inovex

inovex classes

Kubernetes

Johannes M. Scheuermann &
Timo Heinrichs

München, 20. September 2018

Johannes M. Scheuermann

Macht Dinge in der Wolke

- Team ITO since 2014
- Kubernetes since 2014
- Certified Kubernetes Admin
- LF authorized Kubernetes Admin trainer
- Cloud technologies
- @johscheuer

Timo Heinrichs

Macht Dinge in der Wolke

- Team ITO since 2017
- Docker since v0.9
- Kubernetes since 2017
- Cloud technologies
- @theinrichs

Agenda

- A brief overview of history
- Containers
- Kubernetes - a high level overview
- Hands on part

- Hands on part: https://github.com/johscheuer/inovex_classes


What about you?

A brief overview of History

A brief overview of History

- 1967: IBMs first Hypervisor (LPAR)
- 1979: chroot (Unix 1982)
- 1998: A company called VMWare was established
- 2000: FreeBSD Jails
- 2002: Linux namespaces
- 2006: Linux cgroups
- 2007: AIX WPARs
- 2008: LXC
- 2013: Docker
- 2014: Kubernetes

(Linux) Containers - What's inside?

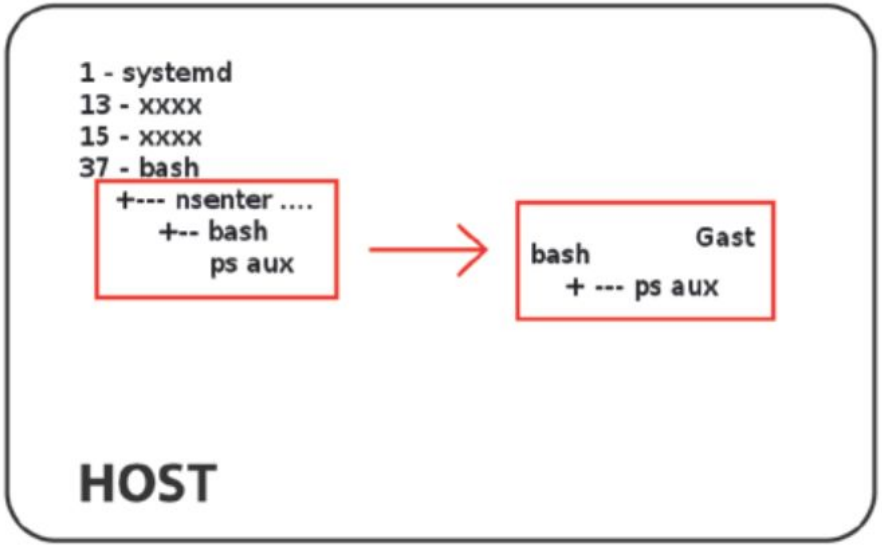
- chroot
- namespaces
- cgroups
- layered filesystem
- capabilities

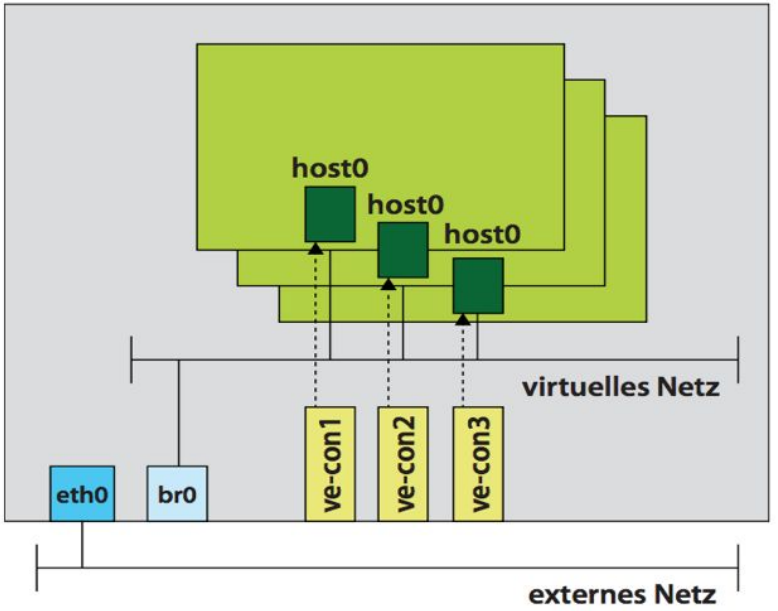
Chroot

```
$ tree /home/vagrant/
/home/vagrant/
├── jail
│   ├── bin
│   │   ├── bash
│   │   ├── ls
│   │   └── tree
│   ├── inside.jail
│   ├── lib
│   │   └── x86_64-linux-gnu
│   │       ├── libacl.so.1
│   │       ├── libattr.so.1
│   │       ├── libc.so.6
│   │       ├── libdl.so.2
│   │       ├── libpcre.so.3
│   │       ├── libselinux.so.1
│   │       └── libtinfo.so.5
│   └── lib64
│       └── ld-linux-x86-64.so.2
└── outside.jail

5 directories, 13 files
$
```

```
$ sudo chroot /home/vagrant/jail /bin/bash
$ tree /
/
|-- bin
|   |-- bash
|   |-- ls
|   `-- tree
|-- inside.jail
|-- lib
|   `-- x86_64-linux-gnu
|       |-- libacl.so.1
|       |-- libattr.so.1
|       |-- libc.so.6
|       |-- libdl.so.2
|       |-- libpcre.so.3
|       |-- libselinux.so.1
|       `-- libtinfo.so.5
`-- lib64
    `-- ld-linux-x86-64.so.2

4 directories, 12 files
$
```

