
KubeCon
CloudNativeCon
OPEN SOURCE SUMMIT
China 2019

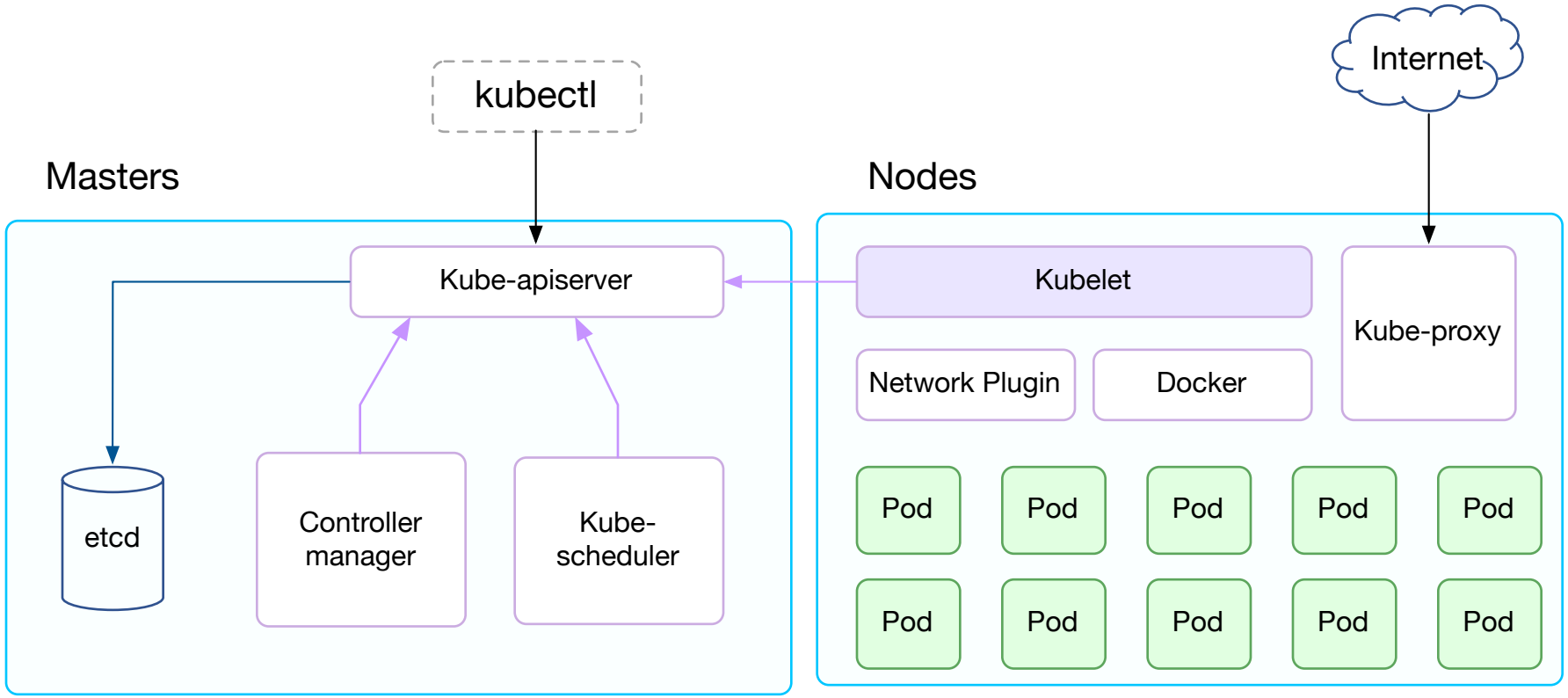
Container Runtime Evolving in Kubernetes

