

Kata Containers: when OCI meets virtualization

彭涛 bergwolf@hyper.sh

Outlines

- Docker and OCI Intro
- Kata Containers
- Kata Containers and OCI
- Kata Containers and CRI
- Q&A

About

- About me
- About hyper.sh
 - Pioneers on virtualized container technology
 - Creator of multiple OSS projects
 - hyperhq/runv
 - hyperhq/hyperd
 - kubernetes/frakti
 - Public cloud service
 - <https://hyper.sh>
 - The only Chinese startup in Forrester Enterprise Container Platforms, 2018, amongst Amazon, Microsoft, Google, Alibaba Cloud, Huawei, Tencent Cloud etc.
- runV -> Kata Containers




Docker

- Hottest container technology
- From container runtime, to cluster management, network/storage plugin, container orchestration, even OS packaging
- Cornerstone of the container ecosystem
- Industry disruptors
 - Lots of OSS projects and startup companies

Docker and OCI

- De facto standard vs. Container industry standard
- Image-spec
 - How to building, transporting, and preparing a container image
 - A manifest, an image index (optional), a set of filesystem layers, and a configuration
- Runtime-spec
 - How to run a file system bundle
 - Configuration, execution environment, and lifecycle of a container

Docker Advantages

 docker =  Container +  Docker Image

轻量级 ✓

快速 ✓

隔离性 ❌

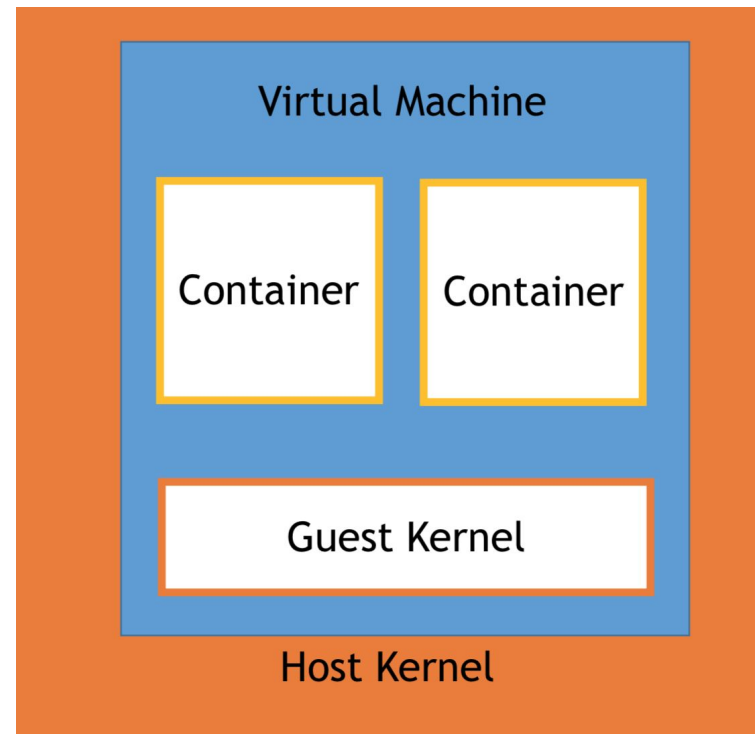
随时随地 ✓

便携 ✓

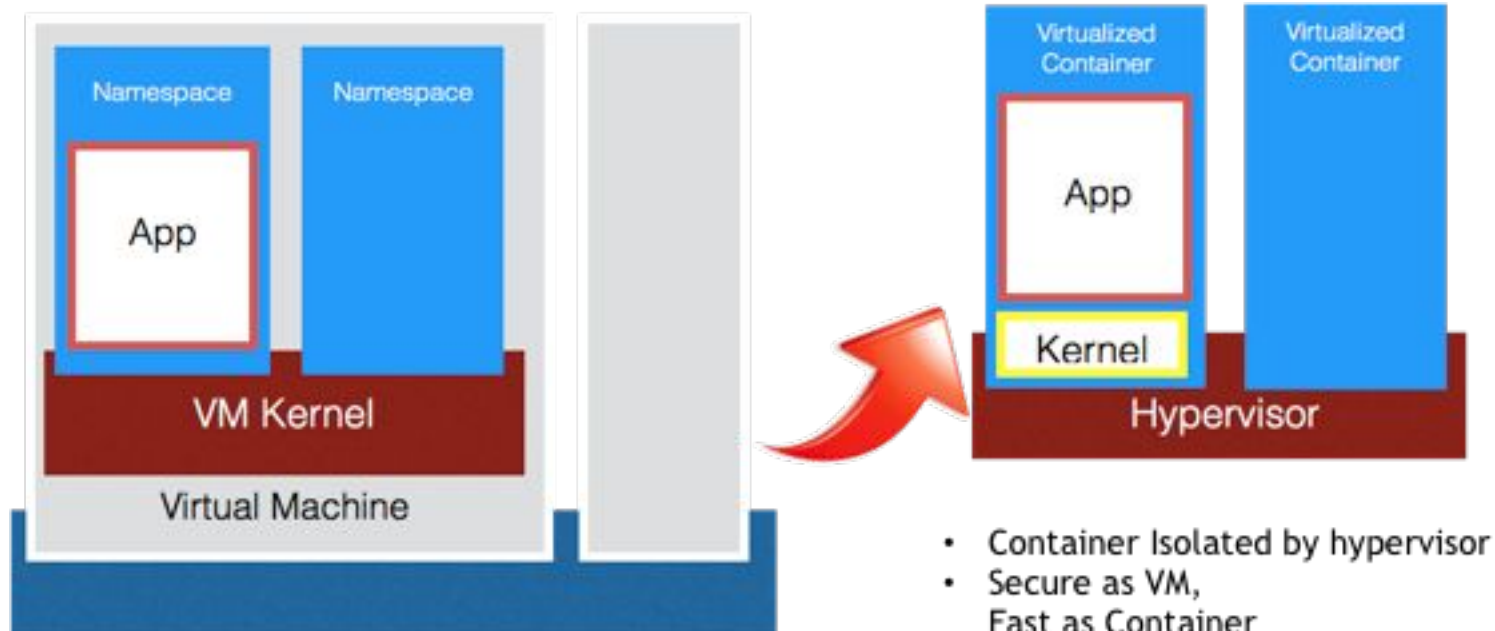
不可变更 ✓

Container Isolation

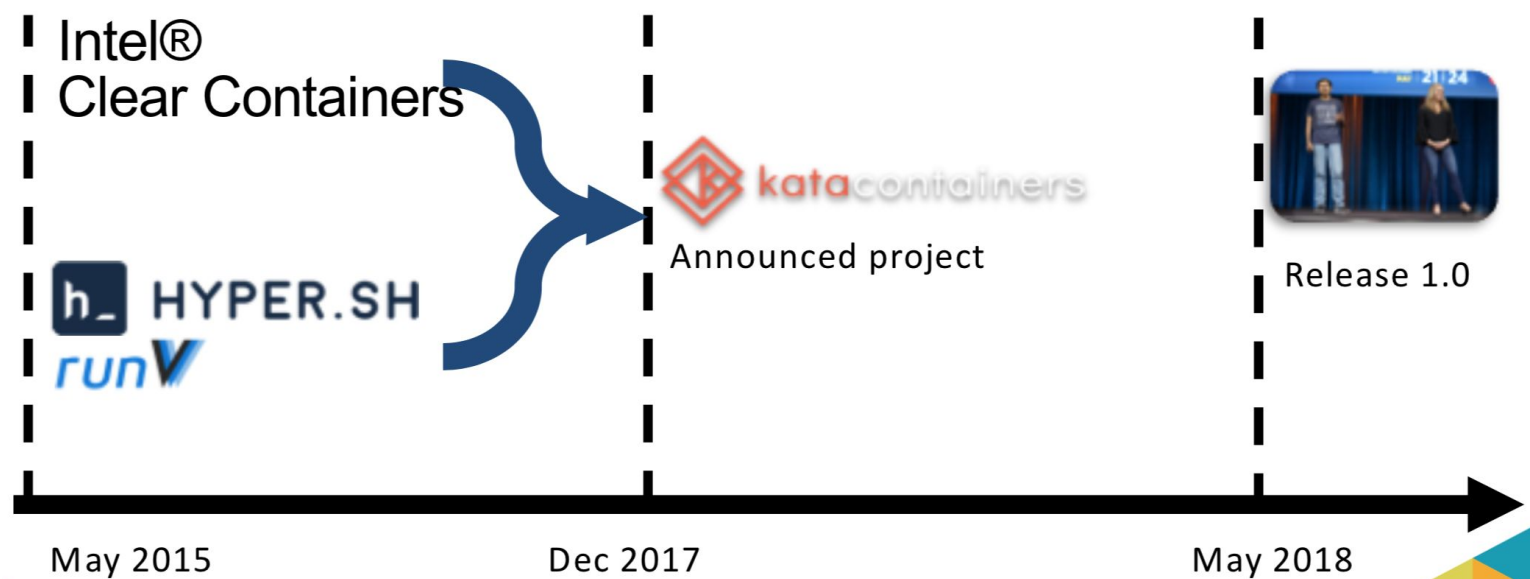
- Container isolation has improved a lot
- Affected by security bugs from time to time
 - Dirty COW (CVE-2016-5195)
 - waitid (CVE-2017-5123)
- Resource utilization are affected too
 - Memcg oom
- Weaker than virtual machines



