What’s new with Kata?

Eric Ernst - Apple
KVM Forum 2021

Agenda

- *Refresher* — What is Kata Containers
- What*’s* new
- What’s next

What is Kata?

Kata Containers

Open Source, Open Design, Open Development, Open Community

Kata Containers

Open Source, Open Design, Open Development, Open Community

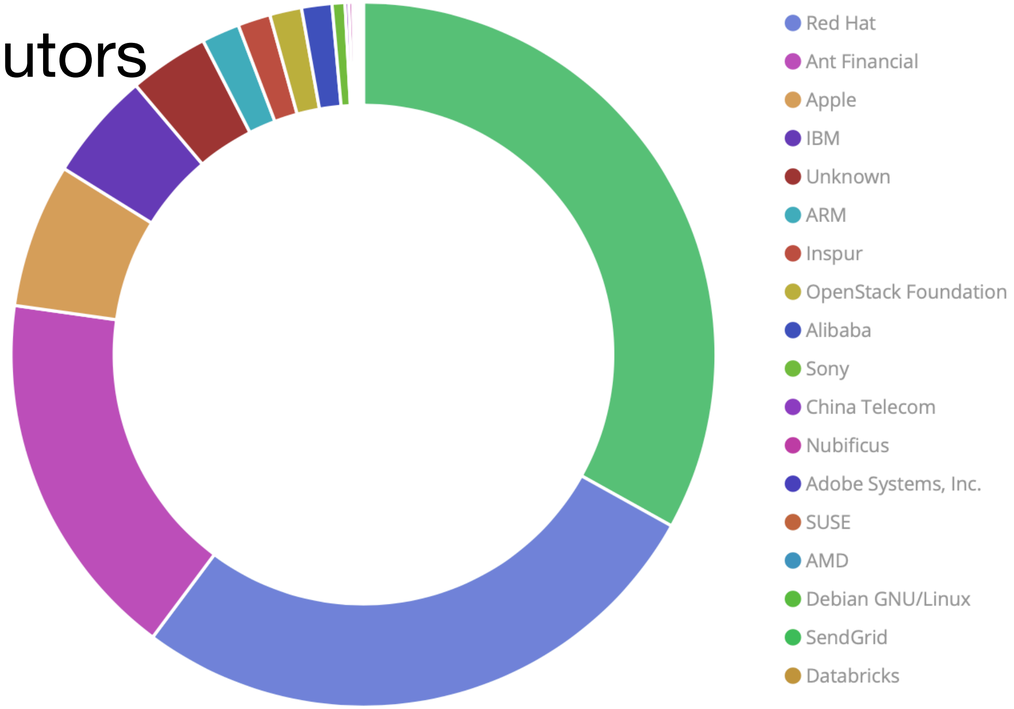
- Open Governance project with Architecture Committee:
  - Archana Shinde, Intel
  - Eric Ernst, Apple
  - Fabiano Fidencio, RedHat
  - Peng Tao, Ant Financial
  - Samuel Ortiz, Apple

Kata Containers

Open Source, Open Design, Open Development, Open Community

- Open Governance project with Architecture Committee:
  - Archana Shinde, Intel
  - Eric Ernst, Apple
  - Fabiano Fidencio, RedHat
  - Peng Tao, Ant Financial
  - Samuel Ortiz, Apple

- Contributors



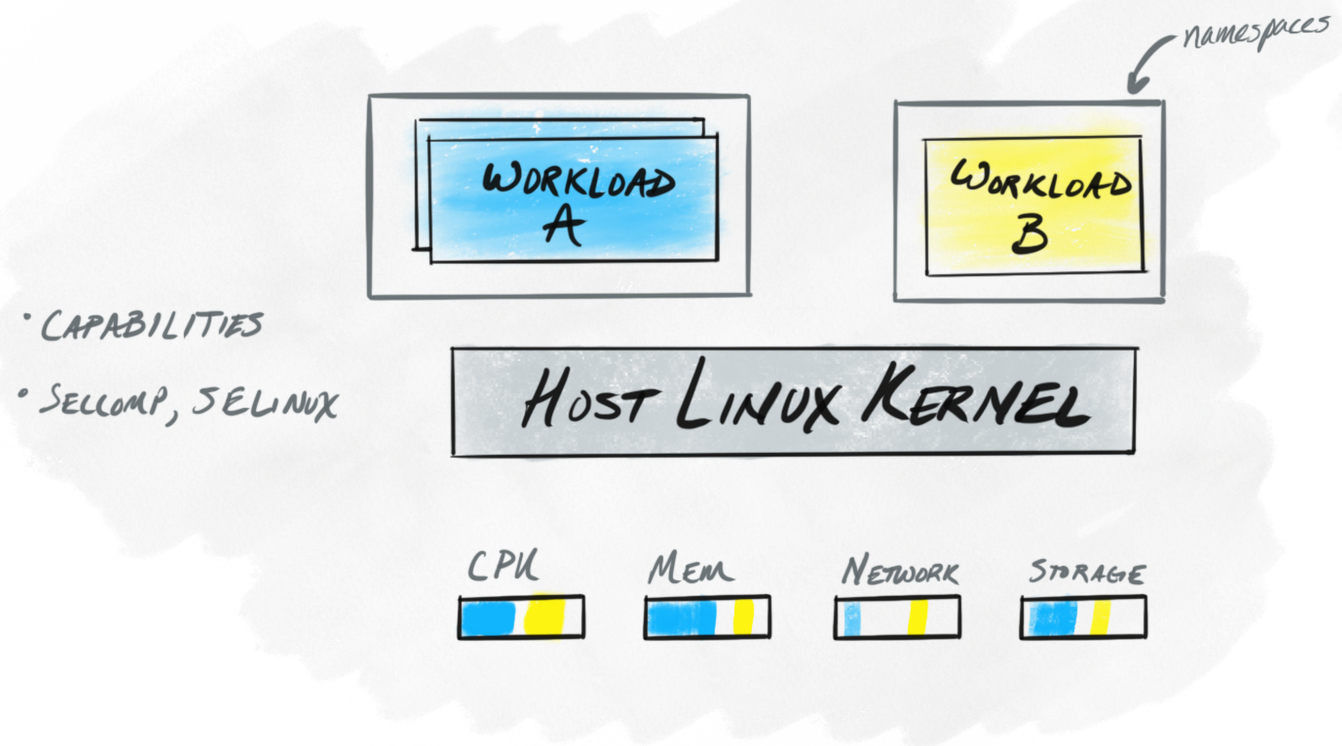
Source: https://katacontainers.biterg.io/

