www.qconferences.com
www.qconbeijing.com
www.qconshanghai.com

QCon

伦敦 | 北京 | 东京 | 纽约 | 圣保罗 | 上海 | 旧金山

London · Beijing · Tokyo · New York · Sao Paulo · Shanghai · San Francisco

QCon全球软件开发大会

International Software Development Conference www.infoq.com/cn

InfoQ ueue

新闻

文章

视频

微信 infoqchina

架构师峰会

@InfoQ

QClub

QCon

架构师

迷你书

infoq.com/cn

@InfoQ infoqchina

软件
正在改变世界！

Docker, DevOps, Security

Chris Swan, CTO

@cpswan

cohesiveFT

Cloud native networking

TL;DR

Dockerfile is awesomely productive

- Great for DevOps

Containers don’t contain

- At least not yet

Images have a manifest problem

- Keep track of your stuff

Why me?

docker

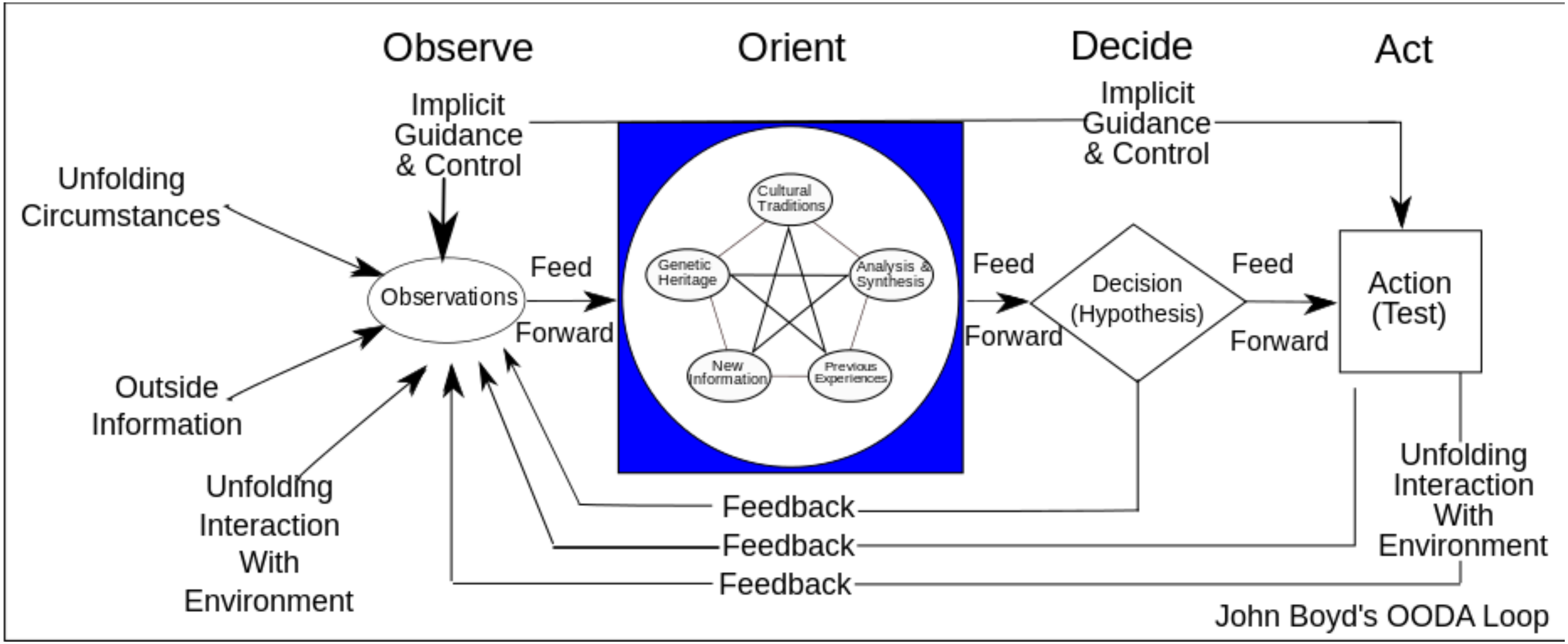
DEV OPS

SECURITY
SECURITY

Let’s start with a demo

What do I mean by ‘DevOps’?

John Boyd's OODA loop



Industrial design maturity - cars

Design for operations

Design for manufacture

Design for purpose

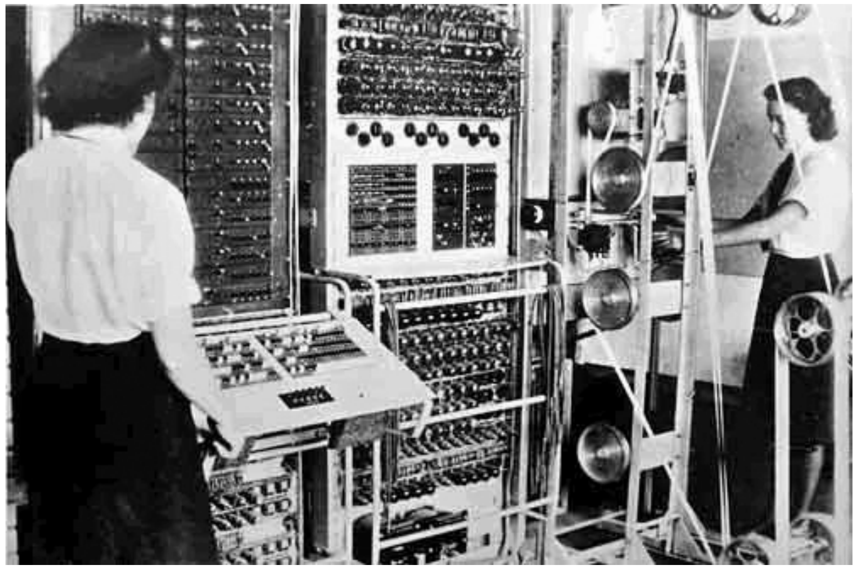
Industrial design maturity - software

Design for operations

↓

DevOps is an artefact

Design for manufacture



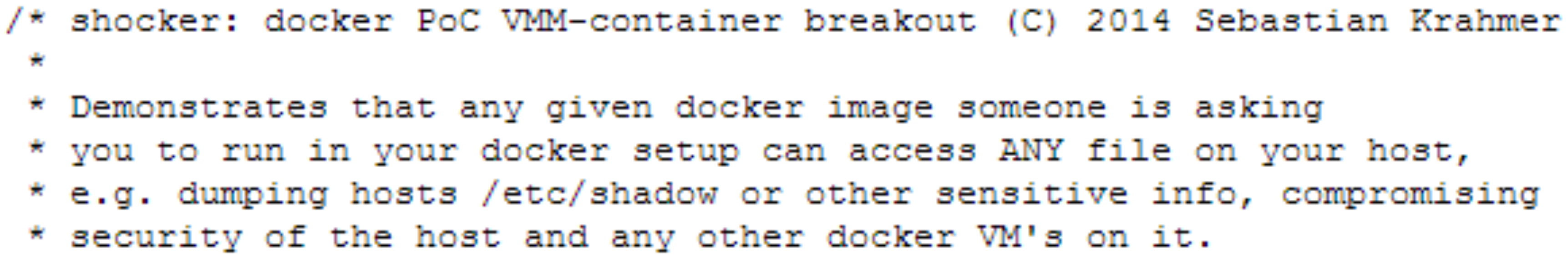
Design for purpose

Containers and containment

With thanks to Dan Walsh @rhatdan
Watch his DockerCon 2014 presentation at http://is.gd/dcrhdw