Pengfei Ni, Microsoft Azure

目录

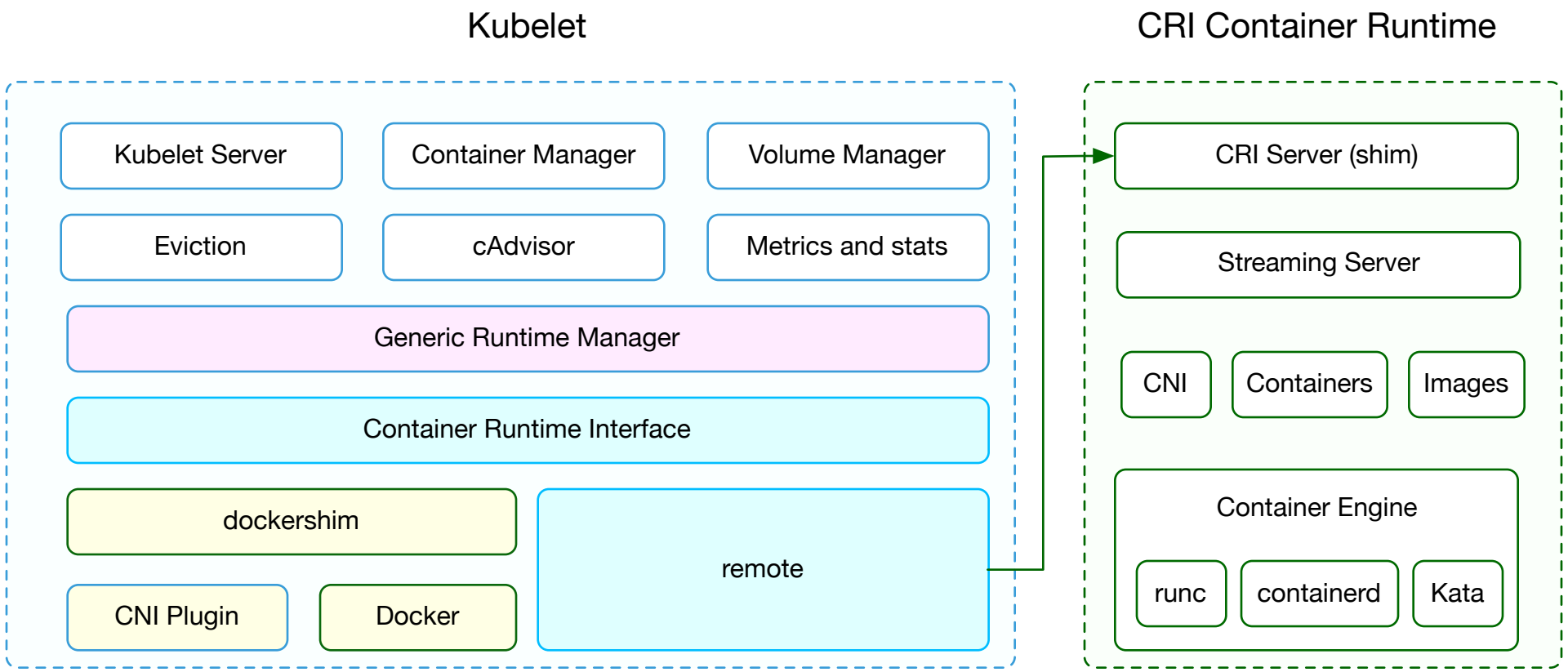
- Kubernetes 架构简介
- 容器运行时接口（CRI）
- 容器运行时演进
- 未来展望

Kubernetes 架构



Kubelet 架构

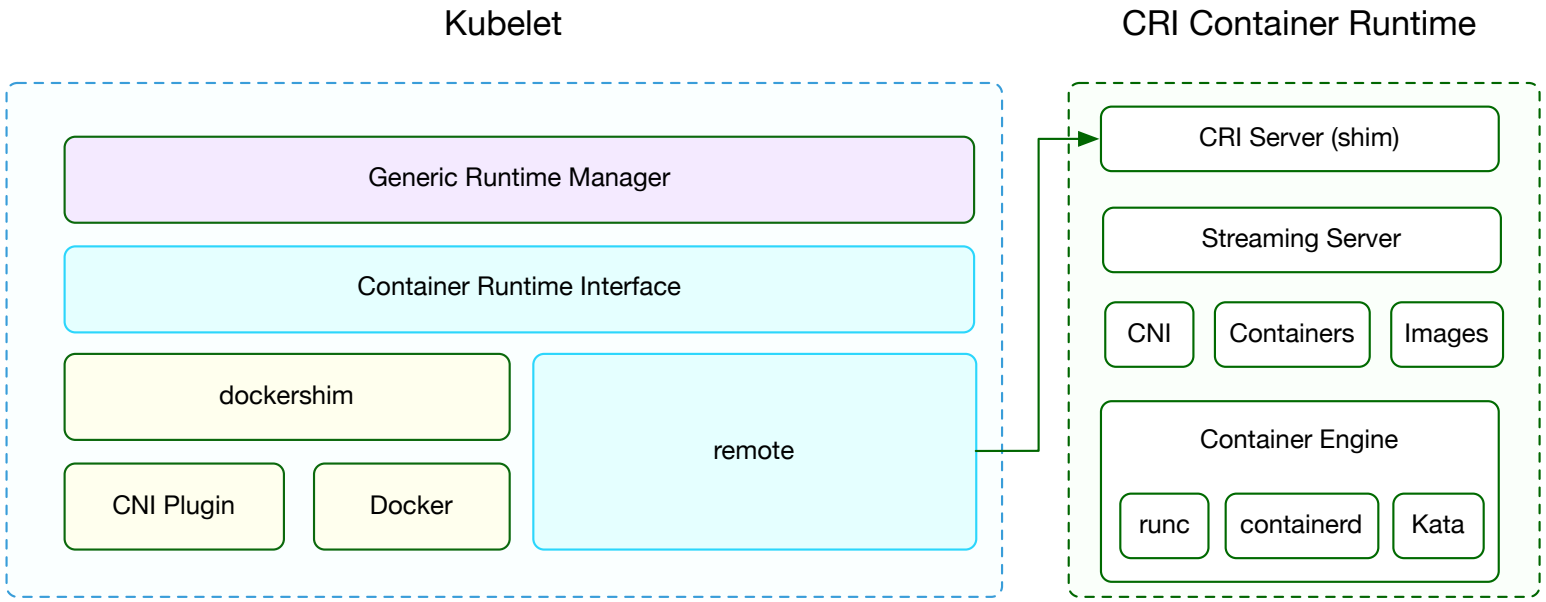
KubeCon CloudNativeCon OPEN SOURCE SUMMIT China 2019



