Tencent 腾讯

Tianus

A Secure Serverless Kubernetes Solution

Jike Song <jikesong@tencent.com>
Yun Wang <yunxwang@tencent.com>

0 Who are we

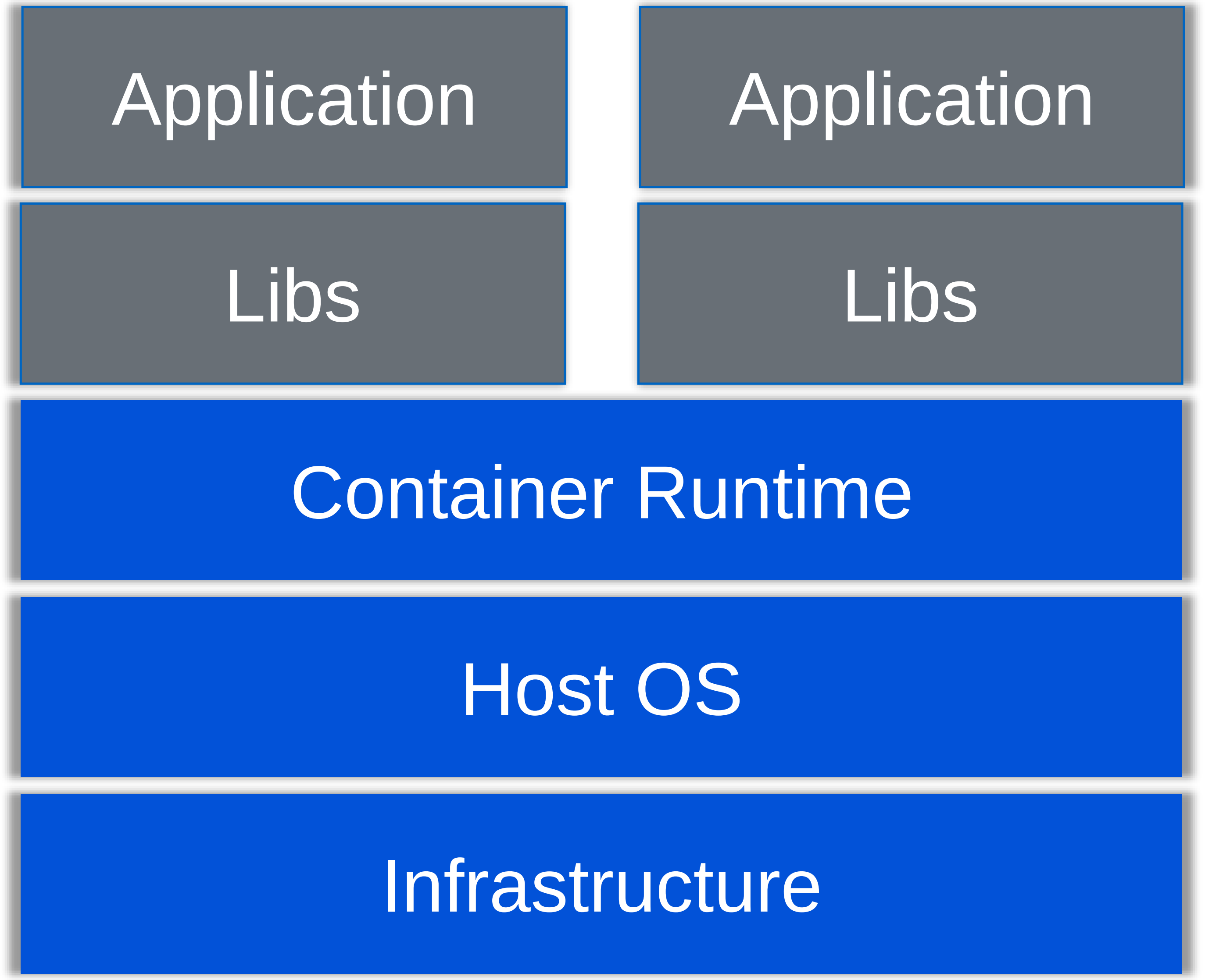
- Tencent Cloud Virtualization team
- Responsible for Tencent Elastic Kubernetes Service (EKS) wrt. Virtualization
- Resumes are welcome

