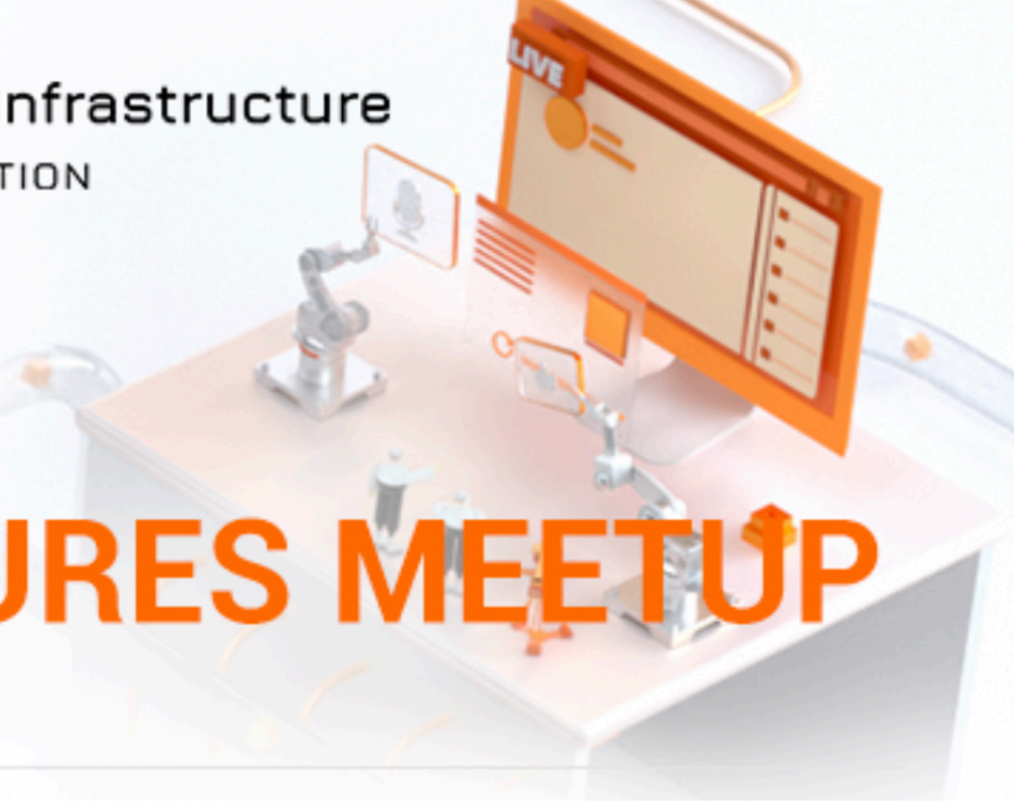
Open Anolis | Open Infrastructure FOUNDATION



CLOUD NATIVE INFRASTRUCTURES MEETUP

活动交流群

关注OpenAnolis.org

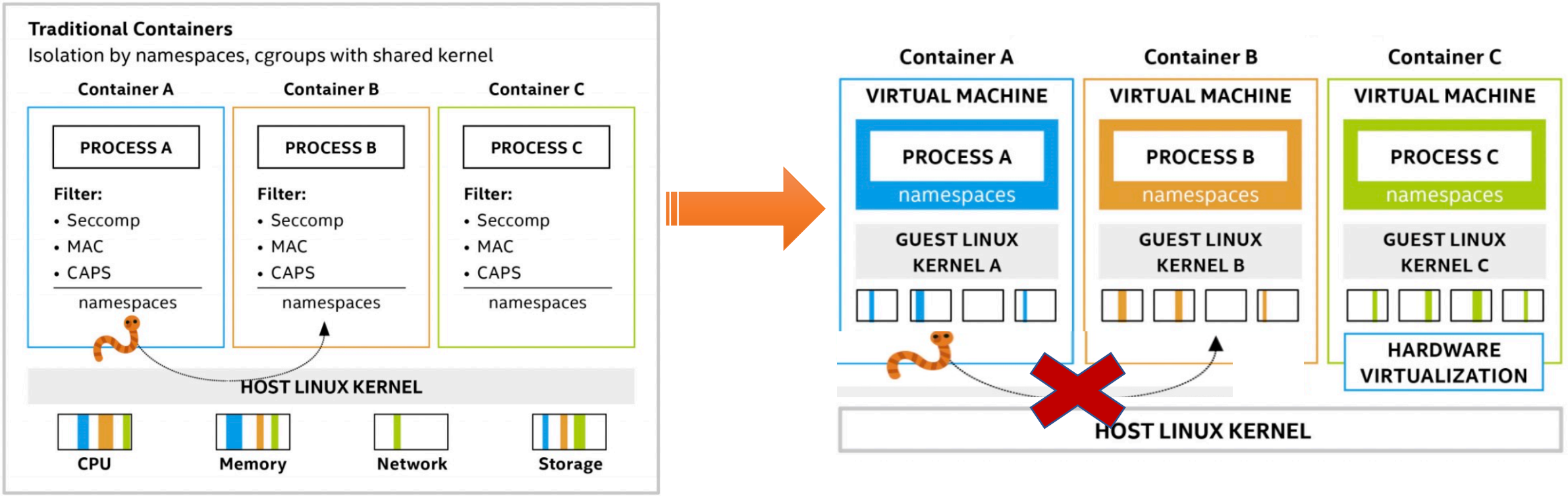
Kata Containers: When Virtualization Meets Cloud Native

