Running Isolated VirtualClusters with Kata & Cluster API

Eric Ernst & Chris Hein

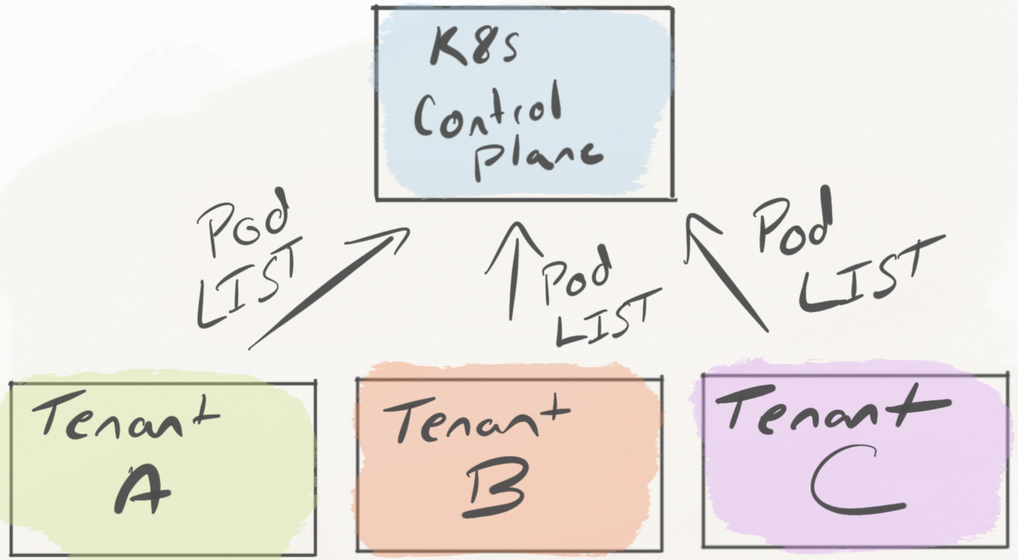
Multi-Tenancy in Kubernetes is hard.

Hard Multi-tenancy is even harder.