Agenda

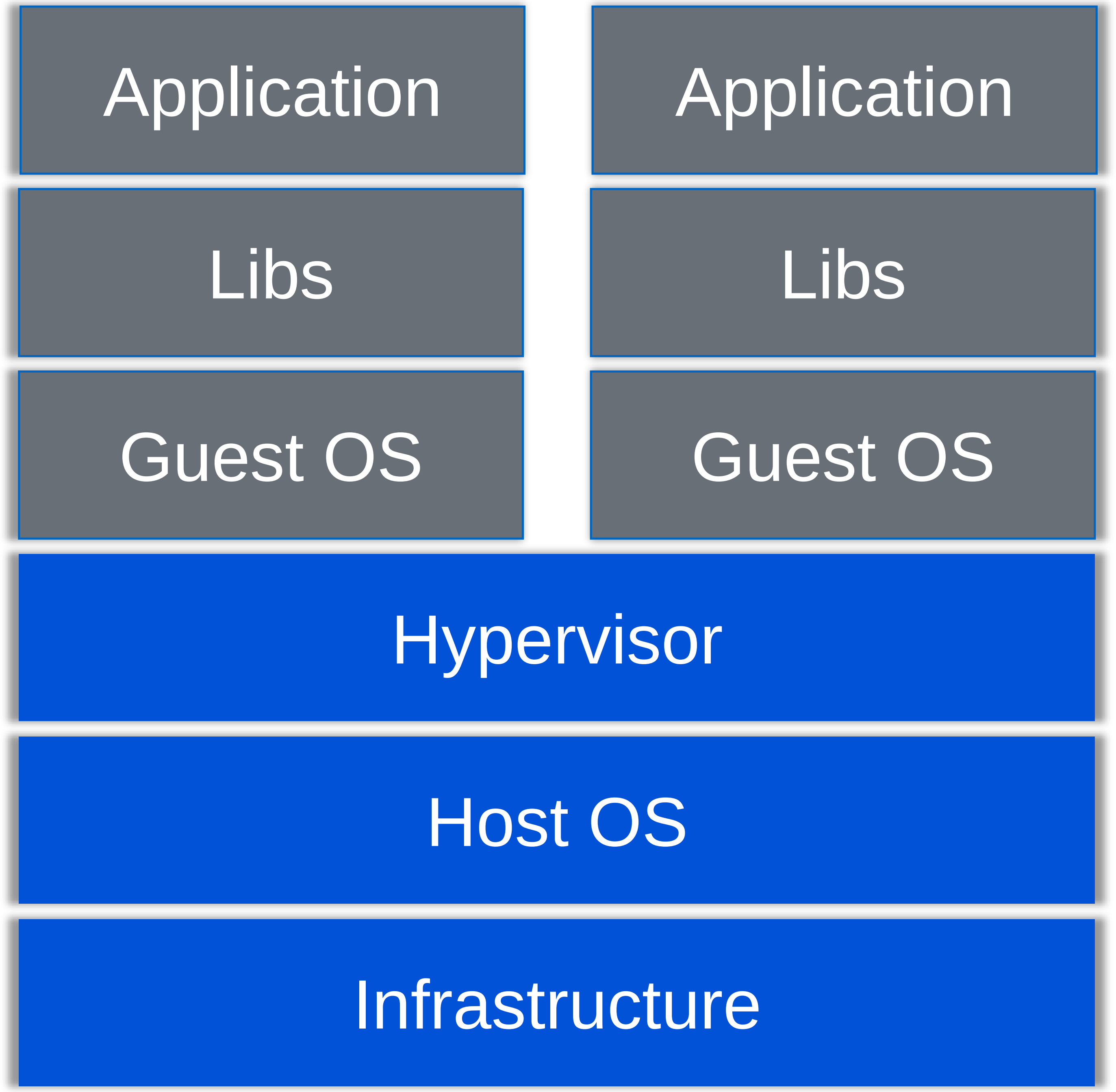
- Background
- Tianus Overview
- Technical Deep Dive

1 Background

Multi-tenant Container vs. VM



Container Deployment



Virtualized Deployment

Container Orchestration

Container

- Placement
- Replication/Scaling
- Upgrades/Downgrades
- …

Resource

- Memory
- CPU
- GPU
- …

Service Management

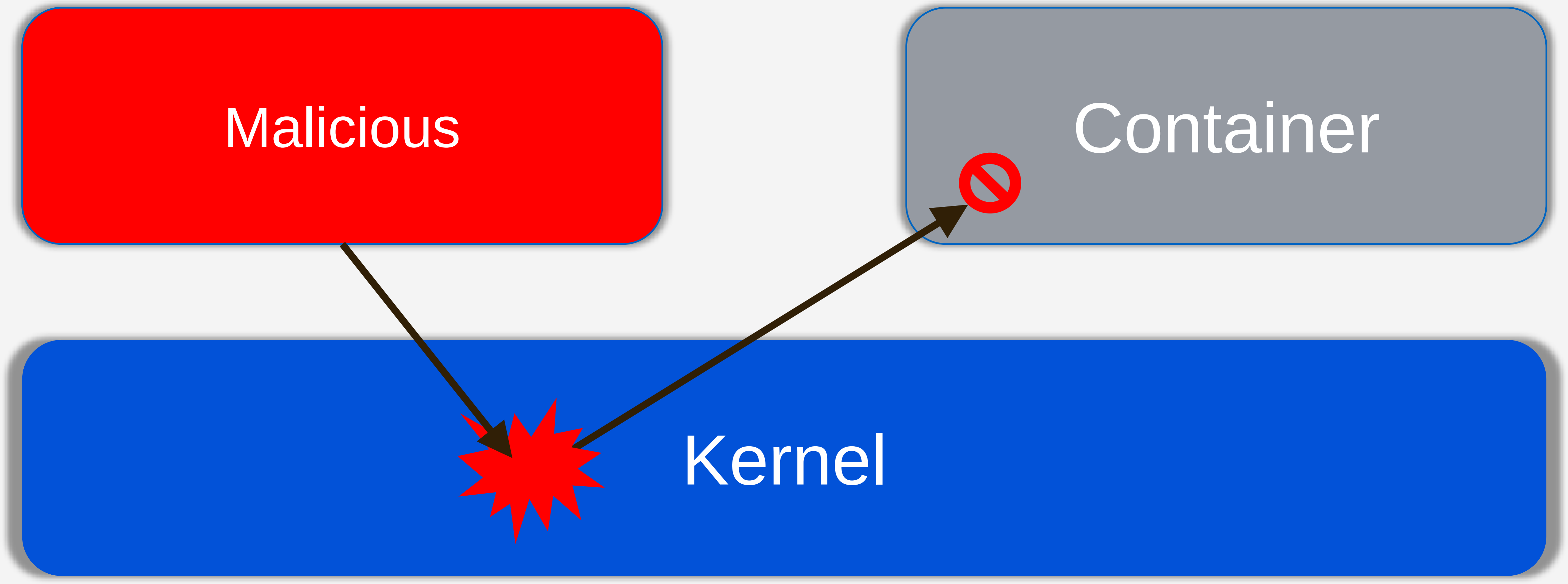
- Labels
- Load Balancing
- Readiness Checking
- …

The De facto standard: Kubernetes

- Kubernetes is a future-proof solution
- One of the top open-source projects of all time
- Has a huge ecosystem and tons of resources around it.

