die Fragen, wie sich Docker als führende Technologie durchsetzen konnte, wie sich andere Technologien im Vergleich zu Docker schlagen und was die Zukunft in Form von Serverless-Technologien mit Bezug zu Containern bereithält.

1.2 Aufbau der Arbeit

Zu Beginn der Arbeit werden benötigte Grundlagen der Technologie erläutert. Dabei werden bestehende Container-Standards betrachtet und alle benötigten Kernel-Funktionen erklärt, die in Container-Runtimes Verwendung finden. Um einen besseren Einblick in die Technologie zu geben wird gezeigt, wie man mit Bash-Befehlen ohne Container-Runtime einen Prozess von einem Host-OS

trennt. Dabei wird darauf eingegangen, wie eine eigene Dateihierarchie isoliert werden kann, wie Namespaces dabei helfen Funktionen des Linux-Kernels zu virtualisieren und wie der isolierte Prozess sicherer ausgeführt werden kann.

Im Anschluss wird die Frage beantwortet, wie Docker am verbreitetste Container-Engine wurde. Dazu wird die Geschichte der Technologie näher betrachtet.

Geschichte,
wie wurde
Docker zu
Docker

Vergleich Do-
cker vs the
World

Aktuelle Pro-
bleme, Or-
chestrierung

Serverless,
wenn zeit
reicht

2 Grundlagen

Container werden häufig als leichtgewichtige VMs beschrieben. Dies ist allerdings nicht richtig. Wie in Abbildung 1 zu erkennen, virtualisieren Container kein vollständiges Betriebssystem (*Operating System*) (OS), sondern lediglich das benötigte Dateisystem. Dabei wird der Kernel des Hosts nicht virtualisiert, sondern mitverwendet. Dies macht Container deutlich leichtgewichtiger als VMs, isoliert allerdings weniger umfangreich als diese.

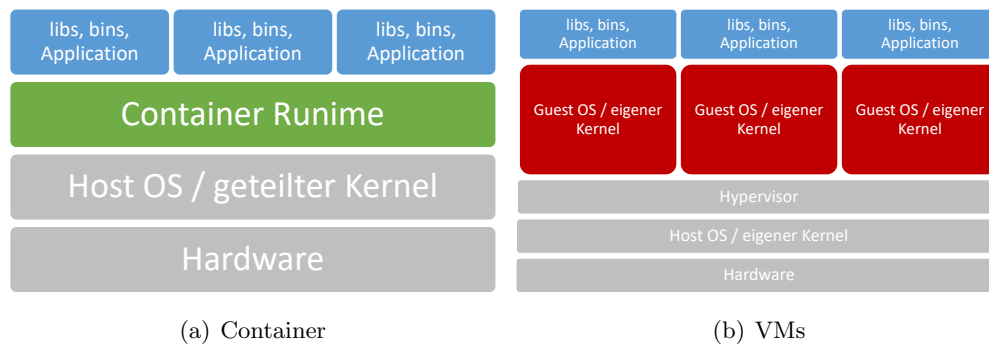


Abbildung 1: Container Isolation im Vergleich zu VMs

Dieses Kapitel behandelt alle benötigten Grundlagen, die zur Isolation eines Prozesses benötigt werden. Es werden vorhandene Standards wie die Open Container Initiative (OCI) und benötigte Systemcalls wie Change Root (chroot) näher erläutert. Zudem wird beschrieben, wie die Isolation, die Container bieten, durch Systemmittel des Linux-Kernels selber erreicht werden kann.

2.1 Standards

Durch die immer größere Verwendung von Containern und die Verbreitung verschiedener Container-Runtimes ist die Standardisierung eine wichtige Aufgabe. Folgend werden Standardisierungsprojekte aufgezählt, die bestehenden Spezifikationen erläutert und aktuelle Aufgaben der Projekte näher betrachtet.

2.1.1 App Container

App Container (appc) ist ein Standard, der viele Aspekte innerhalb der Container-Landschaft behandelt. Dabei lag die Hauptaufgabe darin, eine Laufzeitumgebung wie auch das Image-Format und die Verbreitung von Images zu spezifizieren. Seit 2016 wird das Projekt nicht mehr aktiv weiterentwickelt, da mit der Gründung der OCI ein größeres Standardisierungsprojekt entstand. Bestandteile der appc wurden von der OCI übernommen und dienen als Vorlage für die Spezifikation dieser.

2.1.2 Open Container Initiative

Die OCI ist eine Initiative, die seit 2015 unter der Linux Foundation agiert. Das Ziel der OCI ist es, einen offenen Standard für Container zu schaffen, so dass die Wahl der Container-Laufzeitumgebung nicht mehr zu Inkompatibilität führt. Dabei liegt der Fokus auf eine einfache, schlanke Implementierung (Open Container Initiative, 2018).