VM + Container Made Simple



Kata Containers



Kata Containers

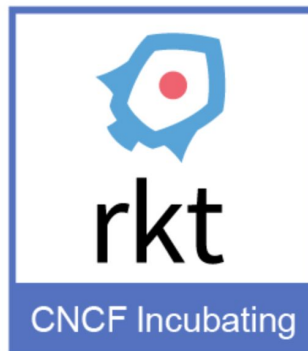
- Founded by Hyper.sh and Intel
- Managed by OpenStack Foundation
- Huawei, Google, MSFT in Arch Committee
- Contributors from Redhat, ARM, IBM etc.
- Apache 2.0 License

Industry Support

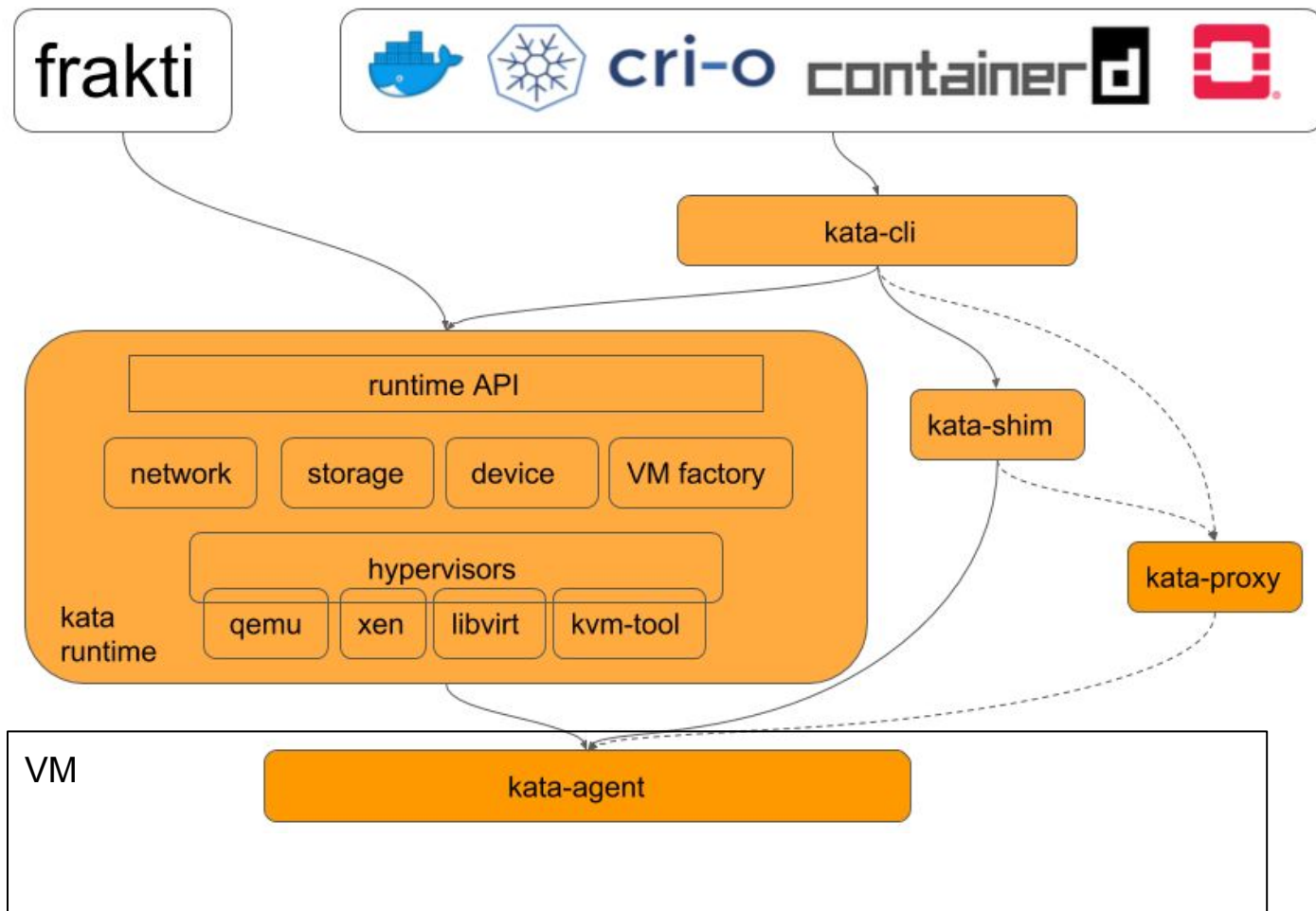


CNCF Landscape