Container Internals

- Container facilities `Namespaces` and `cgroups` to achieve isolation.
- All containers on the machine shares host’s kernel
- Kernel vulnerabilities will compromise the security.

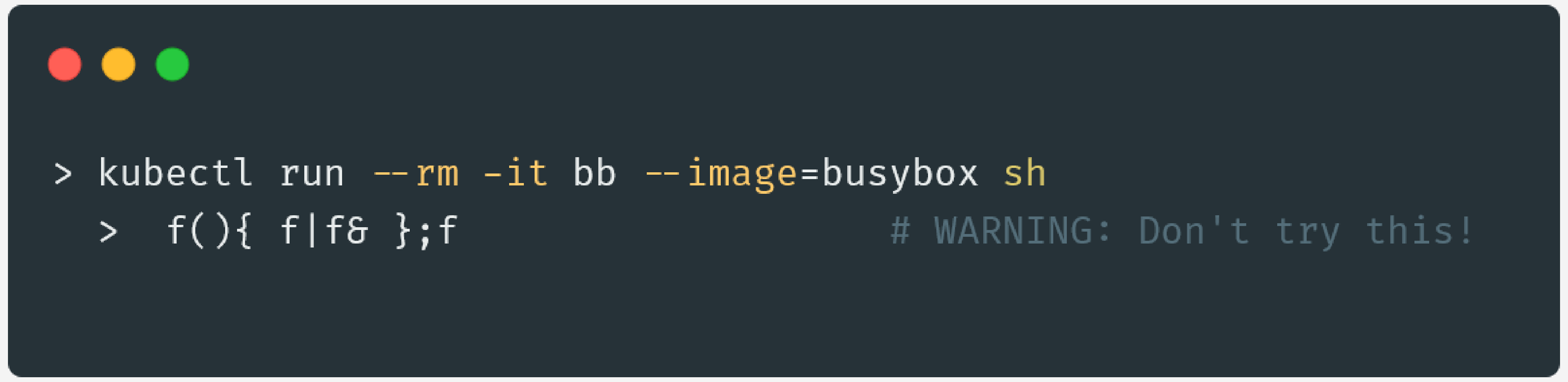


Isolation Challenges

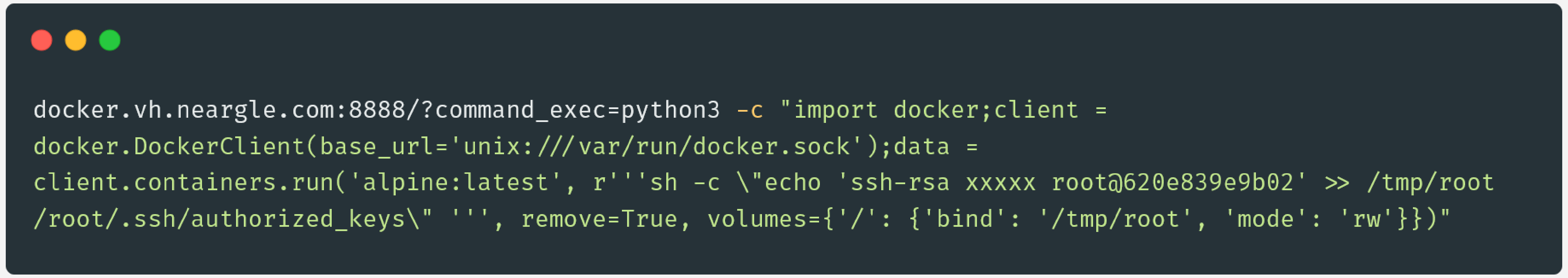
- Container Escape
  - CVE-2018-14634 Integer overflow vulnerability
  - CVE-2016-5195 Dirty COW vulnerability
  - CVE-2019-5736 Docker runc vulnerability
  - CVE-2019-14271 Docker CP vulnerability

- Noisy neighbor

– Impact performance on CPU, Memory, Bandwidth, Buffer IO, PIDs, File descriptors