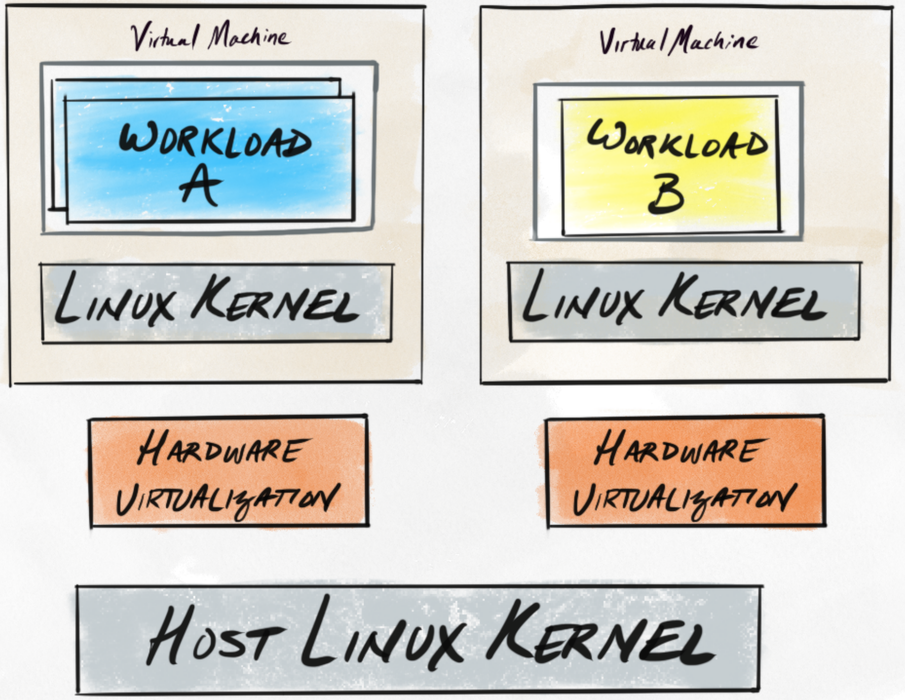
Kata Containers is a *secure container runtime* that provides stronger workload isolation using hardware virtualization technology as a second layer of defense.