Shocker

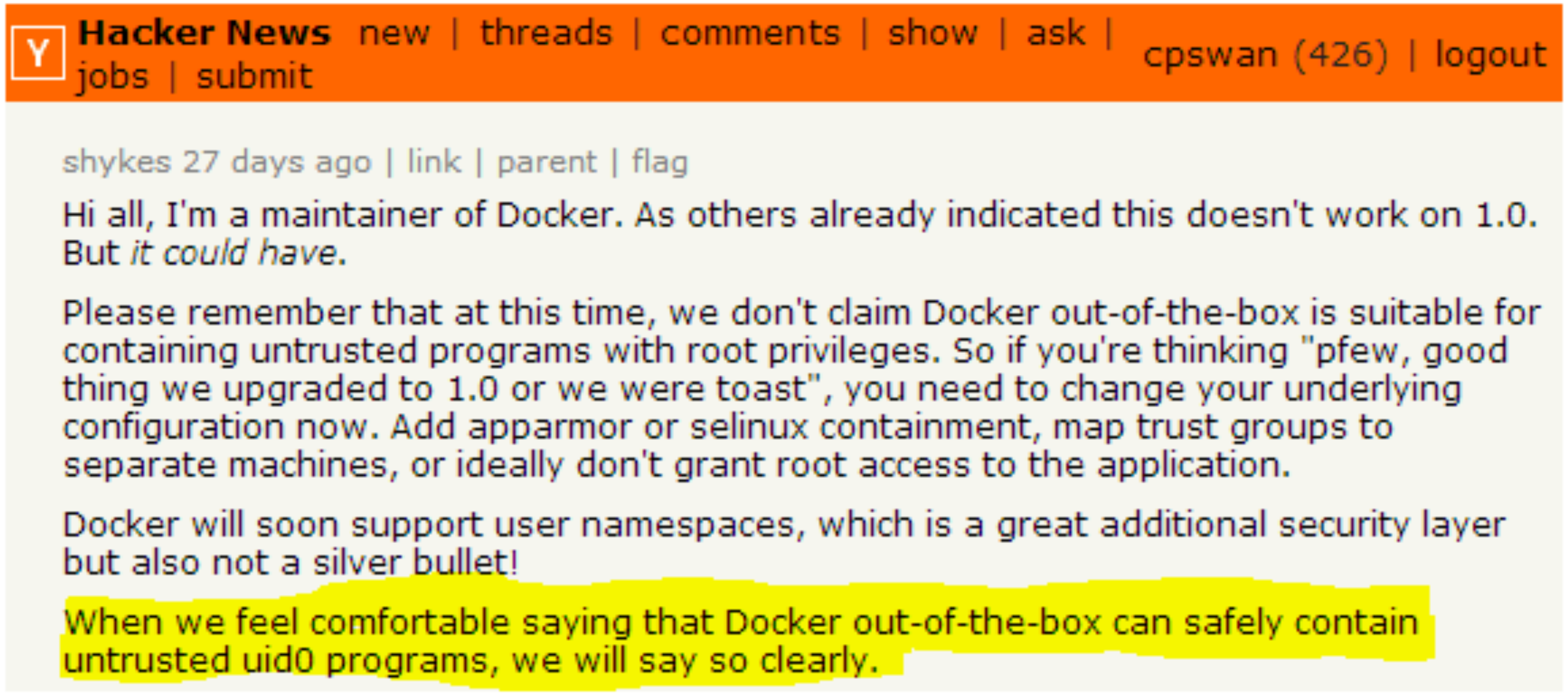
The issue



```
/* shocker: docker PoC VMM-container breakout (C) 2014 Sebastian Krahmer
 *
 * Demonstrates that any given docker image someone is asking
 * you to run in your docker setup can access ANY file on your host,
 * e.g. dumping hosts /etc/shadow or other sensitive info, compromising
 * security of the host and any other docker VM's on it.
```

http://stealth.openwall.net/xSports/shocker.c

The response



Y
Hacker News new | threads | comments | show | ask | jobs | submit
cpswan (426) | logout

shykes 27 days ago | link | parent | flag

Hi all, I'm a maintainer of Docker. As others already indicated this doesn't work on 1.0. But *it could have*.

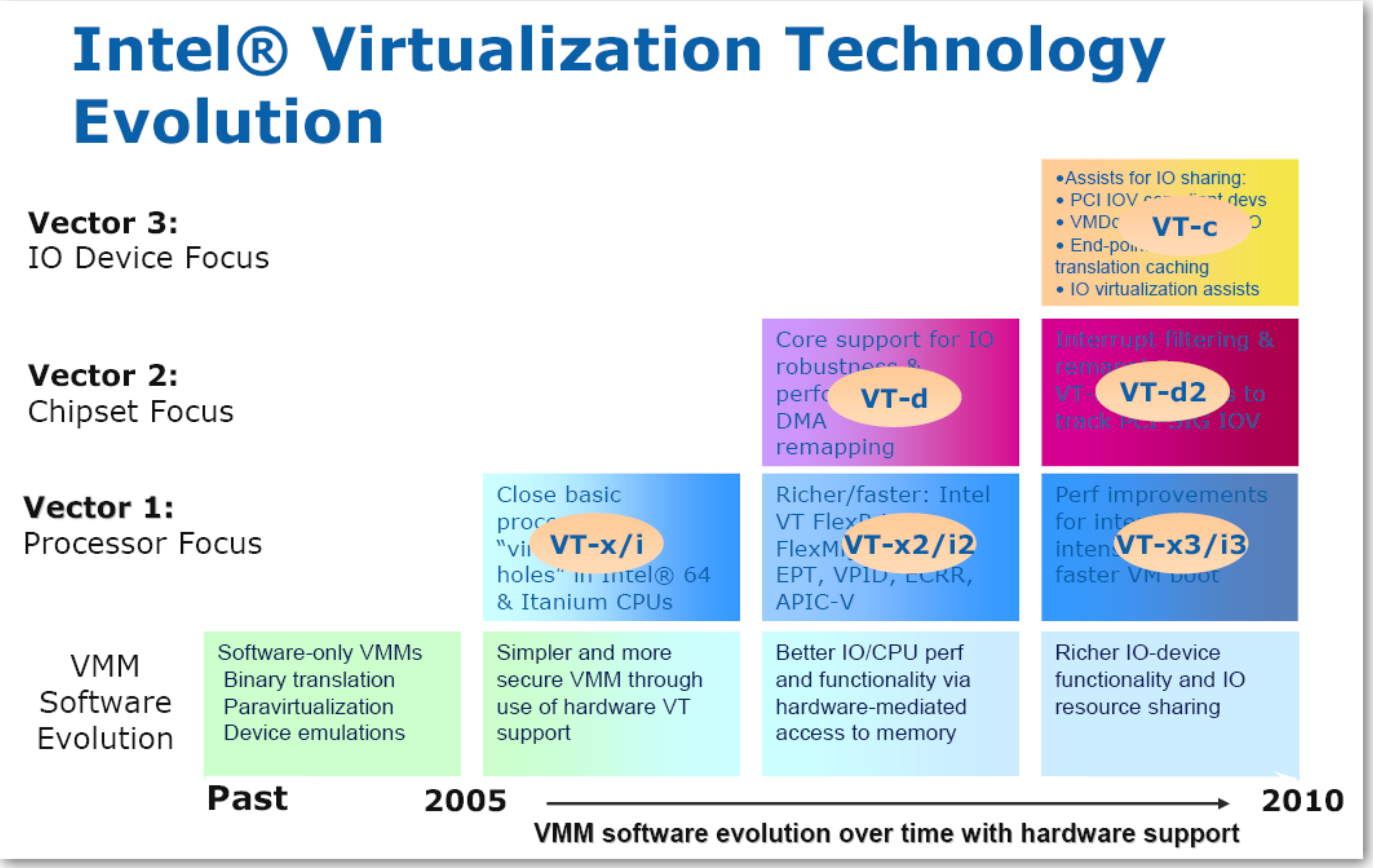
Please remember that at this time, we don't claim Docker out-of-the-box is suitable for containing untrusted programs with root privileges. So if you're thinking "pfew, good thing we upgraded to 1.0 or we were toast", you need to change your underlying configuration now. Add apparmor or selinux containment, map trust groups to separate machines, or ideally don't grant root access to the application.