Multi-Tenant Control Planes

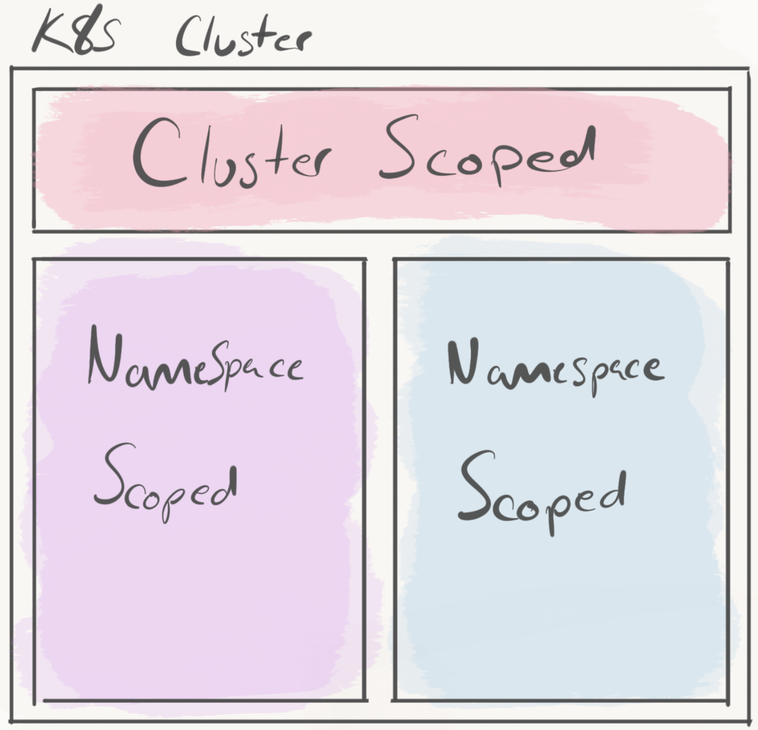
Issues with Multi-Tenant Control Planes

✗ Clumsy Clients



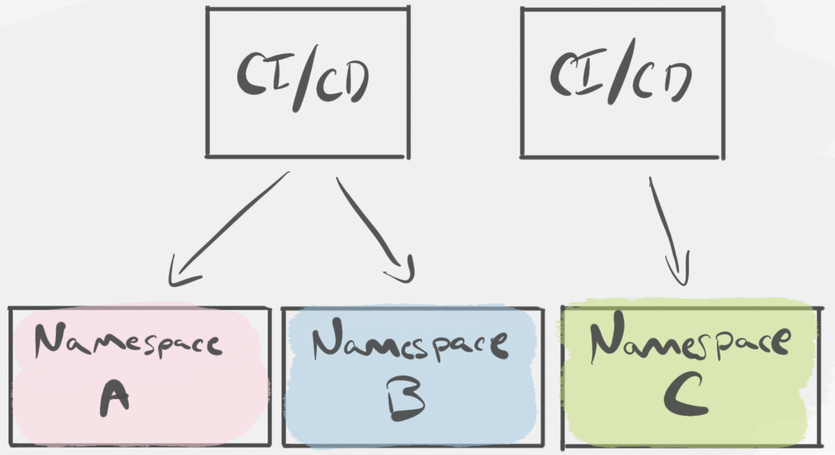
Issues with Multi-Tenant Control Planes

- ✗ Clumsy Clients
- ✗ Access to Cluster Scoped Resources

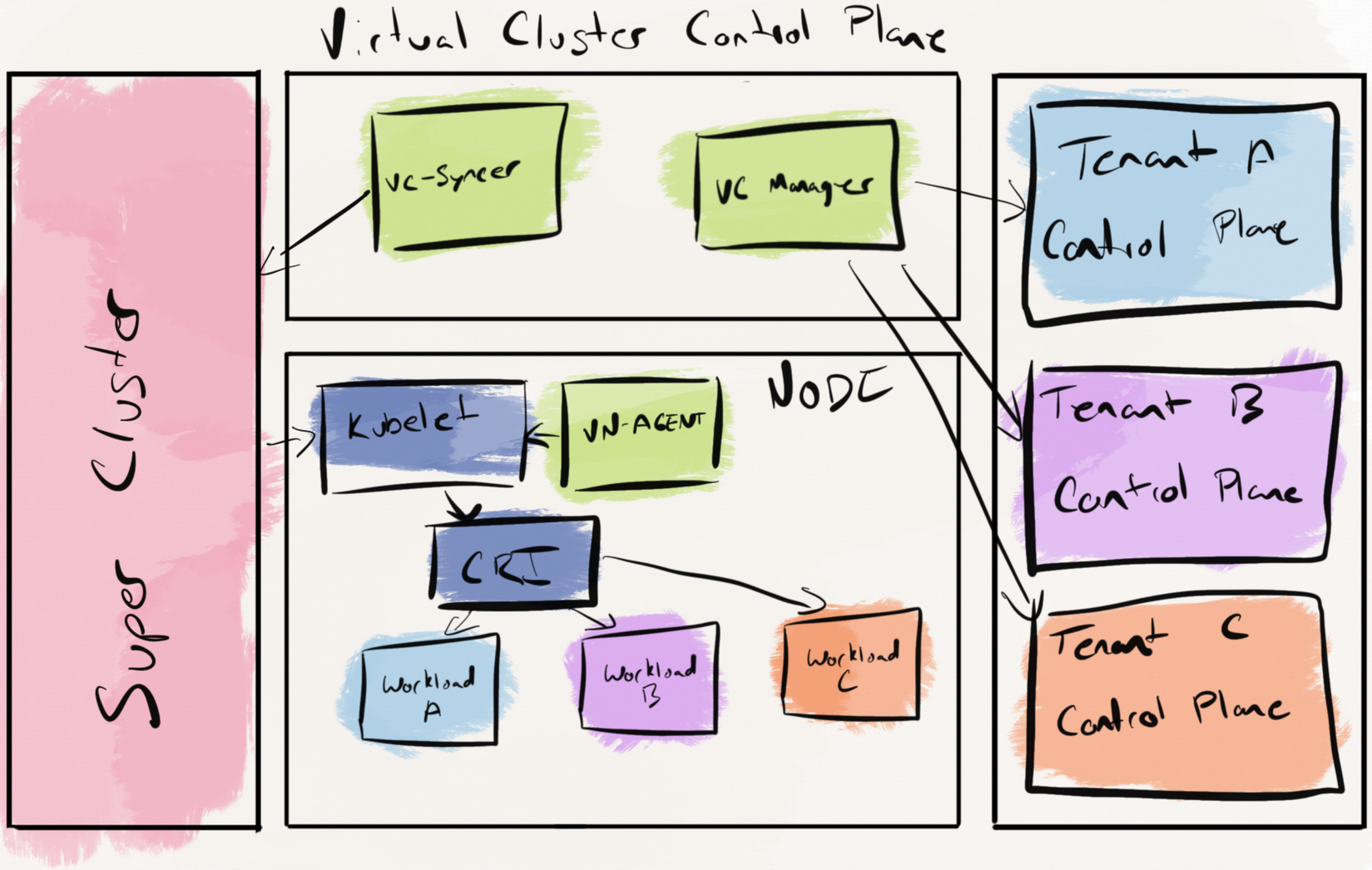