Primary threat model

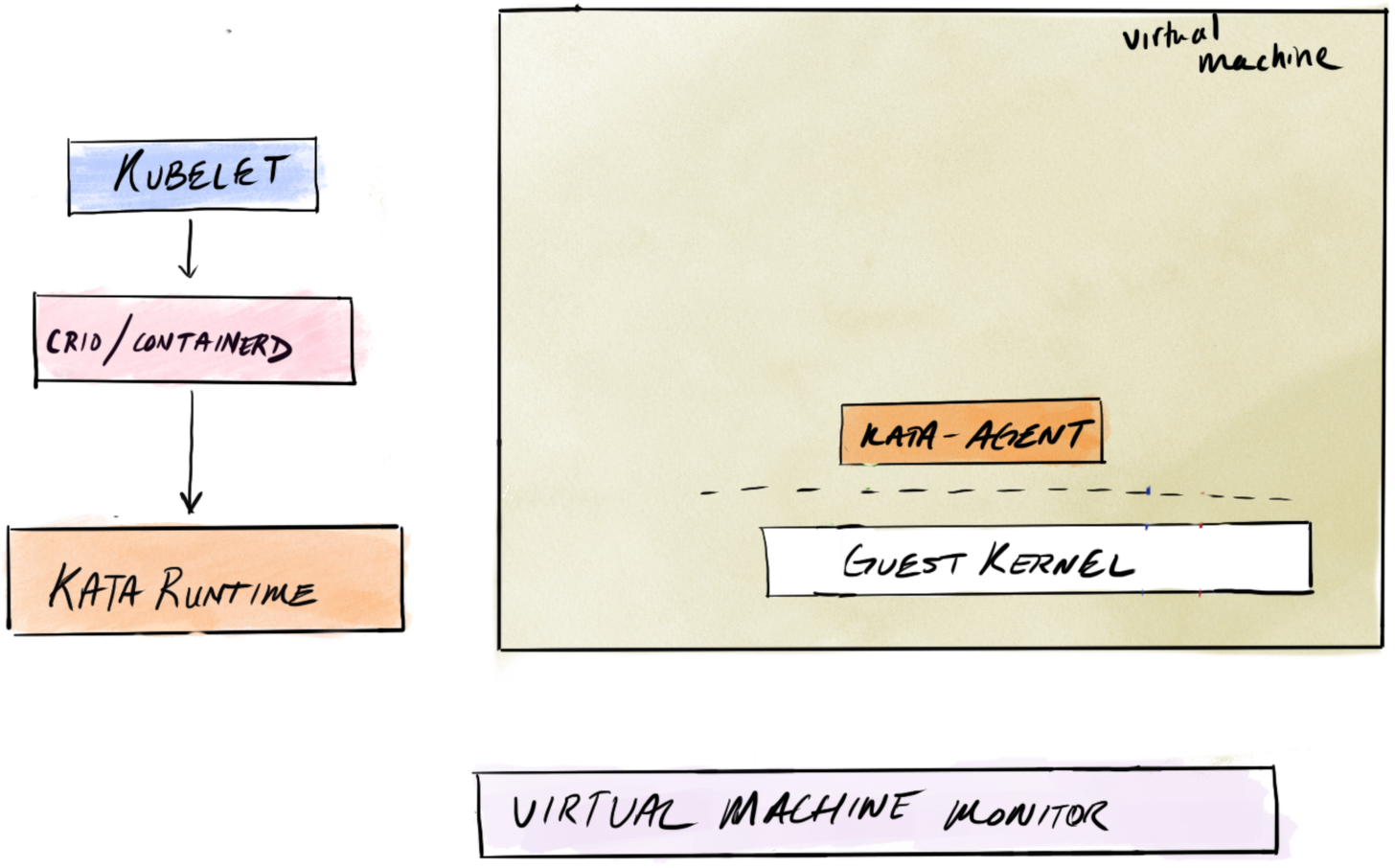
Primary threat model

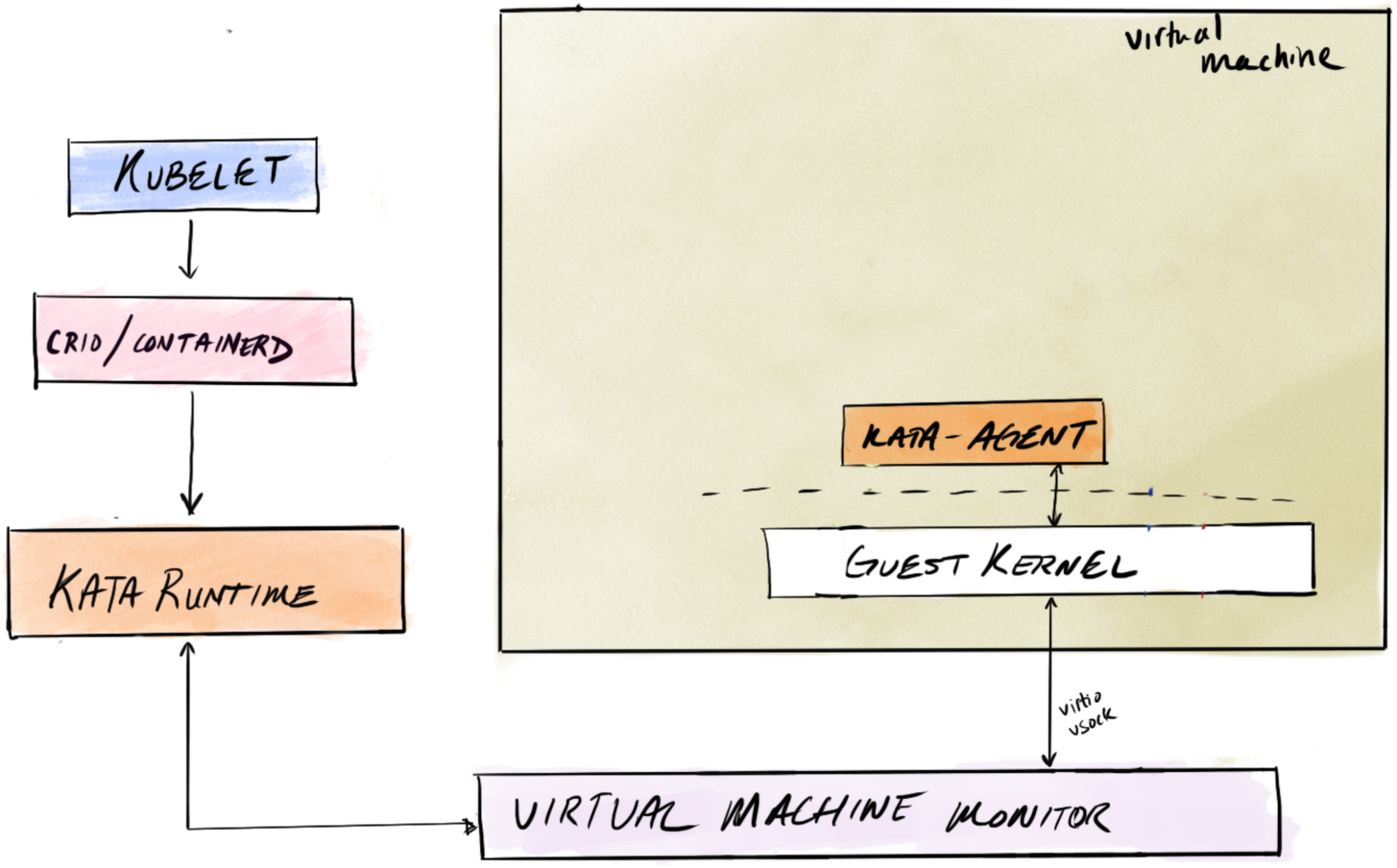
- ✓ Protect host infrastructure, including other workloads
- ✓ Assume each workload is untrusted