Docker will soon support user namespaces, which is a great additional security layer but also not a silver bullet!

When we feel comfortable saying that Docker out-of-the-box can safely contain untrusted uid0 programs, we will say so clearly.

https://news.ycombinator.com/item?id=7910117

Because containers aren't VMs and this has yet to come:



Intel® Virtualization Technology Evolution

Vector 3: IO Device Focus

Vector 2: Chipset Focus

Vector 1: Processor Focus

VMM Software Evolution

•Assists for IO sharing:
• PCI IOV devs
• VMDc O
• End-poi
translation caching
• IO virtualization assists
VT-c
Core support for IO robustness & perf DMA remapping
VT-d
Interrupt filtering & remap VT- s to track PCI SIG IOV
VT-d2
Close basic proc "vi holes" in Intel® 64 & Itanium CPUs
VT-x/i
Richer/faster: Intel VT Flex FlexM EPT, VPID, ECRR, APIC-V
VT-x2/i2
Perf improvements for inte inten faster VM boot
VT-x3/i3

Software-only VMMs
Binary translation
Paravirtualization
Device emulations

Simpler and more secure VMM through use of hardware VT support

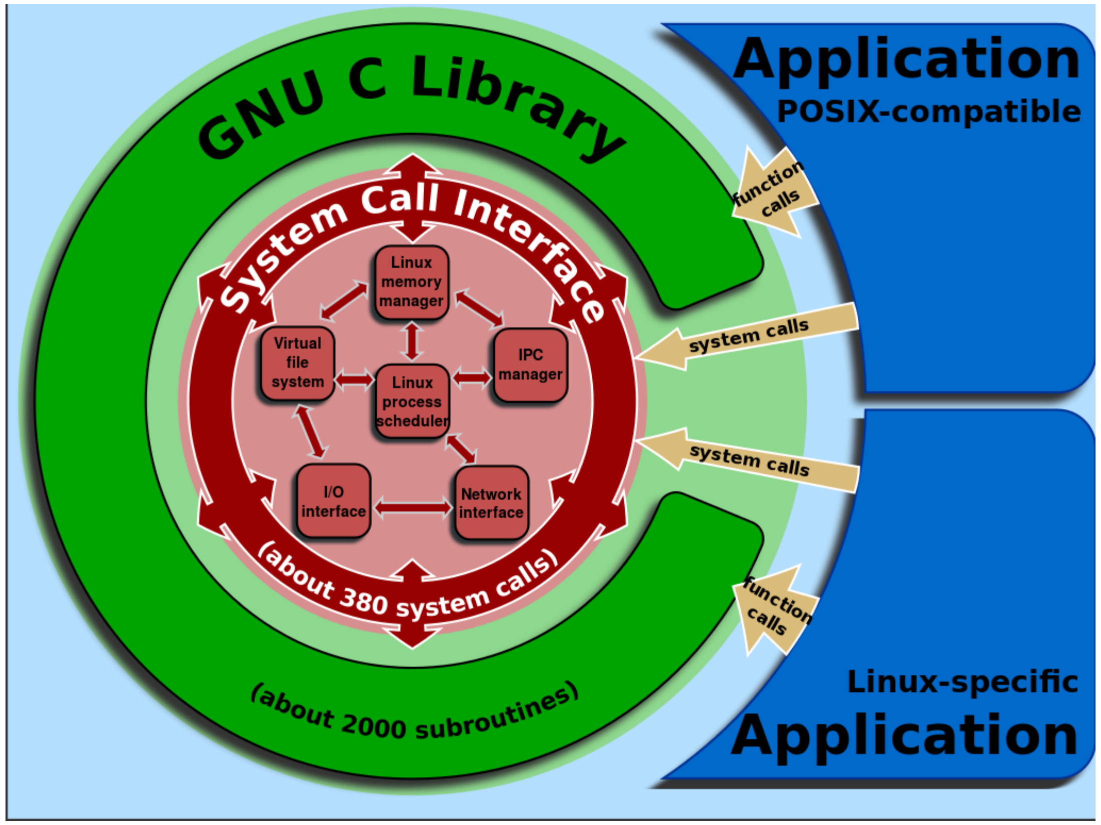
Better IO/CPU perf and functionality via hardware-mediated access to memory

Richer IO-device functionality and IO resource sharing

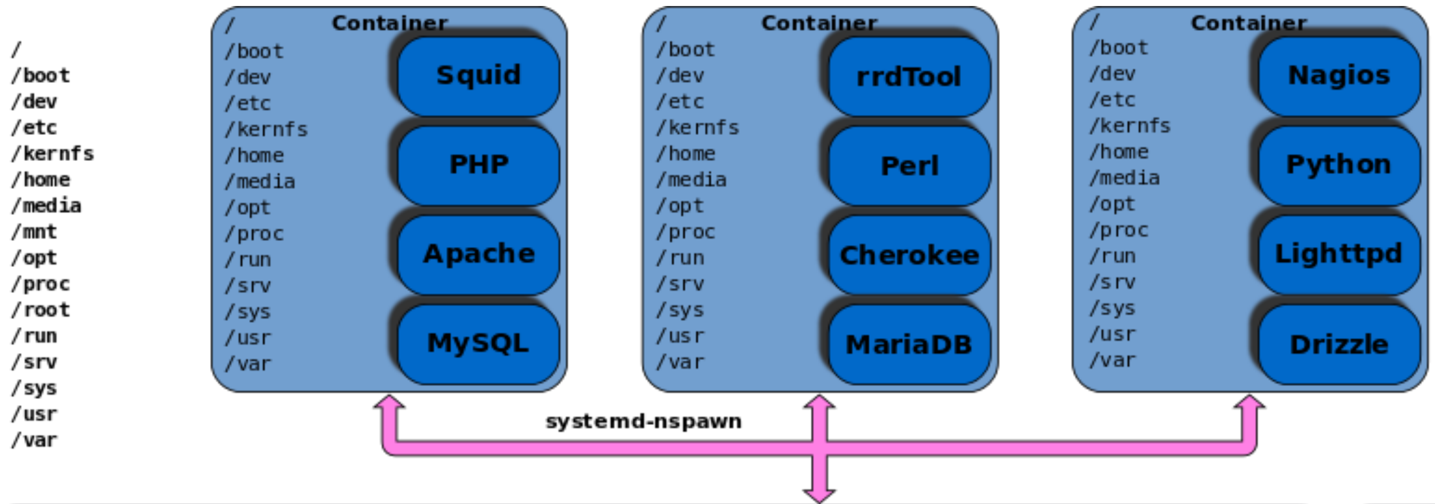
VMM software evolution over time with hardware support