Die OCI arbeitet aktuell an zwei Spezifikationen. Die runtime-spec standardisiert die Laufzeitumgebung von Containern. Dabei wird festgelegt, welche Konfiguration, Prinzipien und Schnittstellen Laufzeitumgebungen stellen müssen. Um die Umsetzung der runtime-spec zu fördern, stellt die OCI eine beispielhafte Implementierung durch runC. Das zweite Projekt der OCI ist die image-spec.

Dieses versucht einen Standard für Images zu definieren. Dabei plant die OCI nicht, vorhandene Image-Formate zu ersetzen, sondern auf diesen aufzubauen und sie zu erweitern (Open Container Initiative, 2018).

Wie in Tabelle 1 zu sehen, wurden einige Konzepte des appc-Projekts in die OCI übernommen. Vor allem die Image-Spezifikation wurde durch die Mitarbeit ehemaliger appc-Maintainer gefördert. Allerdings sind einige Projekte noch nicht übernommen worden. So gibt es keine OCI Spezifikation für die Verbreitung von Images, eines der meistgenutzten Features verschiedener Container-Runtimes. Um die Weiterentwicklung an solchen Projekten zu fördern wurden einige in die Cloud Native Computing Foundation (CNCF) übernommen (Polvi, 2015).

	Standard		Container Runtime	
	OCI	appc	Docker	rkt
Container Image	✗	✓	OCI image-spec	appc Image Format
Image Verbreitung	✗	✓	Docker Registry	appc Discovery Spec
Lokales Speicherformat	✓	✗	keine Spezifikation	keine Spezifikation
Runtime	✓	✓	runC	appc runtime Spec

Tabelle 1: Standards OCI und AppC im Vergleich (Polvi, 2015)

2.1.3 Cloud Native Computing Foundation

Die CNCF beschäftigt sich im Gegensatz zur OCI nicht nur mit Containern, sondern der kompletten Cloud-Native-Landschaft (CNCF, 2018). Projekte wie Kubernetes (K8) und Prometheus werden durch die CNCF weiterentwickelt und publiziert. Da der Cloud-Native Entwicklungsprozess von Containern getragen wird, spielen Technologien wie containerd und rkt eine entscheidende Rolle für die CNCF. Neben Container-Runtimes beinhaltet die CNCF auch Projekte zur Orchestrierung von Containern, Logging und Monitoring dieser, wie auch Spezifikationen, zum Beispiel die TUF, eine Spezifikation die standardisiert, wie Softwarepakete upgedatet werden sollen (CNCF, 2017).

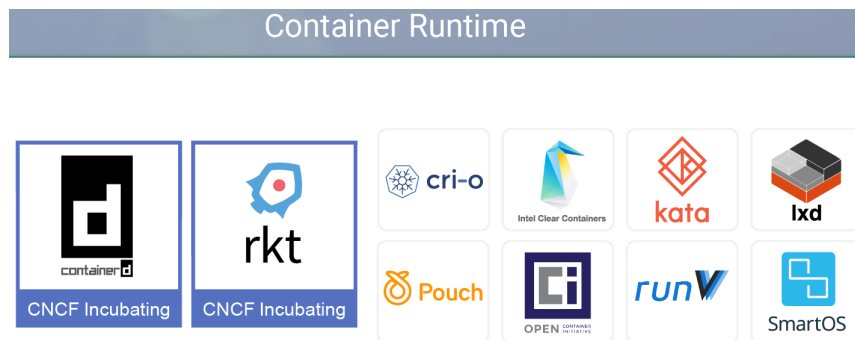


Abbildung 2: CNCF Container Runtime Landschaft (CNCF, 2018)

2.2 Funktionsweise

Container isolieren einzelne Prozesse durch verschiedene Kernel-Technologien, die im Folgenden erklärt werden sollen.

2.2.1 Change Root

Chroot ist ein Unix Systemaufruf, der es erlaubt einen Prozess in einem anderen Wurzelverzeichnis auszuführen (McGrath, 2017). Daraus folgt, dass der Prozess in einer eigenen Verzeichnisstruktur arbeitet und keine Dateien des Host-OS ändern kann. Chroot erlaubt somit die Isolierung des Dateisystems, die Container nutzen.

2.2.2 Control Groups

Control groups (cgroups) dienen dazu, Systemressourcen für einzelne Prozesse zu limitieren. Cgroups sind anders als chroot kein Unix-Feature sondern Teil des Linux-Kernels. Im OS sind cgroups als Dateihierarchie repräsentiert. Das gesamte cgroup-Dateisystem ist unter `/sys/fs/cgroup/` zu finden. Cgroups stellen zur Steuerung verschiedene Controller zur Verfügung.