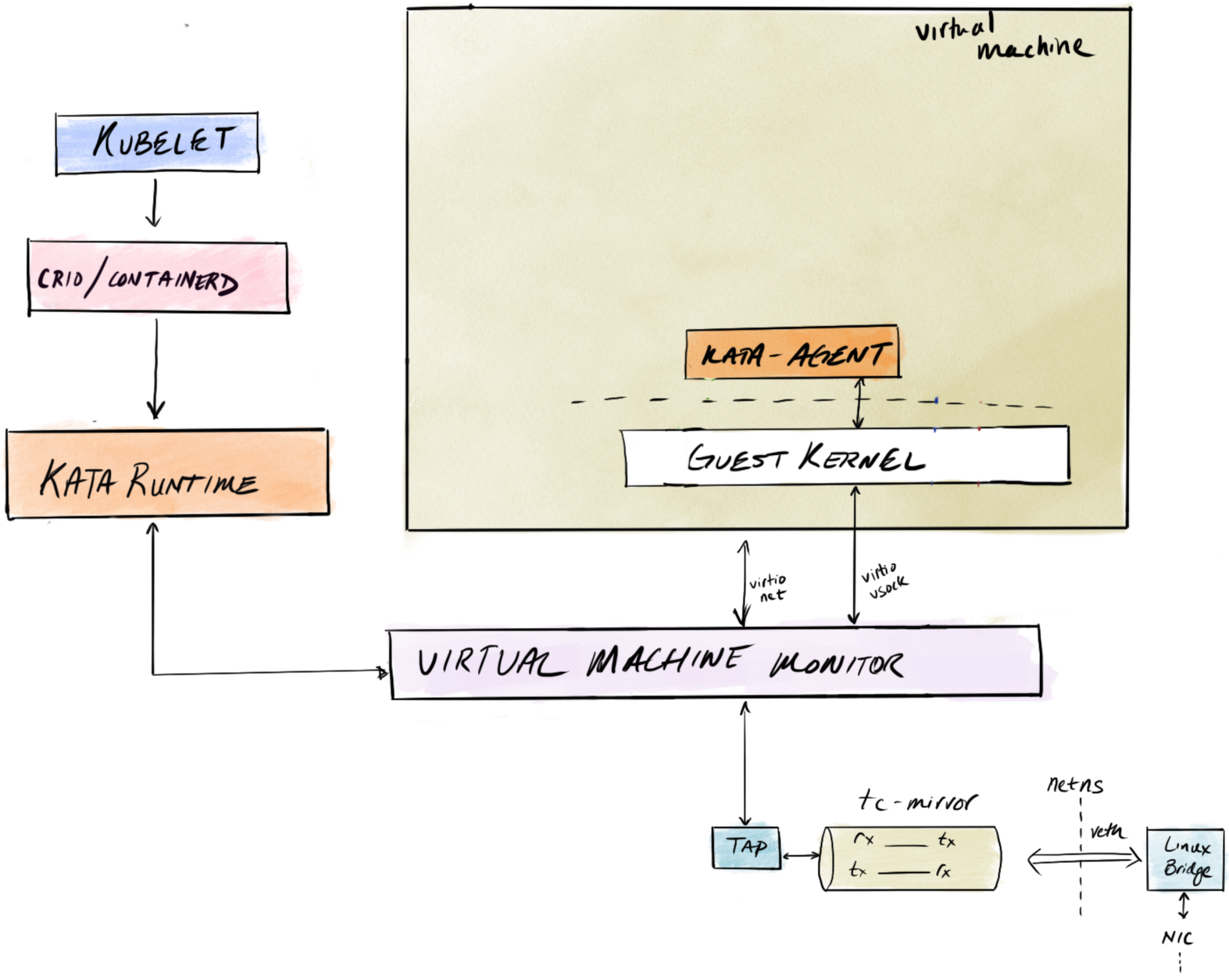

namespaces
WORKLOAD A
WORKLOAD B
• CAPABILITIES
• SECCOMP, SELINUX
HOST LINUX KERNEL
CPU
MEM
NETWORK
STORAGE

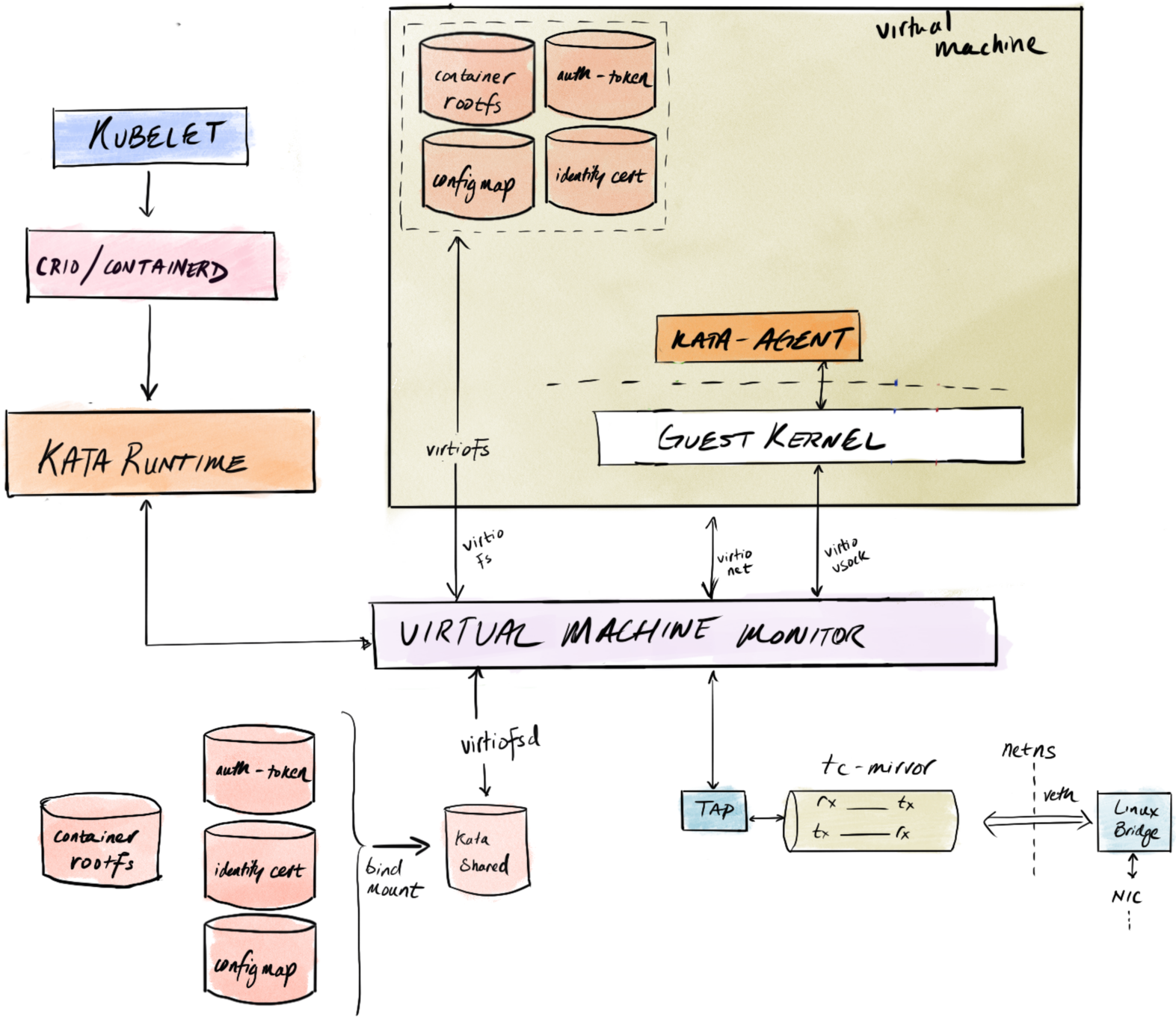

Virtual Machine
WORKLOAD A
LINUX KERNEL
Virtual Machine
WORKLOAD B
LINUX KERNEL
HARDWARE VIRTUALIZATION
HARDWARE VIRTUALIZATION
HOST LINUX KERNEL

