Kata Containers

Story of a container runtime

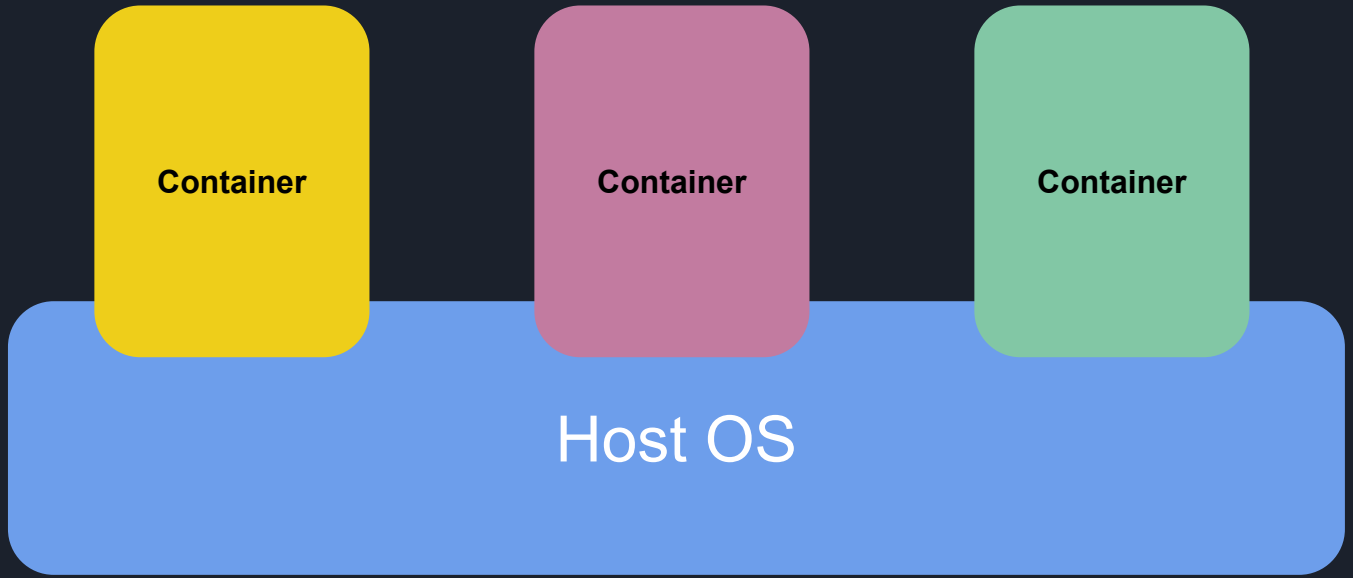
Sébastien Boeuf, Software Engineer
Intel Corporation

Agenda

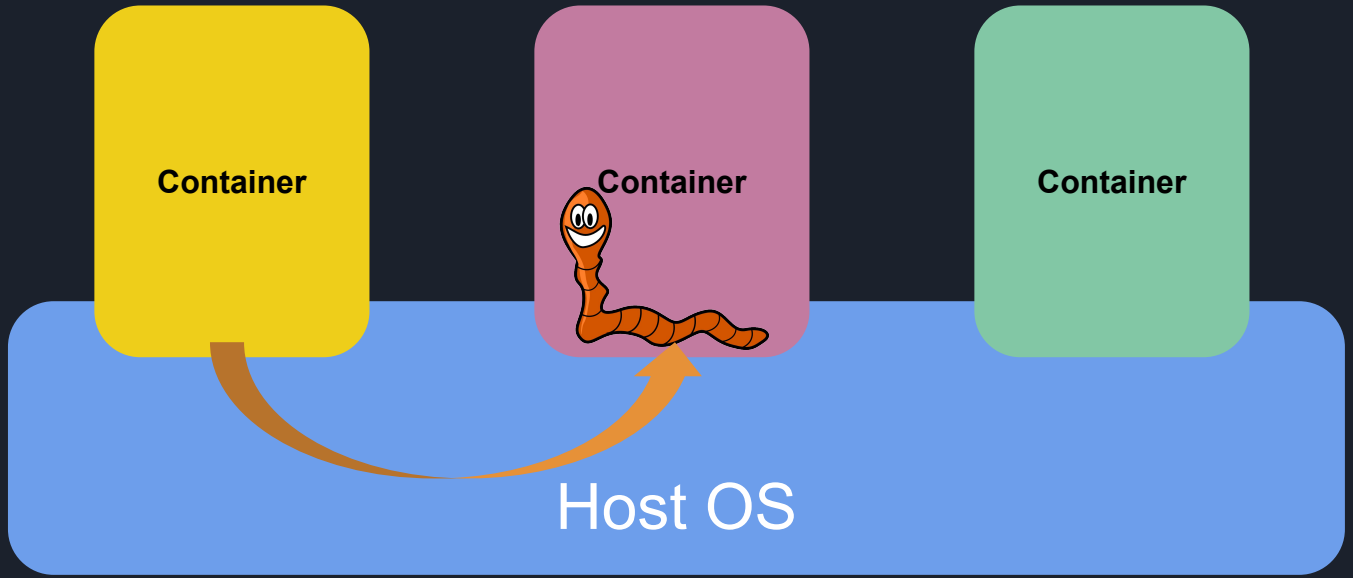
- Why Kata Containers?
- Acceptance
- Community growth
- Ecosystem influence
- Hypervisor flexible

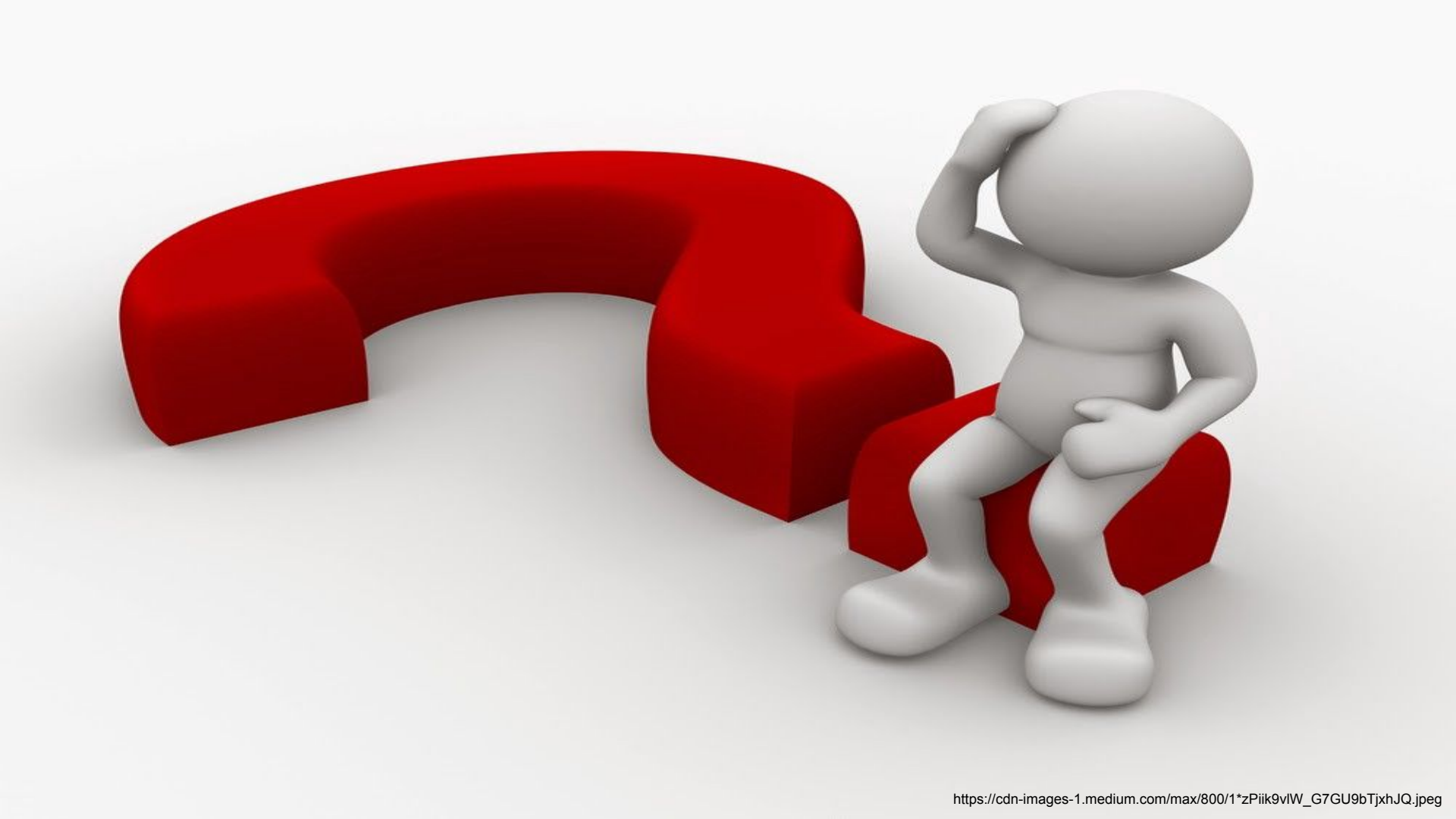
https://regmedia.co.uk/2017/09/11/shutterstock_containers_in_port.jpg