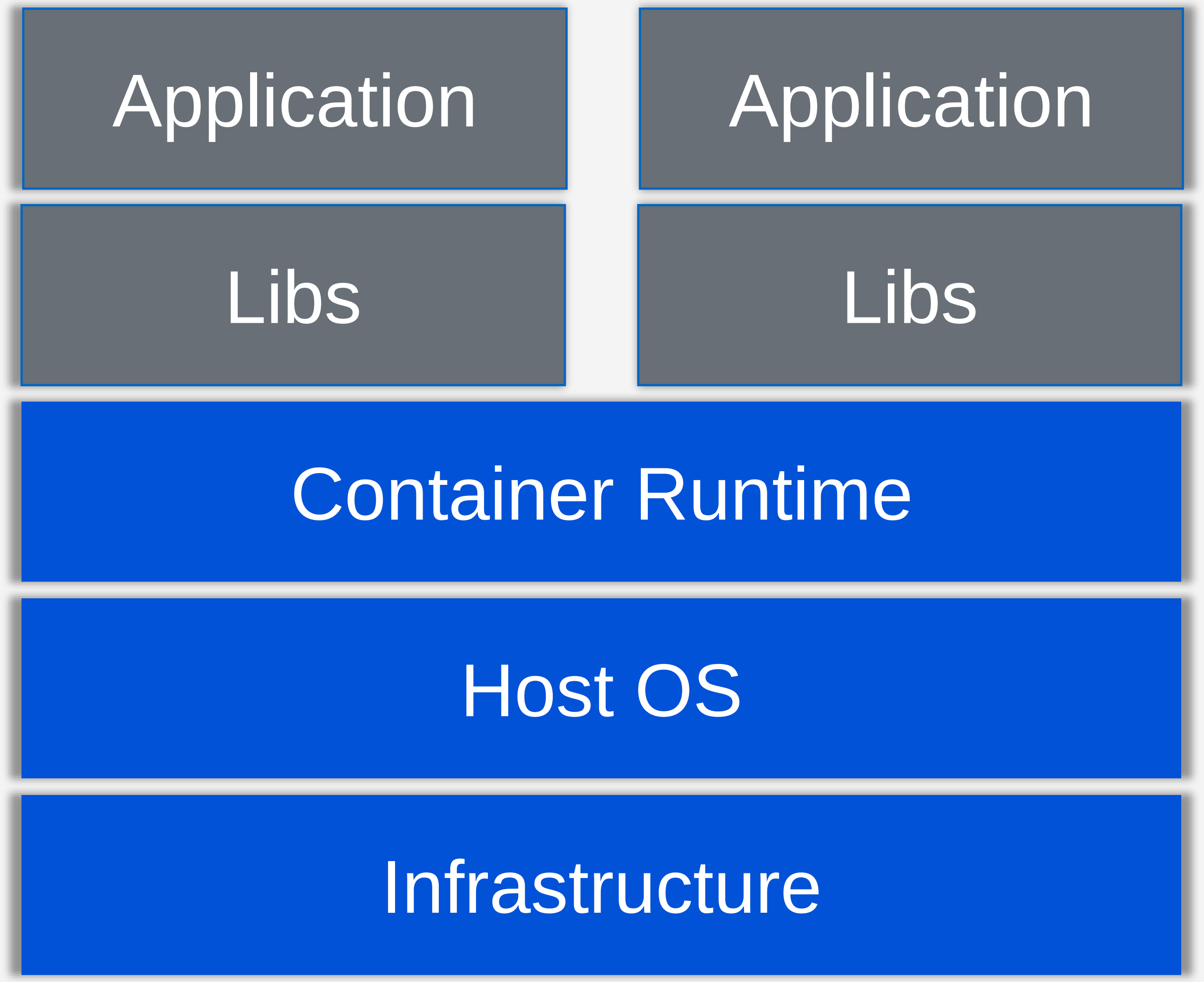


```
> kubectl run --rm -it bb --image=busybox sh
  >  f(){ f|f& };f                      # WARNING: Don't try this!
```

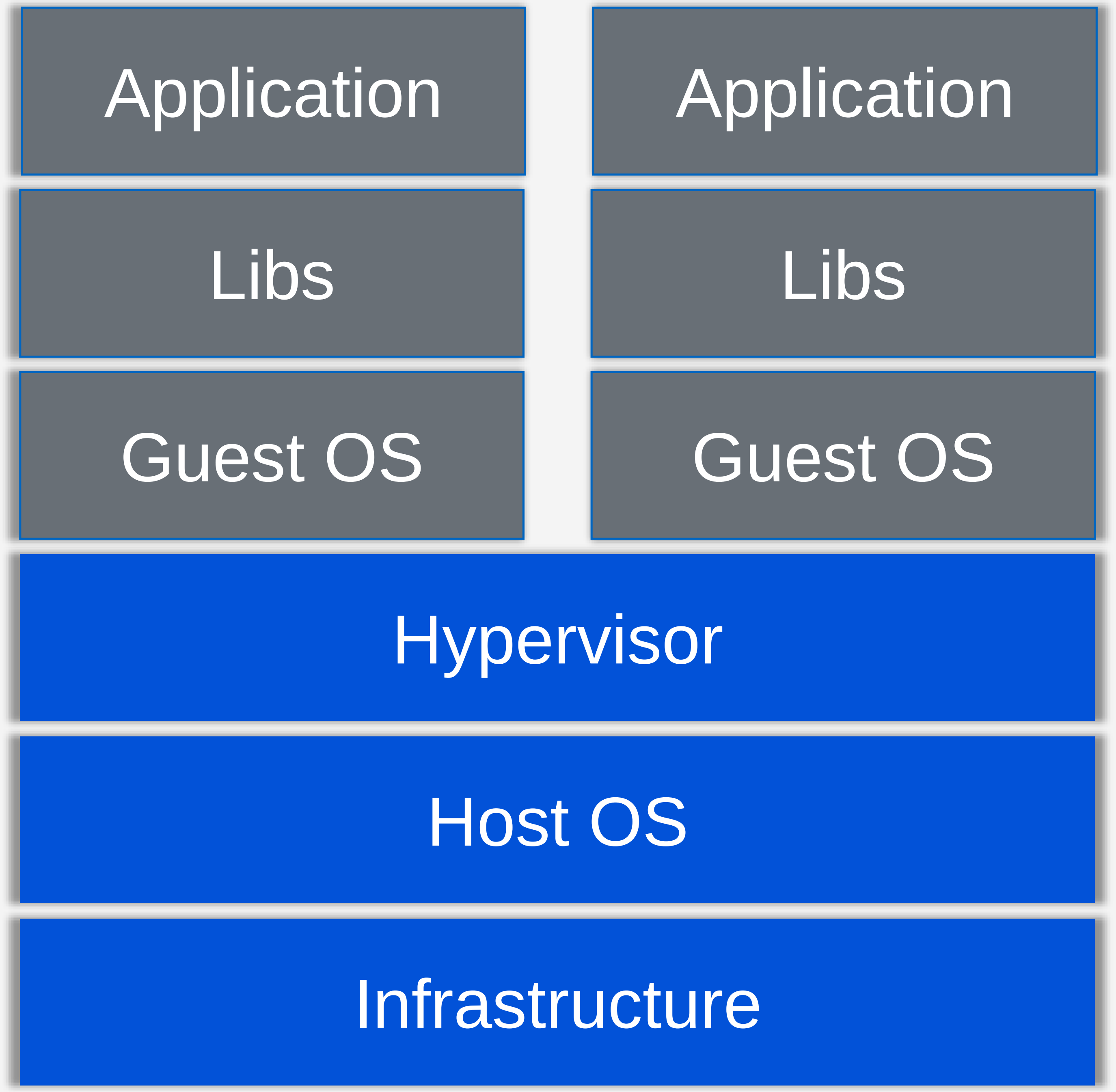


```
docker.vh.neargle.com:8888/?command_exec=python3 -c "import docker;client =
docker.DockerClient(base_url='unix:///var/run/docker.sock');data =
client.containers.run('alpine:latest', r'''sh -c \"echo 'ssh-rsa xxxxx root@620e839e9b02' >> /tmp/root
/root/.ssh/authorized_keys\" ''', remove=True, volumes={'/': {'bind': '/tmp/root', 'mode': 'rw'}})"
```