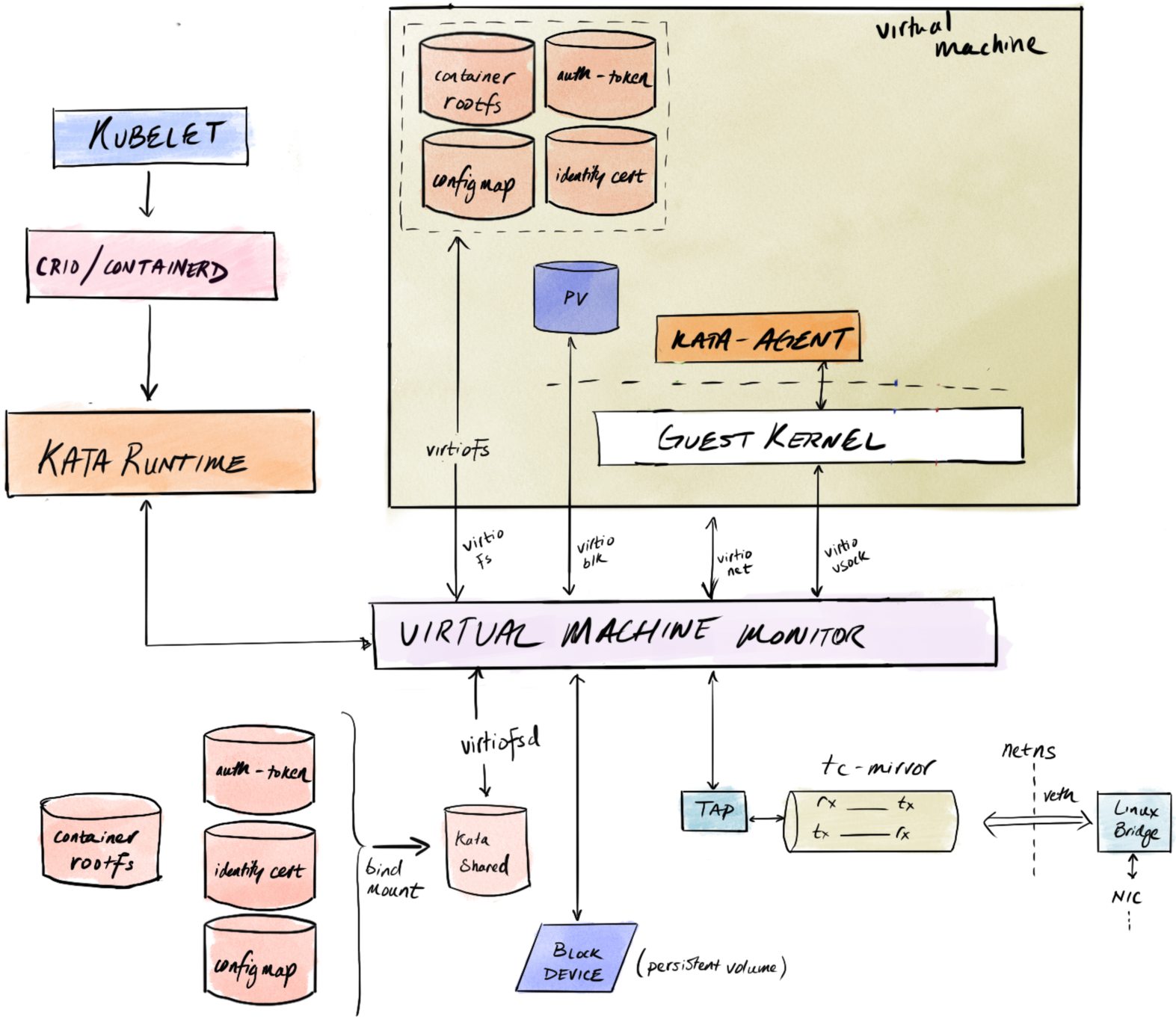
How to wrap a pod in a VM

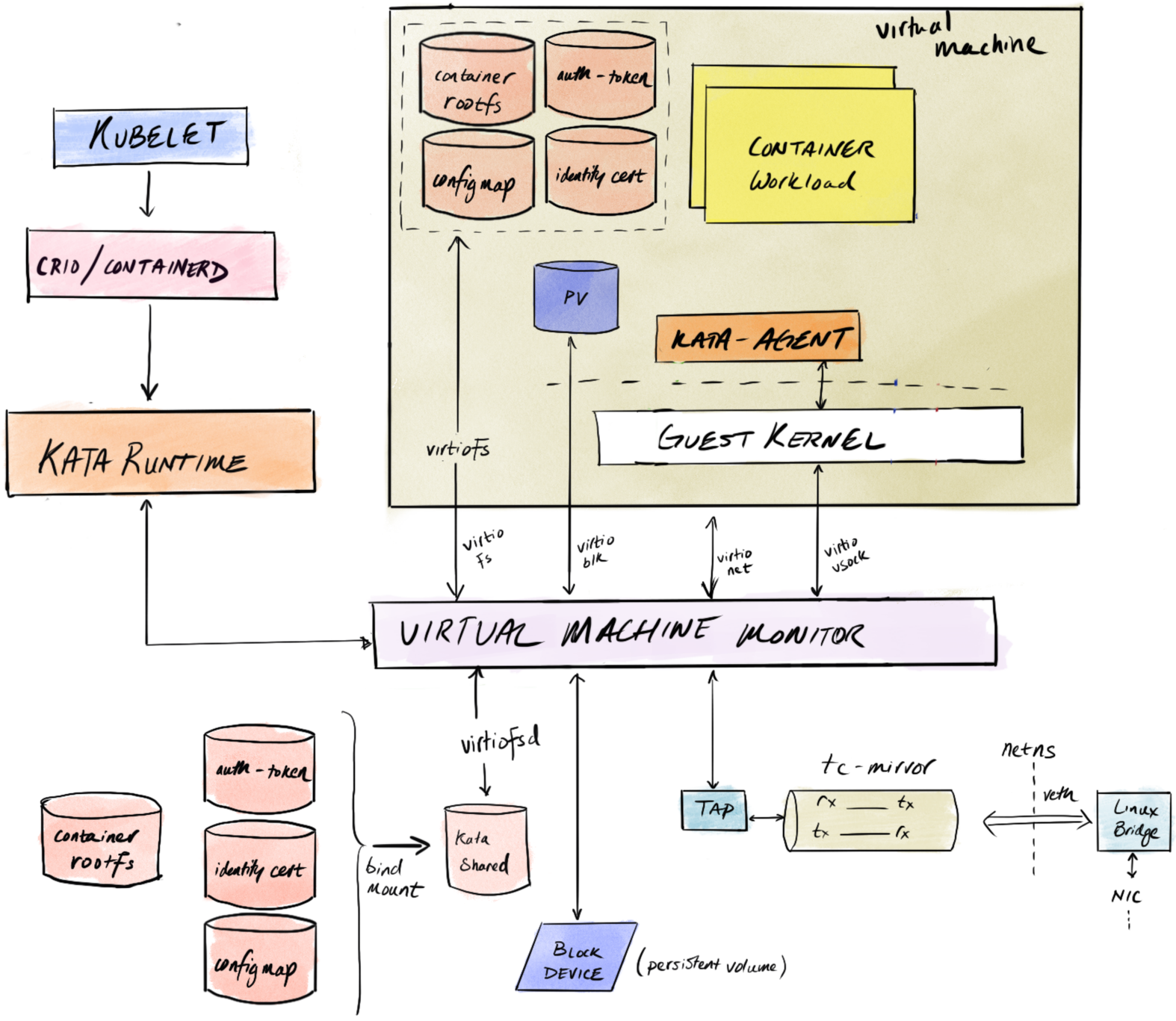

KUBELET
CRIO / CONTAINERD
KATA RUNTIME
virtual machine
KATA-AGENT
GUEST KERNEL
VIRTUAL MACHINE MONITOR

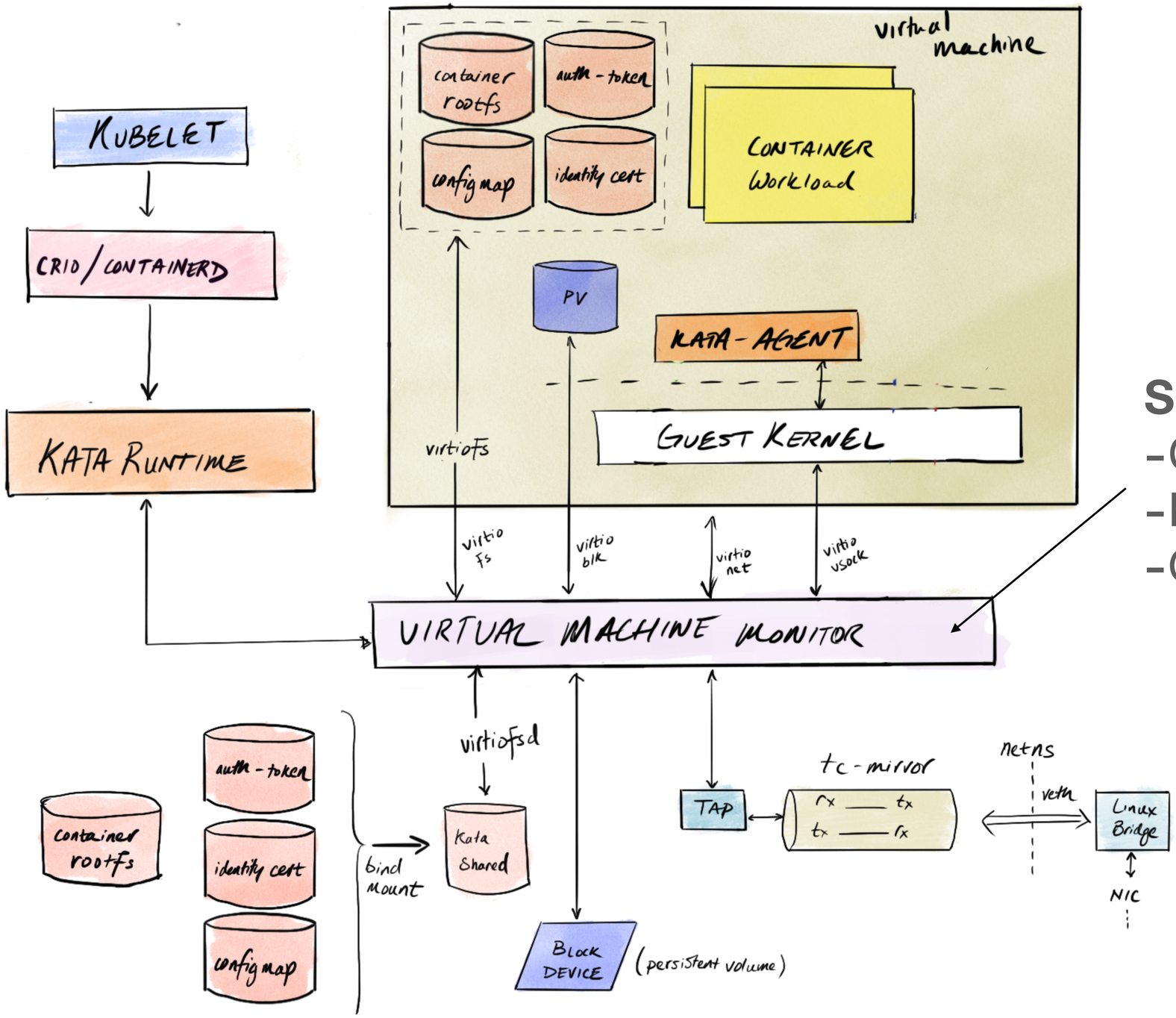

KUBELET
CRIO / CONTAINERD
KATA RUNTIME
virtual machine
KATA-AGENT
GUEST KERNEL
virtio vsock
VIRTUAL MACHINE MONITOR


virtual machine
KUBELET
CRIO / CONTAINERD
KATA RUNTIME
KATA-AGENT
GUEST KERNEL
virtio net
virtio vsock
VIRTUAL MACHINE MONITOR
tc-mirror
rx
tx
tx
rx
TAP
netns
veth
Linux Bridge
NIC


virtual machine
container rootfs
auth-token
config map
identity cert
KUBELET
CRIO / CONTAINERD
KATA RUNTIME
KATA-AGENT
GUEST KERNEL
virtioFs
virtio fs
virtio net
virtio vsock
VIRTUAL MACHINE MONITOR
virtiofsd
container rootfs
auth-token
identity cert
config map
bind mount
Kata Shared
TAP
tc-mirror
rx
tx
tx
rx
netns
veth
Linux Bridge
NIC


virtual machine
container rootfs
auth-token
config map
identity cert
KUBELET
CRIO / CONTAINERD
KATA RUNTIME
PV
KATA-AGENT
GUEST KERNEL