Xu Wang @ Ant Group

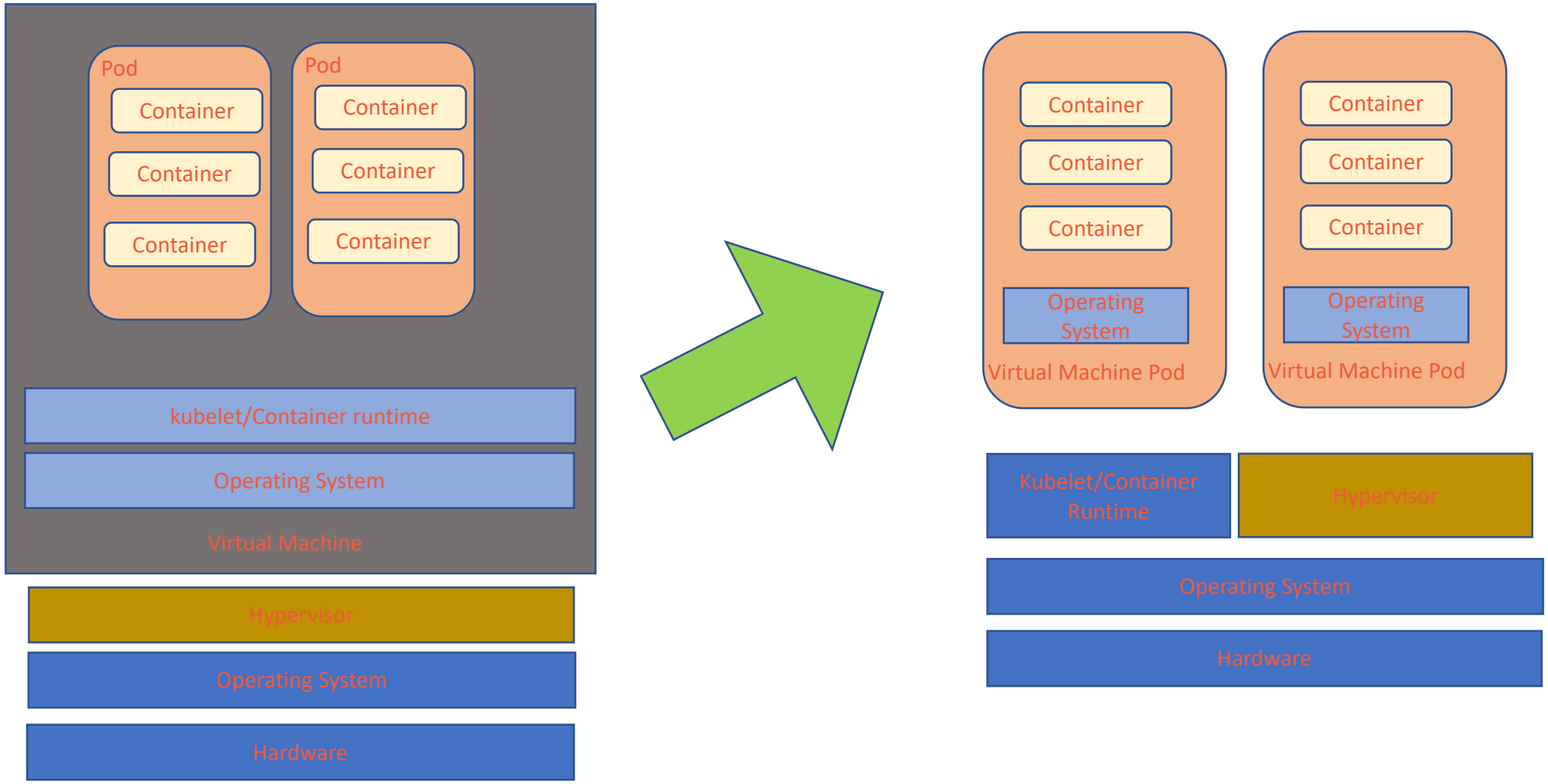
Index

- Kata Containers Introduction
- Kata Containers 2.0
- Kata @ Ant
- Future Works

What Is Kata Containers

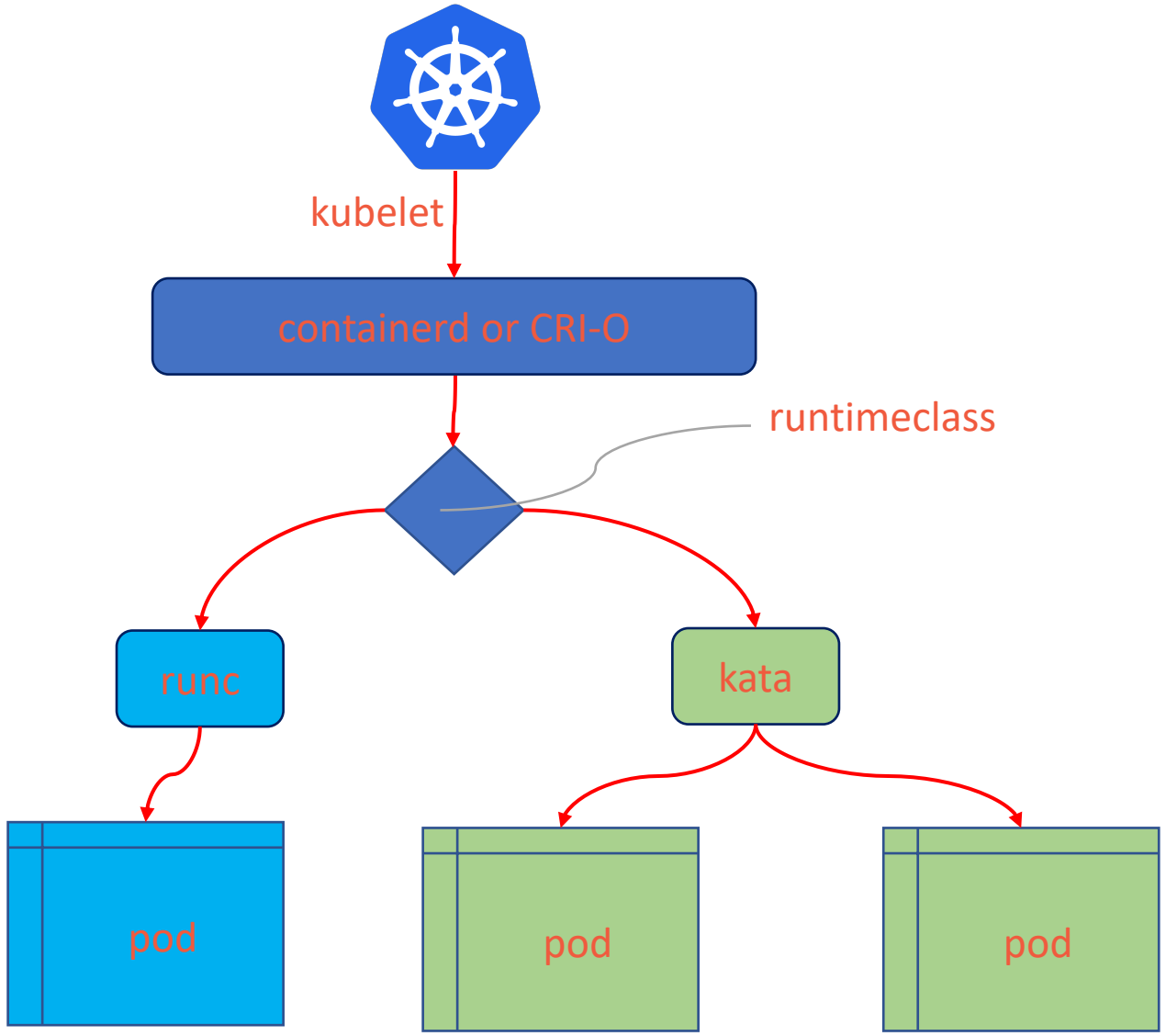


Virtual Machine as a Pod