Possible to have our cake and eat it?

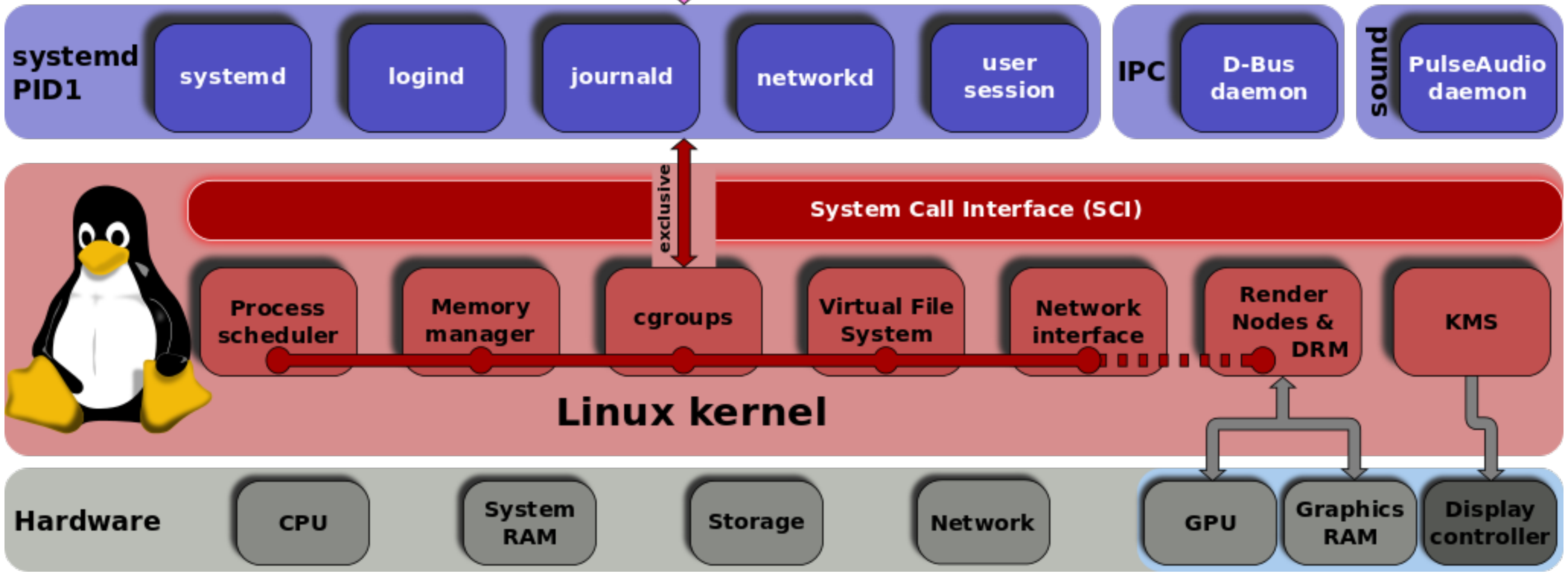
cgroups



GNU C Library
Application
POSIX-compatible
function calls
System Call Interface
Linux memory manager
Virtual file system
IPC manager
Linux process scheduler
I/O interface
Network interface
system calls
system calls
(about 380 system calls)
function calls
(about 2000 subroutines)
Linux-specific
Application



/
/boot
/dev
/etc
/kernfs
/home
/media
/mnt
/opt
/proc
/root
/run
/srv
/sys
/usr
/var
Container
/
/boot
/dev
/etc
/kernfs
/home
/media
/opt
/proc
/run
/srv
/sys
/usr
/var
Squid
PHP
Apache
MySQL
Container
/
/boot
/dev
/etc
/kernfs
/home
/media
/opt
/proc
/run
/srv
/sys
/usr
/var
rrdTool
Perl
Cherokee
MariaDB
Container
/
/boot
/dev
/etc
/kernfs
/home
/media
/opt
/proc
/run
/srv
/sys
/usr
/var
Nagios
Python
Lighttpd
Drizzle
systemd-nspawn



systemd PID1
systemd
logind
journald
networkd
user session
IPC
D-Bus daemon
sound
PulseAudio daemon
exclusive
System Call Interface (SCI)
Process scheduler
Memory manager
cgroups
Virtual File System
Network interface
Render Nodes & DRM
KMS
Linux kernel
Hardware
CPU
System RAM
Storage
Network
GPU
Graphics RAM
Display controller

namespaces

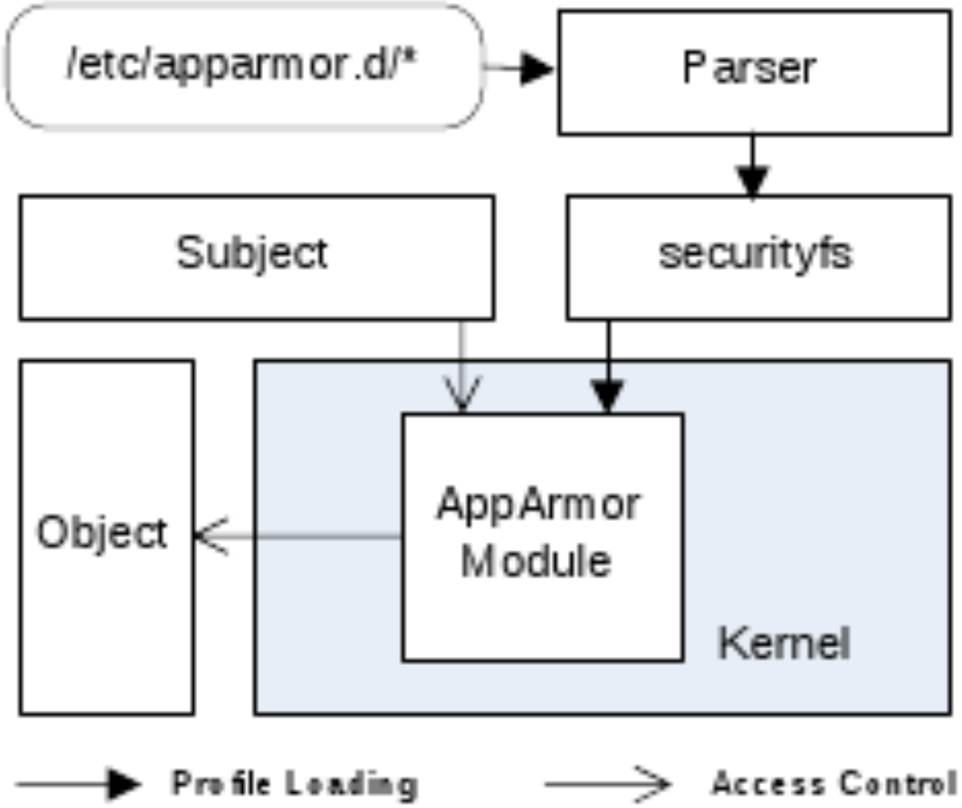
mnt	mount points, filesystems
pid	processes
net	network
ipc	inter process communication
uts	hostname
device	devices
user	UIDs