virtioFs
virtio fs
virtio blk
virtio net
virtio vsock
VIRTUAL MACHINE MONITOR
virtiofsd
bind mount
Kata Shared
container rootfs
auth-token
identity cert
config map
Block DEVICE
(persistent volume)
TAP
tc-mirror
rx
tx
tx
rx
netns
veth
Linux Bridge
NIC


KUBELET
CRIO / CONTAINERD
KATA RUNTIME
virtual machine
container rootfs
auth-token
config map
identity cert
CONTAINER Workload
PV
KATA-AGENT
GUEST KERNEL
virtioFs
virtio fs
virtio blk
virtio net
virtio vsock
VIRTUAL MACHINE MONITOR
virtiofsd
container rootfs
auth-token
identity cert
config map
bind mount
Kata Shared
BLOCK DEVICE
(persistent volume)
TAP
tc-mirror
rx
tx
tx
rx
netns
veth
Linux Bridge
NIC



Supported VMMs:

- Cloud Hypervisor
- Firecracker
- QEMU

What’s new?

Kata 2.x improvements, updates

- Agent reimplemented in rust, asynchronous support added
- Default shared filesystem now virtio-fs
- virtio-fs limitation for inotify
  - Implement workaround in Kata to present watchable mount
- Open-telemetry support
- QEMU: Default machine type now Q35

What’s next?