Kubelet 容器运行时

KubeCon CloudNativeCon
OPEN SOURCE SUMMIT
China 2019

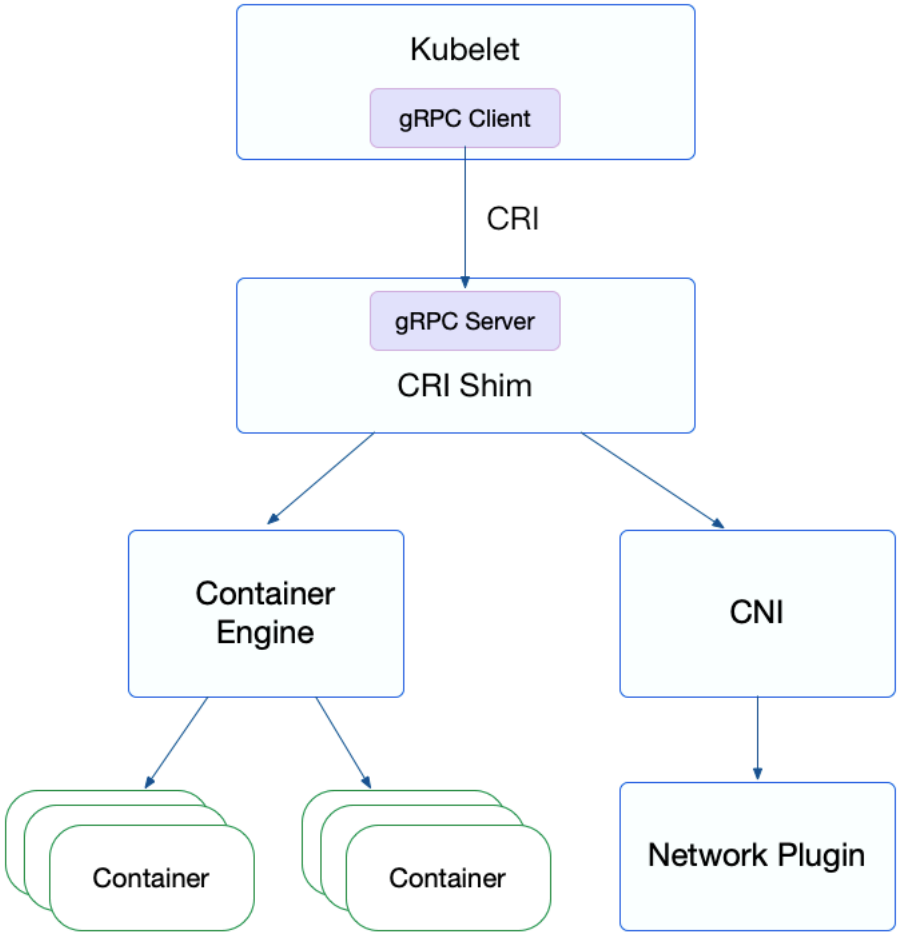
- Kubelet 通过 CRI 接口管理容器运行时
- 容器运行时负责管理容器、镜像和网络
- Kubelet 内置 Docker 和 CNI 网络插件的支持
- 通过容器运行时接口（CRI）支持外部运行时
  - cri-containerd
  - cri-o



容器运行时接口（CRI）

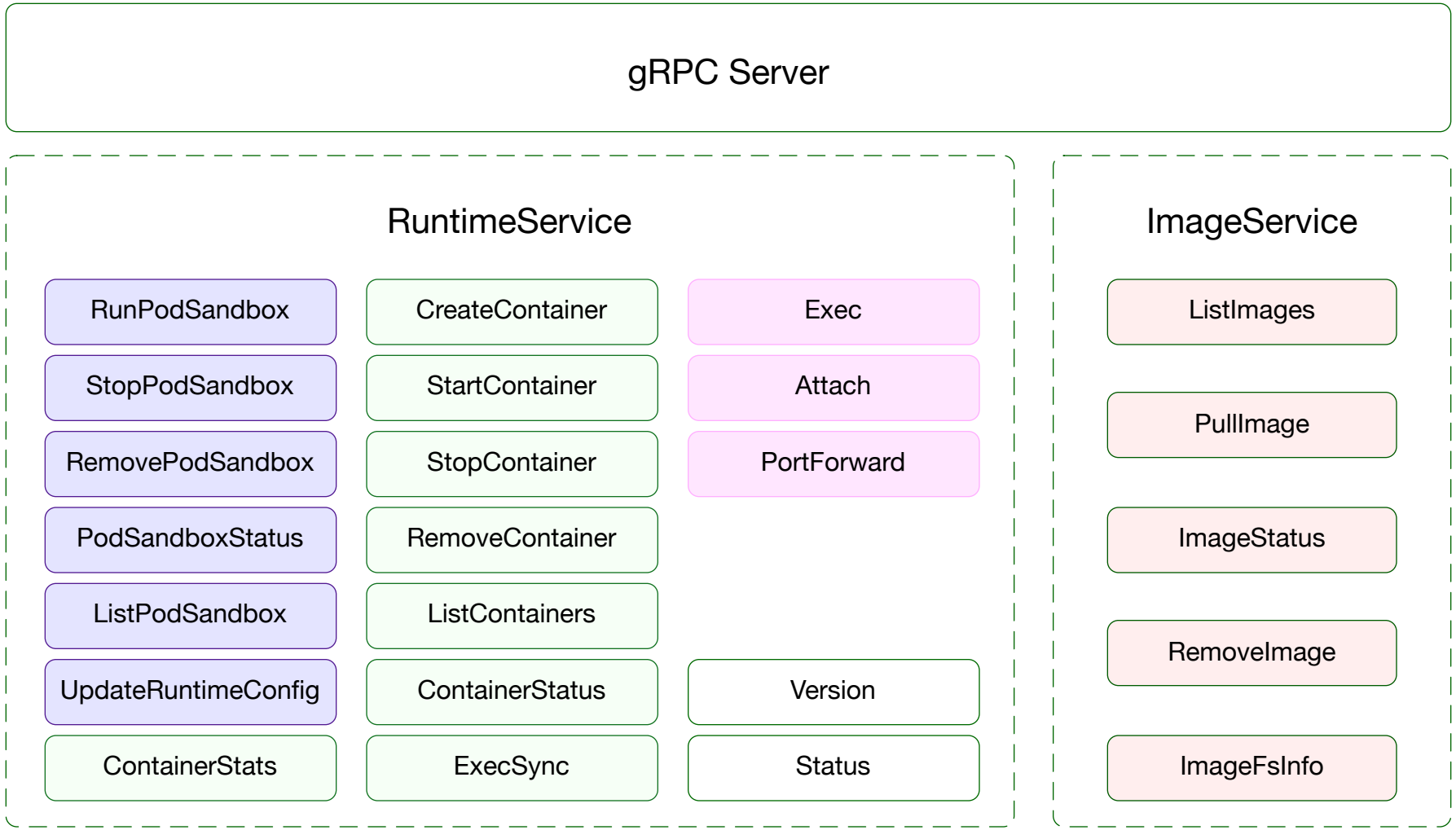
KubeCon CloudNativeCon OPEN SOURCE SUMMIT China 2019