Namespaces



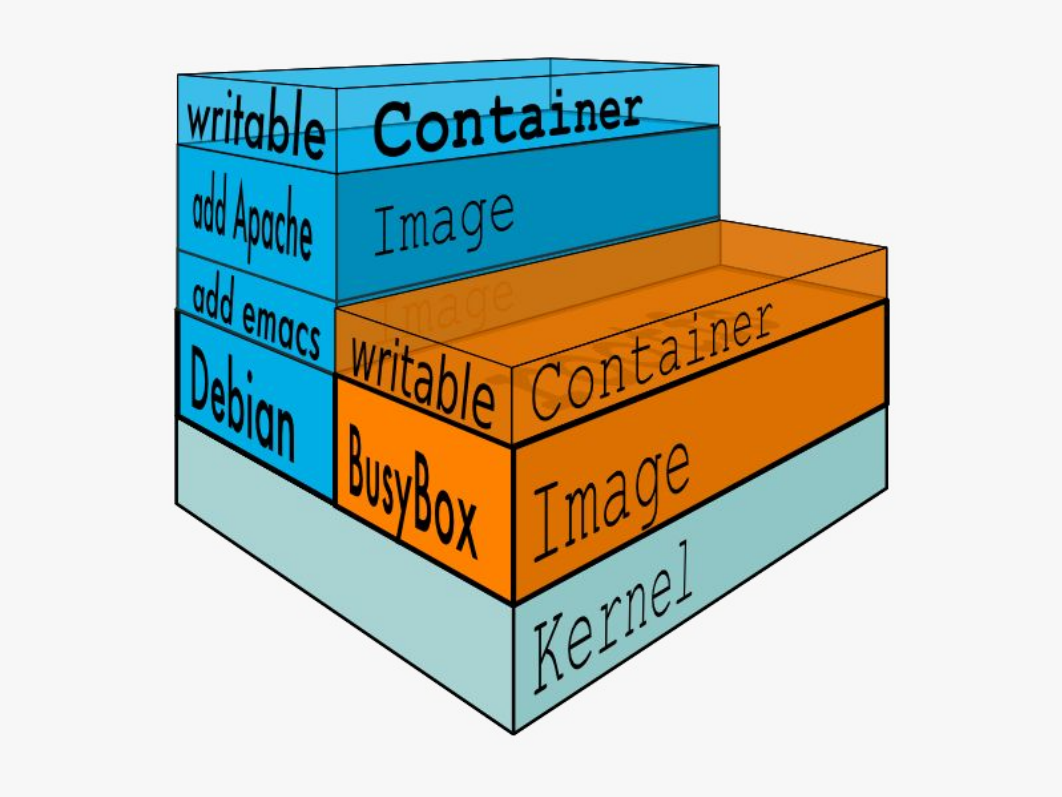


inovex

cgroups

- Hierarchy
- Resource limiting
  - CPU
  - Memory
  - …
- Prioritization
- Accounting
- Control
  - Freeze groups of processes (checkpointing)

Layered filesystem



inovex

Capabilities

- Two categories privileged/unprivileged process
  - privileged => UID == 0
  - unprivileged 0> UID != 0
- Privileged processes bypass kernel permission checks
- Since Linux Kernel 2.2 -> capabilities
  - Per-Thread attribute
- Some capabilities includes other capabilities
  - e.g. “CAP_SYS_ADMIN”

Source: http://man7.org/linux/man-pages/man7/capabilities.7.html

Docker - What's special?