Containers



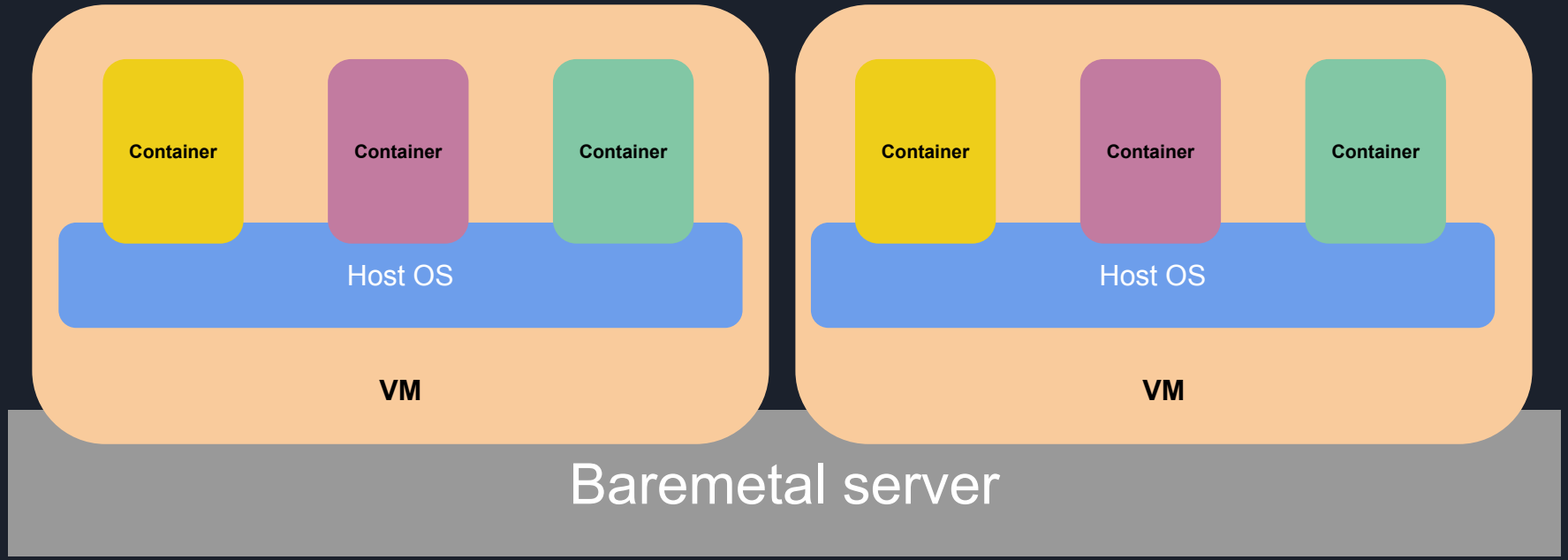
Security threat





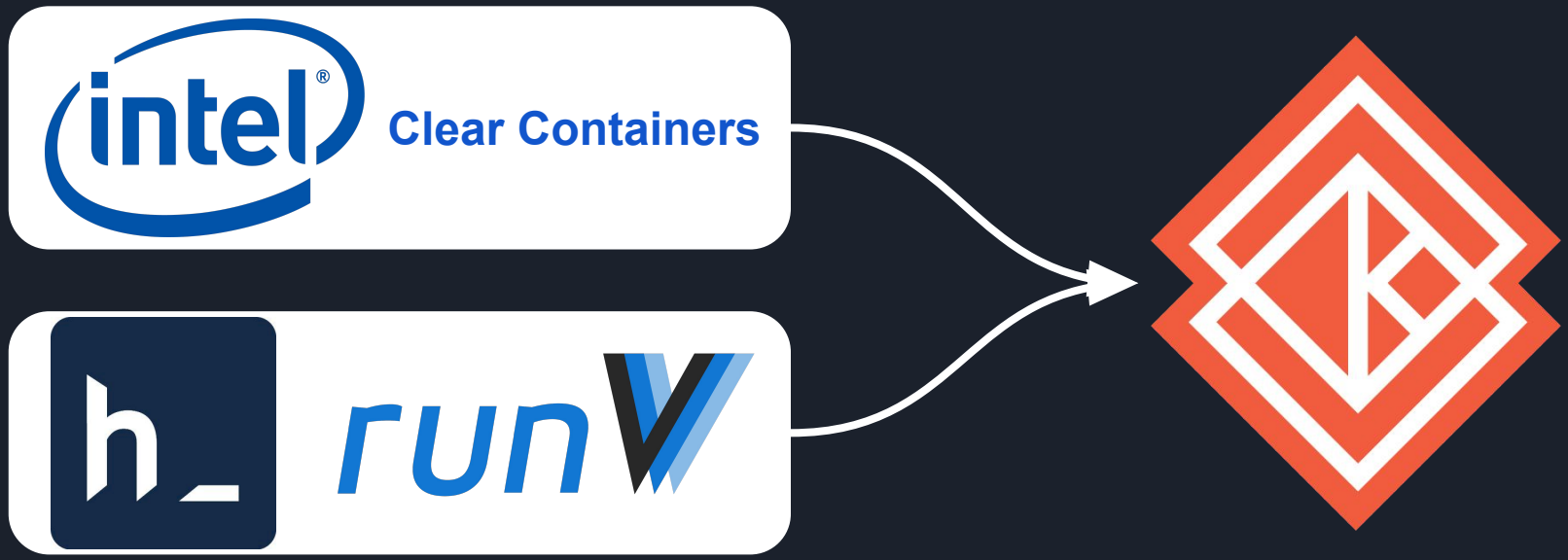
https://cdn-images-1.medium.com/max/800/1*zPiik9vlW_G7GU9bTjxhJQ.jpeg

Manual isolation

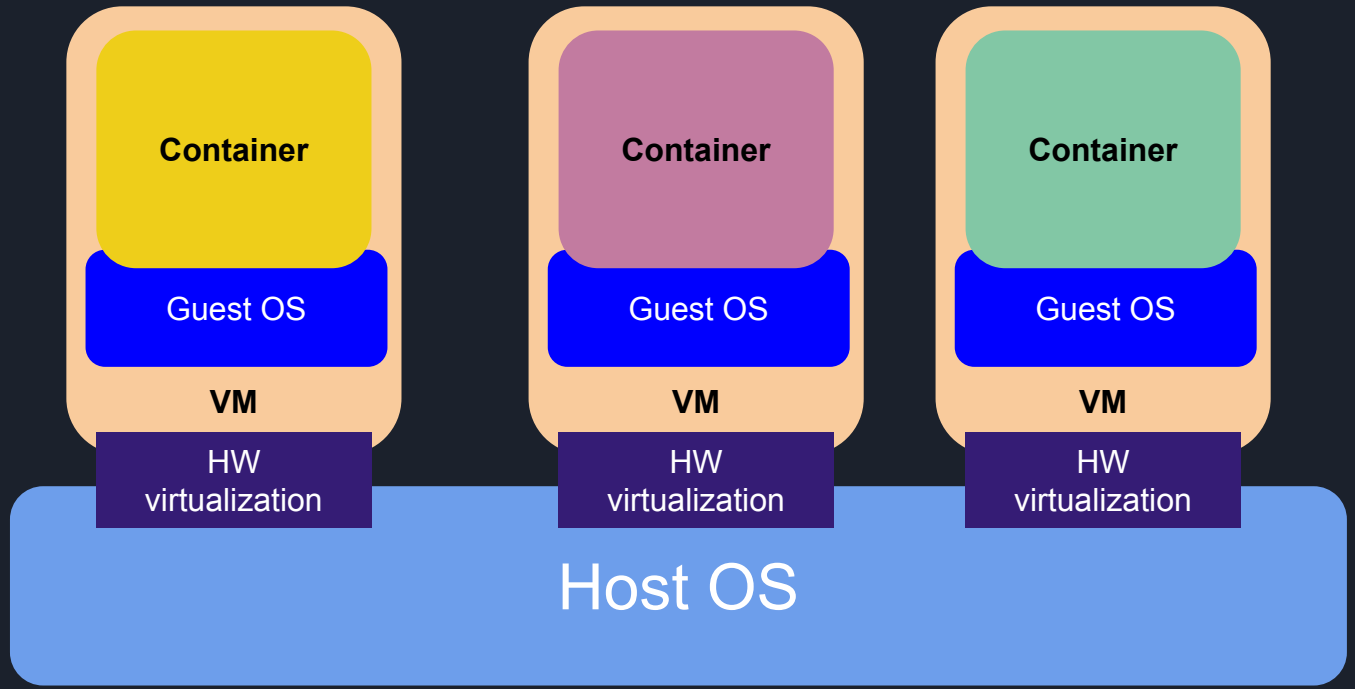


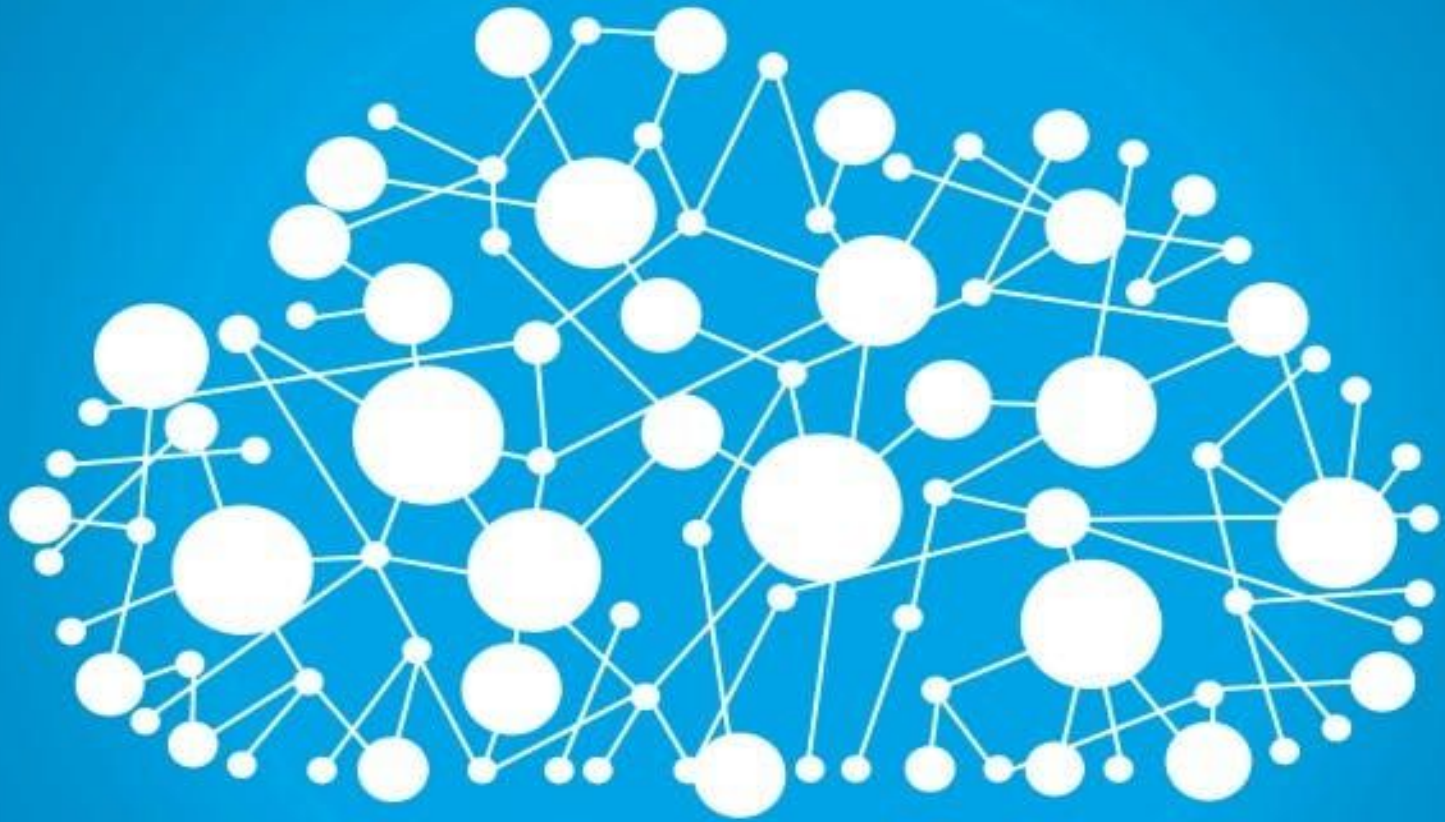
https://s3.amazonaws.com/wordpress-production/wp-content/uploads/2015/12/collaborative-problem-solving.jpg