- Container Runtime Interface (CRI) 组成
  - 基于 gRPC 的容器运行时接口
  - Streaming server 库
  - CRI Tools
- 外部容器运行时实现 gPRC 服务端，kubelet 作为客户端访问
- 无需关心内部通信逻辑，只需实现 RuntimeService 和 ImageService 接口
- 容器运行时负责管理容器网络，如使用 CNI 等
- 社区多种实现，如 cri-containerd、cri-o等
- Kubelet 配置 --container-runtime-endpoint



CRI 接口

KubeCon CloudNativeCon OPEN SOURCE SUMMIT China 2019

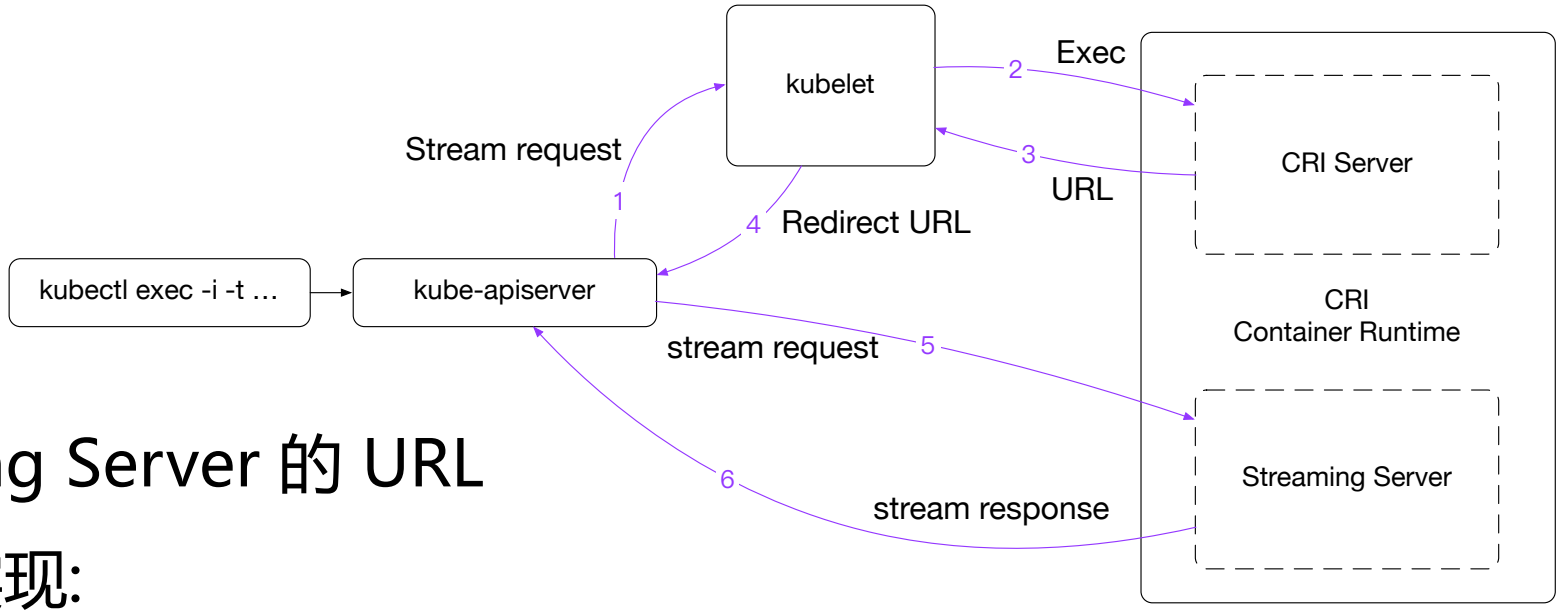


CRI Streaming 接口

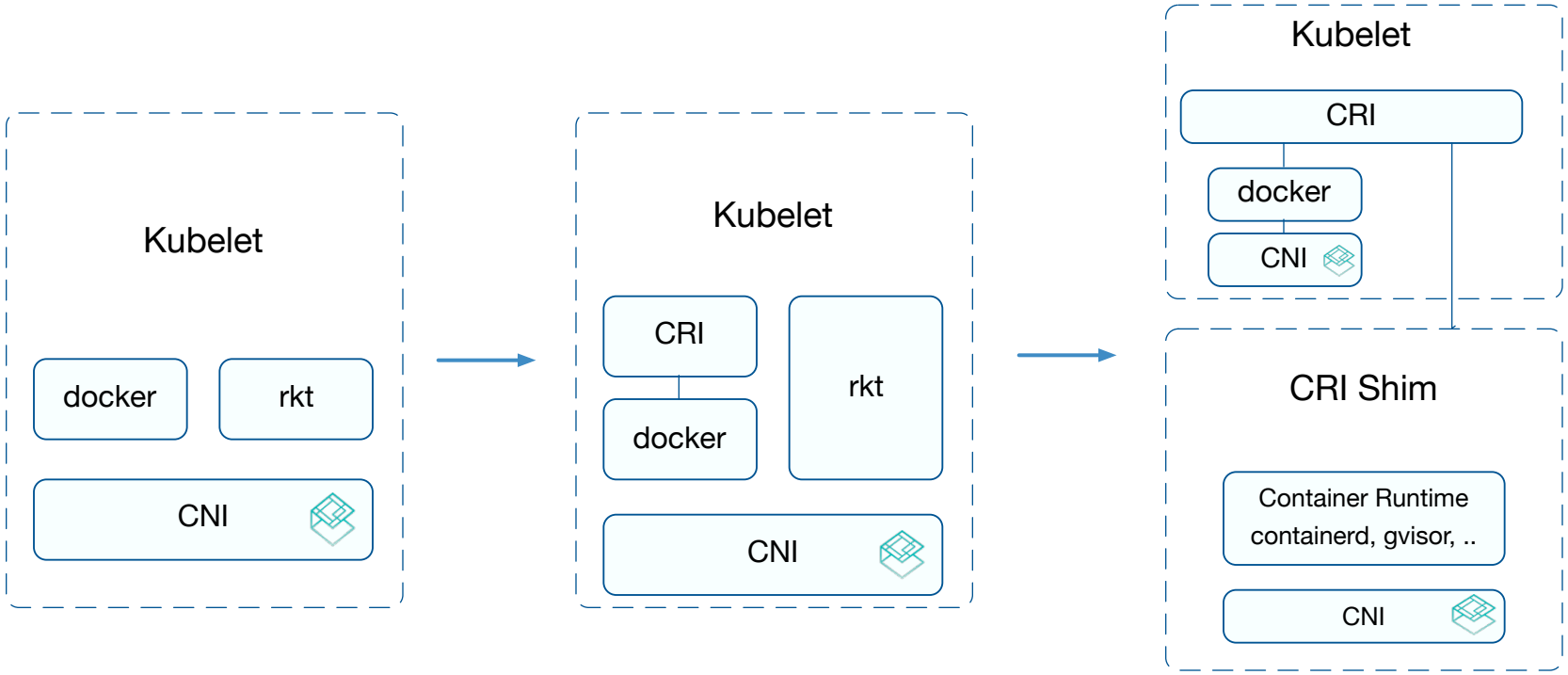
KubeCon CloudNativeCon
OPEN SOURCE SUMMIT
China 2019

- Streaming 接口
  - Exec
  - Attach
  - PortForward
- 运行时返回 Streaming Server 的 URL
- Streaming Server 实现:

  https://github.com/kubernetes/kubernetes/tree/master/pkg/kubelet/server/streaming