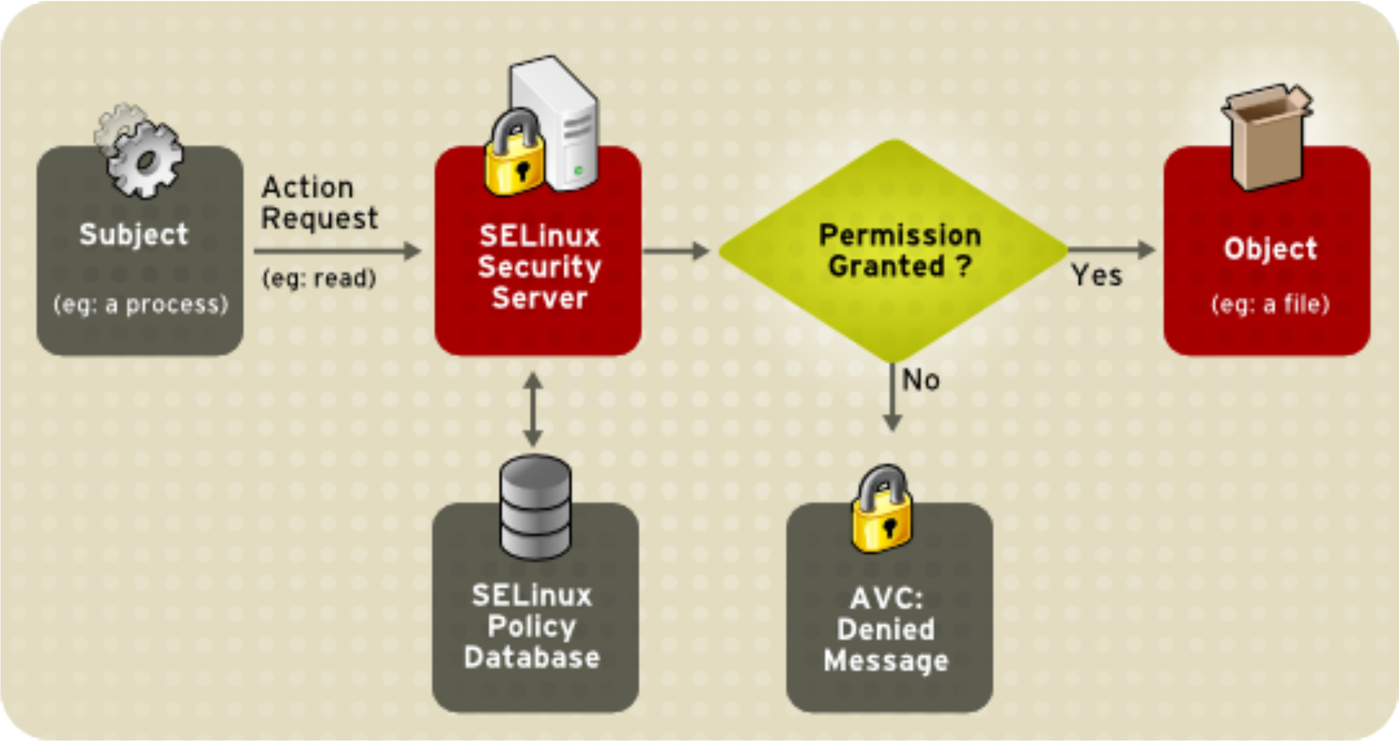
capabilities

Fine grained control over ‘root’ privileges:

- deny all "mount" operations;
- deny access to raw sockets (to prevent packet spoofing);
- deny access to some filesystem operations, like creating new device nodes, changing the owner of files, or altering attributes (including the immutable flag);
- deny module loading;
- etc.