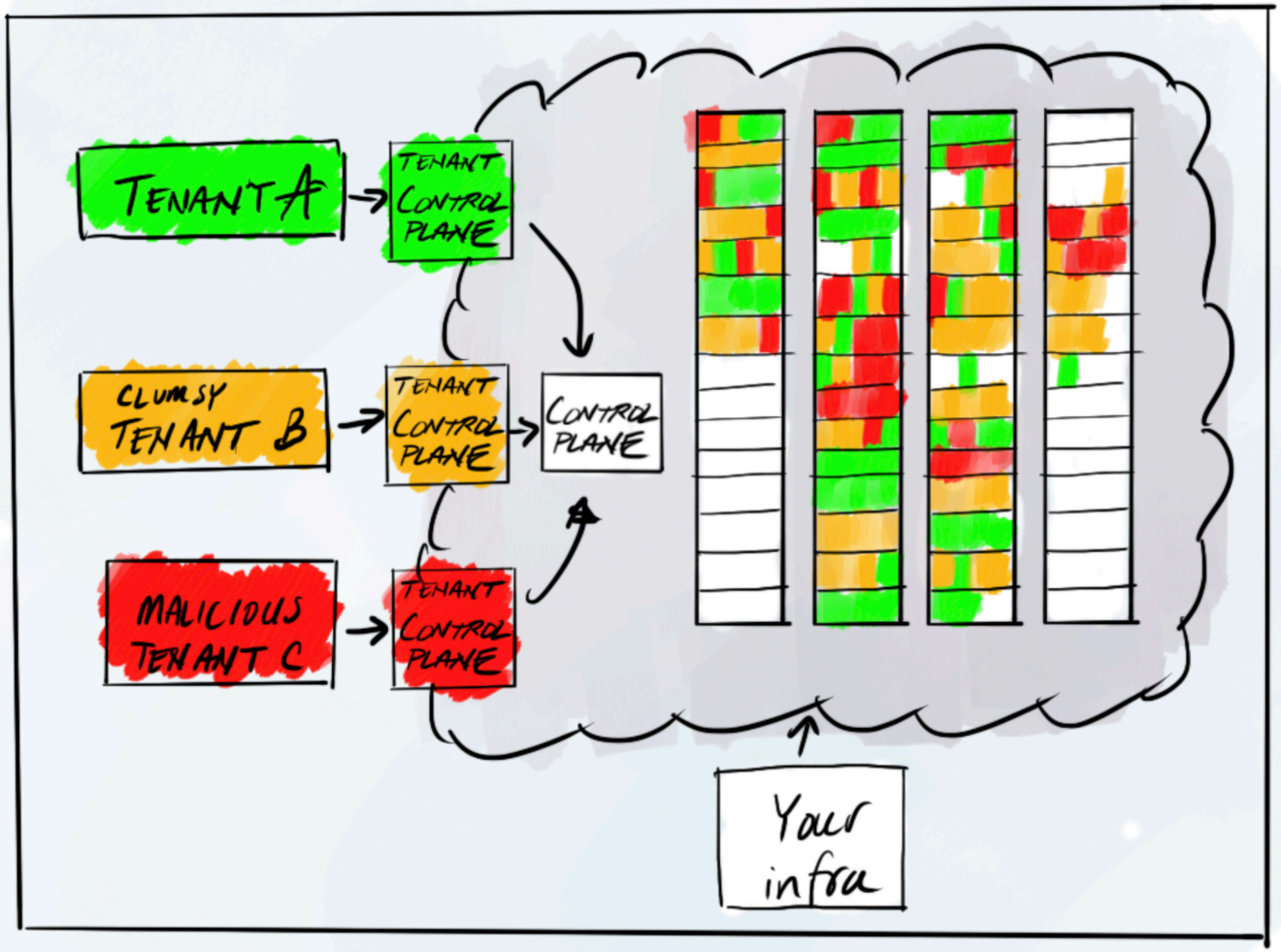
Issues with Multi-Tenant Control Planes

- ✗ Clumsy Clients
- ✗ Access to Cluster Scoped Resources
- ✗ Usage of Cloud Native Tools



Isolating Control Planes


Virtual Cluster Control Plane
Super Cluster
vc-syncer
VC Manager
Tenant A Control Plane
Tenant B Control Plane
Tenant C Control Plane
NODE
Kubelet
VN-AGENT
CRI
Workload A
Workload B
Workload C


TENANT A
TENANT CONTROL PLANE
CLUMSY TENANT B
TENANT CONTROL PLANE
CONTROL PLANE
MALICIOUS TENANT C
TENANT CONTROL PLANE
Your infra

Issues with Multi-Tenant Control Planes

- ✗ Clumsy Clients
- ✗ Access to Cluster Scoped Resources
- ✗ Ability to use off-the-shelf Cloud Native Tools