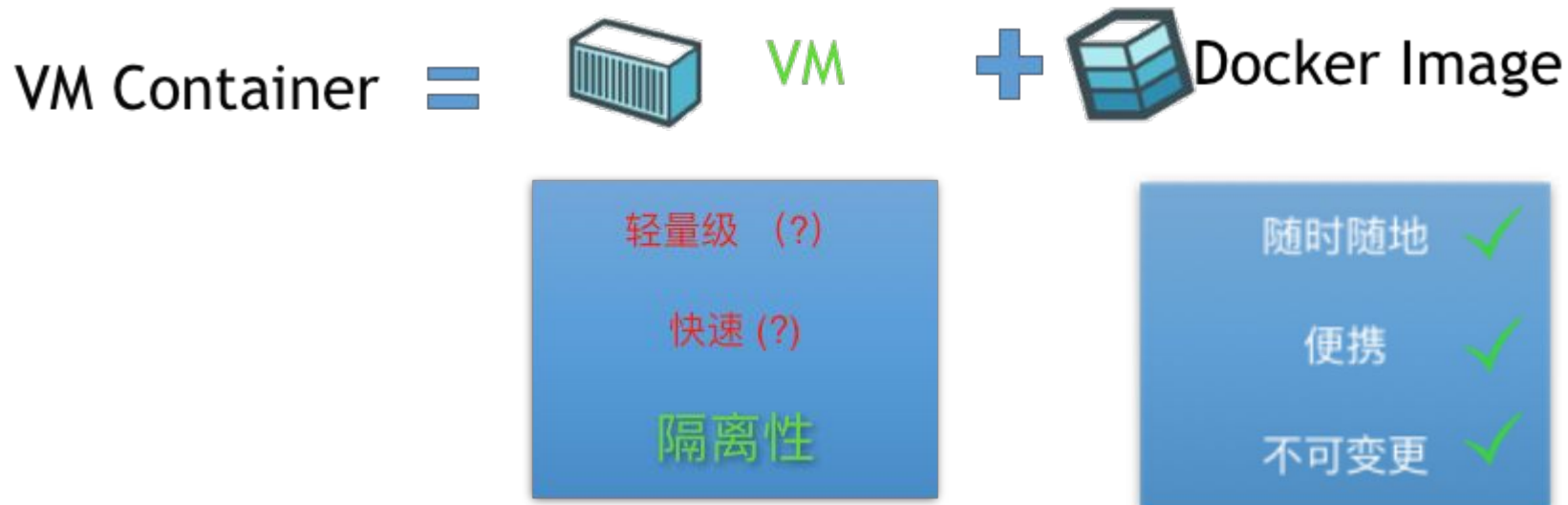
Container Runtime



Kata Containers



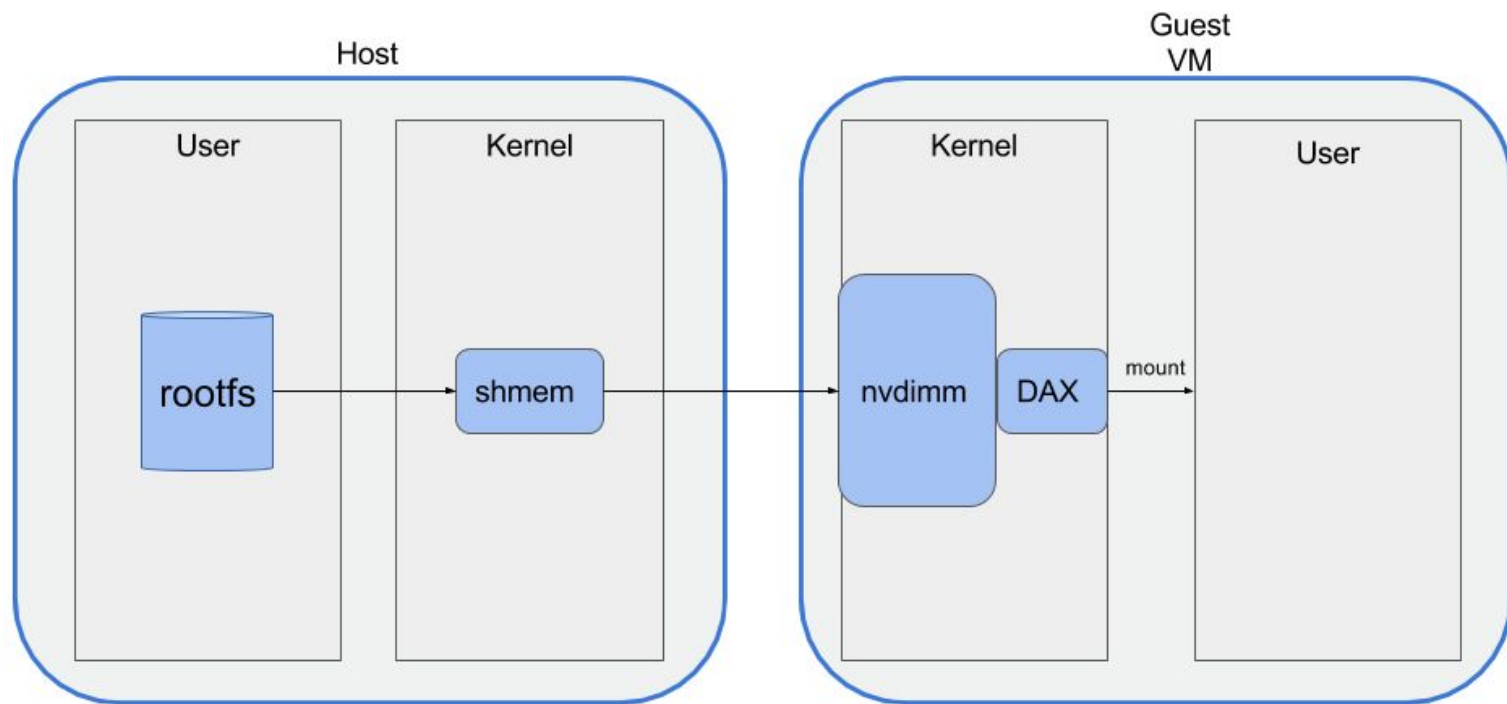
VM-based Containers



Lightweight and Fast

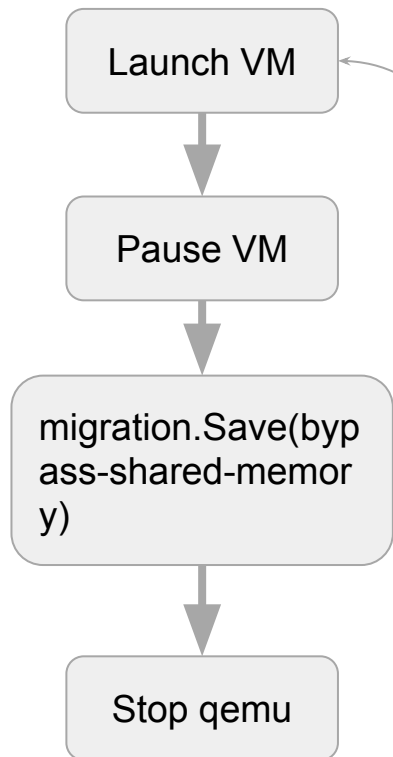
- Customised guest kernel
- Highly optimized qemu-lite
- DAX/nvdimmem: map rootfs image to guest memory