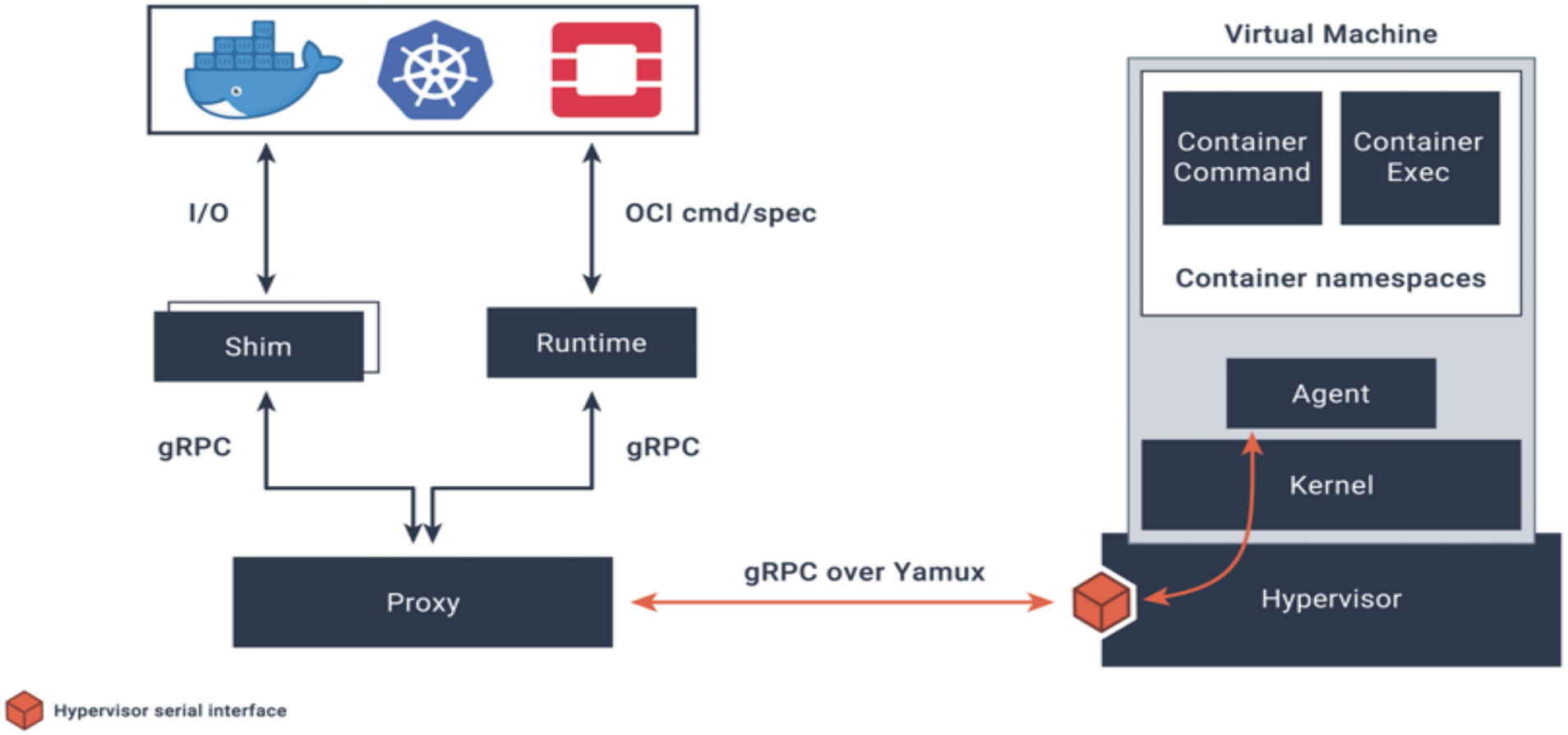


Kata & Kubernetes

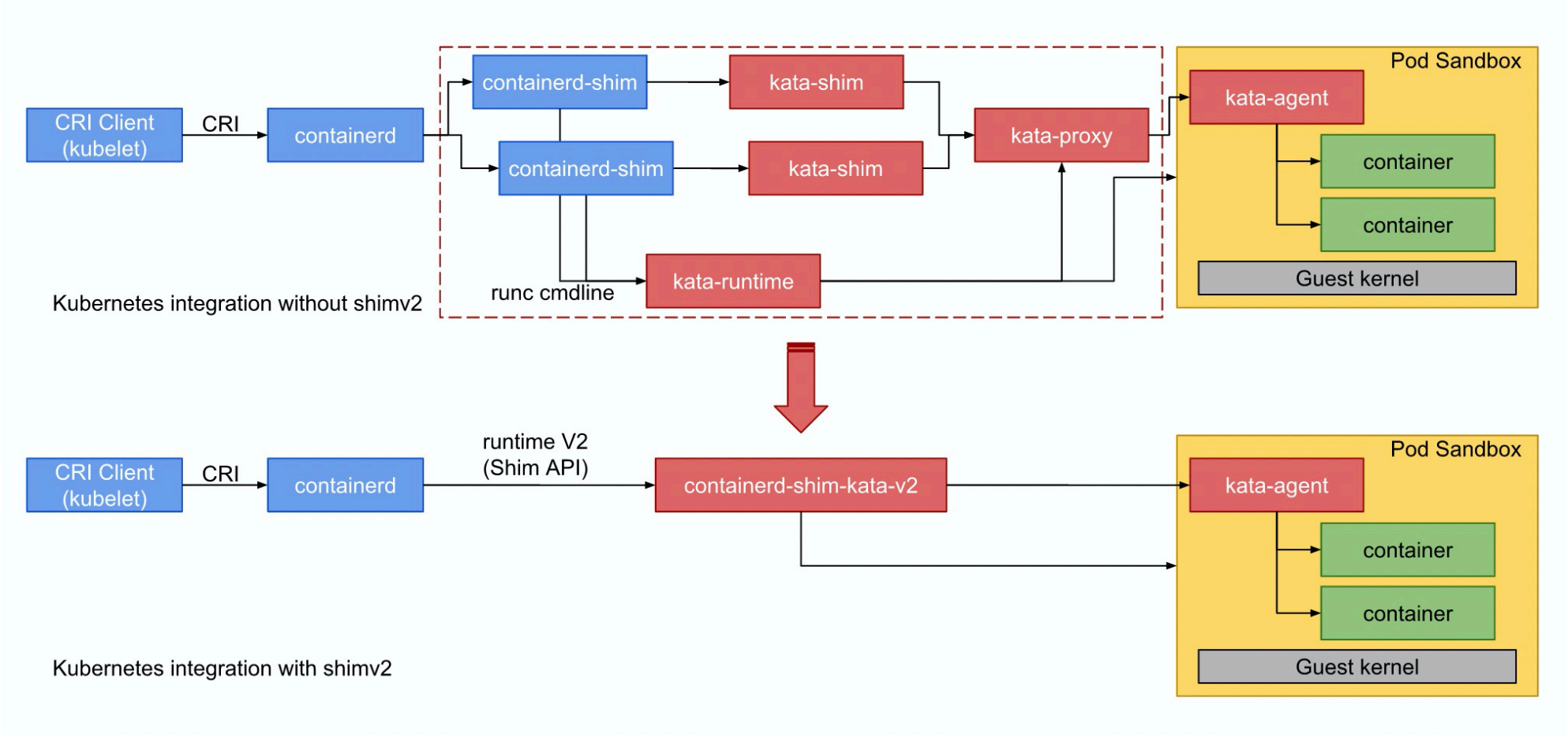
- RuntimeClass Resource
- Run runc and kata containers in the Same cluster/node
- Beta feature since v1.14
- runtimeClassName in pod spec