Mandatory Access Control (MAC): AppArmor and SELinux



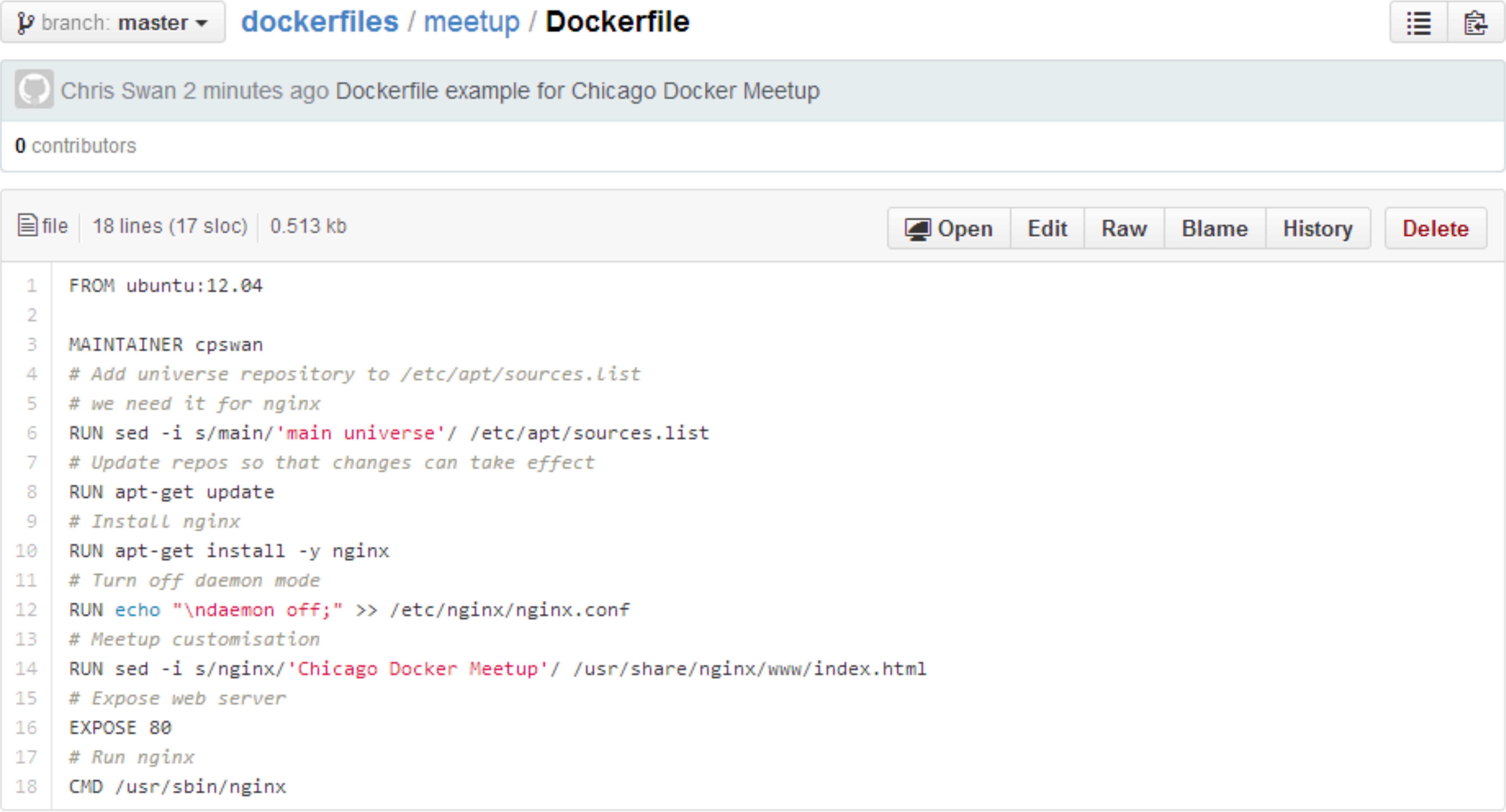


<optimist>Containers will contain</optimist>

- Use of namespaces, capabilities and MAC will improve
  - **Might be a game of 'whack a mole'**
    - Hard to tell when we're done (is @solomonstre's word going to be enough?)
- Libcontainer can drive other mechanisms
  - **More secure options might come**
- Hardware support might come
  - **Existing rings 1 & 2 aren't used much, but aren't really suitable**
  - **VT-x introduced ring -1, do we need a ring 0.5?**

The manifest problem

My Dockerfile from earlier



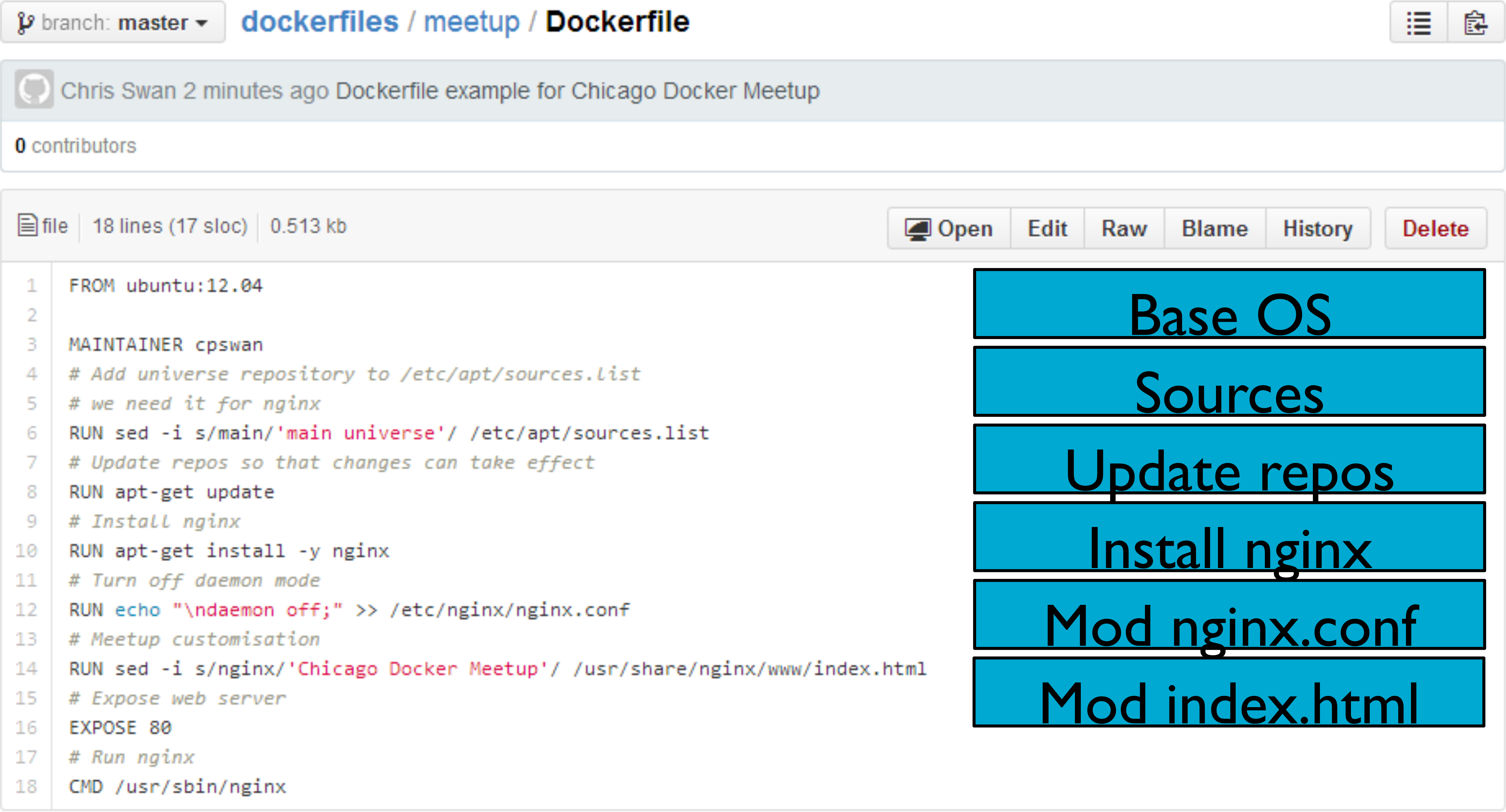
Chris Swan 2 minutes ago Dockerfile example for Chicago Docker Meetup

0 contributors

```
FROM ubuntu:12.04

MAINTAINER cpswan
# Add universe repository to /etc/apt/sources.list
# we need it for nginx
RUN sed -i s/main/'main universe'/ /etc/apt/sources.list
# Update repos so that changes can take effect
RUN apt-get update
# Install nginx
RUN apt-get install -y nginx
# Turn off daemon mode
RUN echo "\ndaemon off;" >> /etc/nginx/nginx.conf
# Meetup customisation
RUN sed -i s/nginx/'Chicago Docker Meetup'/ /usr/share/nginx/www/index.html
# Expose web server
EXPOSE 80
# Run nginx
CMD /usr/sbin/nginx
```

Each active line creates a layer



dockerfiles / meetup / Dockerfile

Chris Swan 2 minutes ago Dockerfile example for Chicago Docker Meetup

0 contributors

file | 18 lines (17 sloc) | 0.513 kb

```
FROM ubuntu:12.04

MAINTAINER cpswan
# Add universe repository to /etc/apt/sources.list
# we need it for nginx
RUN sed -i s/main/'main universe'/ /etc/apt/sources.list
# Update repos so that changes can take effect
RUN apt-get update
# Install nginx
RUN apt-get install -y nginx
# Turn off daemon mode
RUN echo "\ndaemon off;" >> /etc/nginx/nginx.conf
# Meetup customisation
RUN sed -i s/nginx/'Chicago Docker Meetup'/ /usr/share/nginx/www/index.html
# Expose web server
EXPOSE 80
# Run nginx
CMD /usr/sbin/nginx
```