- VM templating: Boot from VM templates to share all guest kernel, initramfs and initial memory states (patch being reviewed by QEMU upstream)
- VM caching: pre-boot guests

DAX/nvdim

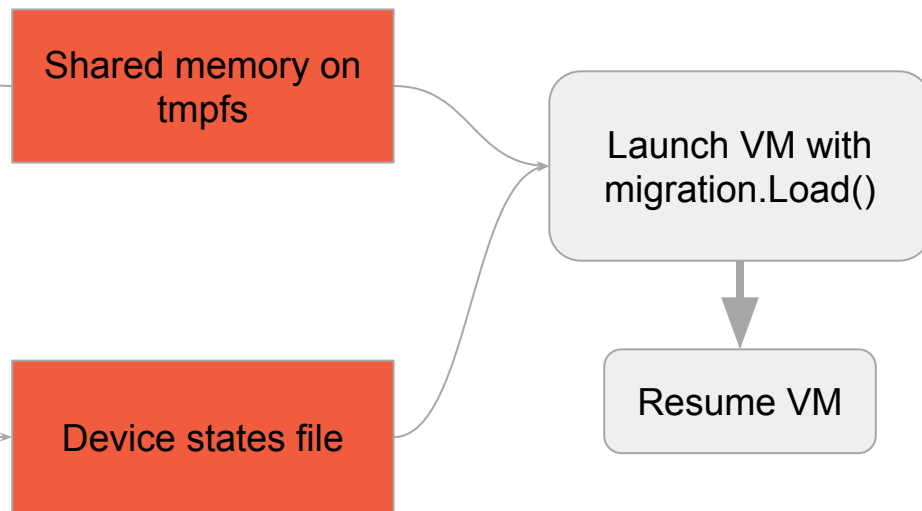


VM Templating

Create VM Template



Create from VM Template



Lightweight and Fast

- Sub-second boot up time
- Very Small memory overhead
 - Small Qemu process footprint
 - Guest kernel/initramfs/rootfs/kata-agent neglectable

Kata Containers and OCI

- OCI runtime-spec compatible
- Replace runc in Docker
 - runc command line compatibility
 - Support docker CNM
 - `Docker run -d --runtime kata nginx`

However...

- OCI runtime-spec has some assumptions based on linux container
- But we are running VMs!!