Legacy



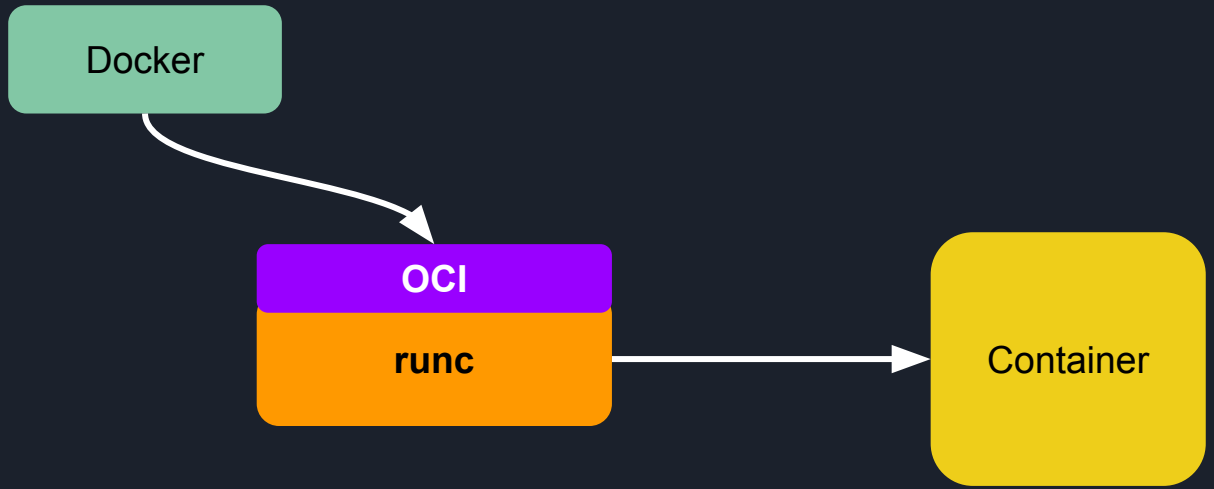
Kata Containers



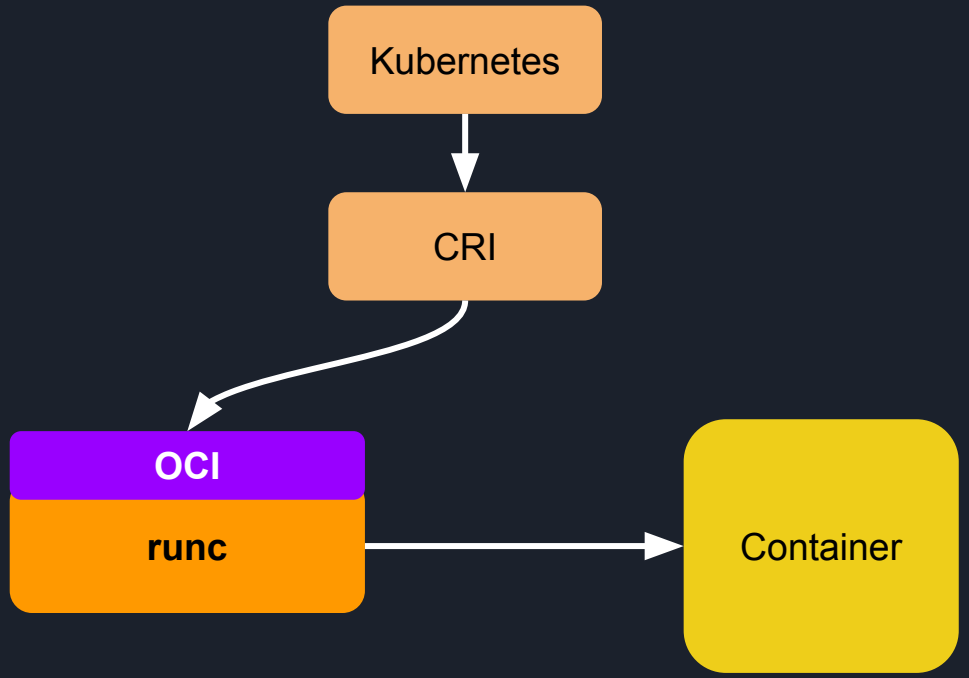


https://marketingweek.imgix.net/content/uploads/2017/06/30121536/Ecosystem-body-image.jpg