- Uses "proven" technologies
- Made them developer friendly
  - Not only for Kernel hackers
  - Still most features are hard to understand
- Made the Container image distribution easier
  - Docker Hub
  - Docker images
  - "docker run -ti hello-world"

Hands on
Container 101

kubernetes

by Google

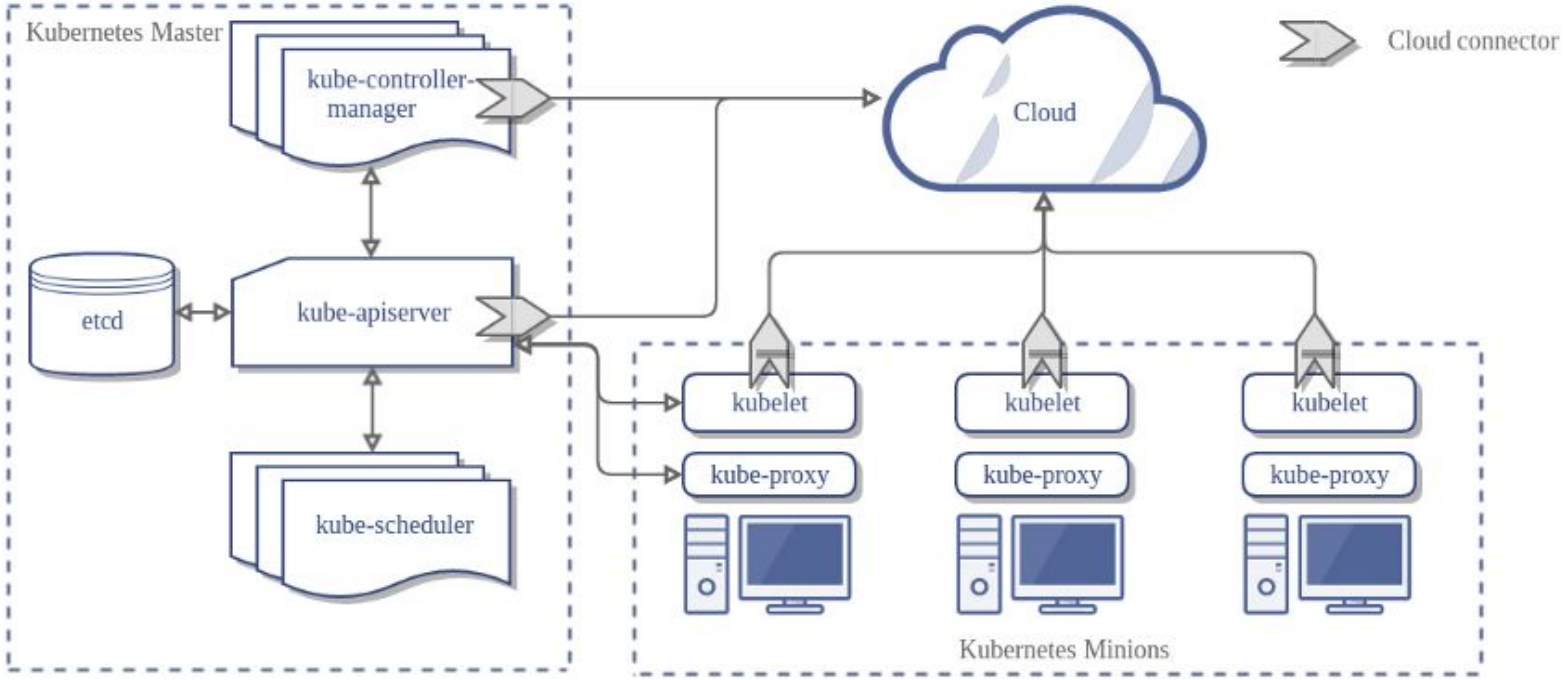
History of Kubernetes

- Built with lessons learned from Google
- Borg (https://research.google.com/pubs/pub43438.html)
- Omega (https://ai.google/research/pubs/pub41684)
- 2014: Open sourced
- 2015: Version 1.0
- Currently: v1.11.3


How to orchestrate a fleet?

“Kubernetes is an open-source system for automating deployment, scaling, and management of containerized applications”

What is Kubernetes?