Controller	Ressource
io	Zugriff und Nutzung von Block Geräten wie Festplatten
memory	Monitoring und Beschränken des Arbeitsspeichers
pids	Limitierung der Anzahl an Unterprozessen
perf_event	Erlaubt Performance Monitoring der Prozesse
rdma	Zugriffe über RDMA limitieren oder sperren
cpu	CPU-Zyklen und maximale CPU-Bandwidth

Tabelle 2: Cgroups-Controller und deren Verwendung (S. H. M. Kerrisk, 2018)

2.2.3 Namespaces

Namespaces abstrahieren einzelne Bereiche des OS. Sie werden genutzt, um globale Ressourcen zu virtualisieren. Ein Namespace kapselt dabei einzelne Ressourcen. Veränderungen an diesen sind für alle Prozesse innerhalb desselben Namespaces sichtbar, allerdings außerhalb dieses unsichtbar (Biederman, 2017).

Namespace	Ressource
Cgroup	Cgroup-Dateisystem
IPC	System V IPC, POSIX Nachrichten
Network	Netzwerk Geräte, Stacks, Ports, ...
Mount	Mount Punkte
PID	Prozess IDs
User	Nutzer und Gruppen IDs
UTS	Hostnamen und Domännennamen

Tabelle 3: Linux Namespaces und verbundene Ressourcen (Biederman, 2017)

2.2.4 Mounting

Durch die Isolation eines Prozesses und der Bedingung, das Container unveränderlich sein sollen, stellt sich die Frage, wie man Containern dynamische

Inhalte aus dem Host-System zur Verfügung stellt. Dies ist vor allem wichtig, wenn bei Veränderung der Umgebung nicht den Container neu gestartet werden soll. Sollte zum Beispiel eine neue Datei durch einen Webserver zur Verfügung gestellt werden, möchte man nicht den Container neu starten. Die Lösung dieses Problems ist der Unix-Systembefehl `mount`.

Mit diesem Befehl wird eine beliebige Dateihierarchie an eine andere Stelle des Dateibaums angeheftet. Durch dieses vorgehen kann man Ordner vom Host-System für das mit `chroot` isolierte Dateisystem des Containers zugänglich machen. Dabei ist zu beachten, dass es sich bei dem gemounteten Ordner nicht um einen symbolischen Link handelt. Diese könnten durch den Aufruf von `chroot` nicht mehr aufgelöst werden.

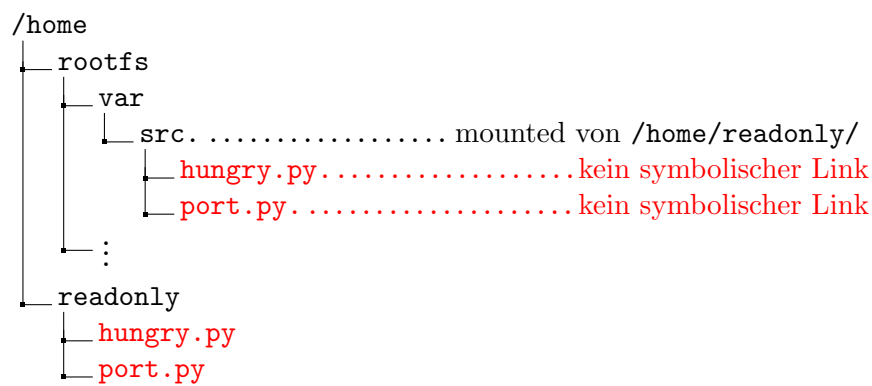


Abbildung 3: Auszug aus Dateisystem mit gemounteten Dateien

2.2.5 Netzwerk

Einen weiteren Aspekt, den Container vom Host-OS isolieren ist das Netzwerk. Dabei kommen virtuelle Ethernet-Adapter zum Einsatz. Diese erlauben es, ein unabhängiges Netzwerk zu erzeugen. Ein Adapter der virtuellen Ethernet-Verbindung wird dabei der Process Identifier (PID) des Containers zugewiesen, das andere dem Host. Zusätzlich wird der Network-Namespace genutzt um eine vollständige Isolation des Netzwerks zu erhalten.



Abbildung 4: Netzwerkseparation zwischen Host und Container

2.2.6 Sicherheit

"Docker is about running random code downloaded from the Internet and running it as root"

—Dan Walsh (Red Hat)

Container haben ein großes Problem. Alle genannten Kernel-Features müssen als Nutzer `root` ausgeführt werden. Dadurch haben die gestarteten Prozesse häufig Berechtigungen, die es erlauben würden, aus der Isolierung des Containers auszubrechen. Um dies zu verhindern, können verschiedene Sicherheitskonzepte verwendet werden.

Das leichteste dieser Konzepte sind Capabilities. Jedem Prozess, sowie jeder Datei kann eine Liste an Capabilities zugeordnet oder genommen werden. Dabei können einzelnen Dateien beispielsweise die Rechte genommen werden, auf Port 80 zu hören. Auch viele Systemaufrufe können über Capabilities gewährt oder verwehrt werden. Ein anderes Konzept ist die Implementation eines *Mandatory Access Control*-Systems wie SELinux oder AppArmor. Diese Implementationen sind granularer als Capabilities, allerdings mit einem höheren Konfigurationsaufwand verbunden.