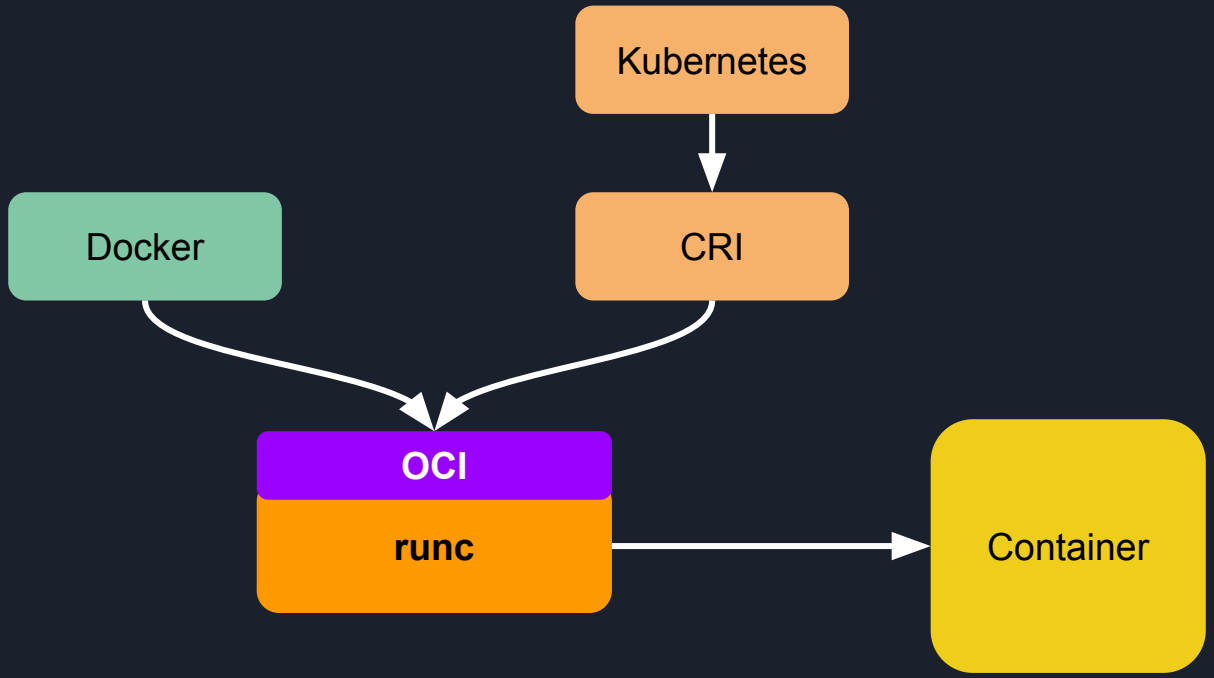
Container ecosystem



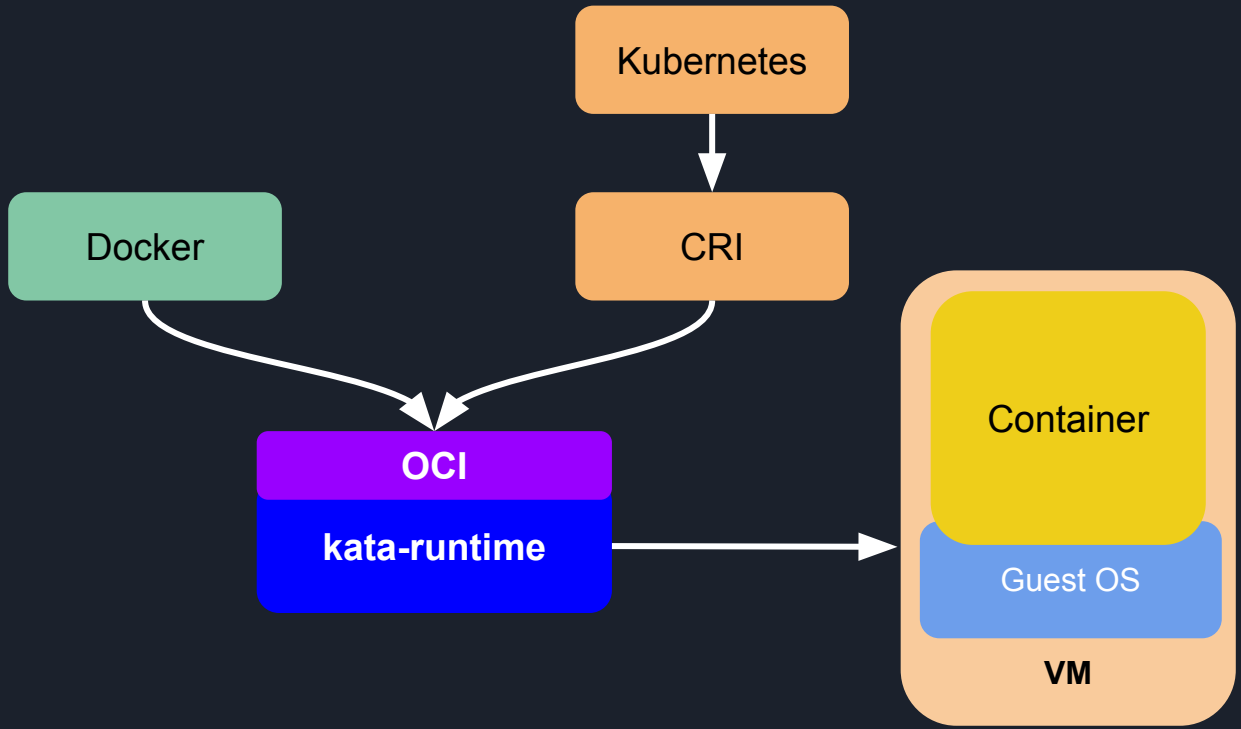
Container ecosystem

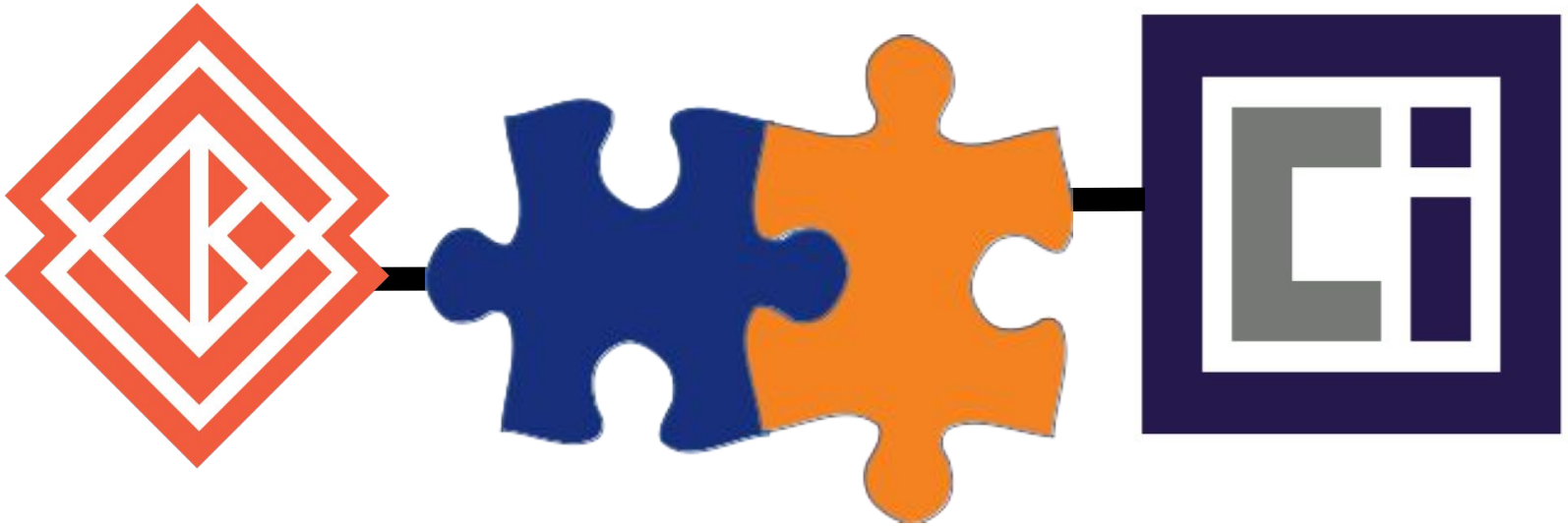


Container ecosystem

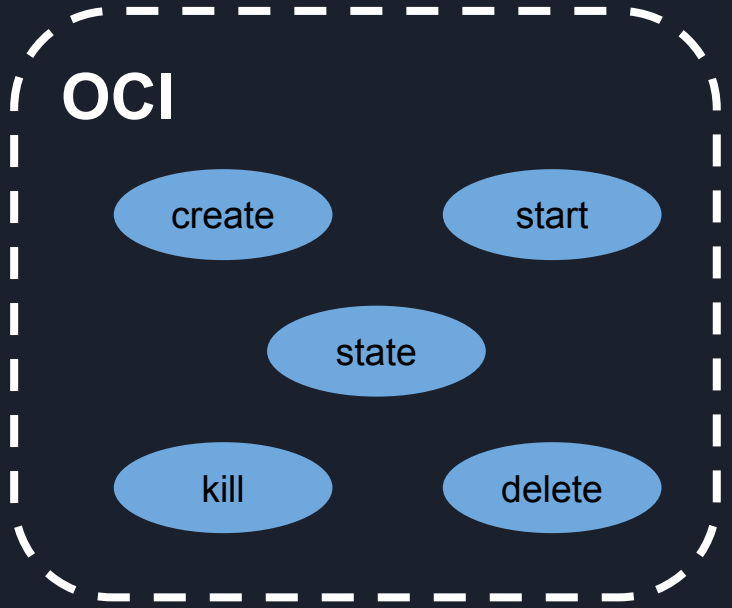


Seamless integration

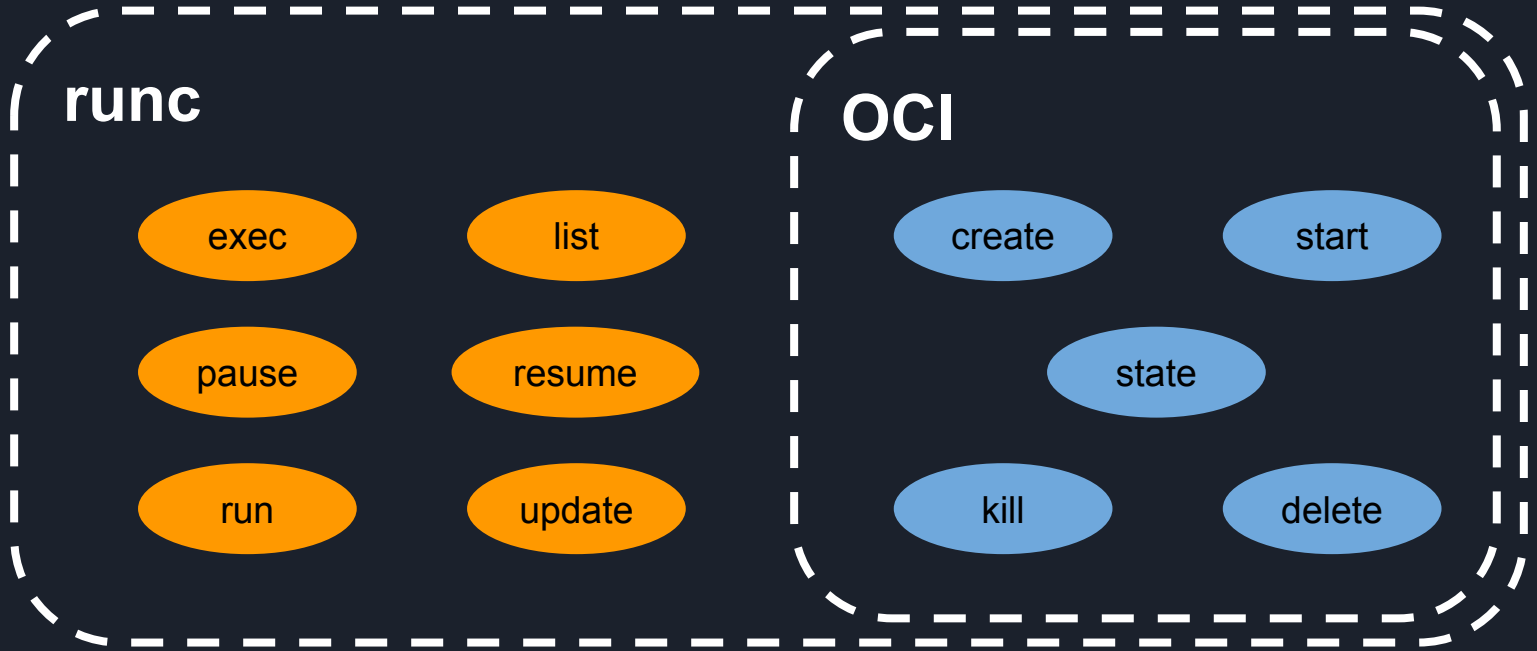




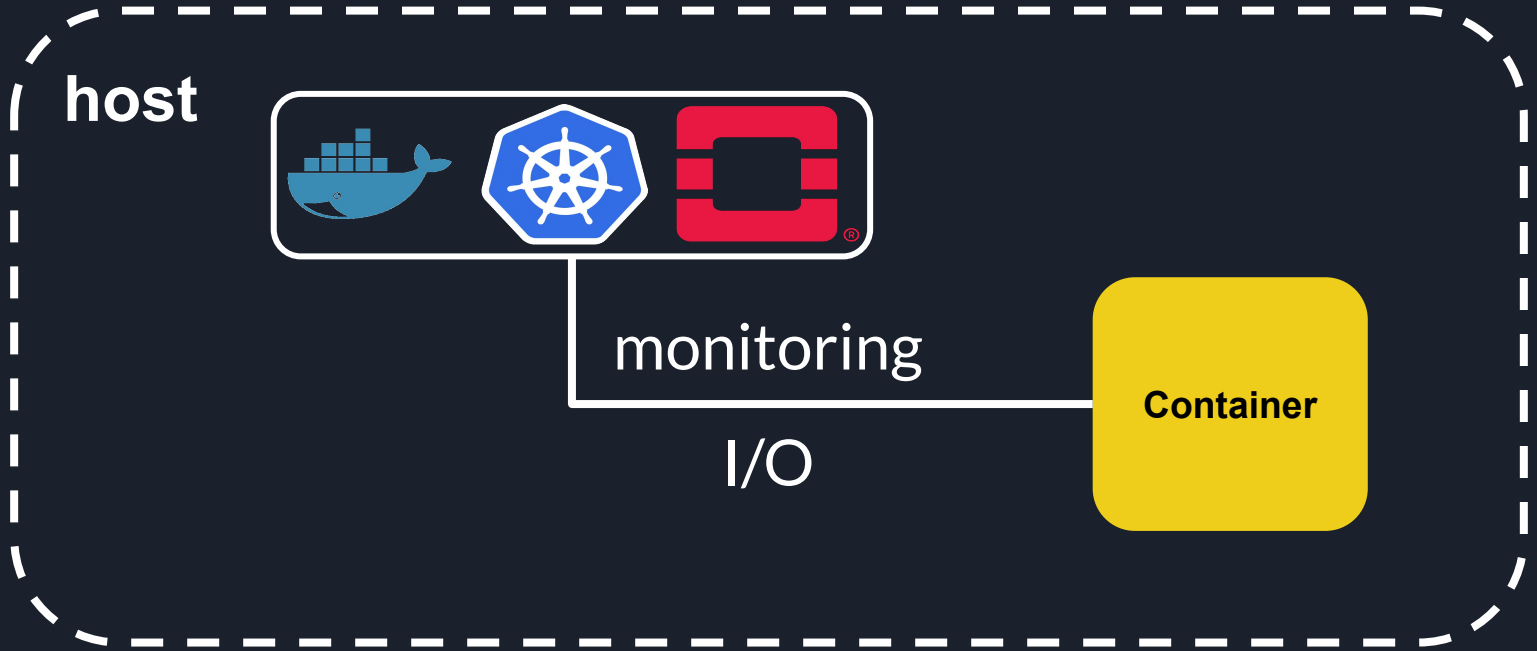
OCI compatible



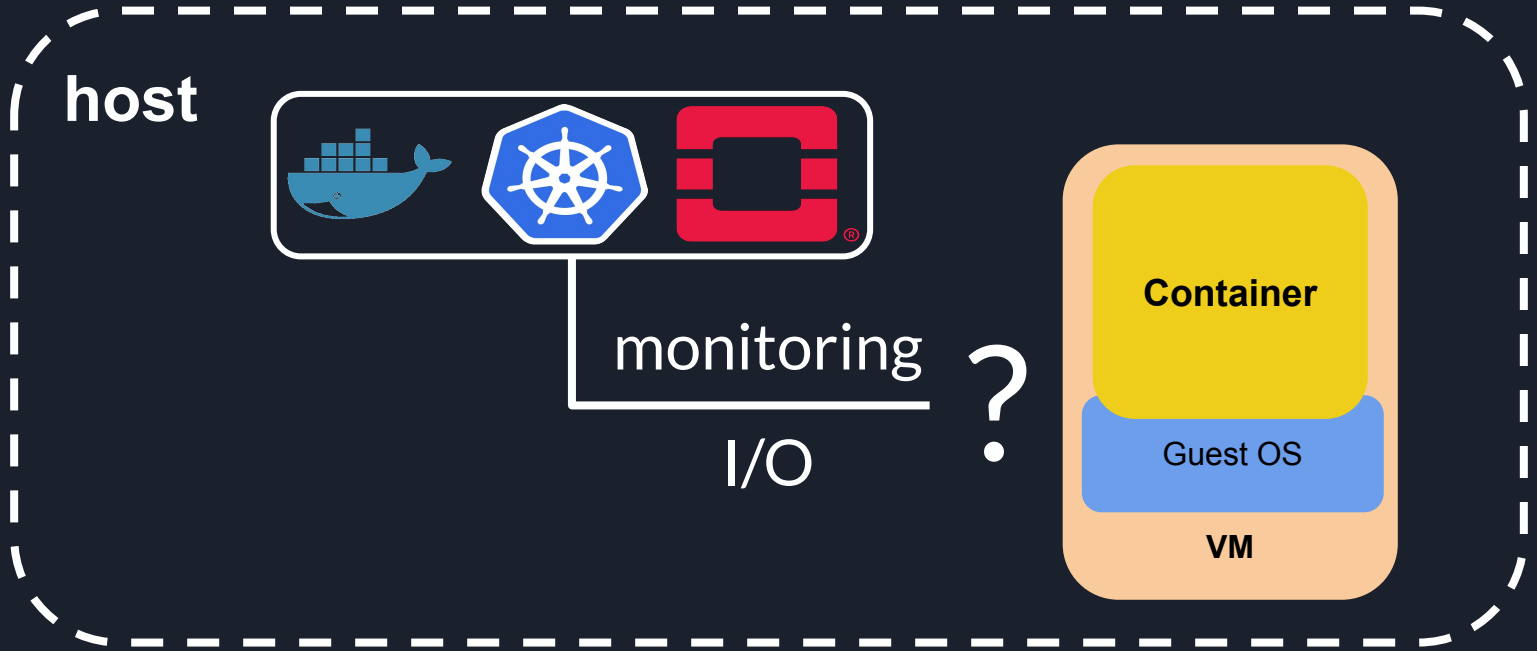
OCI compatible



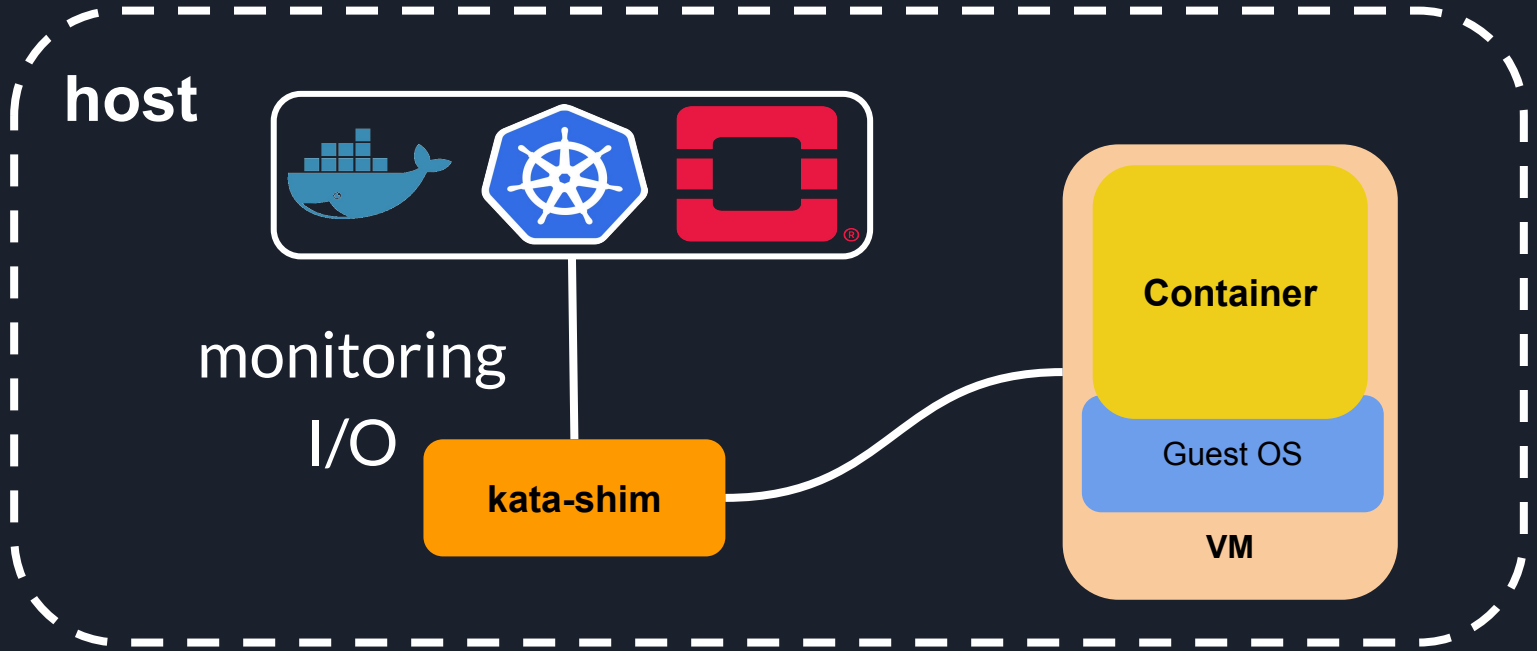
OCI compatible