Architecture

- Each component has a dedicated task
  - Do one thing and do it well
- Each component can be built HA
- Each component can fail without breaking the system
  - except for the API server
  - except for the etcd

API server

- “Front-end” to the cluster
- All components communicate via the API server
- Validates and configures data
- Services REST operations
- Only component connected to etcd

Scheduler

- Uses algorithm to schedule pods
- Checks quota restrictions
- Custom scheduling possible
  - Multiple schedulers can be used
- Affinity rules for placement
- Taints can be used to repel pods
- Pod binding forces scheduling

Controller-Manager

- kube-controller-manager
  - Implements control loops
  - Watches state of the cluster
  - Current state → desired state
- cloud-controller-manager
  - Interacts with the cloud

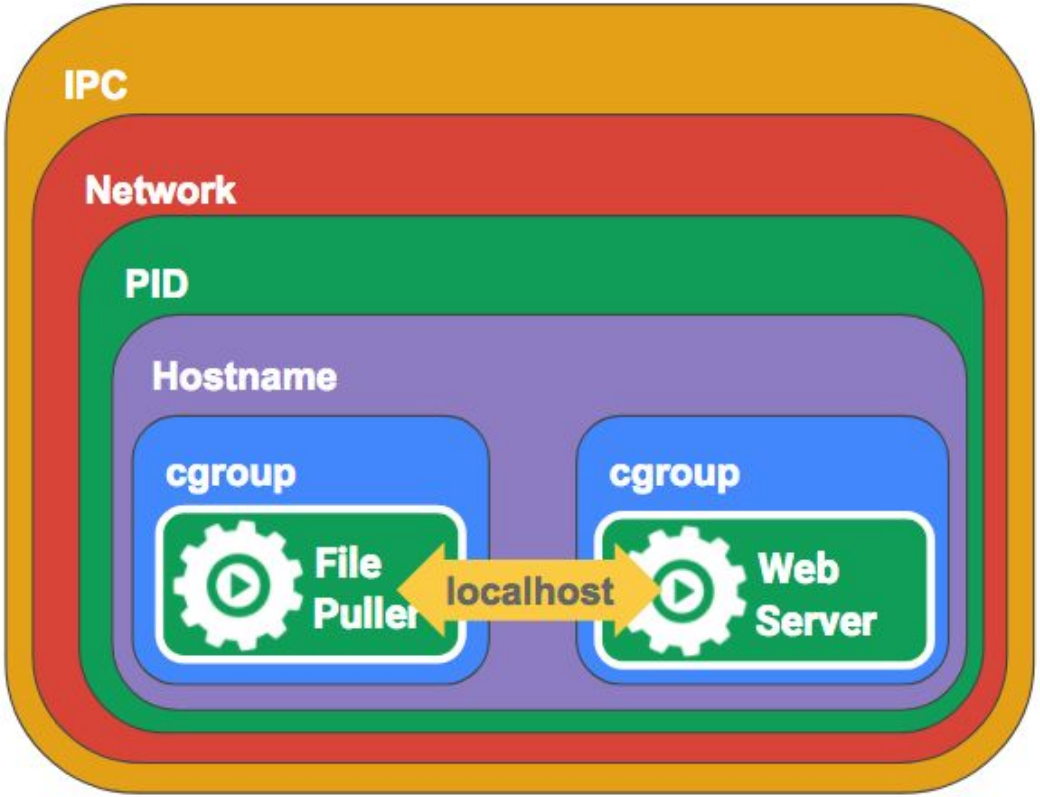
kubelet

- Agent on each node
- Can be used standalone
- Uses PodSpec to start containers
- Mounts volumes and secrets
- Passes requests to underlying container engine
- Report status of pods and nodes to API server

kube-proxy

- Defines rules to enable inter-services communication
- Simple load-balancing
- Can use iptables or ipvs

Pods


IPC
Network
PID
Hostname
cgroup
File
Puller
localhost
cgroup
Web
Server

Terminology

- Controllers
  - Deployment
  - ReplicaSet
  - DaemonSet
  - Jobs
- Labels
- Taints
- Tolerations
- Annotations

etcd - the memory of kubernetes

Consensus

- (Eventually) Agreeing on one result
- Majority must agree on a proposed value
- Eventually known by all members
- Can have the concept of a leader
  - Leader is voted by majority
- Fault-tolerance to unreliable communication
- Often used logical clocks (monotonic counter)
  - Why?