Capability	Systemaufruf	Erklärung
CAP_CHOWN	chown	Ändern des Owners einer Datei
CAP_KILL	kill	Beenden eines Prozesses
CAP_CHROOT	chroot	Wechseln des Root Directories
CAP_NET_BIND_SERVICE		Binden eines Prozesses auf Ports <1024
CAP_SYS_TIME	stime, settimeofday	Setzen der Systemzeit
CAP_SYS_ADMIN	mount, umount, setns, pipe, syslog, ...	Alles, was in keine andere Kategorie passt

Tabelle 4: Einige Capabilities (M. Kerrisk, 2018)

2.2.7 Container unter Windows

Viele Kernel-Features des Linux-Kernels erlauben eine Isolation und wurden teilweise spezifisch für diese entwickelt (Biederman, 2017). Seit 2016 können spezifisch Docker-Container auch unter Windows genutzt werden. Dabei trennt Microsoft Container in zwei verschiedene Isolationen auf.

Windows Server Container 2016 sind ein nativer Windows Ansatz zur Isolation eines Prozesses. Dabei werden, wie bereits unter Linux Systemen, verschiedene Kernel-Technologien verwendet, um einen einzelnen Prozess zu isolieren. Der andere Ansatz sind Hyper-V Container. Diese sind eher VMs als Containers. Dabei ist der größte Unterschied, das Hyper-V Container eine minimale Installation eines Windows Betriebssystems nutzen, um auf diesem einzelne Isolationen zu erstellen, während Windows Server Container ein gemeinsames OS Nutzen, nämlich Windows Server

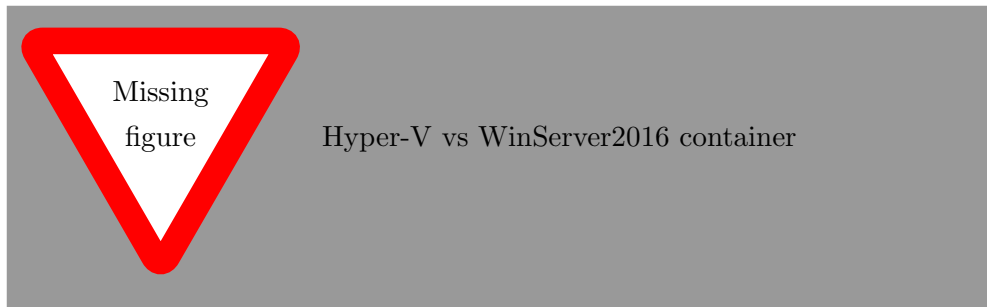


Abbildung 5: Unterschiede der Isolation bei Containern unter Windows

2.3 Eigene Implementierung

Um die in Abschnitt 2.2 erläuterten Kernel-Features näher zu beleuchten wird folgend gezeigt, wie ein Prozess isoliert vom Host-System ausführen kann. Dabei wird darauf eingegangen, wie ein tarball durch das Tool buildroot erstellt werden kann, was am Beispiel der Python-Runtime gezeigt wird. Dieser lässt sich folgend in Container-Runtimes wie rkt importieren. Im Folgenden wird dieser mithilfe der Kernel-Funktionen aus Abschnitt 2.2 isoliert. Das Ergebnis ist ein Dateisystem innerhalb eines Host-OS, indem eine Instanz der Python-Runtime isoliert und ohne Root-Berechtigungen ausgeführt wird.

2.3.1 Erstellen eines tarballs

Bei einem tarball handelt es sich um ein komprimiertes Dateisystem. Dabei wird ein vollwertiges Linux-Dateisystem stark komprimiert, um es leichter zu versenden oder zusichern. Das Erzeugen eines tarballs ist durch das Tool buildroot einfach. Buildroot ist ein Unix-Tool, welches zur Erstellung von minimalistischen Linux-Distributionen für Embedded-Systems entworfen wurde. Es erlaubt aber auch, nur ein Dateisystem zu erzeugen, ohne Kernel oder Init-System. Dies ist entscheidend, da bei Containern der bestehende Kernel des Host-OS mitbenutzt wird (*siehe Abbildung 1*). Das Init-System, welches dazu dient, neue Prozesse zu starten, wird in einem Container ebenfalls nicht benötigt, da diese nur einen Prozess ausführen. Diese Features von buildroot erlauben es, ein Image für Container zu erzeugen.