Multi-tenant Container vs. VM (revisited)



Container Deployment



Virtualized Deployment

Container Runtime

cri-o

containerd

OPEN CONTAINER INITIATIVE runc

docker

LXC

podman

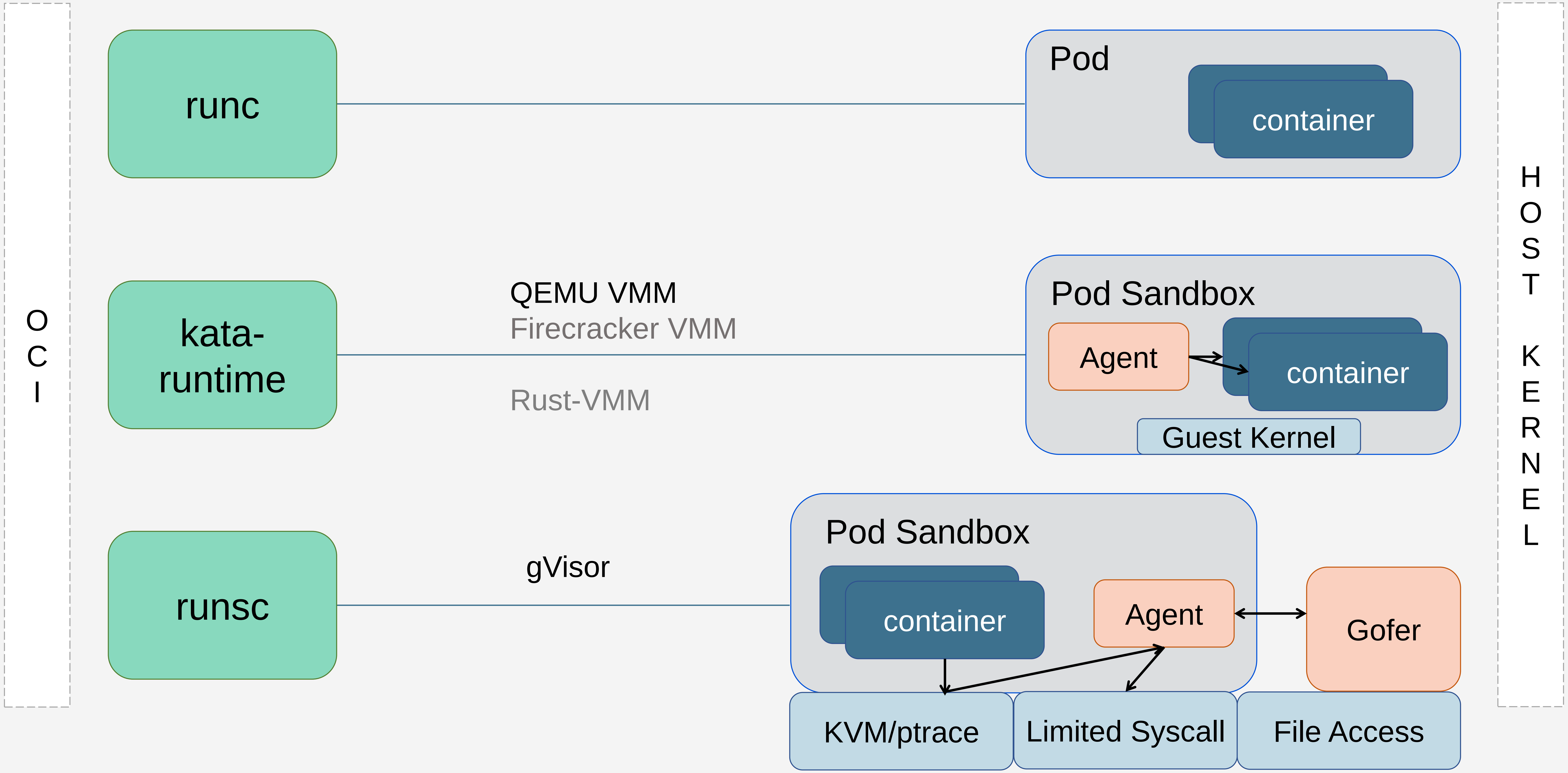
Secure Container Runtime

gVisor

Firecracker

kata containers

OCI Implementation



Comparison

Solution	Typical Software	Pros & Cons
Native Container	RunC	- Fast, low overhead - Little isolation
VM-based Sandbox	Kata-container	- Safest - Performance & Resource overhead
User Space Kernel	gVisor	- Reduced attack interface - Limited Syscalls - Special build

2 Tianus Overview

What is Tianus

- Goal: A Secure Container Runtime for Kubernetes in a Serverless Manner
- Widely used by Tencent Elastic Kubernetes Service (EKS).
- Will be open-source in the near future

2 Overview

Secure | CRI-Compatible | Serverless

KVM + cri-o + Virtual Kubelet = **Tianus**

Tencent + Ianus

Ianus (IPA: / ˈ*ja:nus*/), in ancient Roman religion and myth, is the god of beginnings, gates, transitions, time, duality, doorways, passages, frames, and endings. He is usually depicted as having two faces, since he looks to the future and to the past.