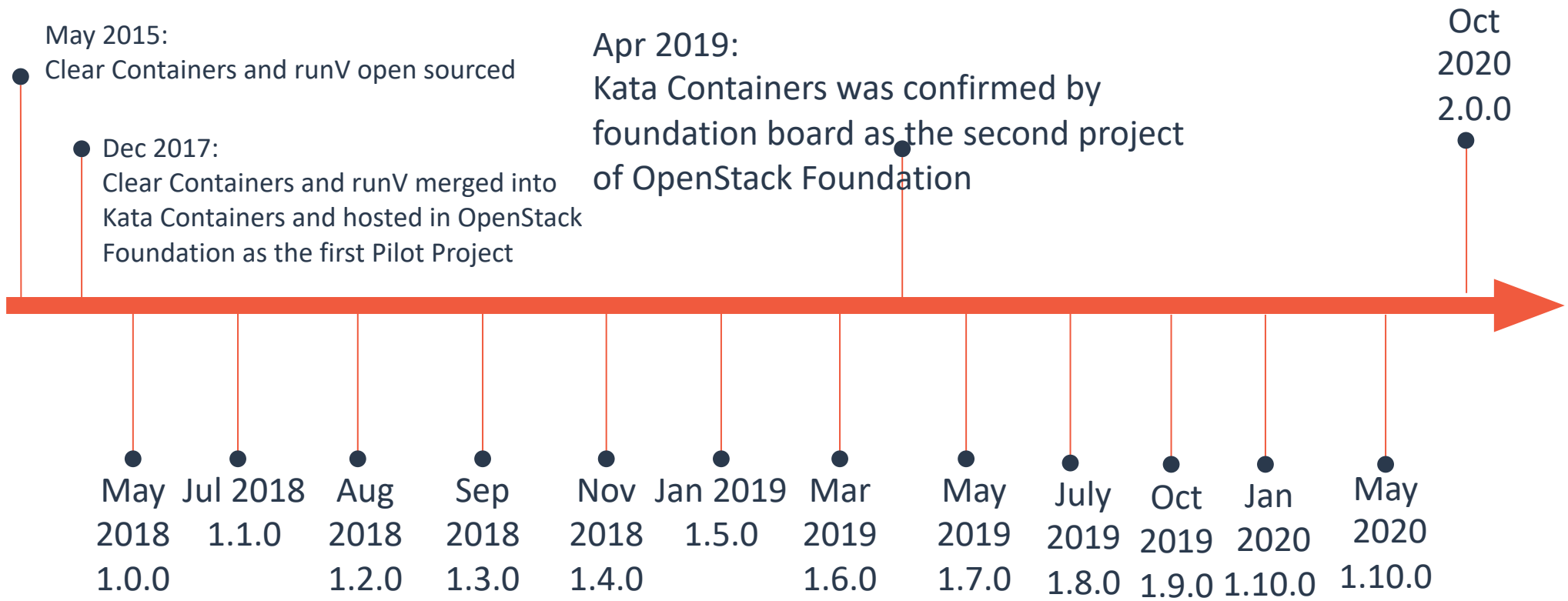
Kata Containers 1.x Architecture



Architecture Evolvement



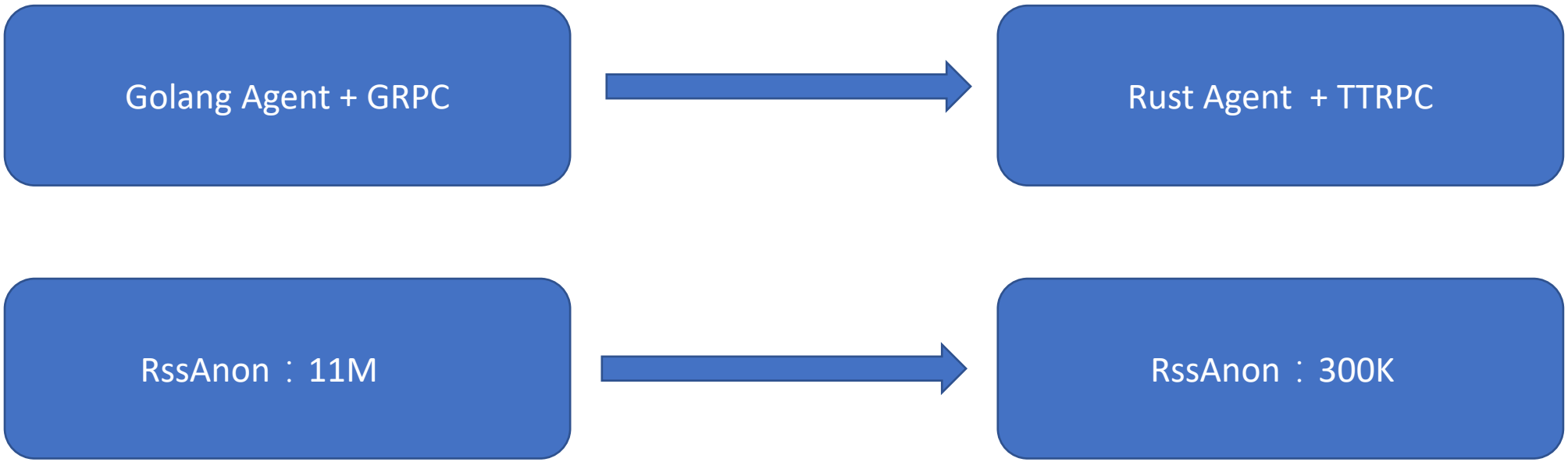
A bit history



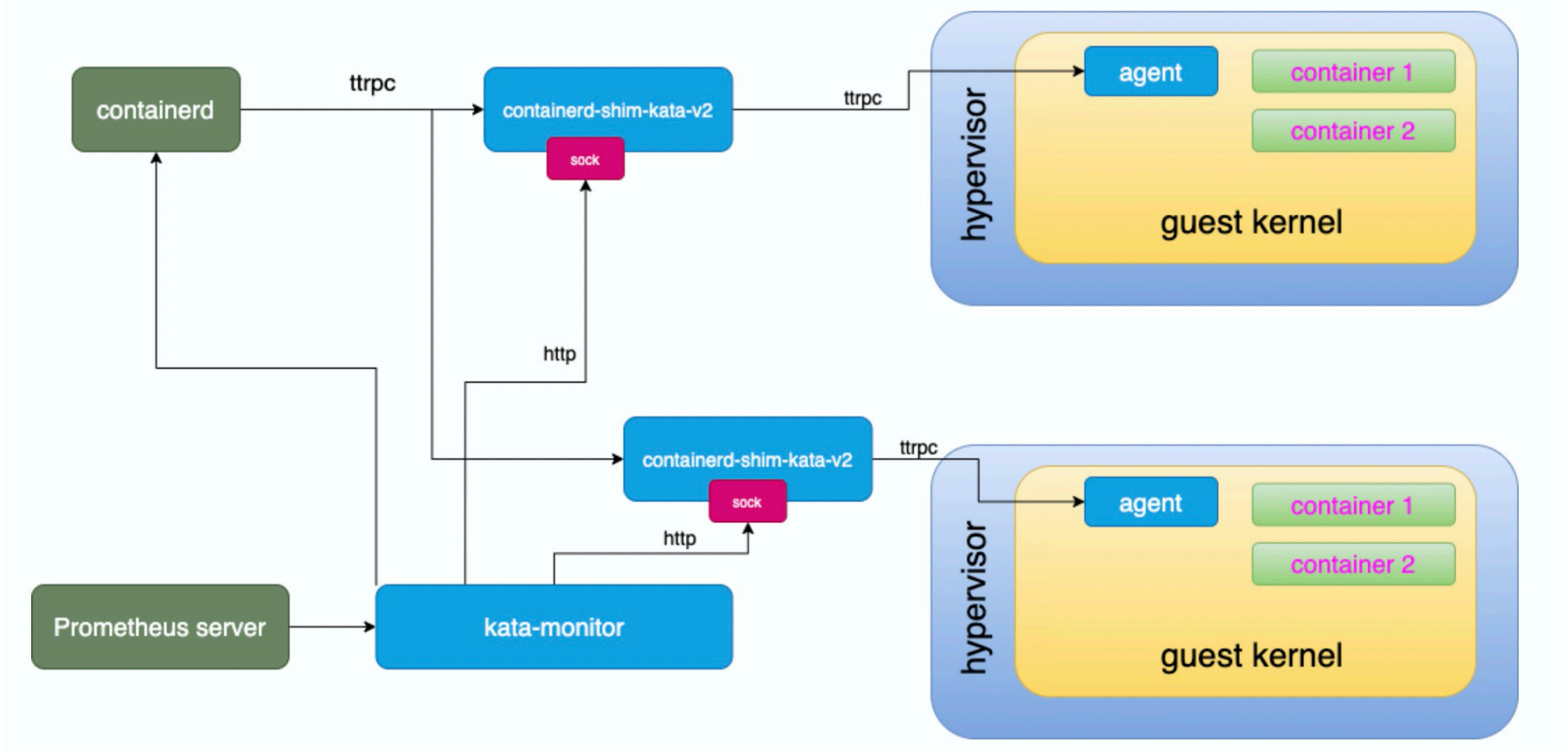
What's New In 2.0

- Kata-agent has been rewritten in rust to significantly reduce memory overhead overall attack surface.
- Agent protocol has been simplified to use ttRPC from grpc.
- Component called Kata-monitor has been introduced to improve observability and manageability
- New component called agent-ctl has been introduced to help validate agent API
- We have moved to a mono-repo model, all code and document repositories consolidated into one single repo.
- Virtio-fs is now the default shared file system type which mean better POSIX compliance compared to virtio-9p
- Latest Cloud Hypervisor support on par with QEMU
- Move to support solely shimv2 api to simplify code and reduce attack surface further. This also meanskata-shim and kata-proxy used in 1.x are no longer required.
- Guest kernel updated to v5.4.71
- QEMU updated to v5.0.0