Base OS

Sources

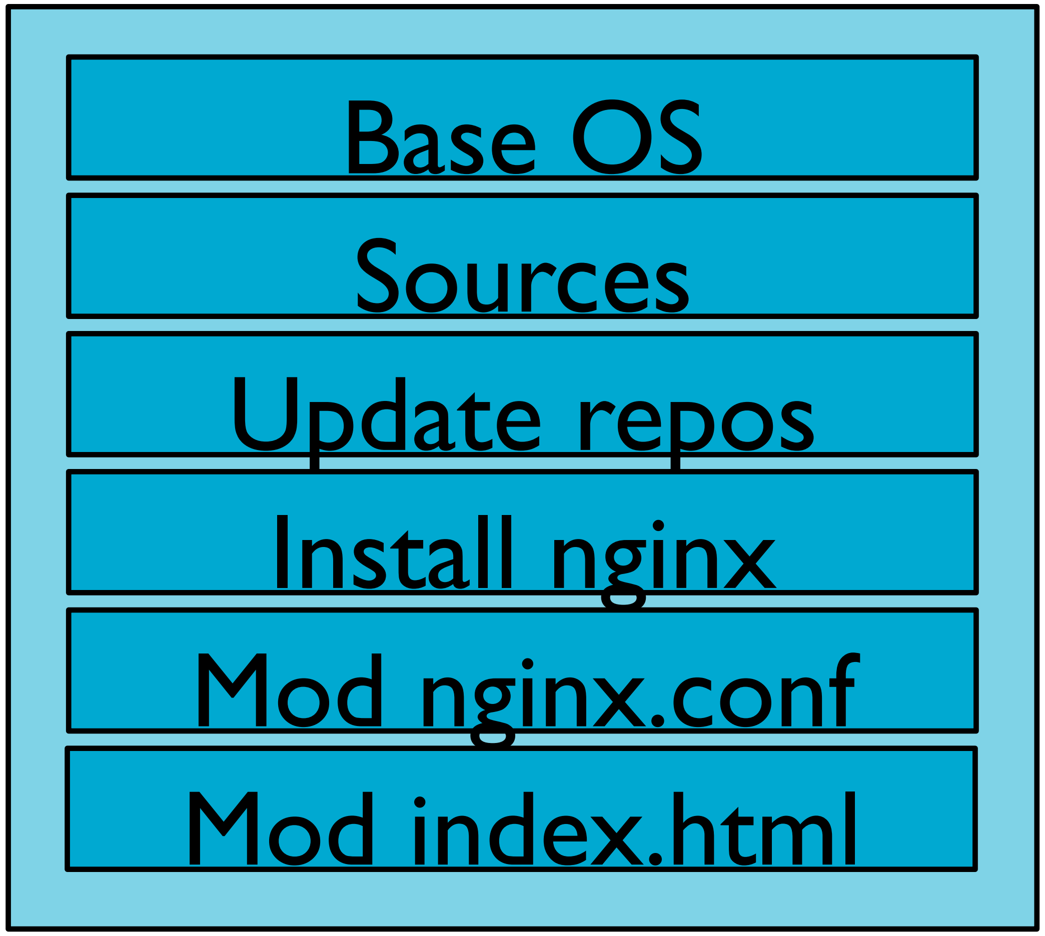
Update repos

Install nginx

Mod nginx.conf

Mod index.html

An image binds layers together



Base OS

Sources

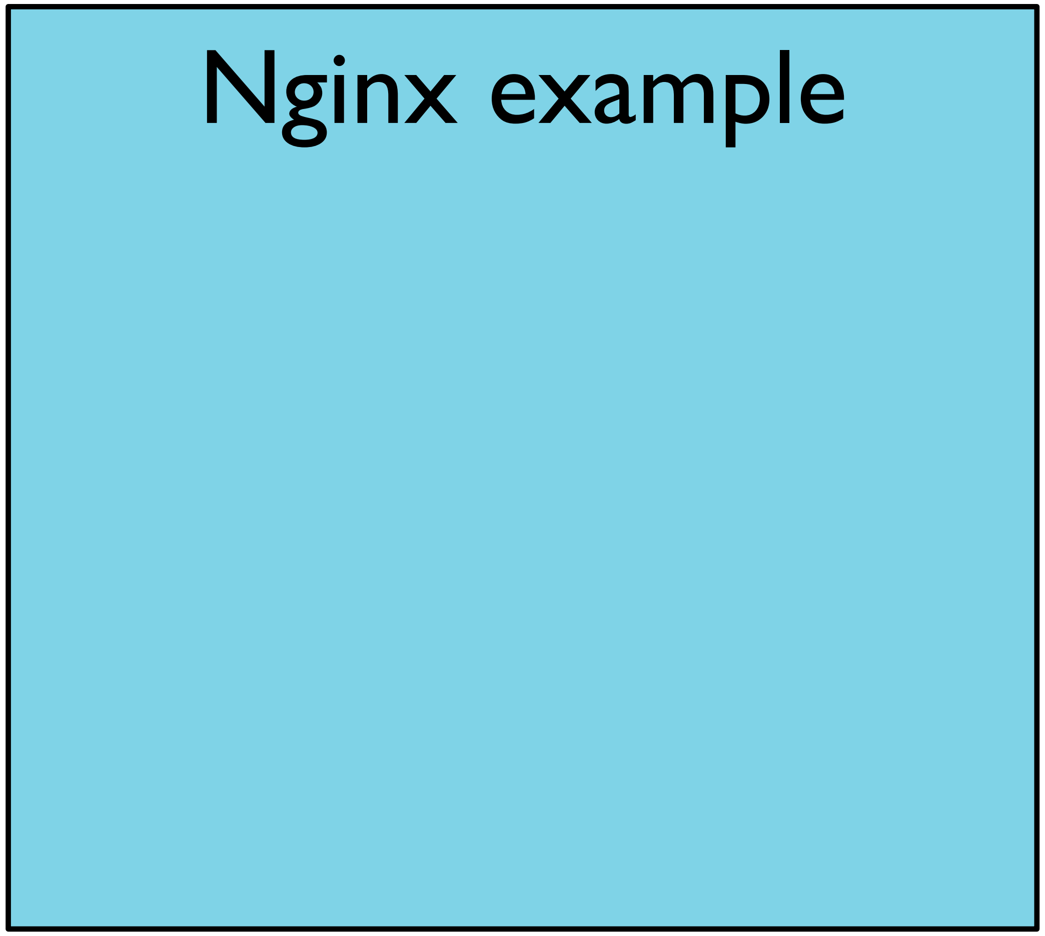
Update repos

Install nginx

Mod nginx.conf

Mod index.html

The image is the unit of deployment

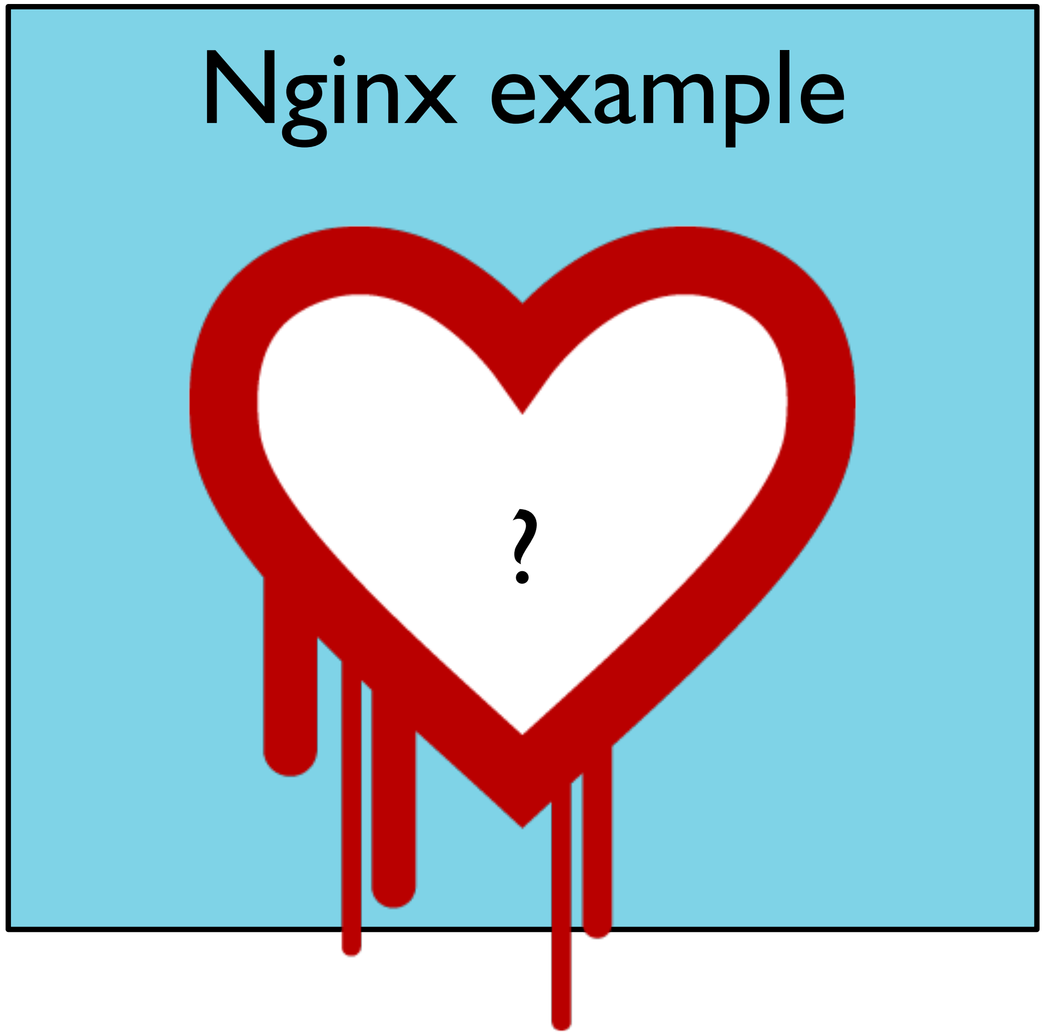


Nginx example

What version of nginx is that?

Nginx example

What version of OpenSSL installed?



Nginx example

Problem 1 – non determinism

Whilst we want this to be cached in the short term:

```
apt-get install nginx
```

We perhaps don’t want it cached in the long term

What are those durations?

2 – the manifest problem

When I run

```
apt-get install nginx