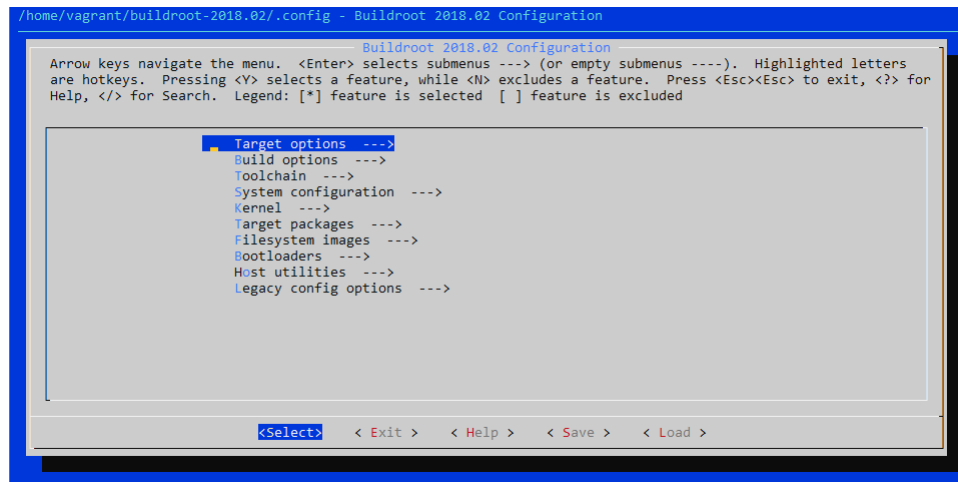


Abbildung 6: Buildroot Menü nach Start des Tools

Zudem erlaubt es buildroot, einzelne Bibliotheken, wie zum Beispiel die Python-Runtime, beim Build-Prozess der Distribution zu integrieren. Nach dem Einstellen der benötigten Bibliotheken und dem deaktivieren der, für Container, unnötigen Features erzeugt Buildroot eine Config-Datei, die alle Änderungen beinhaltet. Durch das Starten des Build-Prozesses mit dem Befehl `make` wird die gewünschte Linux-Distribution erstellt. Am Ende dieses Prozesses liegt im Ordner `/buildroot/out/images/` das gewünschte Dateisystem `rootfs.tar`.


```

/usr/bin/g++ -I/home/vagrant/buildroot-2018.02/output/host/include -O2 -I/home/vagrant/buildroot-2018.02/output/host/inc
lude -DPROGVERSION="1.19" -c -o main.o main.cc
/usr/bin/g++ -L/home/vagrant/buildroot-2018.02/output/host/lib -Wl,-rpath,/home/vagrant/buildroot-2018.02/output/host/li
b -O2 -I/home/vagrant/buildroot-2018.02/output/host/include -o lzip_arg_parser.o file_index.o list.o encoder_base.o enco
der.o fast_encoder.o decoder.o main.o
>>> host-lzip 1.19 Installing to host directory
PATH="/home/vagrant/buildroot-2018.02/output/host/bin:/home/vagrant/buildroot-2018.02/output/host/sbin:/home/vagrant/bin
:/home/vagrant/.local/bin:/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/usr/games:/usr/local/games" PKG_
CONFIG="/home/vagrant/buildroot-2018.02/output/host/bin/pkg-config" PKG_CONFIG_SYSROOT_DIR="/" PKG_CONFIG_ALLOW_SYSTEM_C
FLAGS=1 PKG_CONFIG_ALLOW_SYSTEM_LIBS=1 PKG_CONFIG_LIBDIR="/home/vagrant/buildroot-2018.02/output/host/lib/pkgconfig:/hom
e/vagrant/buildroot-2018.02/output/host/share/pkgconfig" /usr/bin/make -j3 -C /home/vagrant/buildroot-2018.02/output/bui
ld/host-lzip-1.19 install
if [ ! -d "/home/vagrant/buildroot-2018.02/output/host/bin" ] ; then install -d -m 755 "/home/vagrant/buildroot-2018.02/
output/host/bin" ; fi
if [ ! -d "/home/vagrant/buildroot-2018.02/output/host/share/info" ] ; then install -d -m 755 "/home/vagrant/buildroot-2
018.02/output/host/share/info" ; fi
if [ ! -d "/home/vagrant/buildroot-2018.02/output/host/share/man/man1" ] ; then install -d -m 755 "/home/vagrant/buildro
ot-2018.02/output/host/share/man/man1" ; fi
rm -f "/home/vagrant/buildroot-2018.02/output/host/share/info/lzip.info"
install -m 755 ./lzip "/home/vagrant/buildroot-2018.02/output/host/bin/lzip"
rm -f "/home/vagrant/buildroot-2018.02/output/host/share/man/man1/lzip.1"
install -m 644 ./doc/lzip.1 "/home/vagrant/buildroot-2018.02/output/host/share/man/man1/lzip.1"
install -m 644 ./doc/lzip.info "/home/vagrant/buildroot-2018.02/output/host/share/info/lzip.info"
if /bin/sh -c "install-info --version" > /dev/null 2>&1 ; then \
    install-info --info-dir="/home/vagrant/buildroot-2018.02/output/host/share/info" "/home/vagrant/buildroot-2018.0
2/output/host/share/info/lzip.info" ; \
fi
>>> host-gawk 4.1.4 Downloading
--2018-03-28 13:00:26-- http://ftpmirror.gnu.org/gawk/gawk-4.1.4.tar.xz
Resolving ftpmirror.gnu.org (ftpmirror.gnu.org)...