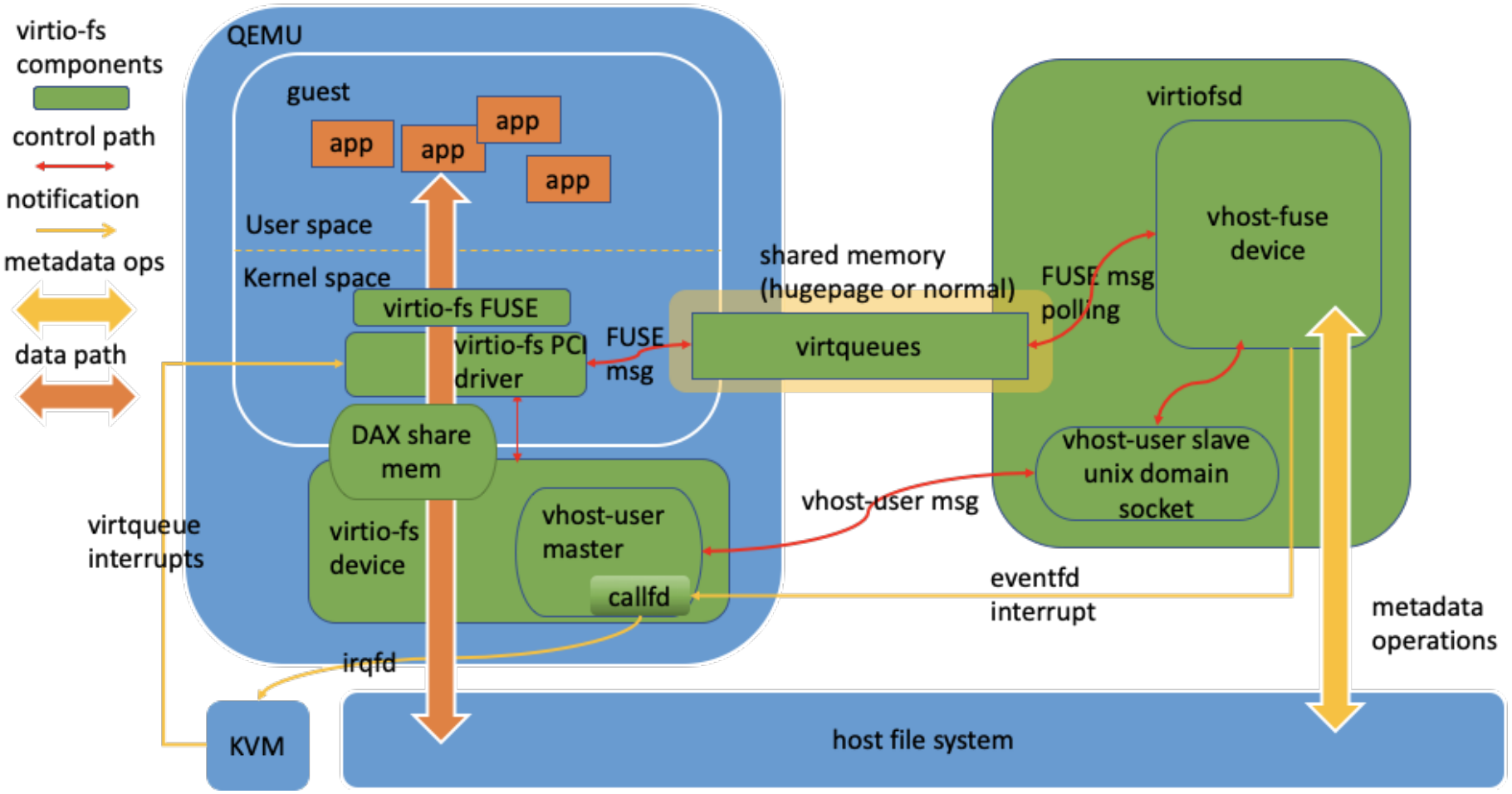
rust-agent + ttRPC



Kata Metrics



virtio-fs

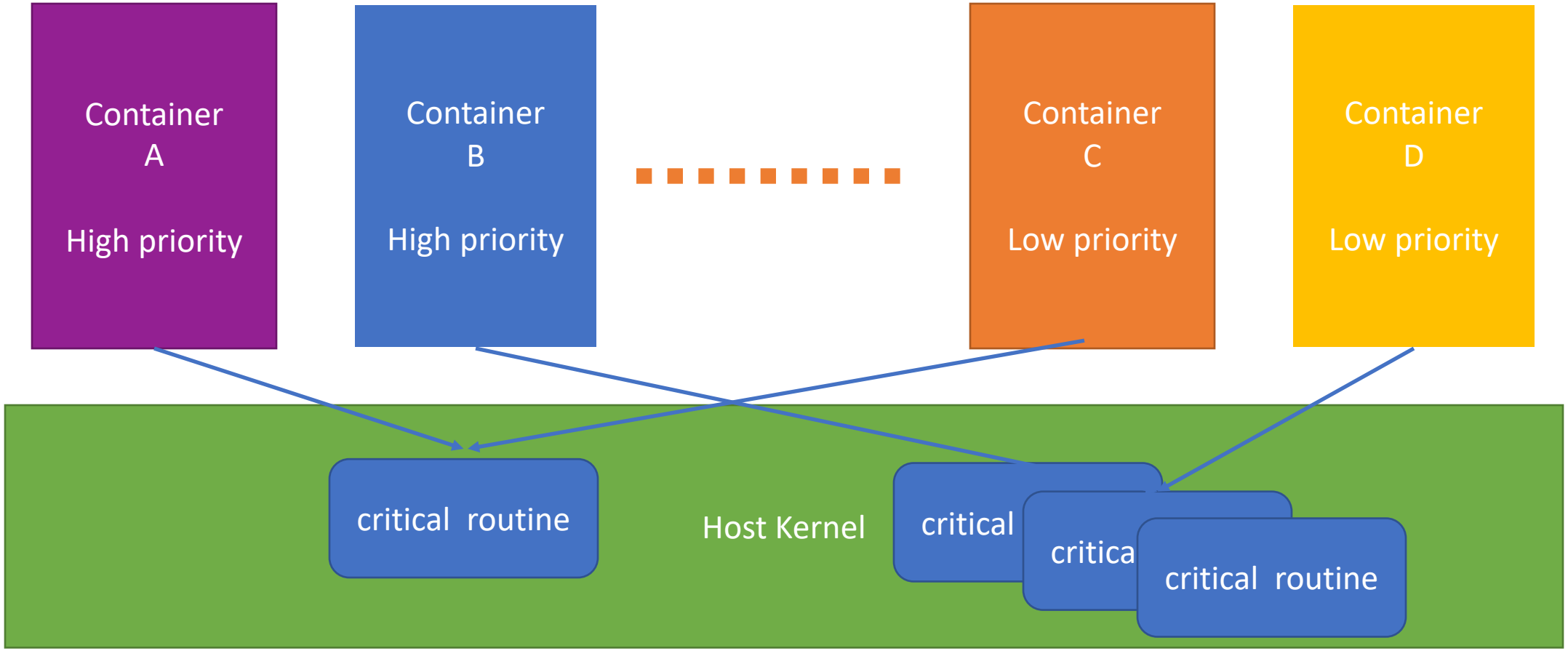

virtio-fs components
control path
notification
metadata ops
data path
QEMU
guest
app
app
app
app
User space
Kernel space
virtio-fs FUSE
virtio-fs PCI driver
FUSE msg
DAX share mem
virtio-fs device
vhost-user master
callfd
virtqueue interrupts
irqfd
KVM
shared memory (hugepage or normal)
virtqueues
virtiofsd
vhost-fuse device
FUSE msg polling
vhost-user slave unix domain socket
vhost-user msg
eventfd interrupt
metadata operations
host file system