Consensus - Algorithms

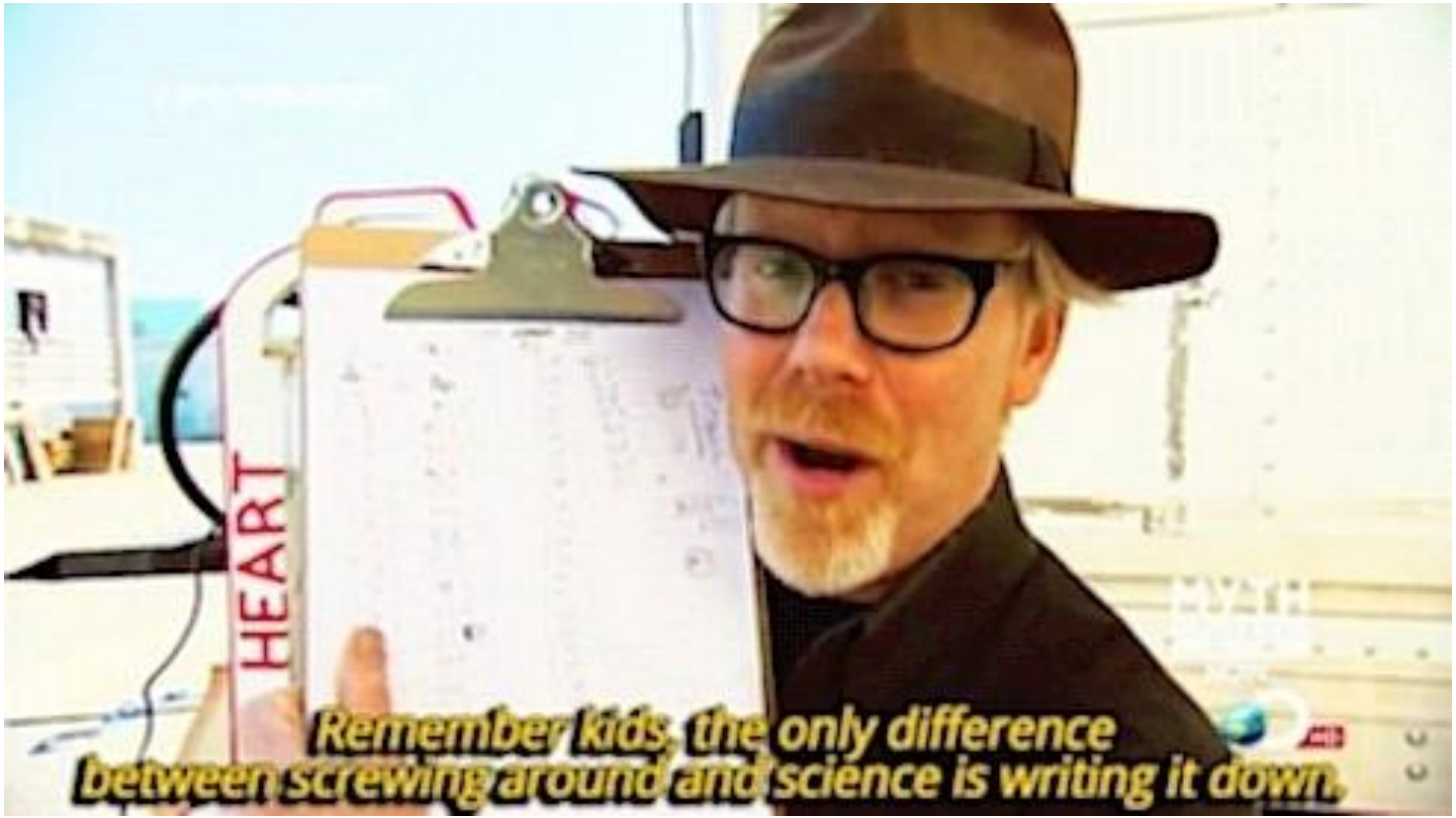
- Paxos (https://lamport.azurewebsites.net/pubs/lamport-paxos.pdf)
- Multi-Paxos
- Raft (https://raft.github.io)
- Zab (https://dl.acm.org/citation.cfm?id=2056409)
- ...

Consensus - Implementations

- Chubby (Google lock-server) (Paxos)
- ZooKeeper
- etcd
- ..