```

Abbildung 7: Ausgabe während des Erstellens eines tarballs mit buildroot

2.3.2 Isolieren der Python-Runtime

In Abschnitt 2.3.1 wurde ein Dateisystem mit der Python-Runtime erstellt. Dieses muss nun isoliert, ein Pythonprogramm in das Dateisystem gemounted und ausgeführt werden.

Der erstellte tarball wird durch folgende Bash-Befehle entpackt.

```

cd /home
mkdir rootfs
cp rootfs.tar rootfs/
cd rootfs
sudo tar xvf rootfs.tar
sudo rm rootfs.tar

```

Listing 1: Entpacken des buildroot tarballs nach /home/rootfs

Um einen Prozess mit dem Wurzelverzeichnis /home/rootfs/ auszuführen, ist lediglich der folgende Aufruf nötig.

Durch diesen wird ein Webserver auf Adresse <http://0.0.0.0:8000> ausgeführt,

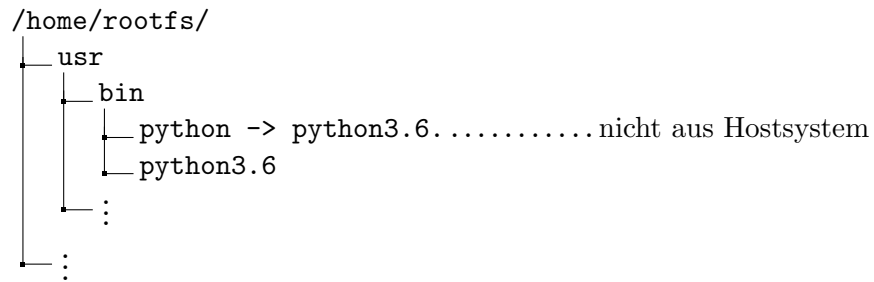


Abbildung 8: Dateibaum nach entpacken der `rootfs.tar`

```
sudo chroot rootfs /usr/bin/python3.6 -m http.server
```

Listing 2: Shell-Command um Webserver mit definierter Wurzel zu starten

der alle ihm zugänglichen Dateien zum Download bereitstellt. Beim Aufrufen dieser Adresse erkennt man, dass der Webserver nur Zugriff auf die in `/home/rootfs/` liegenden Dateien hat.

Directory listing for /

- [bin/](#)
- [dev/](#)
- [etc/](#)
- [lib/](#)
- [lib64/](#)
- [linuxrc/](#)
- [media/](#)
- [mnt/](#)
- [opt/](#)
- [proc/](#)
- [root/](#)
- [run/](#)
- [sbin/](#)
- [sys/](#)
- [tmp/](#)
- [usr/](#)
- [var/](#)

Abbildung 9: Python Webserver mit festgesetztem Root-Verzeichnis

Um weiterhin Zugriff auf dynamische Inhalte aus dem Host-System zu haben, kann man, wie in Abschnitt 2.2.4 erklärt, entsprechende Verzeichnisse in das neue rootfs des Prozesses mounten.

```
nsenter --mount=/proc/<PID isolierter Prozess>/ns/mnt \
  mount --bind -o ro \
    $PWD/readonly \
    $PWD/rootfs/var/src
```

Listing 3: Mounten von Verzeichnis /readonly/ zu /rootfs/var/src/

Durch die Dateitrennung und den Aufruf von `chroot` tritt allerdings ein großes Problem auf. Der Python-Webserver wird mit erhöhten Rechten ausgeführt, da diese für den Aufruf von `chroot` benötigt werden.

```
vagrant@vagrant:/home$ sudo chroot rootfs/ /bin/sh
/ # whoami
root
/ #
```

Abbildung 10: Root-Eskalation durch Aufruf von `sudo chroot`

Dies führt zu vielen Problemen. Ein Prozess, der nur auf diese Weise isoliert wird, könnte beispielsweise Prozesse auf dem Hostsystem mit `kill <pid>` beenden. Die Lösung dieses Problems sind die in Abschnitt 2.2.3 beschriebenen namespaces.

Um alle Prozesse des Hostsystems vor dem Container zu verstecken, muss der PID-Namespace des Container-Prozesses neu gemounted werden.

```
sudo unshare -p --mount-proc=$PWD/rootfs/proc -f chroot
↪ rootfs /bin/sh
```

Listing 4: Remount des PID-Namespaces und Chroot einer Shell

Beim Aufruf von „`ps aux`“ wird nur noch der Prozess `/bin/sh` angezeigt, der die PID 1 bekommen hat. Dieses Vorgehen löst allerdings nicht die Ursache des Problems. Der gestartete Prozess läuft auch weiterhin unter dem Nutzer `root`. Ein auf diese Weise isolierter Prozess, kann zum Beispiel auf Port 80 hören. Um diese Berechtigungen zu entfernen werden die in Abschnitt 2.2.6 angesprochenen Capabilities verwendet.