Kata Containers & Ant

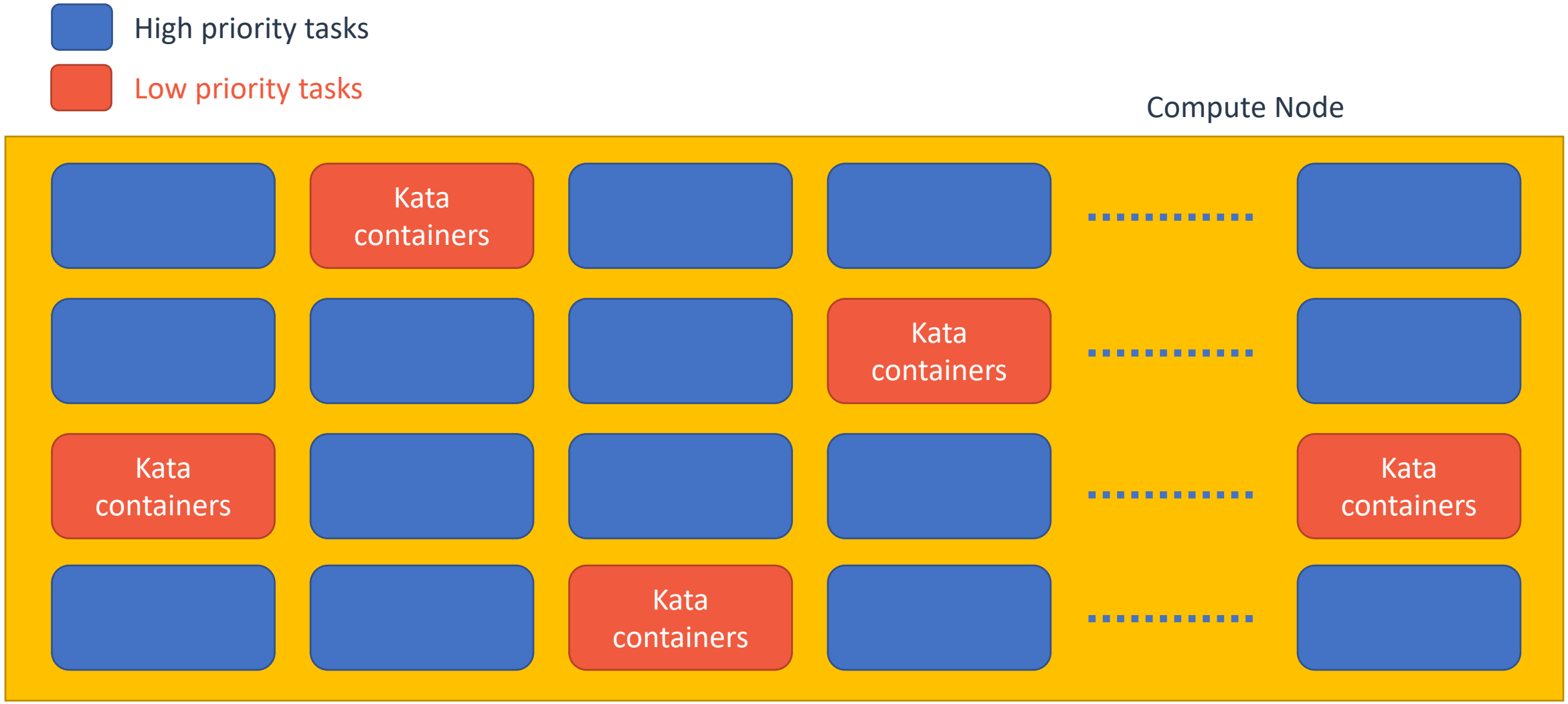
katacontainers

Conventional container workload isolation

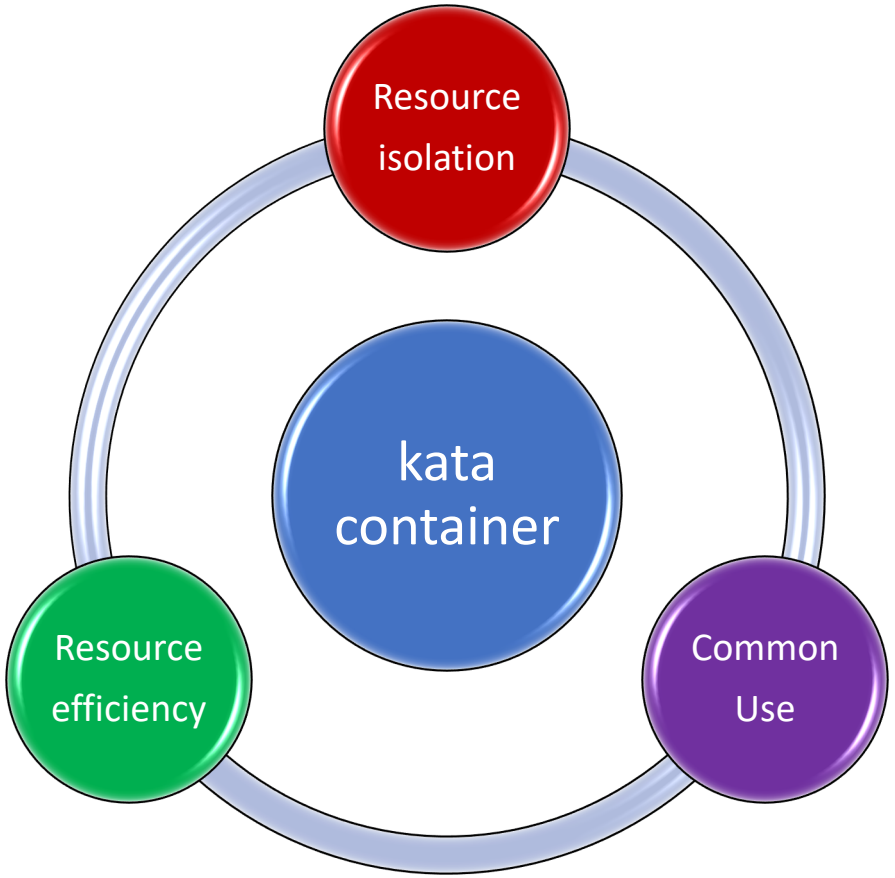
- Container A's response time would be delayed