Statue of Ianus

Arch of Ianus

Container Runtime

cri-o

containerd

OPEN CONTAINER INITIATIVE
runc

docker

LXC

podman

Secure Container Runtime

gVisor

Firecracker

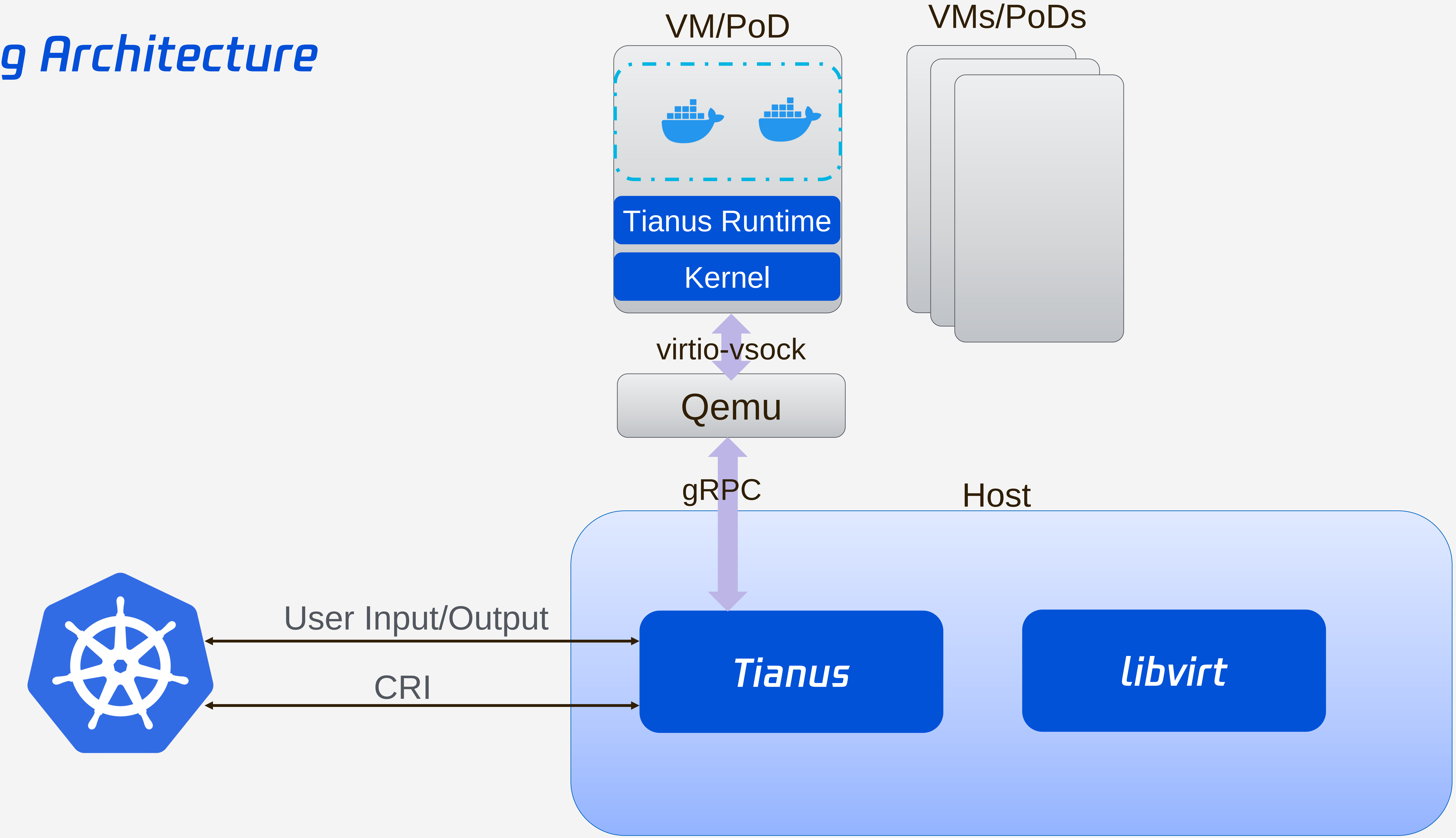
kata
containers

Why Tianus

- Existing Secure Runtime, e.g. Kata container, is not suitable for Cloud.
  - I/O is slow.
  - The isolation in not good enough. (e.g.: privileged container, host-guest file sharing)
  - GPU is so important in the cloud computing, but GPU hot-plug is problematic.
  - Live Migration support for rich containers is absent.
  - Hybrid Deployment with VM is difficult