OCI compatible



OCI compatible




https://www.incimages.com/uploaded_files/image/1940x900/getty_524541622_2000133320009280310_370635.jpg

Community growth

Additional architectures

- aarch64 (**ARM**)
- ppc64 and s390 (**IBM**)

Enhanced stability and production ready

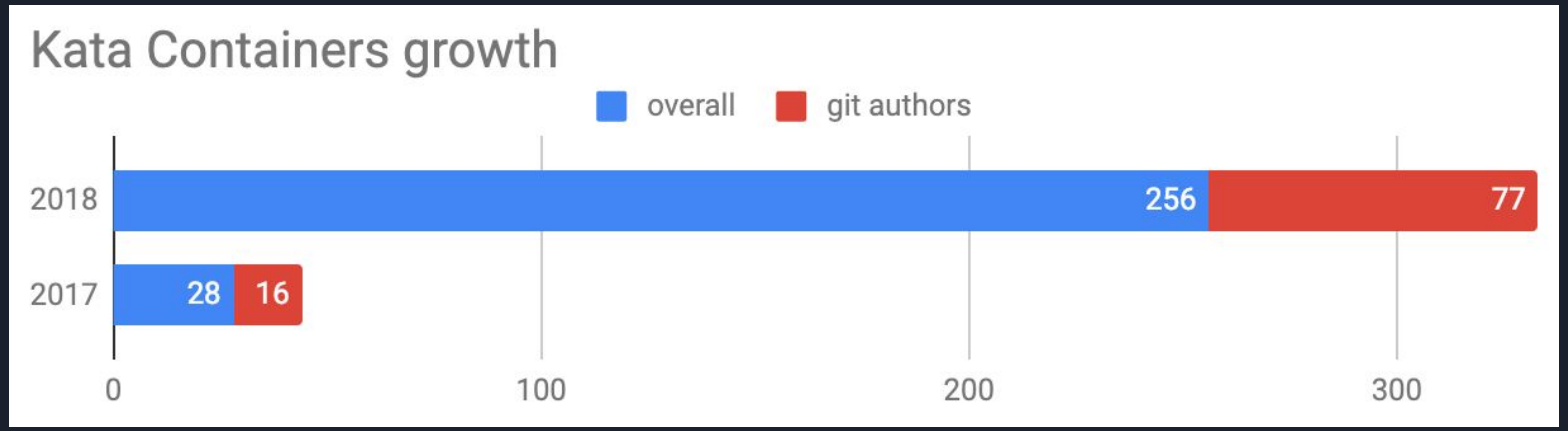
- **Huawei**
- **Baidu**
- **Alibaba**

Community growth

CI resources

- Vexxhost (**Vexxhost**)
- Azure (**Microsoft**)
- AWS (**Amazon**)
- GCE (**Google**)