Performance Improvements

Performance Improvements

Networking

Provide a VM native interface when feasible

Storage

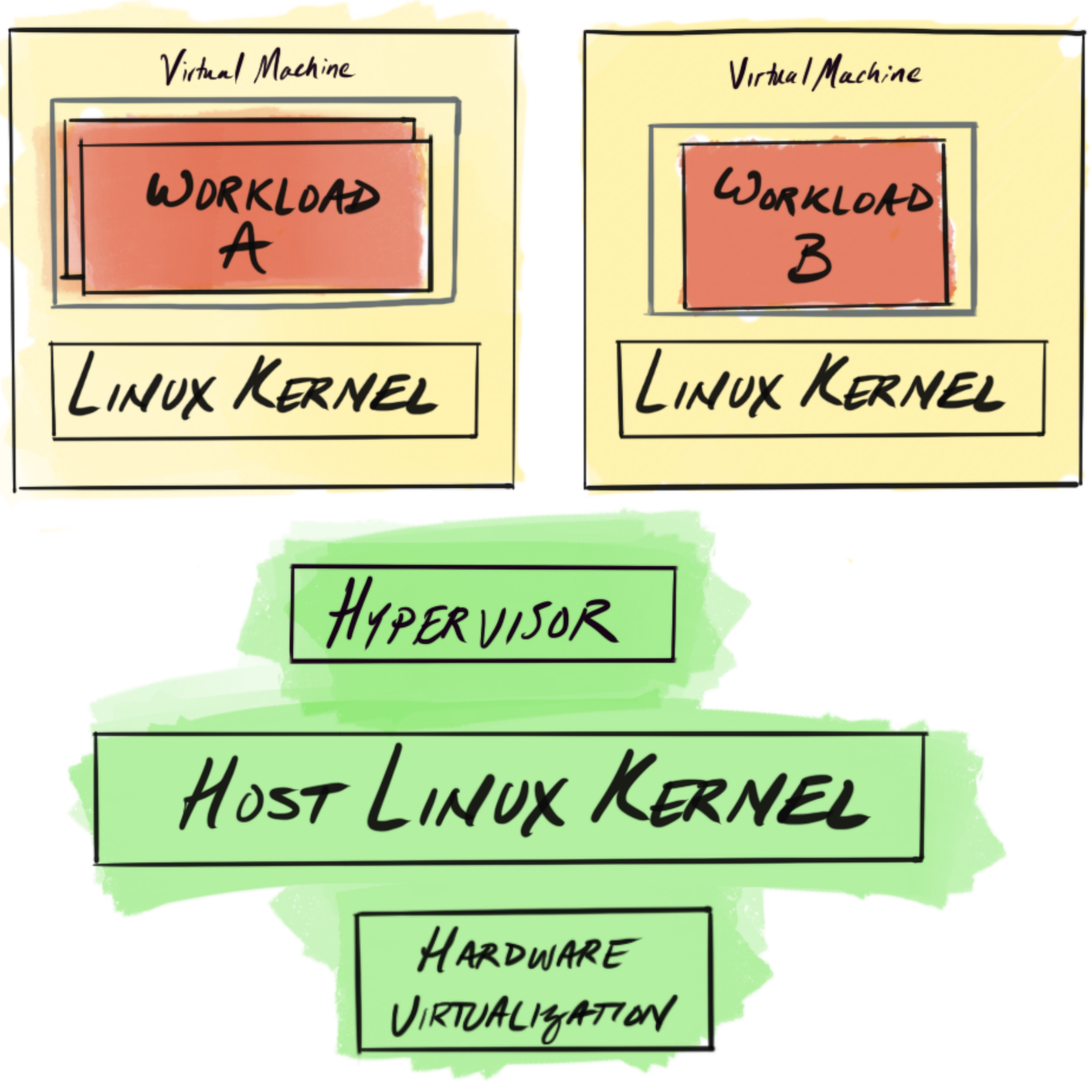
Directly consume block device at virtual machine layer