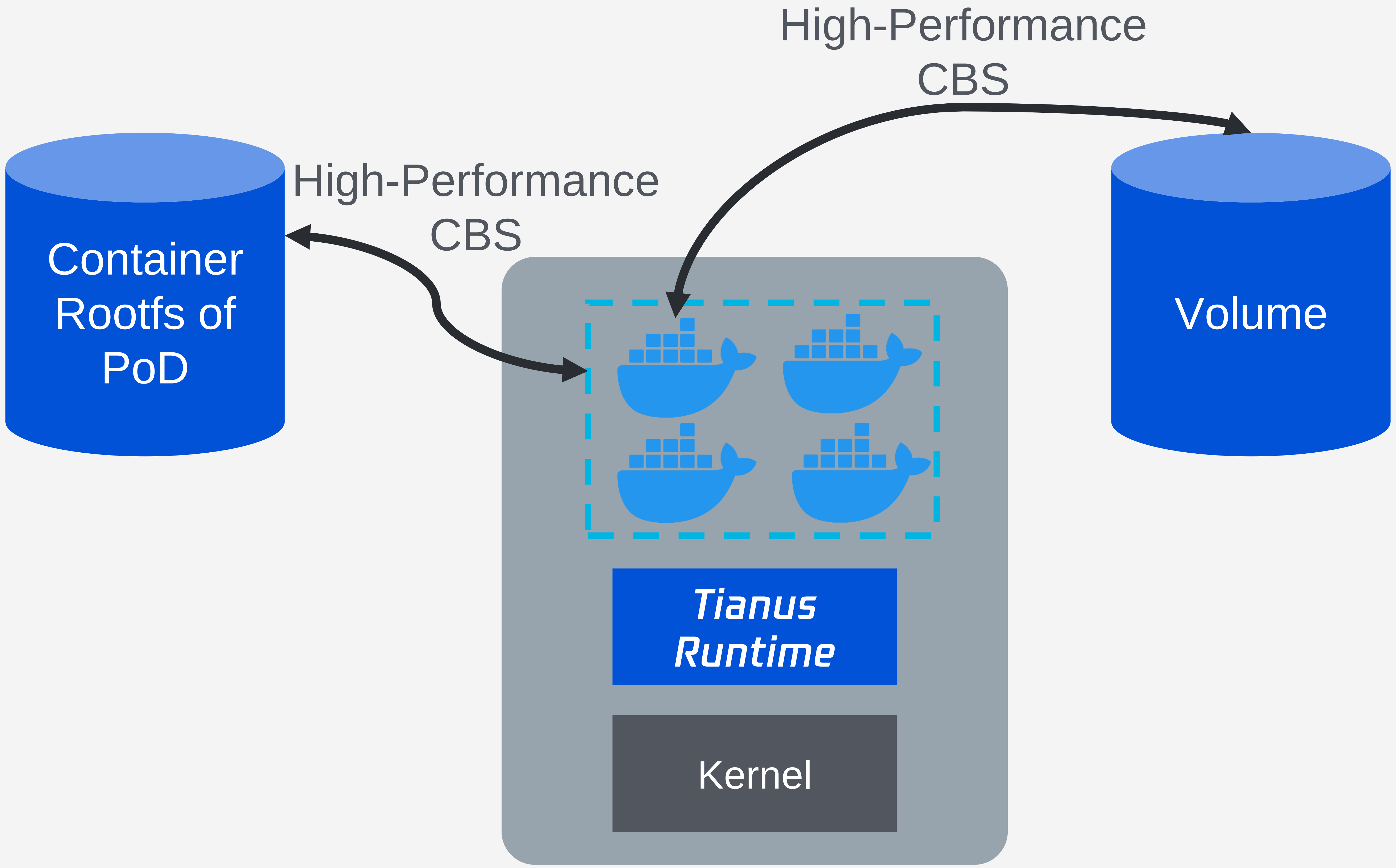
3 Technical Deep Dive

Big Architecture



Speed

- No fs/blk-device-sharing mechanism needed.
- Great I/O performance benefits from high-performance CBS.
- Reuse CBS for every tenant to speed up boot time.
- No file-sharing between host and guest reduces the attack surface.

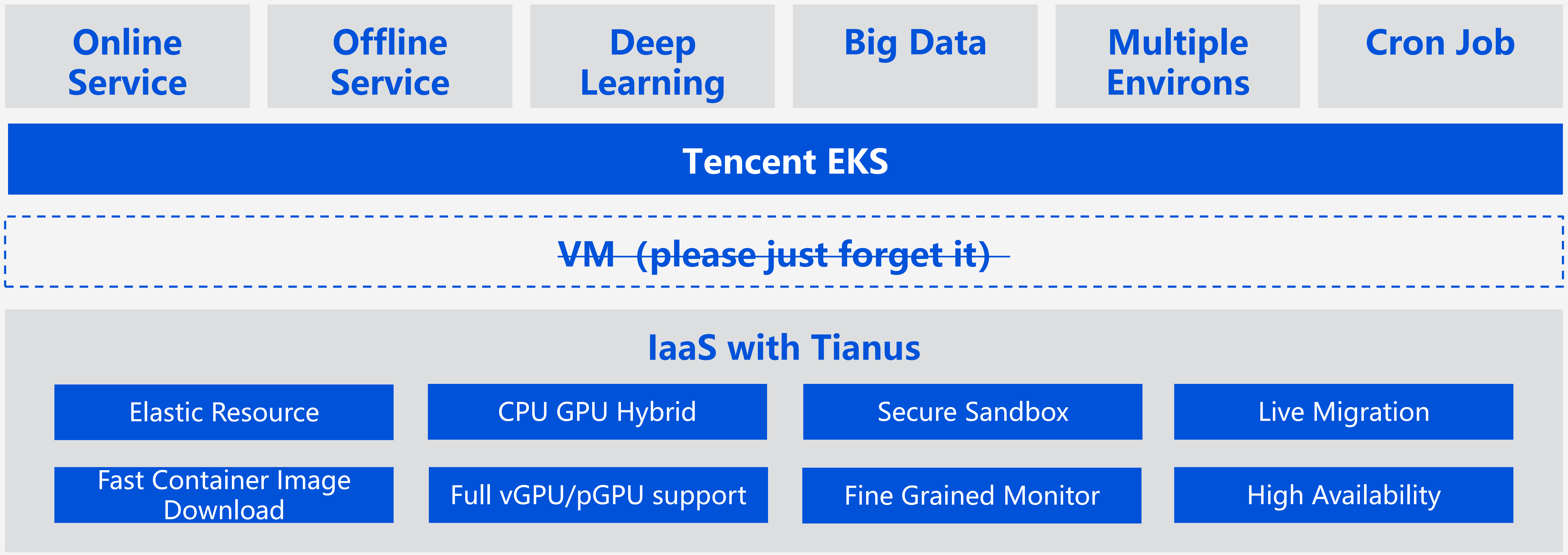


Tianus Runtime

- Tlinux OS based rootfs
- Customized, Optimized container runtime
- Listening via VSOCK instead of Unix-Domain/TCP socket