Community growth



2000 pull requests / 100 contributors

https://hbr.org/resources/images/article_assets/2015/05/MAY15_19_686097-001.jpg

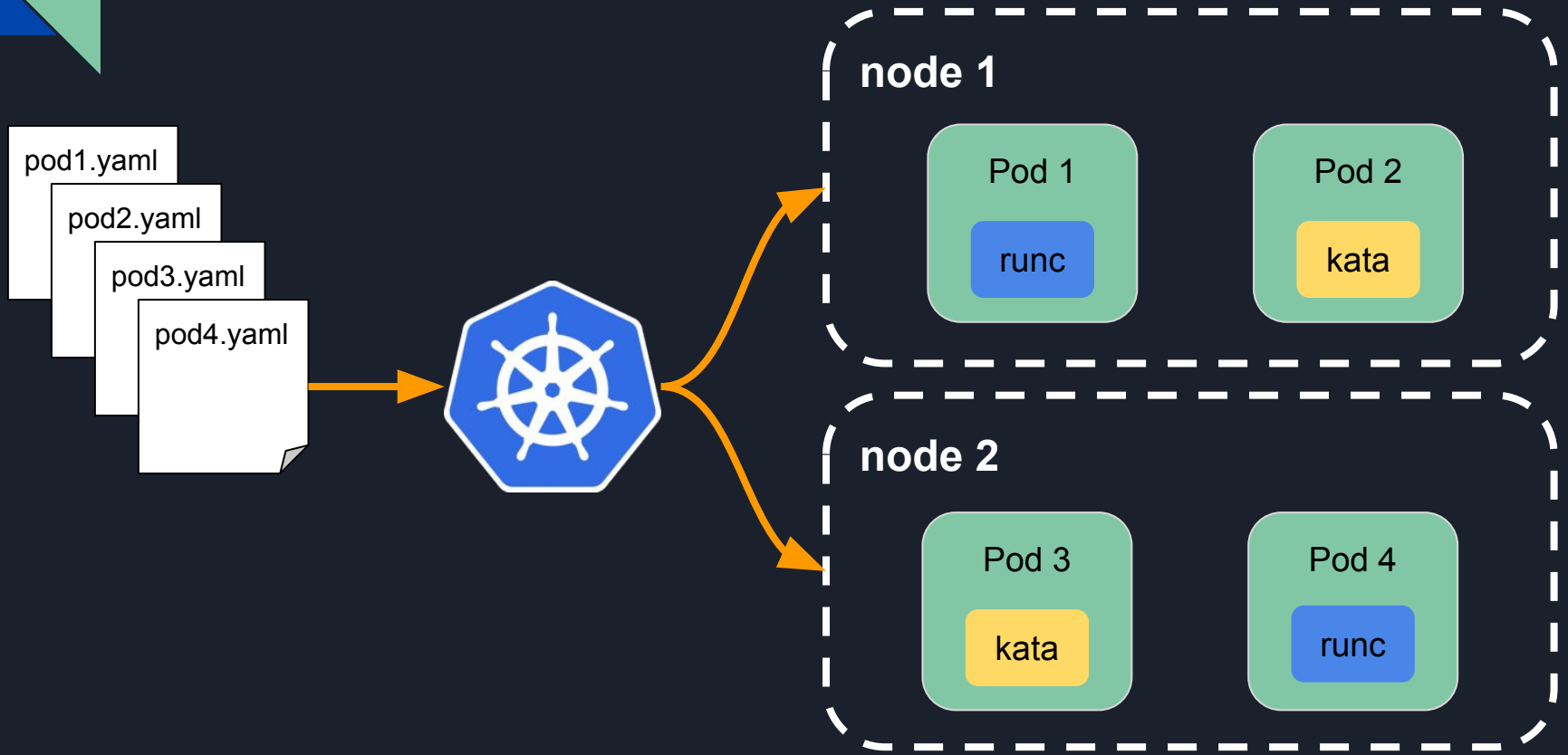
Extend OCI

```
// VM contains information for virtual-machine-based containers.
type VM struct {
        // Hypervisor specifies hypervisor-related configuration for virtual-machine-based containers.
        Hypervisor VMHypervisor `json:"hypervisor,omitempty"`
        // Kernel specifies kernel-related configuration for virtual-machine-based containers.
        Kernel VMKernel `json:"kernel"`
        // Image specifies guest image related configuration for virtual-machine-based containers.
        Image VMImage `json:"image,omitempty"`
}
```

RuntimeClass

```
message RunPodSandboxRequest {
    // Configuration for creating a PodSandbox.
    PodSandboxConfig config = 1;
    // Named runtime configuration to use for this PodSandbox.
    // If the runtime handler is unknown, this request should be rejected.  An
    // empty string should select the default handler, equivalent to the
    // behavior before this feature was added.
    // See https://git.k8s.io/enhancements/keps/sig-node/runtime-class.md
    string runtime_handler = 2;
}
```

RuntimeClass



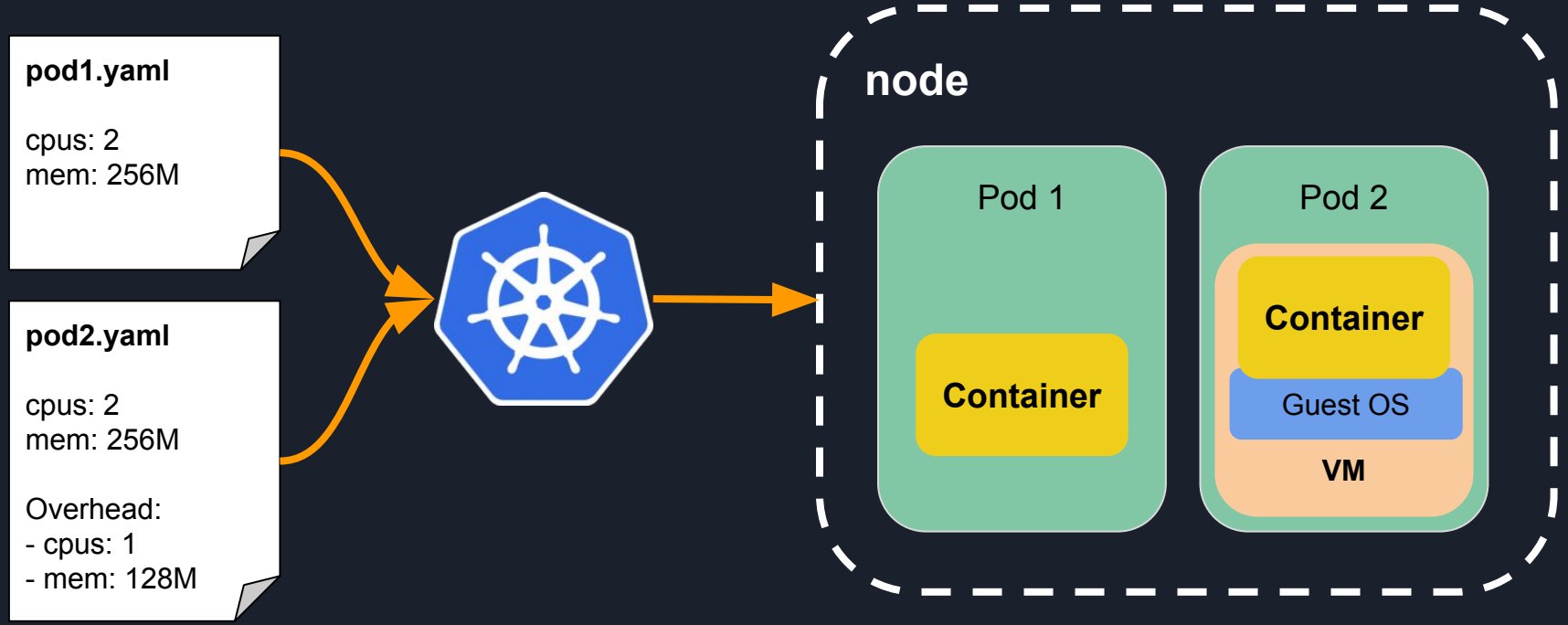
Pod overhead

Introduce a Pod.Spec.Resources field on the pod to specify the pods overhead.

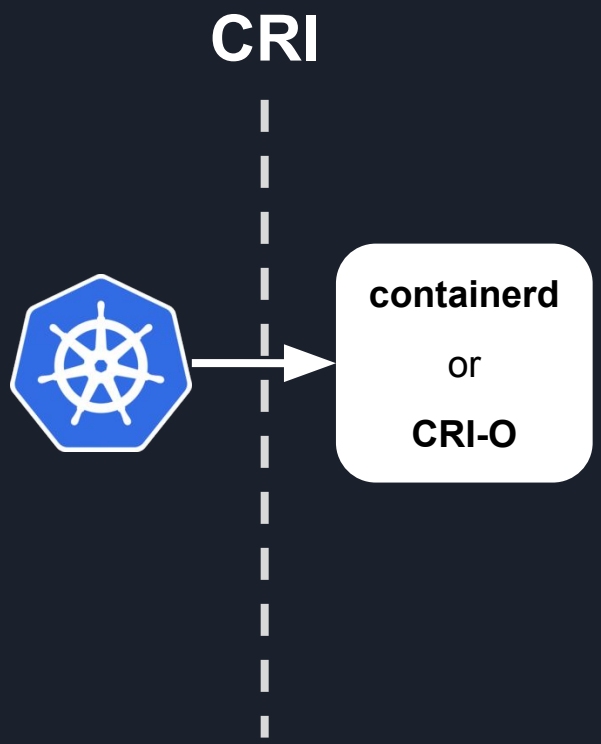
```
Pod {
  Spec PodSpec {
    // Overhead is the resource overhead consumed by the Pod, not including
    // container resource usage. Users should leave this field unset.
    // +optional
    Overhead *ResourceRequirements
  }
}
```

For scheduling, the pod resource requests are added to the container resource requests.