Durch den folgenden Aufruf hat, die in `rootfs` gestartete `/bin/bash` nicht mehr die Möglichkeit, auf niedrigere Ports, wie Port 80 zu hören.

```
capsh --drop=cap_net_bind_service --chroot=rootfs/ --
```

Listing 5: Entfernen der Capability um auf Port 80 zu hören

Um vollständige Isolation des Containers zu erreichen, müssen Systemressourcen, wie Arbeitsspeicher oder CPU-Zyklen limitiert werden. Dazu dienen die

in Abschnitt 2.2.2 beschriebenen cgroups.

Um eine cgroup zu erstellen, muss ein Ordner unterhalb des Wurzelverzeichnisses erstellt werden. Um einen Prozess einer cgroup zuzuordnen, wird die PID des Prozesses in die Datei `/sys/fs/cgroup/<controller>/<cgroupname>/tasks` geschrieben.

```
mkdir /sys/fs/cgroup/memory/container  
echo $ContainerPID > /sys/fs/cgroup/memory/container/tasks
```

Listing 6: Erzeugen einer memory cgroup namens container

Um festzusetzen, wie viel Arbeitsspeicher der isolierte Prozess nutzen darf, kann man innerhalb der cgroup einzelne Limits festlegen.

```
echo "0" >  
→ /sys/fs/cgroup/memory/container/memory.swappiness  
echo "100000000" >  
→ /sys/fs/cgroup/memory/container/memory.limit_in_bytes
```

Listing 7: Limitieren des Arbeitsspeichers und Memory-Swap deaktivieren

Um zu testen, ob die Zuweisung funktioniert und durch den isolierten Prozess maximal ~100Mb Arbeitsspeicher belegt werden können, kann folgendes Python Programm ausgeführt werden. Abbildung 11 zeigt die Ausgabe des Prozesses, der sich in der cgroup container befindet.

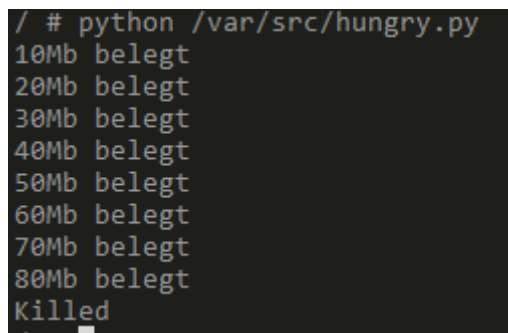
```
#hungry.py - Eating up memory in 10Mb blocks
import time

TEN_MEGABYTE = 10000000

f = open("/dev/urandom", "rb")
data = bytearray()
i = 0

while True:
    data.extend(f.read(TEN_MEGABYTE))
    i += 1
    print("%dMb belegt" % (i*10,))
    time.sleep(1)
```

Listing 8: Python Programm hungry.py um Arbeitsspeicher zu verbrauchen



```
/ # python /var/src/hungry.py
10Mb belegt
20Mb belegt
30Mb belegt
40Mb belegt
50Mb belegt
60Mb belegt
70Mb belegt
80Mb belegt
Killed
```

Abbildung 11: Ausgabe des Pythonprogramms hungry.py

Abbildungsverzeichnis

1	Container Isolation im Vergleich zu VMs	3
2	CNCF Container Runtime Landschaft (CNCF, 2018)	6
3	Auszug aus Dateisystem mit gemounteten Dateien	8
4	Netzwerkseparation zwischen Host und Container	9
5	Unterschiede der Isolation bei Containern unter Windows . . .	11
6	Buildroot Menü nach Start des Tools	12
7	Ausgabe während des Erstellens eines tarballs mit buildroot . .	13
8	Dateibaum nach entpacken der <code>rootfs.tar</code>	14
9	Python Webserver mit festgesetztem Root-Verzeichnis	15
10	Root-Eskalation durch Aufruf von <code>sudo chroot</code>	16
11	Ausgabe des Pythonprogramms <code>hungry.py</code>	18

Tabellenverzeichnis

1	Standards OCI und AppC im Vergleich (Polvi, 2015)	5
2	Cgroups-Controller und deren Verwendung (S. H. M. Kerrisk, 2018)	7
3	Linux Namespaces und verbundene Ressourcen (Biederman, 2017)	7
4	Einige Capabilities (M. Kerrisk, 2018)	10

Listings

1	Entpacken des buildroot tarballs nach /home/rootfs	13
2	Shell-Command um Webserver mit definierter Wurzel zu starten	14
3	Mounten von Verzeichnis /readonly/ zu /rootfs/var/src/ .	15
4	Remount des PID-Namespaces und Chroot einer Shell	16
5	Entfernen der Capability um auf Port 80 zu hören	16
6	Erzeugen einer memory cgroup namens container	17
7	Limitieren des Arbeitsspeichers und Memory-Swap deaktivieren	17
8	Python Programm hungry.py um Arbeitsspeicher zu verbrauchen	18

Glossar

Bash Bash ist eine freie, umfangreiche Shell, die in den meisten Unix-Systemen Standard ist.

