KubeCon CloudNativeCon Europe 2021

Virtual

CRI-O Still <3s Kubernetes

Urvashi Mohnani, Peter Hunt, Mrunal Patel, Sascha Grunert

CRI-O still <3 Kubernetes

What we cover in today's talk

- Latest development updates
- Moving from dockershim to CRI-O
- Heavy load scenarios
- User Namespaces and CRI-O
- Recording seccomp profiles directly from Kubernetes

Development Updates!

- More metrics
- Drop infra getting closer to stable
- Short name alias

  https://www.redhat.com/sysadmin/container-image-short-names
- pprof over unix socket

```
curl --unix-socket /var/run/crio/crio.sock http://localhost/debug/pprof/goroutine?debug=2
```

- More experimental features using annotations including user namespaces, shmsize, devices

Migrating from dockershim

- dockershim slated to be removed in 1.24

  https://github.com/kubernetes/enhancements/tree/master/keps/sig-node/2221-remove-dockershim
- Install and start the CRI-O service
- Point your kubelet to use CRI-O as the runtime
- --container-runtime-endpoint “unix://var/run/crio/crio.sock”

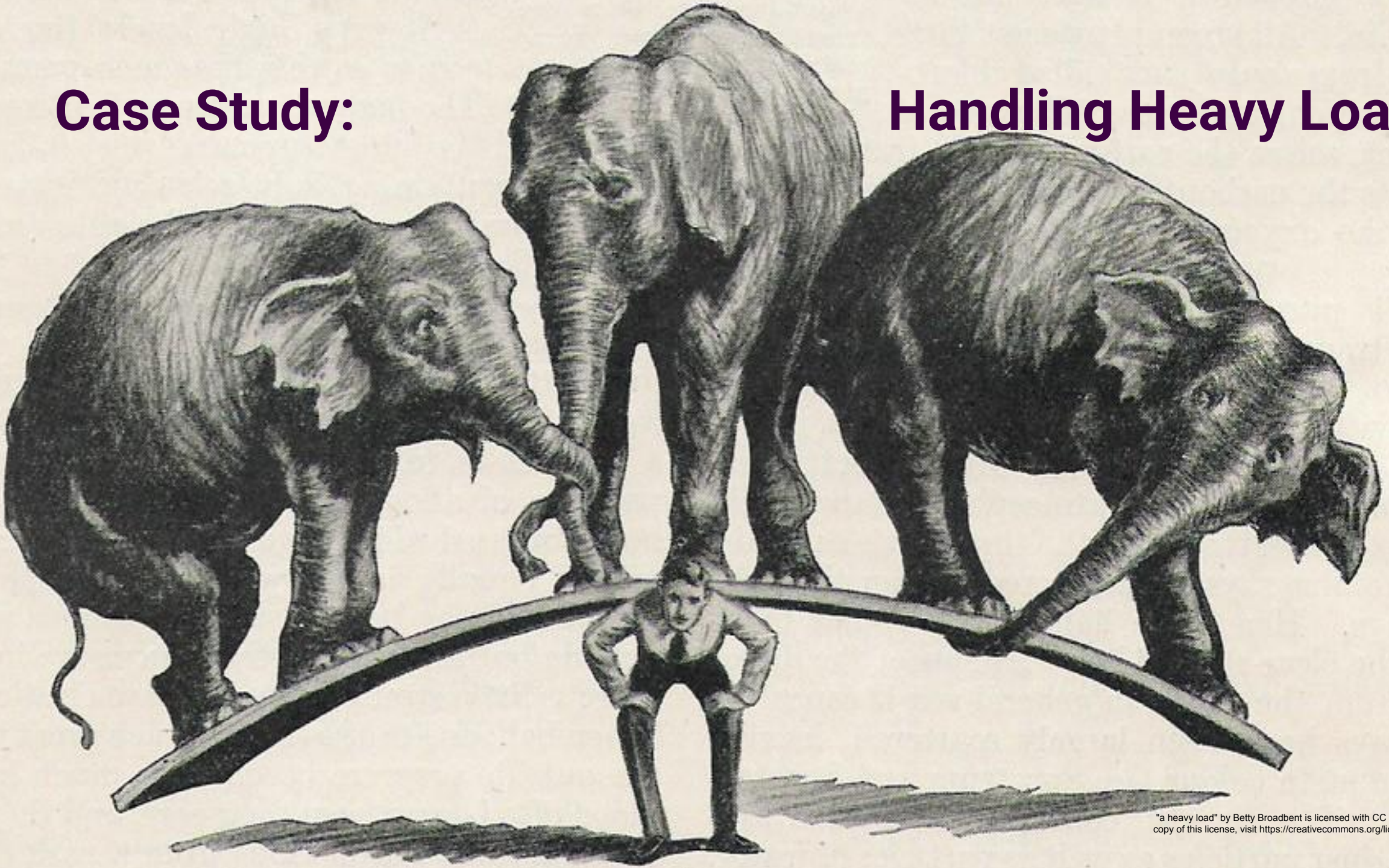
Migrating from dockershim

- crictl for debugging

  https://kubernetes.io/docs/tasks/debug-application-cluster/crictl/
- podman for managing images

  http://docs.podman.io/en/latest/

Case Study: Handling Heavy Load

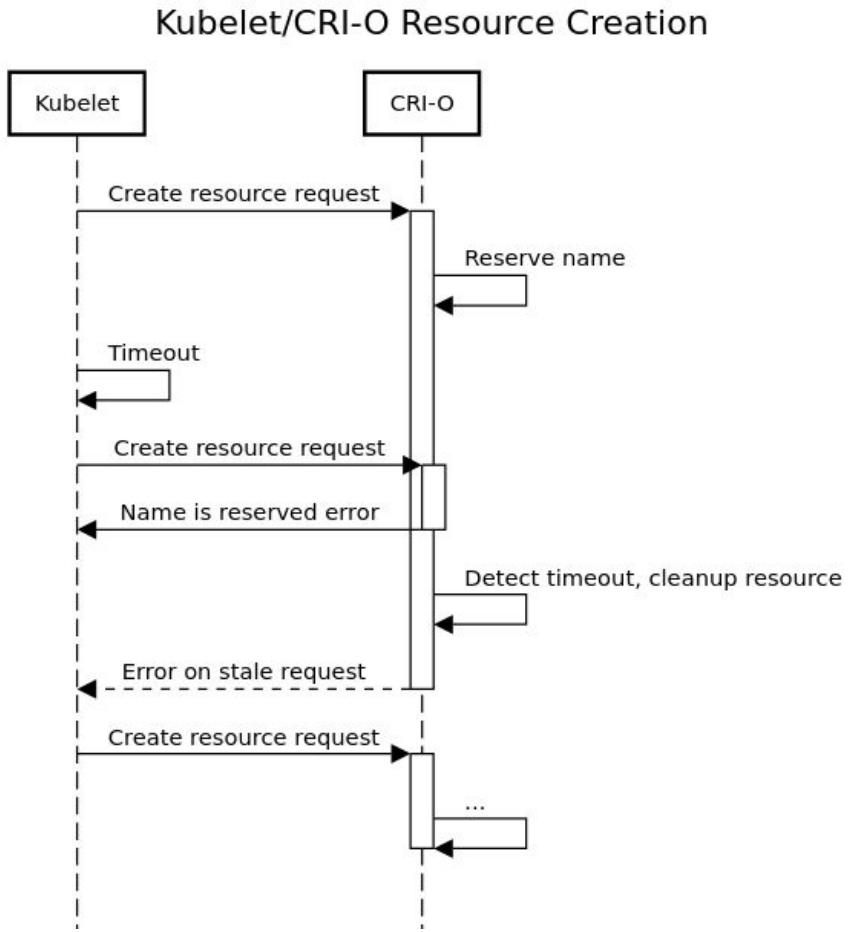


Problem:

Client/Server relationship requires timeouts which causes consistency issues

Problem:

Client/Server relationship requires timeouts which causes consistency issues

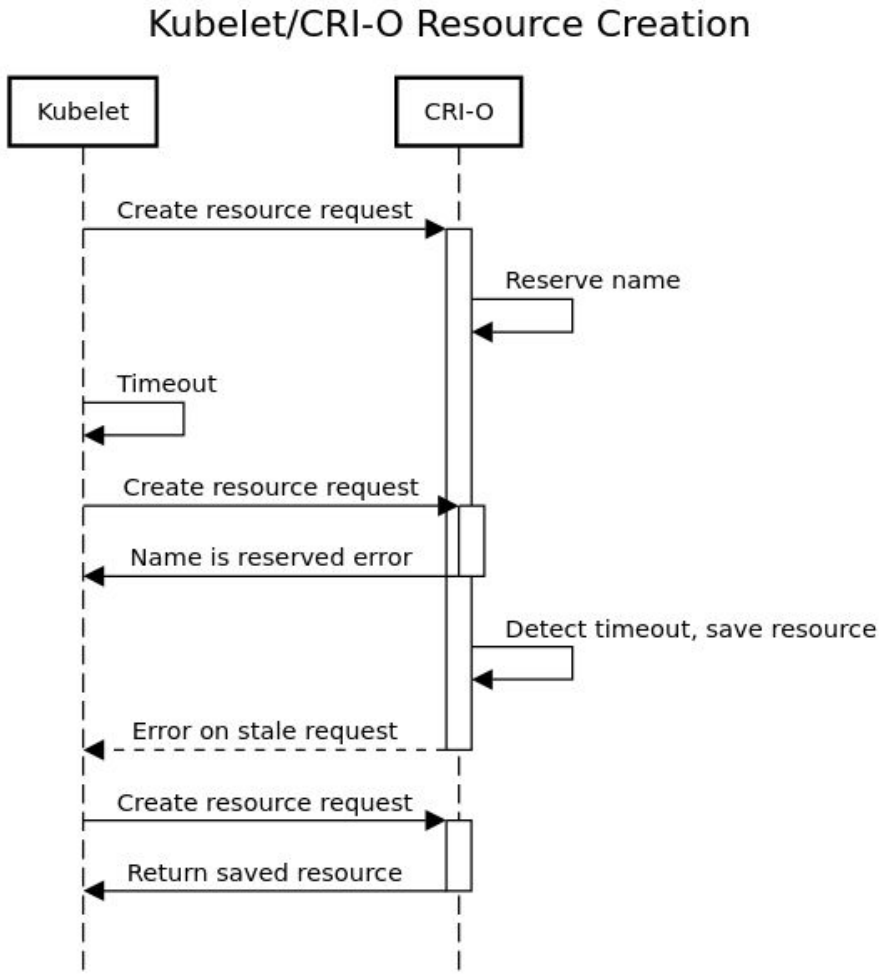


https://sequencediagram.org/

Solution:

Tune behavior to that of the Kubelet

- Finish creating the resource, and save it until the Kubelet asks again

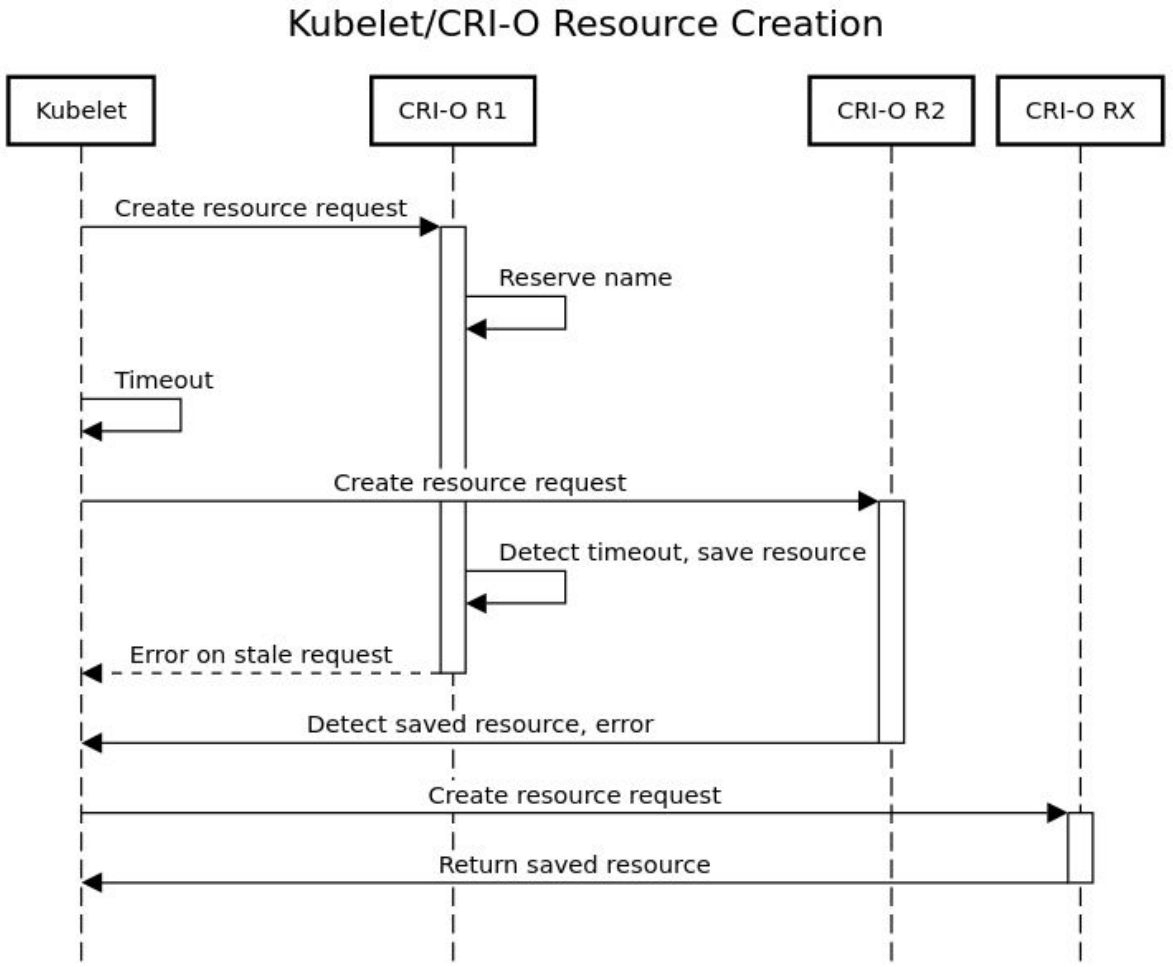


https://sequencediagram.org/

Tune Behavior to Kubelet

Tune behavior to that of the Kubelet

- Finish creating the resource, and save it until the Kubelet asks again
- Throttle new requests



https://sequencediagram.org/

Demo

User Namespaces & CRI-O

Source: https://imgflip.com/i/2cx48p

User Namespaces & CRI-O

- User namespace support in Kubernetes with CRI-O
- Currently in experimental phase
- Enabled by "io.kubernetes.cri-o.userns-mode" annotation
- Discussions going on in the upstream KEP

```
[crio.runtime.runtimes.runc-userns]
runtime_path = ""
runtime_type = "oci"
runtime_root = "/run/runc"
allowed_annotations = ["io.kubernetes.cri-o.userns-mode"]
```

```
apiVersion: v1
kind: Pod
metadata:
  name: userns-pod
  annotations:
    io.kubernetes.cri-o.userns-mode: "auto:size=65536;map-to-root=true"
spec:
  containers:
    - name: userns-ctr
      image: registry.fedoraproject.org/fedora
      command: ["sleep", "10000"]
```

Recording seccomp profiles

seccomp profiles for Kubernetes have to be available as JSON file on disk

```
{
  "defaultAction": "SCMP_ACT_ERRNO",
  "architectures": ["SCMP_ARCH_X86_64"],
  "syscalls": [
    {
      "names": ["clone"],
      "action": "SCMP_ACT_ALLOW"
    }
  ]
}
```

Recording seccomp profiles

Different OCI runtimes have a different minimal syscall footprint to be able to run a container:

- **runc:** capget, capset, chdir, close, epoll_ctl, epoll_pwait, execve, fchown, fcntl, fstat, fstatfs, futex, getdents64, getpid, getppid, nanosleep, newfstatat, openat, prctl, read, sched_yield, setgid, setgroups, setuid, write
- **crun:** arch_prctl, capset, close, execve, exit_group, prctl, read, rt_sigaction, rt_sigprocmask, select, set_tid_address, setresgid, setresuid, write

Recording seccomp profiles

Creating seccomp profiles requires detailed knowledge about:

- The underlying OCI runtime used by CRI-O
- The application syscalls it may execute (in all code paths)

Solution: Record seccomp profiles in a test environment

Recording seccomp profiles

- CRI-O targets to make Kubernetes more secure by default
- Idea: Make `RuntimeDefault` the preferred seccomp profile for all workloads
- Right now all workloads run `Unconfined` by default
- There is a KEP for that:

  https://github.com/kubernetes/enhancements/issues/2413

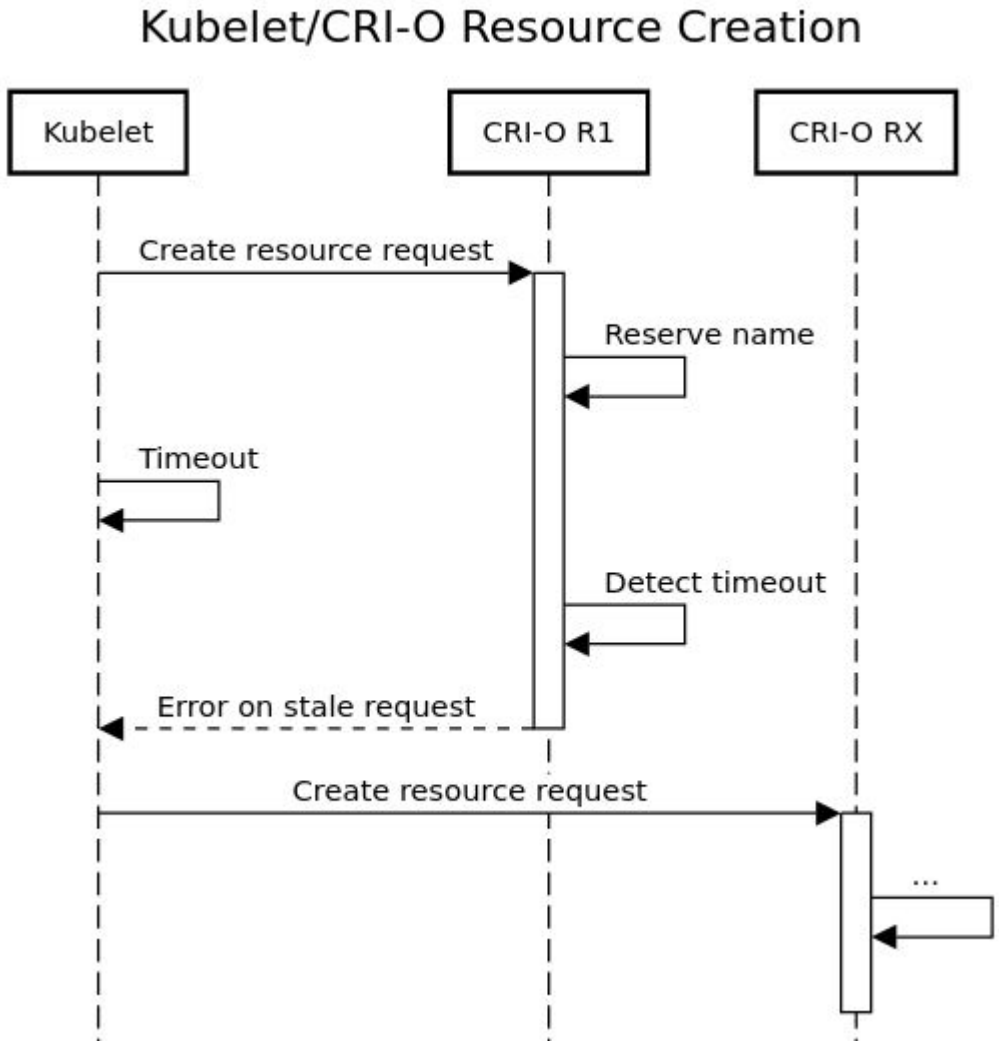
  https://github.com/kubernetes/enhancements/pull/2414


KubeCon
CloudNativeCon
Europe 2021
Virtual
Forward Together »

Solutions:

One could…

- Remove the resource after timed-out creation
- Short-circuit the resource creation



https://sequencediagram.org/