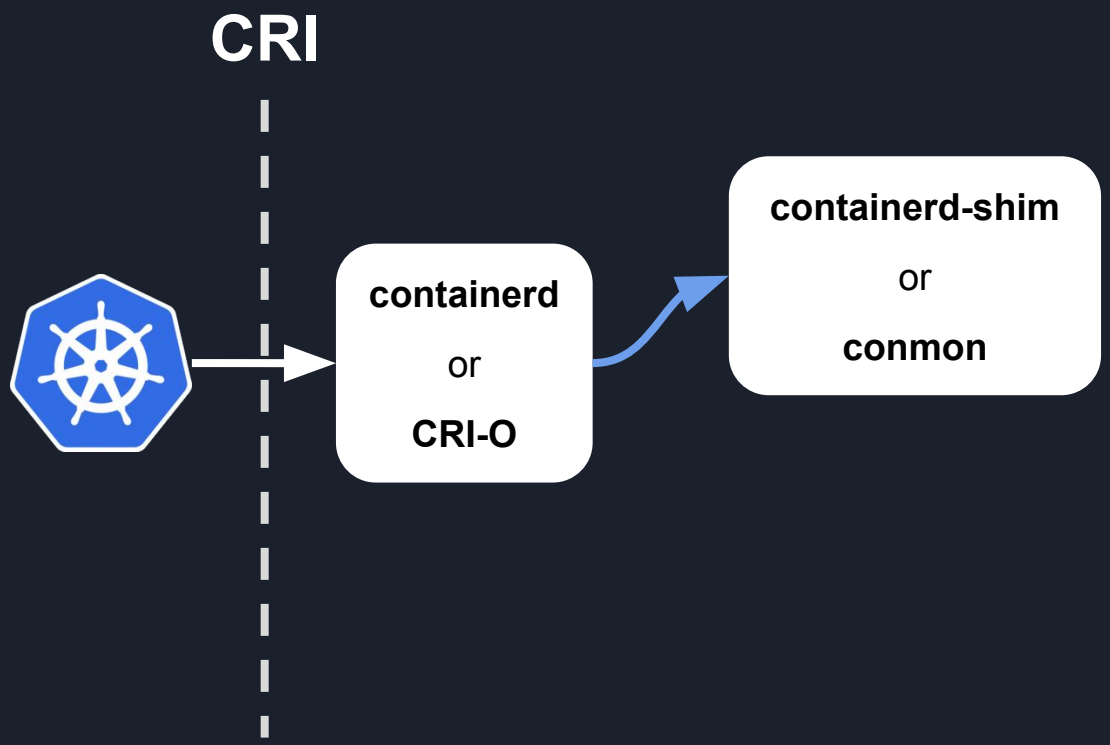
Pod overhead



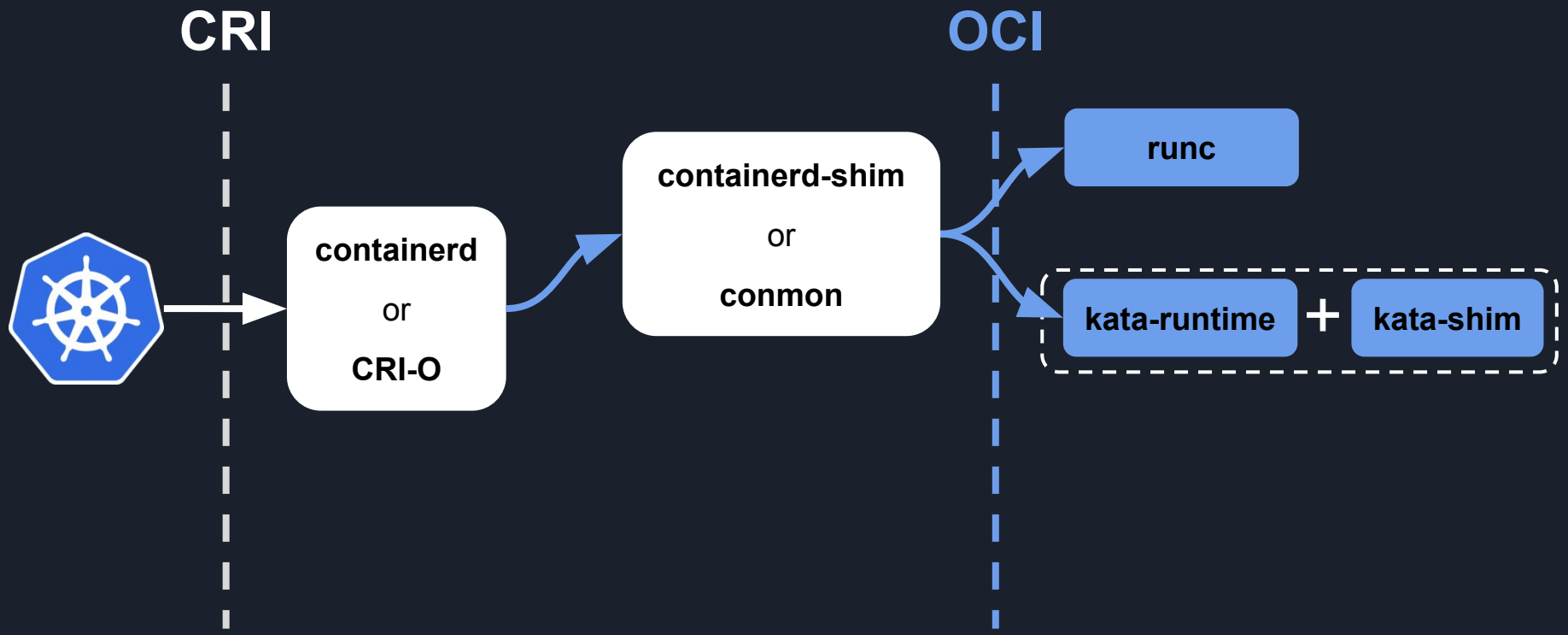
Shim v2



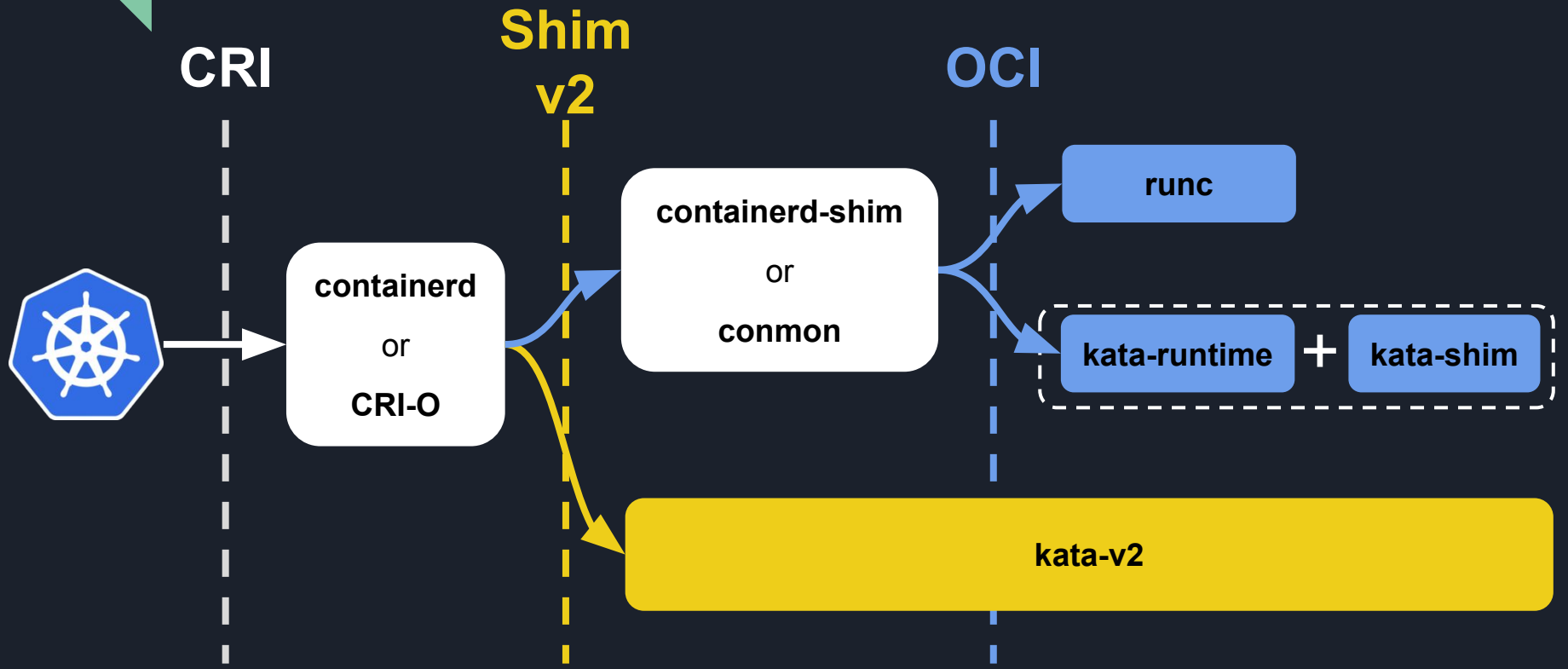
Shim v2



Shim v2

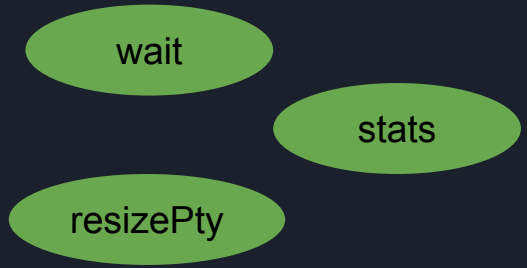

CRI
OCI
containerd
or
CRI-O
containerd-shim
or
conmon
runc
kata-runtime
+
kata-shim