Serverless

Focus on Applications/Containers, not the container platform



Online Service

Offline Service

Deep Learning

Big Data

Multiple Environs

Cron Job

Tencent EKS

~~VM (please just forget it)~~

IaaS with Tianus

Elastic Resource

CPU GPU Hybrid

Secure Sandbox

Live Migration

Fast Container Image Download

Full vGPU/pGPU support

Fine Grained Monitor

High Availability

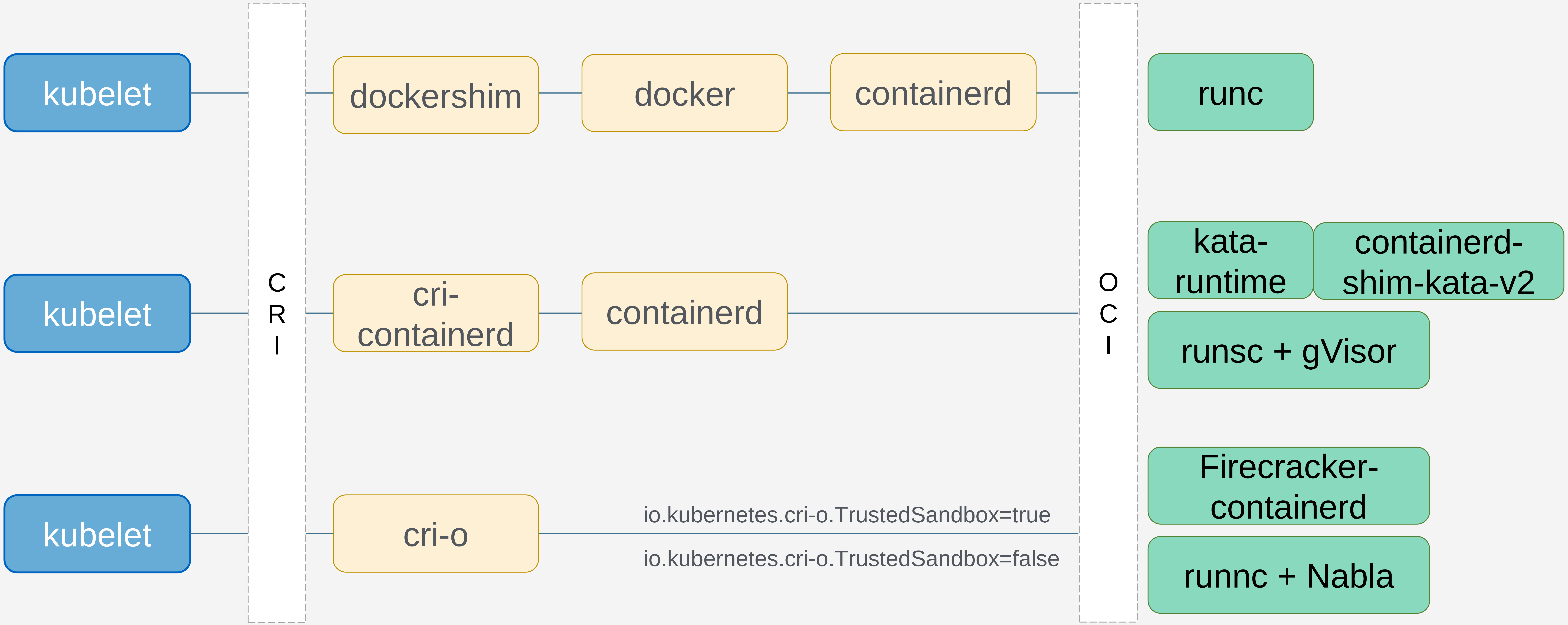
Conclusion

- In essence, Tianus is a high-performance CRI router.
- It provides VM based, fast and secure containers
- Fully fledged with all features one might want:
    - K8S compatibility
    - pGPU/vGPU support
    - Live Migration
    - Hybrid Deployment: { online, offline } x { container, VM }
    - Open-Source

Thanks

Backup

K8s Runtime


kubelet
C
R
I
dockershim
docker
containerd
O
C
I
runc
kubelet
cri-
containerd
containerd
kata-
runtime
containerd-
shim-kata-v2
runsc + gVisor
kubelet
cri-o
io.kubernetes.cri-o.TrustedSandbox=true
io.kubernetes.cri-o.TrustedSandbox=false
Firecracker-
containerd
runnc + Nabla

Comparison

OCI Solution	OCI Compatible	Dedicated Docker Image	Implementation Language	Open source	Hot-plug	Direct access to HW	Required Hypervisors	Backed by
Runc	Yes	Yes	Golang	Yes	No	Yes	None	Docker
gVisor+runsc	Yes	Yes	Golang	Yes	No	No	None or KVM	Google
Kata+qemu	Yes	Yes	Golang, C	Yes	Yes	Yes	KVM	Hyper
Firecracker+ Firecracker-containerd	No	Yes	Rust, Golang	Yes	No	No	KVM	Amazon
Nabla+runnc	Yes	No	C, Golang	Yes	No	No	None	IBM