KubeCon | CloudNativeCon
OPEN SOURCE SUMMIT
China 2019
Container Runtime Evolving in Kubernetes

Pengfei Ni, Microsoft Azure
Agenda

- Kubernetes Architecture
- Container Runtime Interface (CRI)
- Kubelet Container Runtime Evolving
- Future
Kubernetes Architecture

Masters
- etcd
- Controller manager
- Kube-apiserver
- kubectl

Nodes
- Kubelet
- Kube-proxy
- Network Plugin
- Docker
- Pod

Internet
Kubelet Architecture (continued)

Kubelet

- Kubelet Server
- Container Manager
- Volume Manager
- Eviction
- cAdvisor
- Metrics and stats
- Generic Runtime Manager
- Container Runtime Interface
- dockershim
- Docker
- remote

CRI Container Runtime

- CRI Server (shim)
- Streaming Server
- CNI
- Containers
- Images
- Container Engine
  - runc
  - containerd
  - Kata
Kubelet Container Runtime

- Kubelet manages container runtime via CRI
- Container runtime manages images, containers and networking
- Container runtimes
  - Built-in: docker
  - Remote: cri-o, …
• Container Runtime Interface (CRI)
  • gRPC interfaces for container and image mgt
  • Streaming server library
  • CRI Tools
• Kubelet as the client for CRI server
• Easy to integrate based on gRPC framework
• Container runtime should take care of container networking (e.g. CNI)
• Kubelet config: --container-runtime-endpoint
CRI Interface

RuntimeService

- RunPodSandbox
- StopPodSandbox
- RemovePodSandbox
- PodSandboxStatus
- ListPodSandbox
- UpdateRuntimeConfig
- ContainerStats

- CreateContainer
- StartContainer
- StopContainer
- RemoveContainer
- ListContainers
- ContainerStatus
- ExecSync
- Exec
- Attach
- PortForward
- Version
- Status

ImageService

- ListImages
- PullImage
- ImageStatus
- RemoveImage
- ImageFsInfo

gRPC Server
CRI Streaming Interface

- Streaming interface
  - Exec
  - Attach
  - PortForward
- Runtime returns URL for Streaming Server
- Streaming server library
  
  https://github.com/kubernetes/kubernetes/tree/master/pkg/kubelet/server/streaming
Container Runtime Evolving

- **Kubelet**
  - docker
  - rkt
  - CNI

- **Kubelet**
  - CRI
  - docker
  - CNI

- **CRI Shim**
  - Container Runtime: containerd, gvisor, ...
  - CNI
# CRI Container Runtimes

<table>
<thead>
<tr>
<th>CRI Container Runtime</th>
<th>Community</th>
<th>Features</th>
<th>Container Engine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dockershim</td>
<td>Kubernetes</td>
<td>Kubelet built-in</td>
<td>docker</td>
</tr>
<tr>
<td><strong>cri-o</strong></td>
<td>Kubernetes</td>
<td>OCI-compatible and no need of docker</td>
<td>OCI ( runc, kata, gVisor… )</td>
</tr>
<tr>
<td><strong>cri-containerd</strong></td>
<td>Containerd</td>
<td>OCI-compatible and no need of docker</td>
<td>OCI ( runc, kata, gVisor… )</td>
</tr>
<tr>
<td>Frakti</td>
<td>Kubernetes</td>
<td>Virtualization</td>
<td>hyperd, docker</td>
</tr>
<tr>
<td>rktlet</td>
<td>Kubernetes</td>
<td>rkt supported</td>
<td>rkt</td>
</tr>
<tr>
<td>PouchContainer</td>
<td>Alibaba</td>
<td>Rich container engine</td>
<td>OCI ( runc, kata… )</td>
</tr>
<tr>
<td>Virtlet</td>
<td>Mirantis</td>
<td>VM and QCOW2</td>
<td>Libvirt ( KVM )</td>
</tr>
</tbody>
</table>
Container Runtime landscape

CNCF Graduated
- containerd

CNCF Incubating
- cri-o
- rkt
- Firecracker
- gVisor
- kata
- lxd
- Nacula
- Pouch
- runc
- Singularity
- SmartOS
- unik
CRI TOOLS - CRICTL

- Sub-project of SIG Node
- CLI for CRI-compatible container runtimes
- Facilitates container runtime inspecting and debugging
- Simple UX
  - `crictl pods`
  - `crictl ps`
  - `crictl images`
  - `crictl exec`
  - ...
CRI TOOLS - CRITEST

- Test framework for CRI container runtimes
- Validate whether container runtime meets CRI requirements
- Benchmark CRI interfaces
  - `critest -benchmark`
- Integrated into devops is encouraged
- Publish validation results to [TestGrid](https://TestGrid) is encouraged
Future of Container Runtime

Out-of-tree dockershim

- Kubelet CRI
  - dockershim
    - docker
    - CNI
  - cri-containerd
    - containerd
    - CNI
    - runc
    - kata
Security hardened container engine

Application

System calls

Limited system calls

Host Kernel

Hardware

Firecracker

Firecracker scales to thousands of simultaneously validated containers.
Multiple Container Runtimes

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

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

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

- Virtual Kubelet
  - Masquerade kubelet
  - Virtual nodes
- Serverless container providers
  - Azure Container Instance
  - AWS Fargate
  - Service Fabric
  - IoT Edge
- Serverless Kubernetes