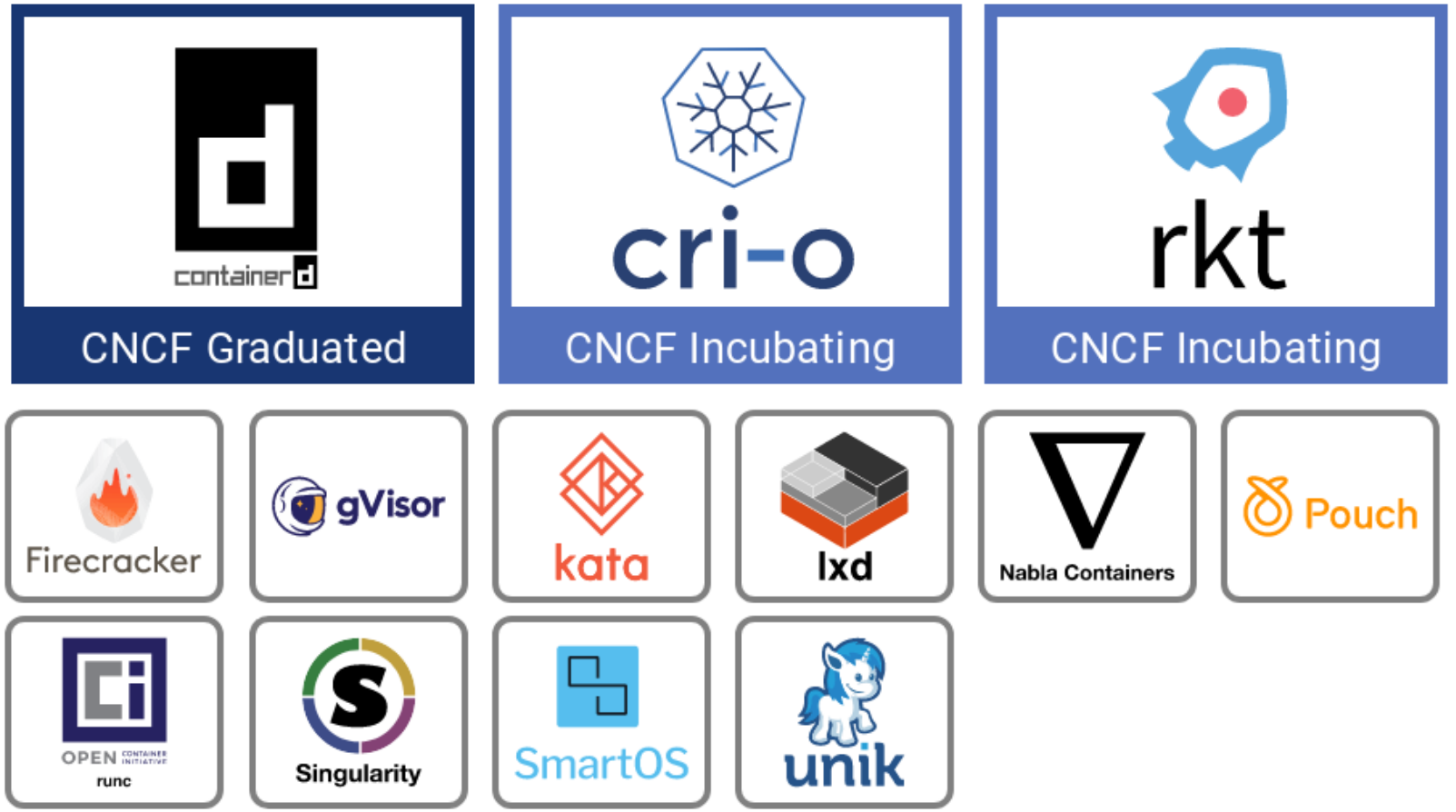
容器运行时演进


Kubelet
docker
rkt
CNI
Kubelet
CRI
docker
rkt
CNI
Kubelet
CRI
docker
CNI
CRI Shim
Container Runtime
containerd, gvisor, ..
CNI

CRI 容器运行时

CRI 容器运行时	维护者	主要特性	容器引擎
Dockershim	Kubernetes	内置实现、特性最新	docker
cri-o	Kubernetes	OCI标准不需要Docker	OCI（runc、kata、gVisor...）
cri-containerd	Containerd	基于containerd 不需要Docker	OCI（runc、kata、gVisor...）
Frakti	Kubernetes	虚拟化容器	hyperd、docker
rktlet	Kubernetes	支持rkt	rkt
PouchContainer	Alibaba	富容器	OCI（runc、kata...）
Virtlet	Mirantis	虚拟机和QCOW2镜像	Libvirt（KVM）

容器运行时蓝图


containerd
CNCF Graduated
cri-o
CNCF Incubating
rkt
CNCF Incubating
Firecracker
gVisor
kata
lxd
Nabla Containers
Pouch
OPEN CONTAINER INITIATIVE
runc
Singularity
SmartOS
unik

CRI TOOLS - CRICTL

KubeCon CloudNativeCon
OPEN SOURCE SUMMIT
China 2019

- CRI 容器运行时命令行工具（SIG Node 子项目）
- 调试容器运行时的利器，辅助用户排查容器问题
- 简单易用
  - *crictl pods*
  - *crictl ps*
  - *crictl images*
  - *crictl exec*
  - *...*