Colocation of high/low priority workloads



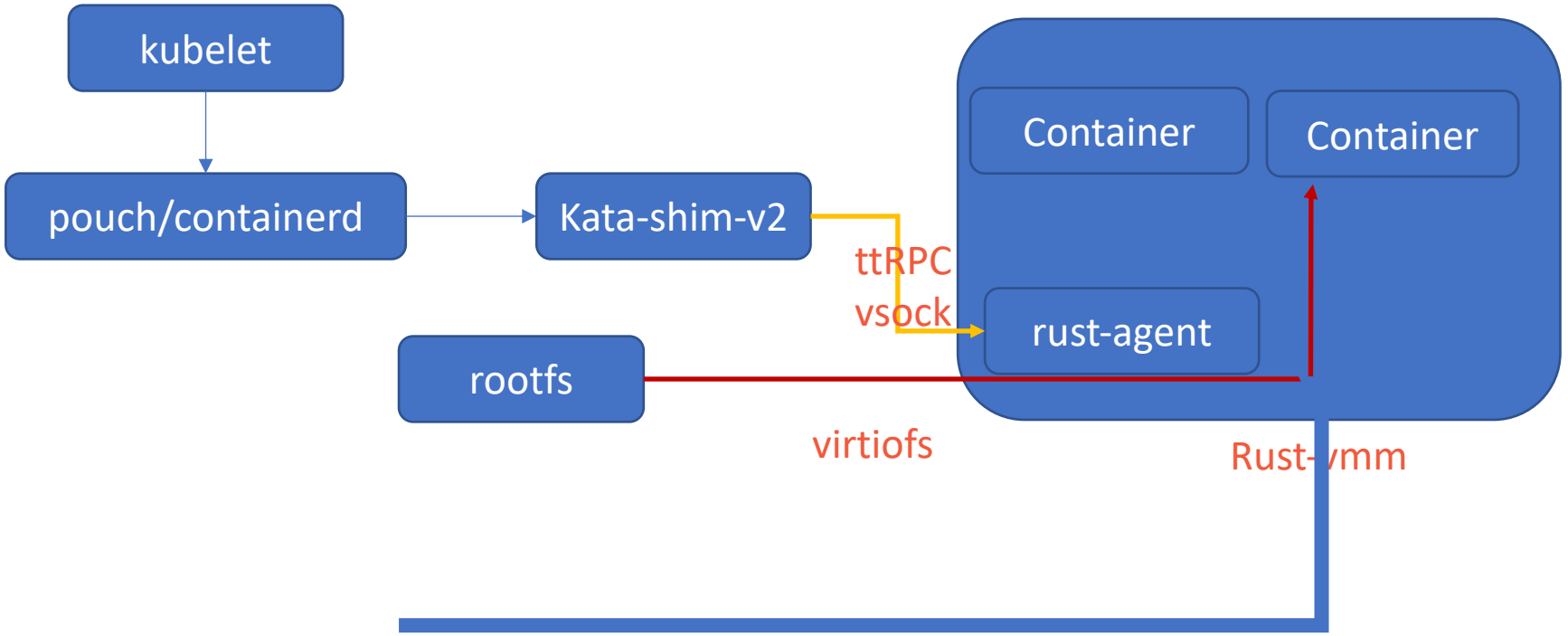
Art of Ballance



Kata Resource Improvements

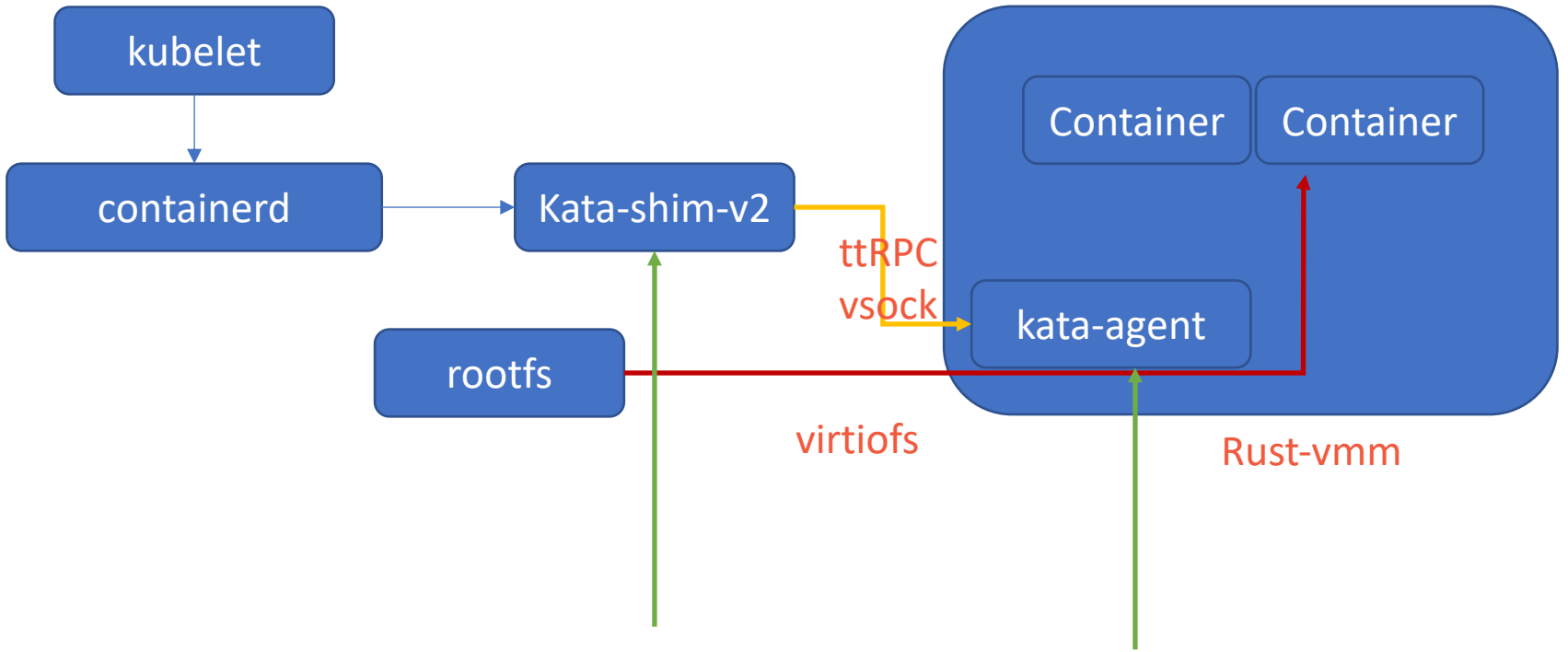
- Rust-VMM Hypervisor
- Rewrite kata-agent from Go to rust
- Replace GRPC with ttRPC
- Replace virtio-serial with vsock
- Replace 9pfs with virtiofs

Extra Ops Channel



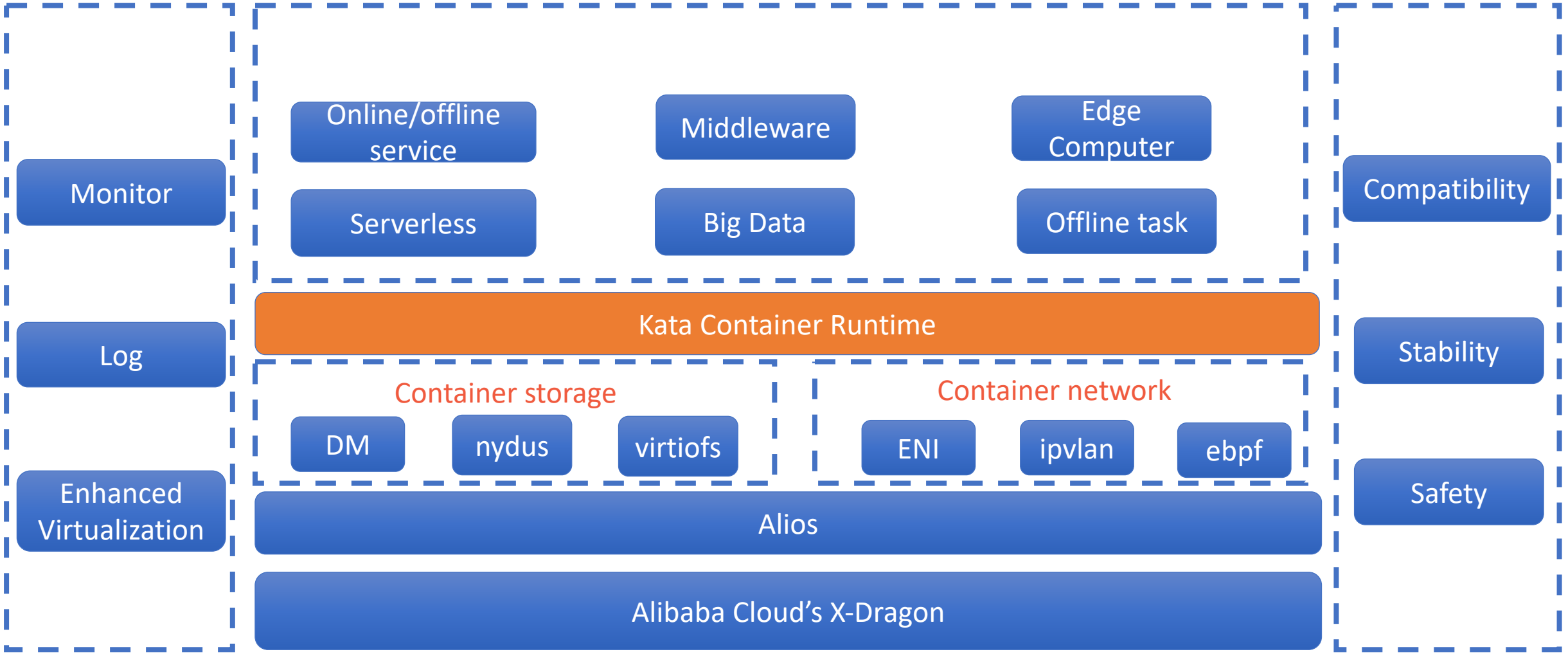
Add an Ops Channel

Hot Upgrade Kata Components



Kata-shim-v2 and kata-agent can be hot upgraded without stopping container tasks

Trust Native Infrustracture Conerstone



What's Next

- More isolation enhancements
- Container image operation improvements
- IO stream improvements
- Networking (port forwarding) improvements
- Debugbility improvements

Links

- website: http://katacontainers.io/
- github: https://github.com/kata-containers
- IRC freenode: #kata-dev
- Slack (bridged to IRC): http://bit.ly/katacontainersslack
- Twitter: @katacontainers
- Mailing list: lists.katacontainers.io