Shim v2


CRI
Shim v2
OCI
containerd or CRI-O
containerd-shim or conmon
runc
kata-runtime + kata-shim
kata-v2

Shim v2

```
service Task {
        rpc State(StateRequest) returns (StateResponse);
        rpc Create(CreateTaskRequest) returns (CreateTaskResponse);
        rpc Start(StartRequest) returns (StartResponse);
        rpc Delete(DeleteRequest) returns (DeleteResponse);
        rpc Pids(PidsRequest) returns (PidsResponse);
        rpc Pause(PauseRequest) returns (google.protobuf.Empty);
        rpc Resume(ResumeRequest) returns (google.protobuf.Empty);
        rpc Checkpoint(CheckpointTaskRequest) returns (google.protobuf.Empty);
        rpc Kill(KillRequest) returns (google.protobuf.Empty);
        rpc Exec(ExecProcessRequest) returns (google.protobuf.Empty);
        rpc ResizePty(ResizePtyRequest) returns (google.protobuf.Empty);
        rpc CloseIO(CloseIORequest) returns (google.protobuf.Empty);
        rpc Update(UpdateTaskRequest) returns (google.protobuf.Empty);
        rpc Wait(WaitRequest) returns (WaitResponse);
        rpc Stats(StatsRequest) returns (StatsResponse);
        rpc Connect(ConnectRequest) returns (ConnectResponse);
        rpc Shutdown(ShutdownRequest) returns (google.protobuf.Empty);
}
```



No host PID assumption!

k8s pod scaling!

Shared filesystem

Virtio-9p

- Not fully POSIX compliant ⇒ Workload functional issues
- Not performant
- Production should use virtio-blk ⇒ devicemapper

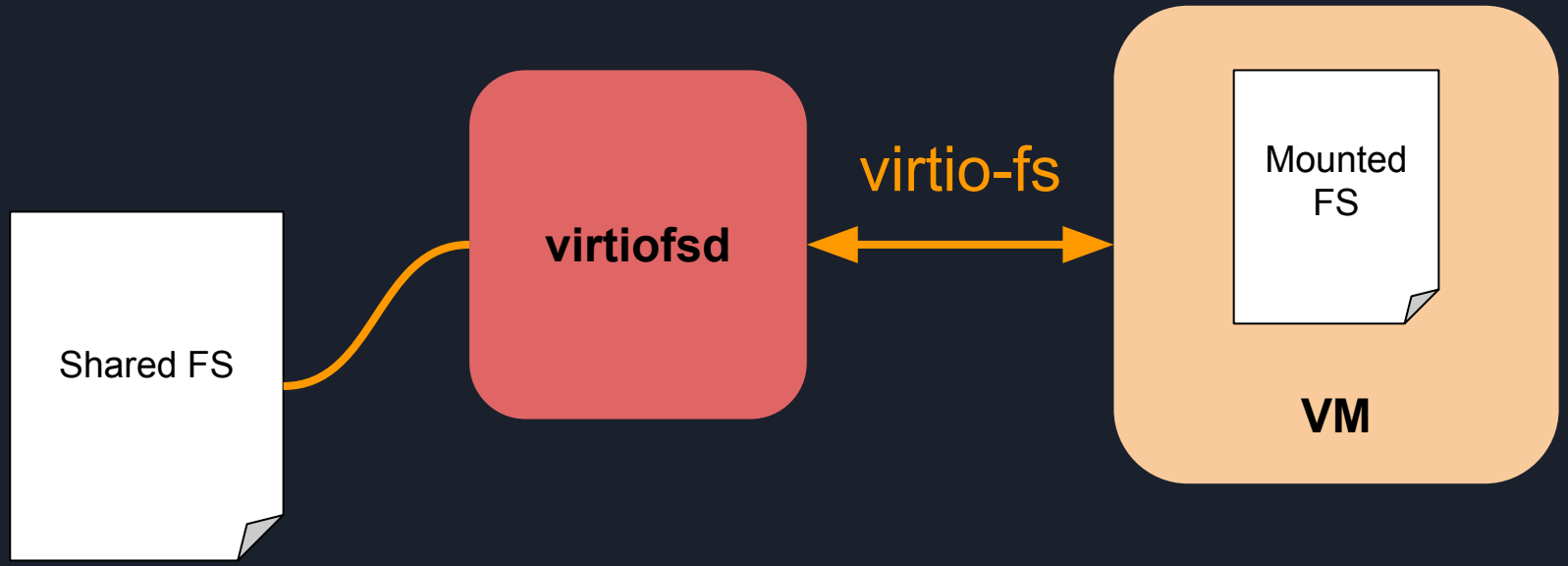
Shared filesystem

Redhat developed replacement for virtio-9p ⇒ **virtio-fs**

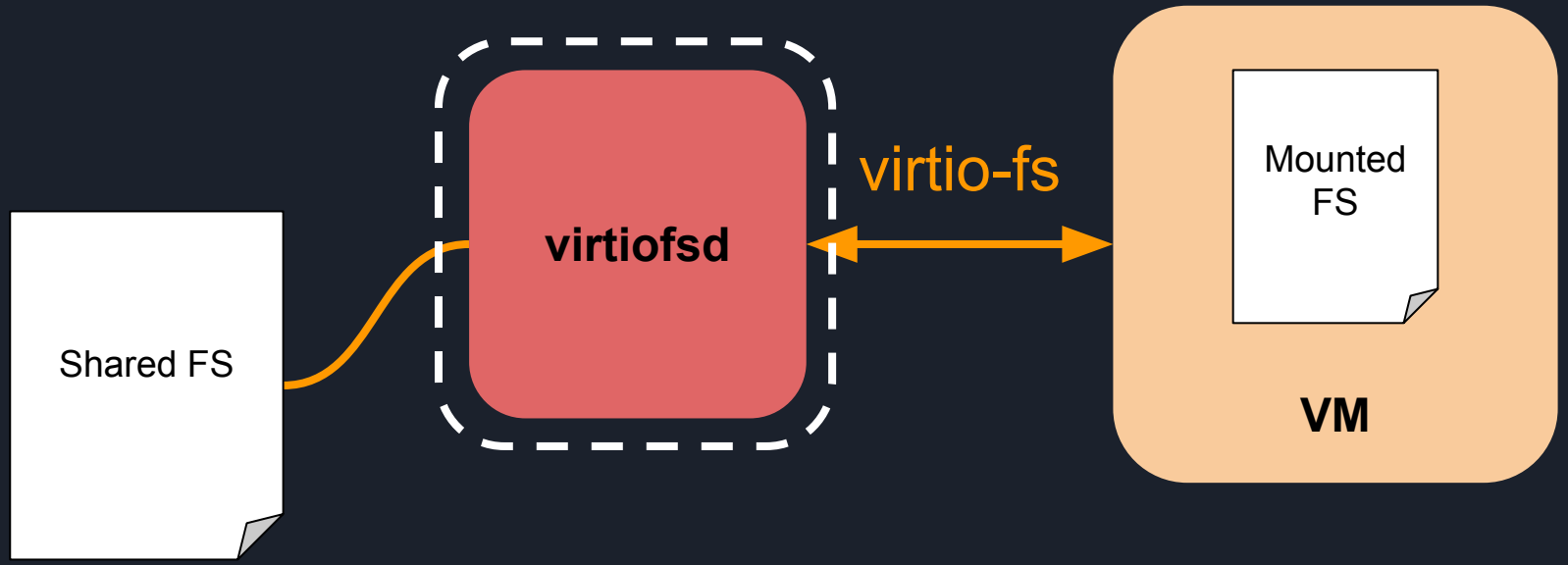
- Fully POSIX compliant ⇒ Solve workload functional issues
- As performant as virtio-blk (with DAX optimization)
- Overlay back into the picture for production

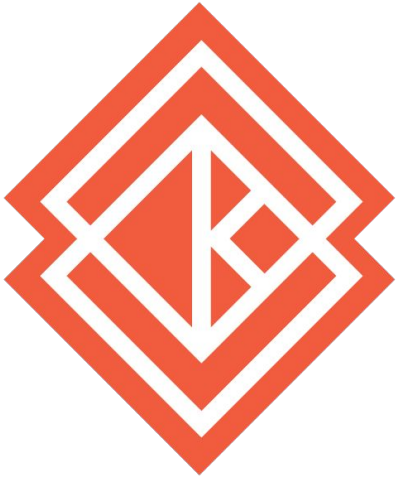
Shared filesystem

Shared filesystem



Shared filesystem





ACRN

QEMU/NEMU

- Swiss army knife hypervisor ⇒ Default for Kata
  - Type 2 (KVM)
  - Multi-purpose
  - Extensive device model (virtio-gpu, virtio-crypto, ...)
  - Direct Device Assignment (VFIO)
- Wide codebase in C ⇒ Potential attack surface
- NEMU reduces the attack surface

Firecracker

- Lightweight hypervisor
  - Type 2 (KVM)
  - Narrow focus: container workloads and FaaS
  - Reduced device model
- Small codebase in Rust ⇒ Highly secure

ACRN (in progress)

- Lightweight hypervisor
  - Type 1
  - Focus on Automotive and IoT
  - Industry standard **FuSa** (Functional Safety)
- Small codebase in C ⇒ Highly secure

http://www.lifeafterlondon.com/wp-content/uploads/2014/07/pick-your-own.jpg

Takeaways

Join the fun!

Sources: https://github.com/kata-containers/runtime

Get started: https://github.com/kata-containers/documentation/blob/master/Developer-Guide.md

Slack: katacontainers.slack.com

IRC: #kata-dev@freenode

Mailing list: kata-dev@lists.katacontainers.io

Thank you