Issues with Multi-Tenant Control Planes

✔ Clumsy Clients

✗ Access to Cluster Scoped Resources

✗ Ability to use off-the-shelf Cloud Native Tools

Issues with Multi-Tenant Control Planes

✔ Clumsy Clients

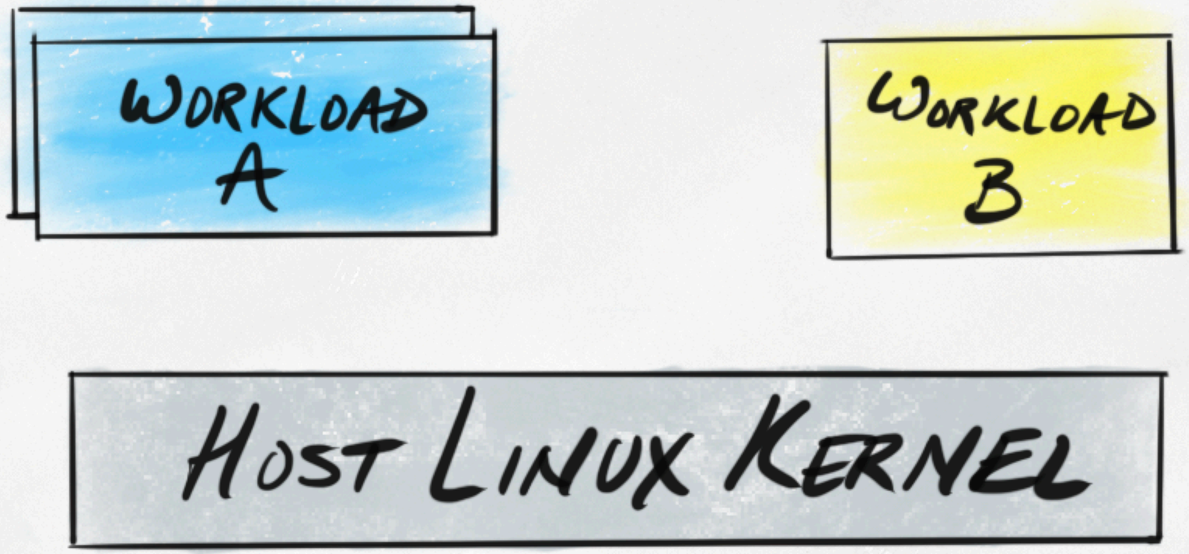
✔ Access to Cluster Scoped Resources

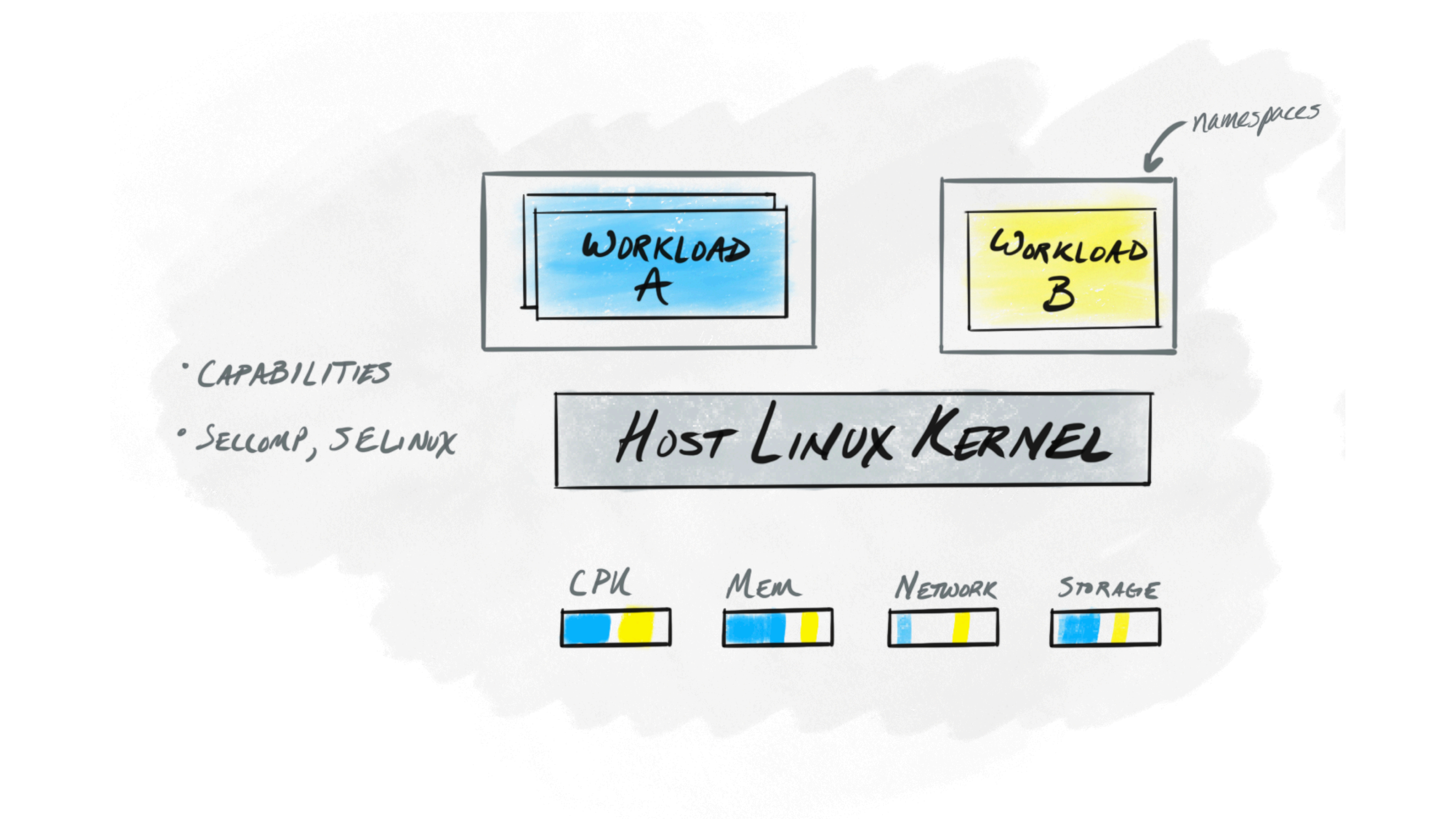
✗ Ability to use off-the-shelf Cloud Native Tools