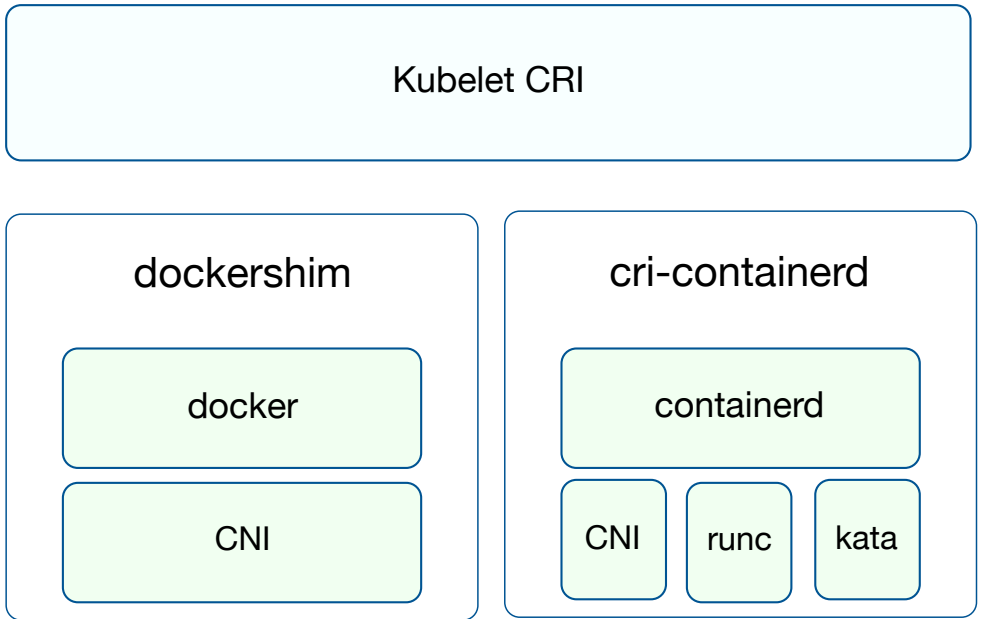
CRI TOOLS - CRITEST

KubeCon CloudNativeCon
OPEN SOURCE SUMMIT
China 2019

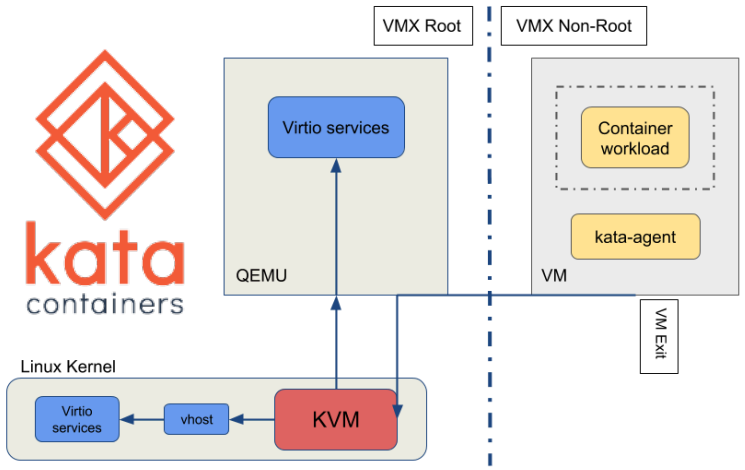
- CRI 容器运行时验证测试工具
- 验证容器运行时是否符合 CRI 要求
- 为 CRI 接口提供性能测试
  - *critest -benchmark*
- 推荐集成到容器运行时 Devops 流程中
- 发布测试结果到 TestGrid

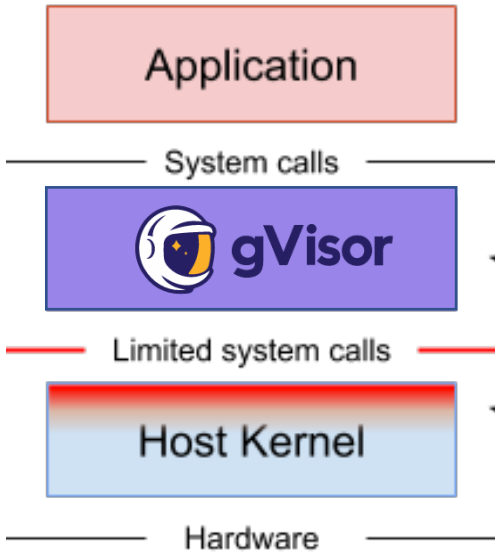
未来展望

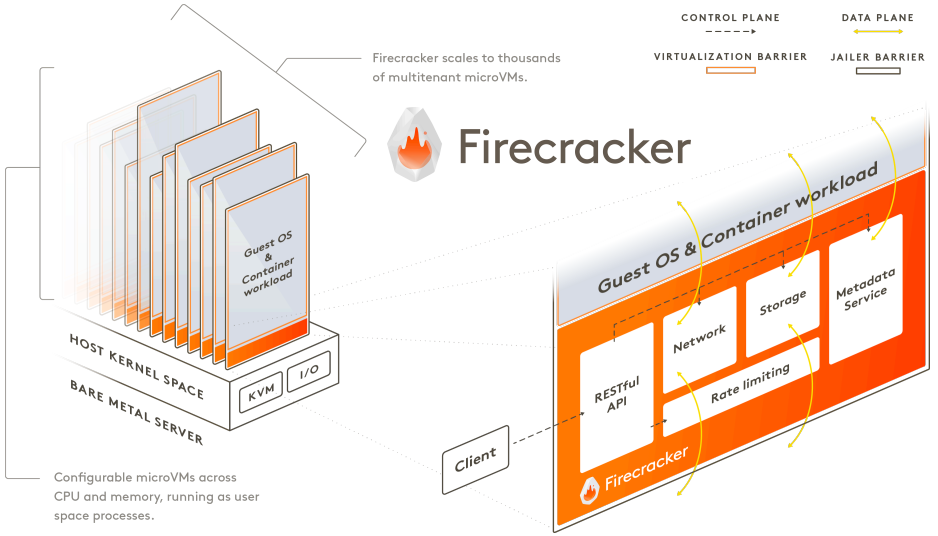
Dockershim 拆分



强隔离容器引擎


VMX Root
VMX Non-Root
kata
containers
Virtio services
QEMU
Container workload
kata-agent
VM
VM Exit
Linux Kernel
Virtio services
vhost
KVM