```

I don't know which version of nginx I just got

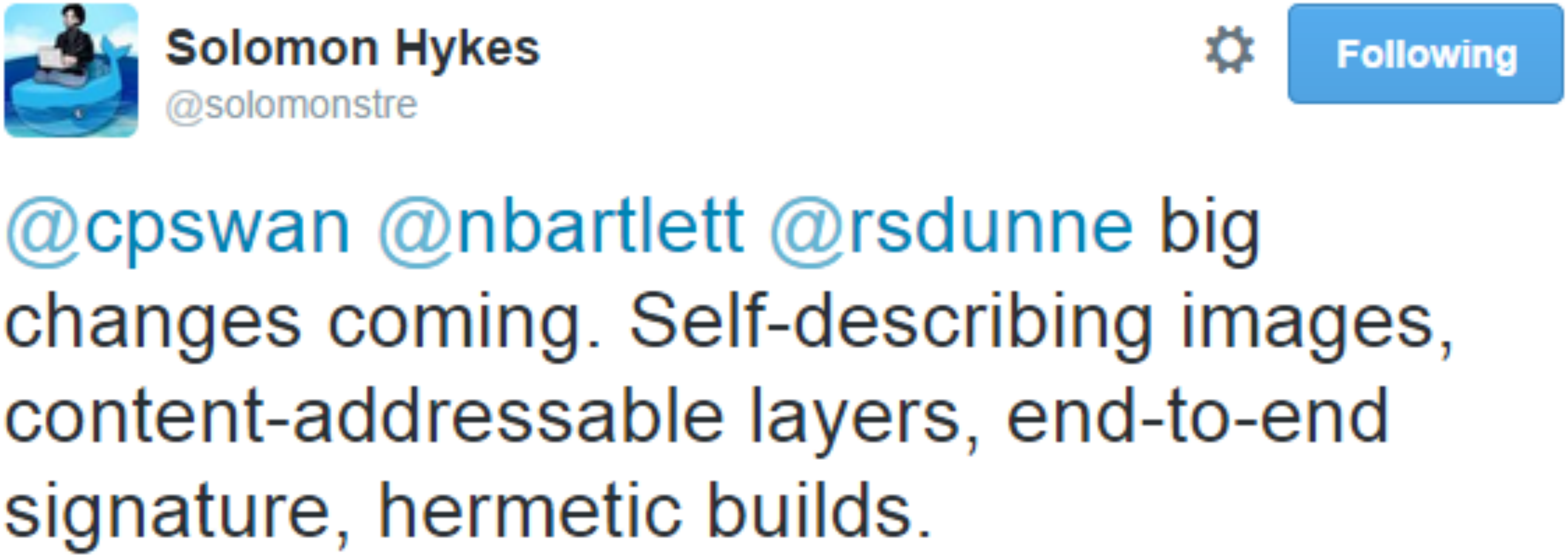
Should I?

```
nginx -v > some_log.txt
```

Or maybe?

```
apt-cache policy nginx > some_log.txt
```

Again, Solomon promises to fix things



Solomon Hykes
@solomonstre

@cpswan @nbartlett @rsdunne big changes coming. Self-describing images, content-addressable layers, end-to-end signature, hermetic builds.

There is another way

TL;DR

Dockerfile is awesomely productive

- Great for DevOps

Containers don’t contain

- At least not yet

Images have a manifest problem

- Keep track of your stuff

Please give me feedback:
http://is.gd/chdmsf

Questions?

cohesiveFT

Chicago, US

ContactMe@cohesiveft.com

☎ +1 888 444 3962