Hands on
Kubernetes high-level overview

Gentle reminder



inovex

Outlook

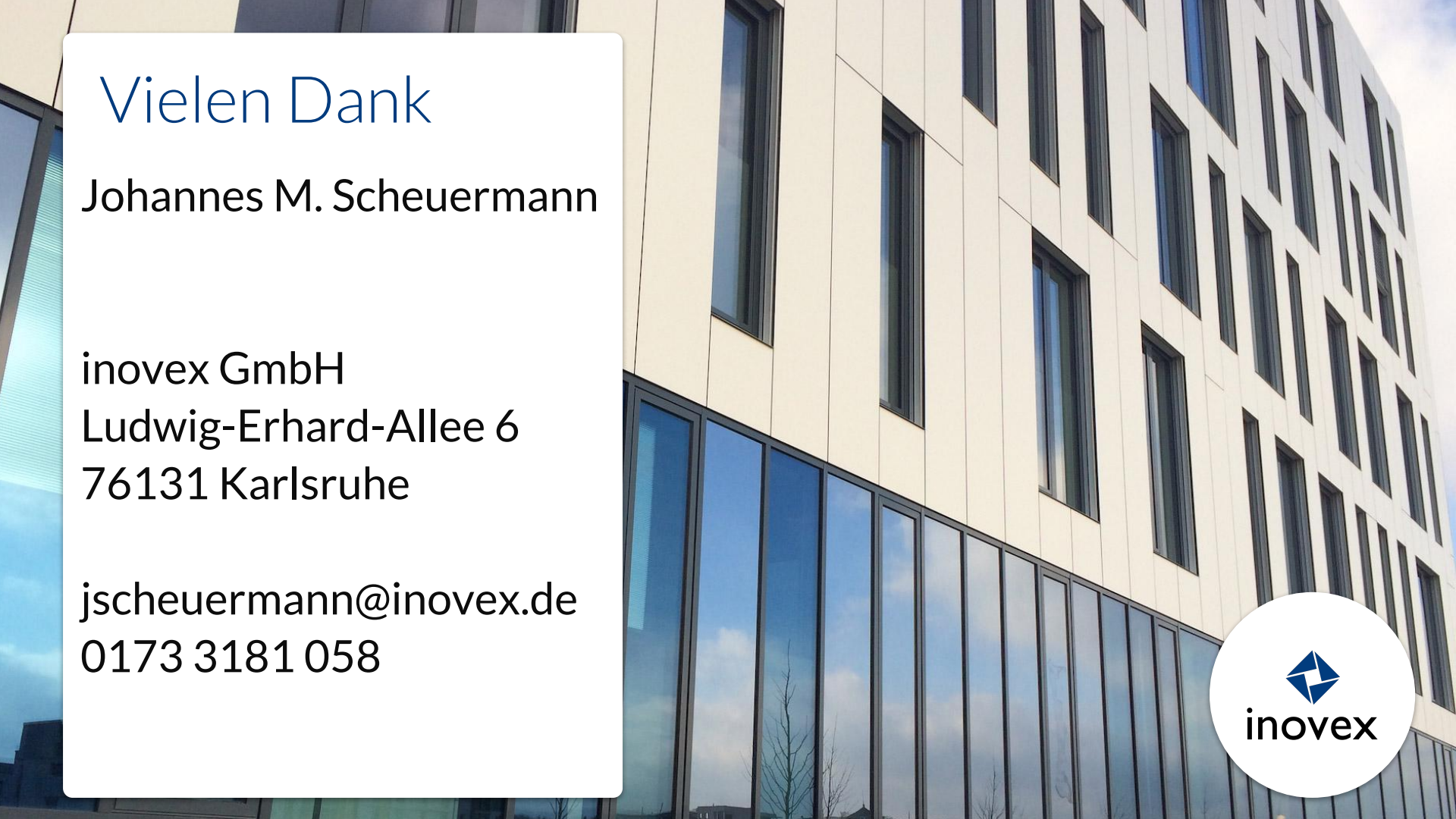
Questions?

Vielen Dank

Johannes M. Scheuermann

inovex GmbH
Ludwig-Erhard-Allee 6
76131 Karlsruhe

jscheuermann@inovex.de
0173 3181 058



Reading list

- https://lwn.net/Articles/689856
- http://man7.org/linux/man-pages/man7/namespaces.7.html
- https://www.ianlewis.org/en/almighty-pause-container
- https://www.kernel.org/doc/Documentation/cgroup-v1/cgroups.txt
- https://kubernetes.io/docs/concepts/architecture/cloud-controller/
- https://coreos.com/etcd/docs/latest/learning/api_guarantees.html
- https://www.ianlewis.org/en/what-are-kubernetes-pods-anyway
- https://github.com/johscheuer/inovex_classes
- https://www.booleanworld.com/depth-guide-iptables-linux-firewall