Cloud-Native Cloud-Native Anwendungen sind spezifisch für Cloud-Architekturen entwickelt. Sie spalten meistens große Funktionalitäten in kleine Micro-services, die mittels API miteinander kommunizieren und sind somit skalierbar, lose gekoppelt und ausfallsicherer als Full-Client Anwendungen.

Image Ein Container Image ist eine Datei, die spezifiziert, wie ein Container von der Laufzeitumgebung ausgeführt werden soll.

Akronyme

appc App Container.

cgroup control group.

chroot Change Root.

CNCF Cloud Native Computing Foundation.

K8 Kubernetes.

OCI Open Container Initiative.

OS Betriebssystem (*Operating System*).

PID Process Identifier.

VM Virtuelle Machine.

Literatur

- BIEDERMAN, Michael Kerrisk; Eric W., 2017. *namespaces(7) - Linux Manual Page*.
- CHIANG, Eric, 2017. *Containers from Scratch*. Auch verfügbar unter: <https://ericchiang.github.io/post/containers-from-scratch/>.
- CNCF, 2017. *Cloud Native Computing Foundation*. Auch verfügbar unter: <https://www.cncf.io/>.
- CNCF, 2018. *CNCF Cloud Native Interactive Landscape*. Auch verfügbar unter: <https://landscape.cncf.io/>.
- CORE OS, 2017. *rkt 1.29.0 Documentation*. Auch verfügbar unter: <https://coreos.com/rkt/docs/latest/>.
- CREQUY, Simone Gotti; Luca Bruno; Iago López Galeiras; Derek Gonyeo; Alban, 2018. *App Container*. Auch verfügbar unter: <https://github.com/appc>.
- DOCKER INC, 2018. *Docker security*. Auch verfügbar unter: <https://docs.docker.com/engine/security/security/>.
- DOCKER INC., 2018. *Docker Documentation*. Auch verfügbar unter: <https://docs.docker.com/>.
- HARRINGTON, Brian "Readbeard", 2015. *Building minimal Containers: Getting Weird with Containers*. Auch verfügbar unter: https://github.com/brianredbeard/minimal_containers.
- HEO, Tejun, 2015. *Control Group v2*.
- KERRISK, Michael, 2018. *capabilities(7) - Linux manual page*.
- KERRISK, Serge Hallyn; Michael, 2018. *cgroups(7) - Linux Manual Page*.
- KNULST, Cornell, 2016. *Deep dive into Windows Server Containers and Docker*. Auch verfügbar unter: <http://blog.xebia.com/deep-dive-into-windows-server-containers-and-docker-part-1-why-should-we-care/>.

- MAS, Raphaël Hertzog; Roland, 2015. *The Debian Administrator's Handbook, Debian Jessie from Discovery to Mastery*. Freexian. ISBN 9791091414043.
- MATTHIAS, Karl; KANE, Sean P., 2015. *Docker: Up & Running: Shipping Reliable Containers in Production*. O'Reilly Media. ISBN 978-1-491-91757-2.
- MCGRATH, Roland, 2017. *chroot(1) - Linux Manual Page*.
- OPEN CONTAINER INITIATIVE, 2018. *Open Container Initiative*. Auch verfügbar unter: <https://www.opencontainers.org/>.
- PETAZZONI, Jérôme, 2013. *Create Lightweight Containers with Buildroot*. Auch verfügbar unter: <https://blog.docker.com/2013/06/create-light-weight-docker-containers-buildroot/>.
- POLVI, Alex, 2015. *Making Sense of Container Standards and Foundations: OCI, CNCF, appc and rkt*. Auch verfügbar unter: <https://coreos.com/blog/making-sense-of-standards.html>.
- ROUMELIOTIS, Rachel, 2016. The Open Container Essentials Video Collection. In: *The Open Container Essentials Video Collection. O'Reilly Open-Source Conference*. ISBN 9781491968253.
- RUSSINOVICH, Mark, 2015. *Containers in Windows Server, Hyper-V and Azure*. Hrsg. von MECHANIC, Microsoft. Auch verfügbar unter: https://www.youtube.com/watch?v=YoA_MM1GPRc.
- ZAK, Karel, 2015. *mount(8) - Linux Manual Page*.

Todo list

Geschichte, wie wurde Docker zu Docker	2
Vergleich Docker vs the World	2
Aktuelle Probleme, Orchestrierung	2
Serverless, wenn zeit reicht	2
Figure: Netzwerk Graph für Container und Host?	9
Figure: Hyper-V vs WinServer2016 container	11