Kata Containers and OCI

- Missing VM related configuration
 - Hypervisor, kernel, initramfs, rootfs image etc.
 - Default config file loaded on each invocation
 - VM description merged to OCI runtime-spec

Kata Containers and OCI

- Missing storage description -- rootfs and volumes are host directories
 - Map through shared file system -- 9pfs
 - Detects device mapper devices and passthrough to the guest if found
 - Still problematic

Kata Containers and OCI

- Device description based on device major/minor
 - Host device major/minor makes no sense to guests
 - Look for the device with major/minor
 - Hotplug to the guest

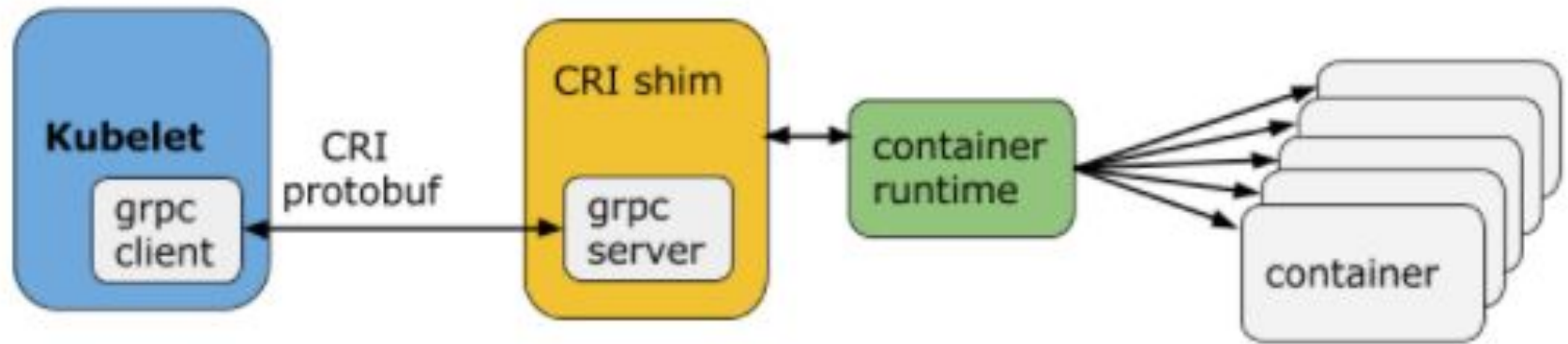
Kata Containers and OCI

- CPU and Memory limits description based on Linux cgroups
 - Approximate conversion
 - vCPU: $(\text{quota} + \text{period} - 1) / \text{period}$
 - Memory: `Memory.Limit >> 20`

Kubernetes CRI

- Plugin interface between kubelet and container runtime
- Allow multiple competing container runtime implementations
 - docker-shim
 - cri-o
 - containerd cri plugin
 - frakti

Kubernetes CRI



Kata Containers and Kubernetes

- Kubernetes pod
 - Multiple containers in a sandbox
 - Smallest schedule unit in a cluster
 - Boundary for resource isolation and sharing
- VM matches pod perfectly in Kata containers

Kata Containers and CRI

- Integration with runc compatible CLI
 - cri-o, containerd cri plugin, docker-shim
 - Missing sandbox abstraction
 - CRI CreateSandbox converted to pause container creation
- OCI runtime-spec limitations
 - Missing storage description

Kata Containers and CRI

- Integration with CRI native runtime APIs
 - frakti
 - No restriction from runc CLI compatibility
 - No pause container
 - CRI CreateSandbox creates VM
- Supports different storage types
 - Local block devices, Ceph rbd, iSCSI, NFS etc.

How to Contribute

- Code
 - <https://github.com/kata-containers>
- Slack
 - <https://katacontainers.slack.com/>
- Mailing List
 - <http://lists.katacontainers.io/>
- Freenode
 - #kata-dev
 - #kata-general

Future

- GPGPU for machine learning
- Edge and IoT
- Linuxd (“Run Linux Kernel as a Daemon”)

Thank YOU!

Q & A

We are hiring!!!
jobs@hyper.sh