CPU performance Isolation

Better cpuset support: use cpuset in guest as well as host

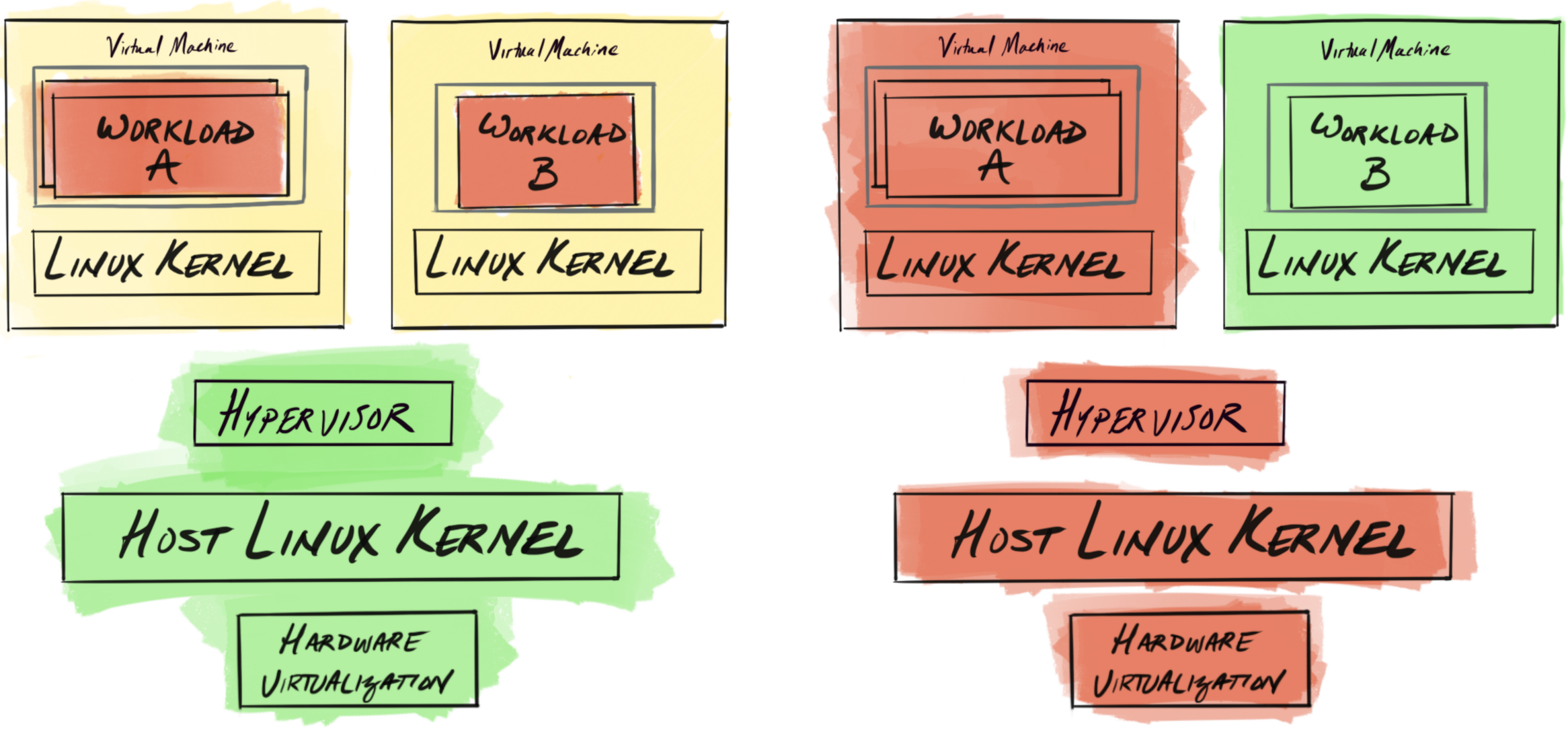
Isolating user vCPU from VMM/IO threads

Confidential Computing



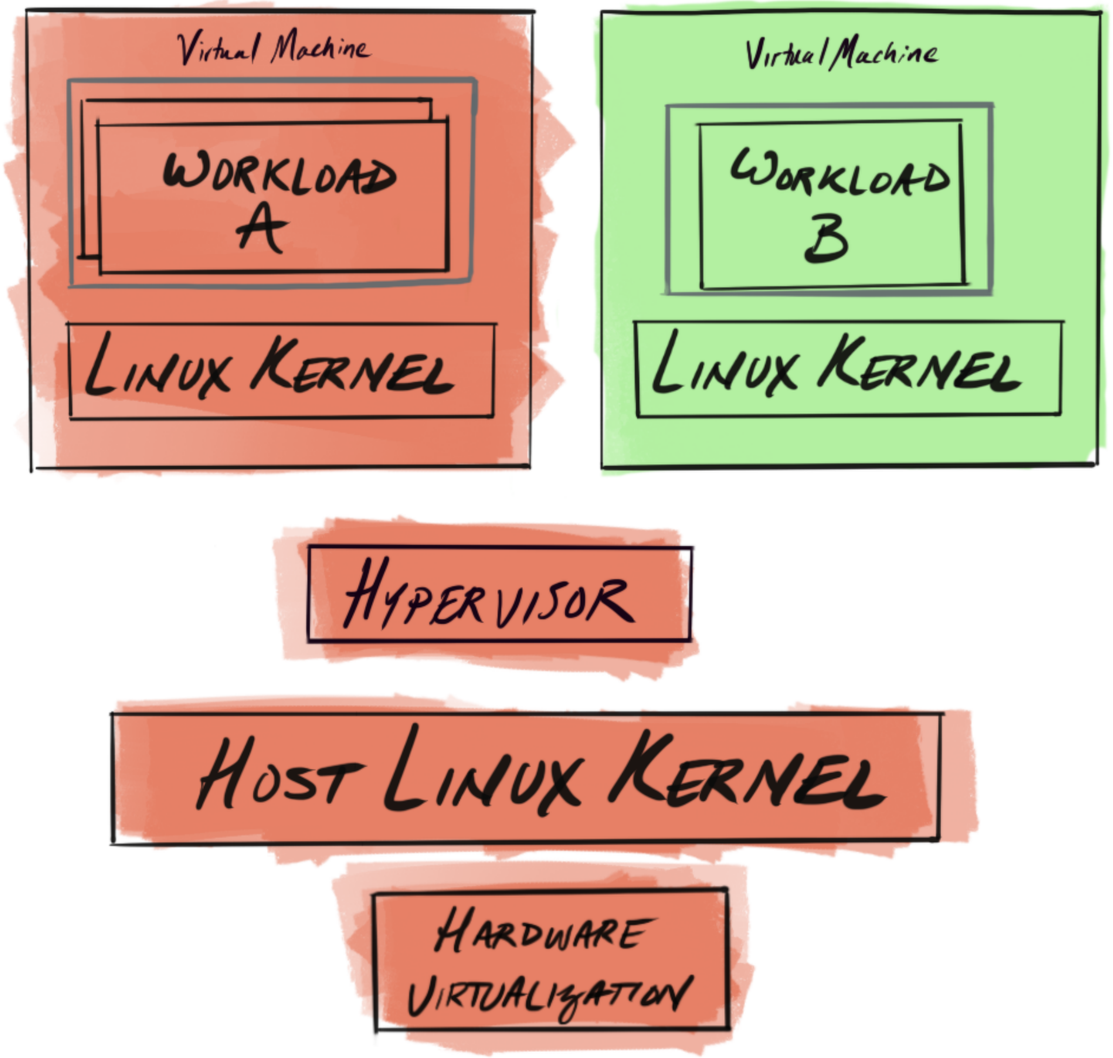
Existing threat model

- Don’t trust the workload.
- Prevent workload <—> workload attacks
- Prevent workload —> host attacks
- Workload is forced to trust host / provider


Virtual Machine
WORKLOAD A
LINUX KERNEL
Virtual Machine
WORKLOAD B
LINUX KERNEL
HYPERVISOR
HOST LINUX KERNEL
HARDWARE VIRTUALIZATION
Virtual Machine
WORKLOAD A
LINUX KERNEL
Virtual Machine
WORKLOAD B
LINUX KERNEL
HYPERVISOR
HOST LINUX KERNEL
HARDWARE VIRTUALIZATION

Extended threat model

- Workload does not trust the cloud service provider
- Host is outside the trust boundary
- Prevent host —> workload attacks



Kata, Confidential Computing

Protect data while it is being used/processed

- Memory and CPU state encryption and integrity checking

Attest the tenant software and hardware stack

- Hardware based quoting
- Software stack measurement

Hardware implementations, support

AMD SEV, IBM PEF and secure execution, Intel TDX

Hypervisor Dependencies

Intel TDX KVM upstreaming WIP

Hypervisor Dependencies

Get involved!

- github.com/kata-containers
- Apache 2.0 license
- katacontainers.io
- Slack: #kata-dev on bit.ly/kataslack