Application
System calls
gVisor
Limited system calls
Host Kernel
Hardware


Firecracker scales to thousands of multitenant microVMs.
CONTROL PLANE
DATA PLANE
VIRTUALIZATION BARRIER
JAILER BARRIER
Firecracker
Guest OS & Container workload
HOST KERNEL SPACE
BARE METAL SERVER
KVM
I/O
Configurable microVMs across CPU and memory, running as user space processes.
Client
RESTful API
Network
Storage
Metadata Service
Rate limiting
Firecracker

多容器运行时

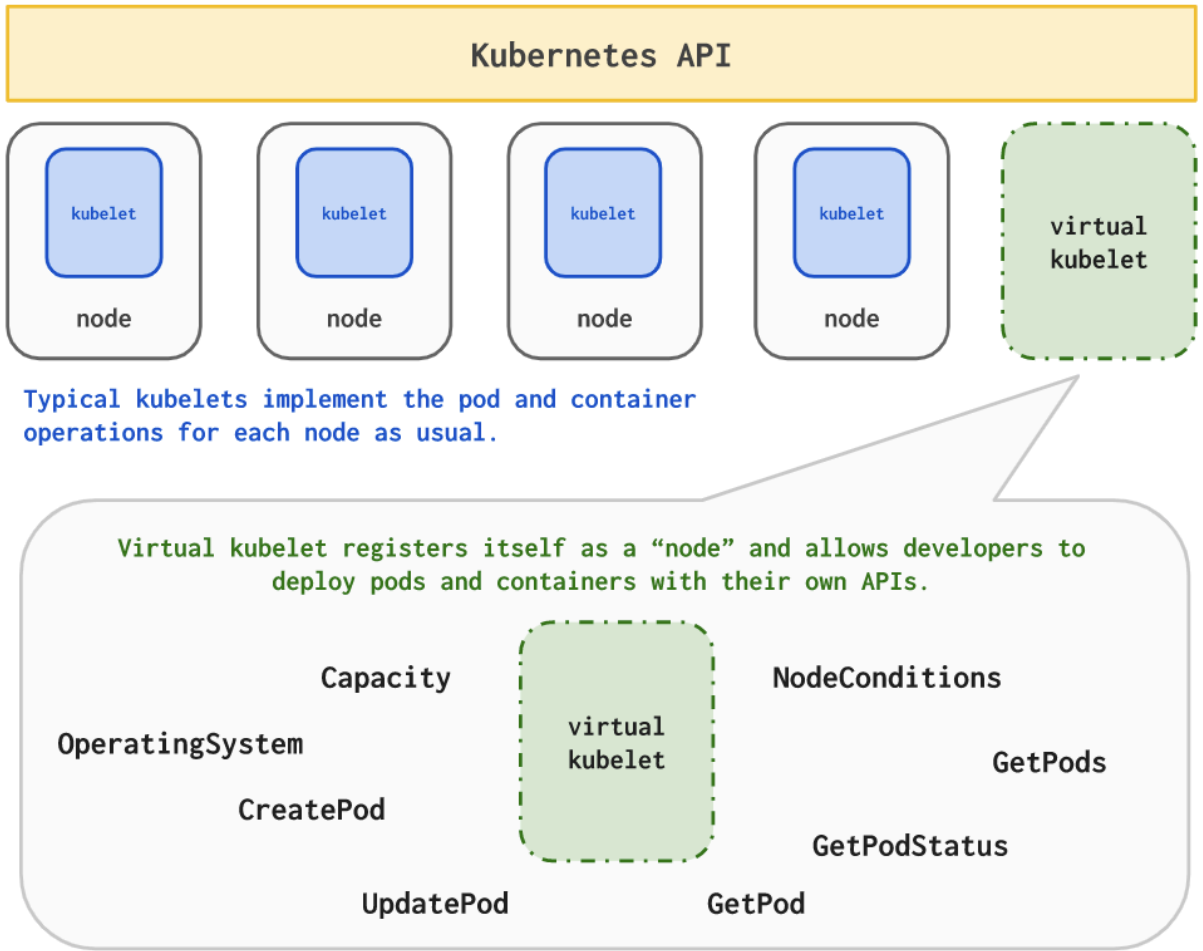
- RuntimeClass (v1.14 beta)
  - runc + kata
  - runc + gVisor
  - Windows server containers + Hyper-V containers

```
apiVersion: node.k8s.io/v1beta1
kind: RuntimeClass
metadata:
  name: myclas
# RuntimeClass is a non-namespaced
resource
handler: myconfiguration
```

```
apiVersion: v1
kind: Pod
metadata:
  name: mypod
spec:
  runtimeClassName: myclass
  # ...
```

Serverless 容器

- Virtual Kubelet
  - 模拟 Kubelet 功能
  - 无限资源的虚拟 Node
- Serverless 容器平台
  - Azure Container Instance
  - AWS Farget
  - Service Fabric
  - IoT Edge
- Serverless Kubernetes



QA