Issues with Multi-Tenant Control Planes

- ✔ Clumsy Clients
- ✔ Access to Cluster Scoped Resources
- ✔ Ability to use off-the-shelf Cloud Native Tools

Workload isolation


WORKLOAD A
WORKLOAD B
HOST LINUX KERNEL


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

Mutually untrusting tenants

Securing the data plane: options

- Don’t provide extra isolation
- Create node pools for each tenant
- Provide stronger workload isolation

Securing the data plane: options

✗ Don’t provide extra isolation

- Create node pools for each tenant
- Provide stronger workload isolation

Securing the data plane: options

✗ Don’t provide extra isolation

✗ Create node pools for each tenant

- Provide stronger workload isolation

Securing the data plane: options

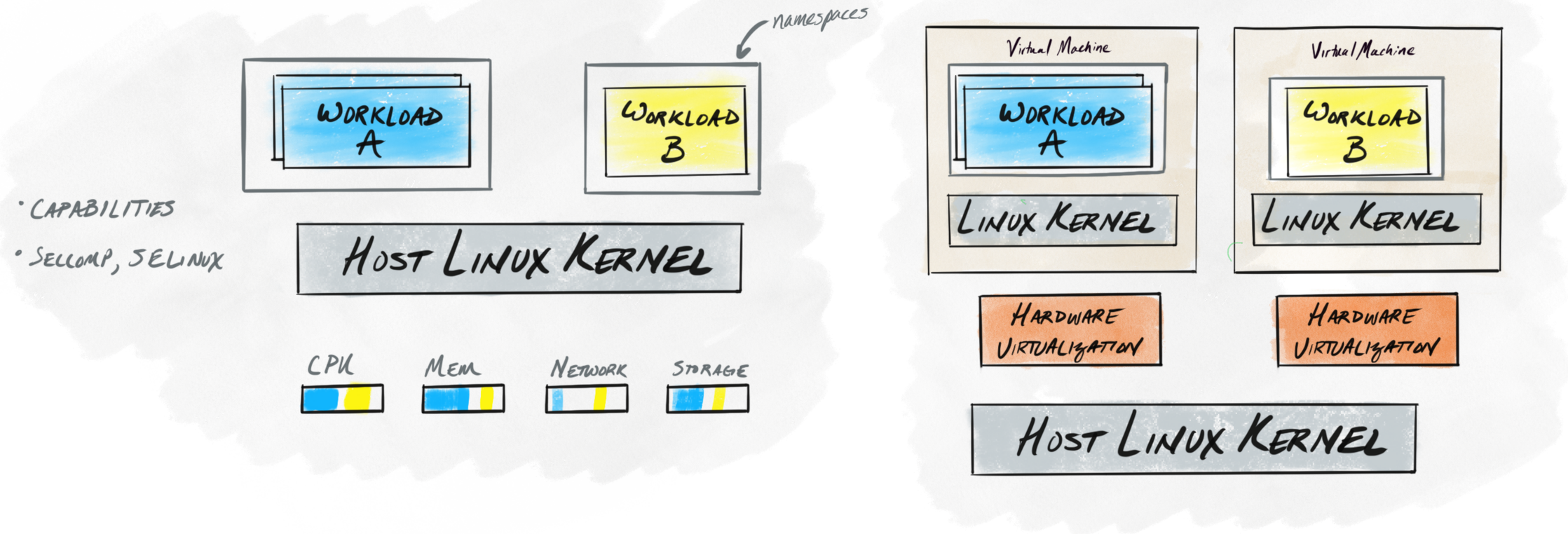
✗ Don’t provide extra isolation

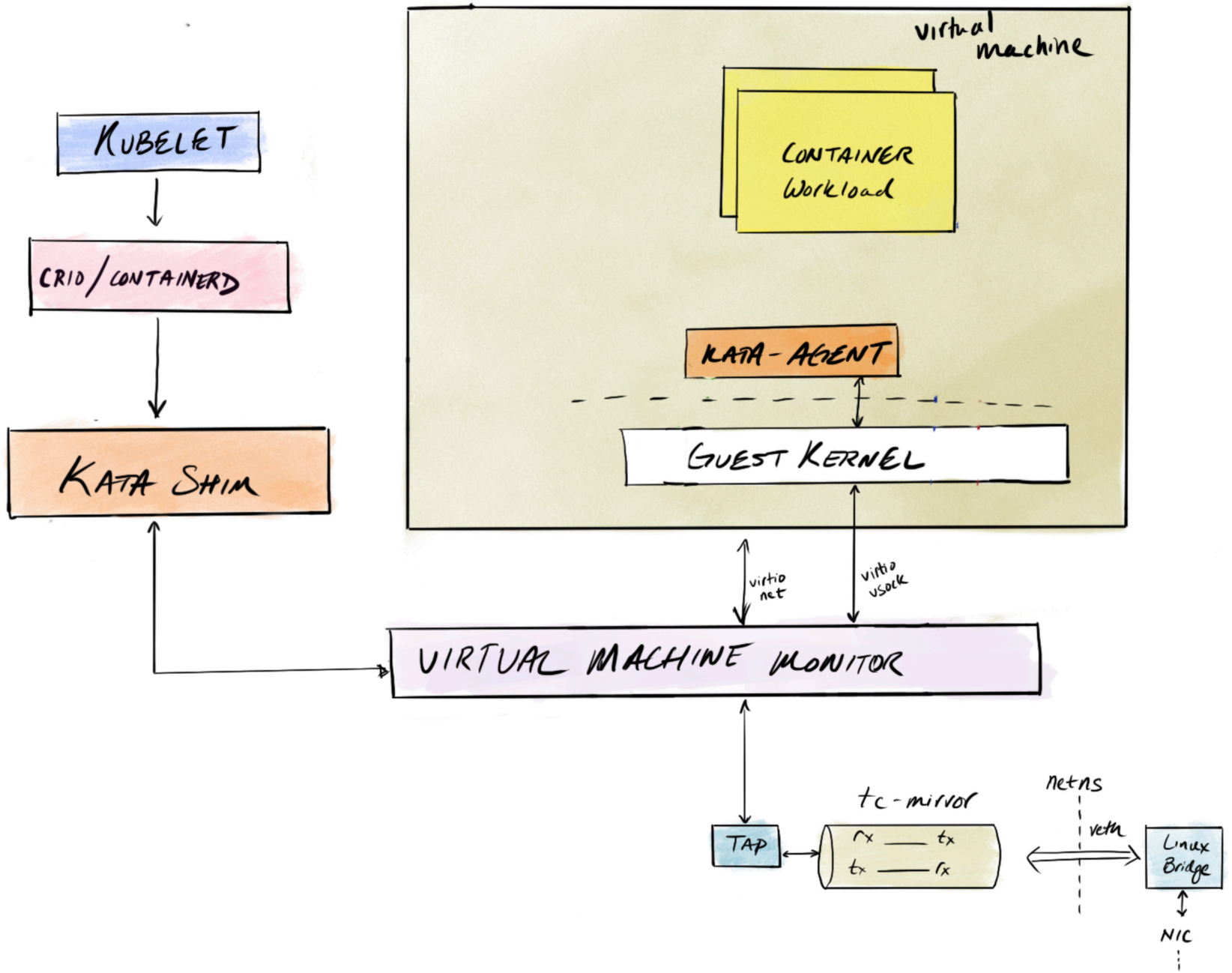
✗ Create node pools for each tenant

✔ Provide stronger workload isolation

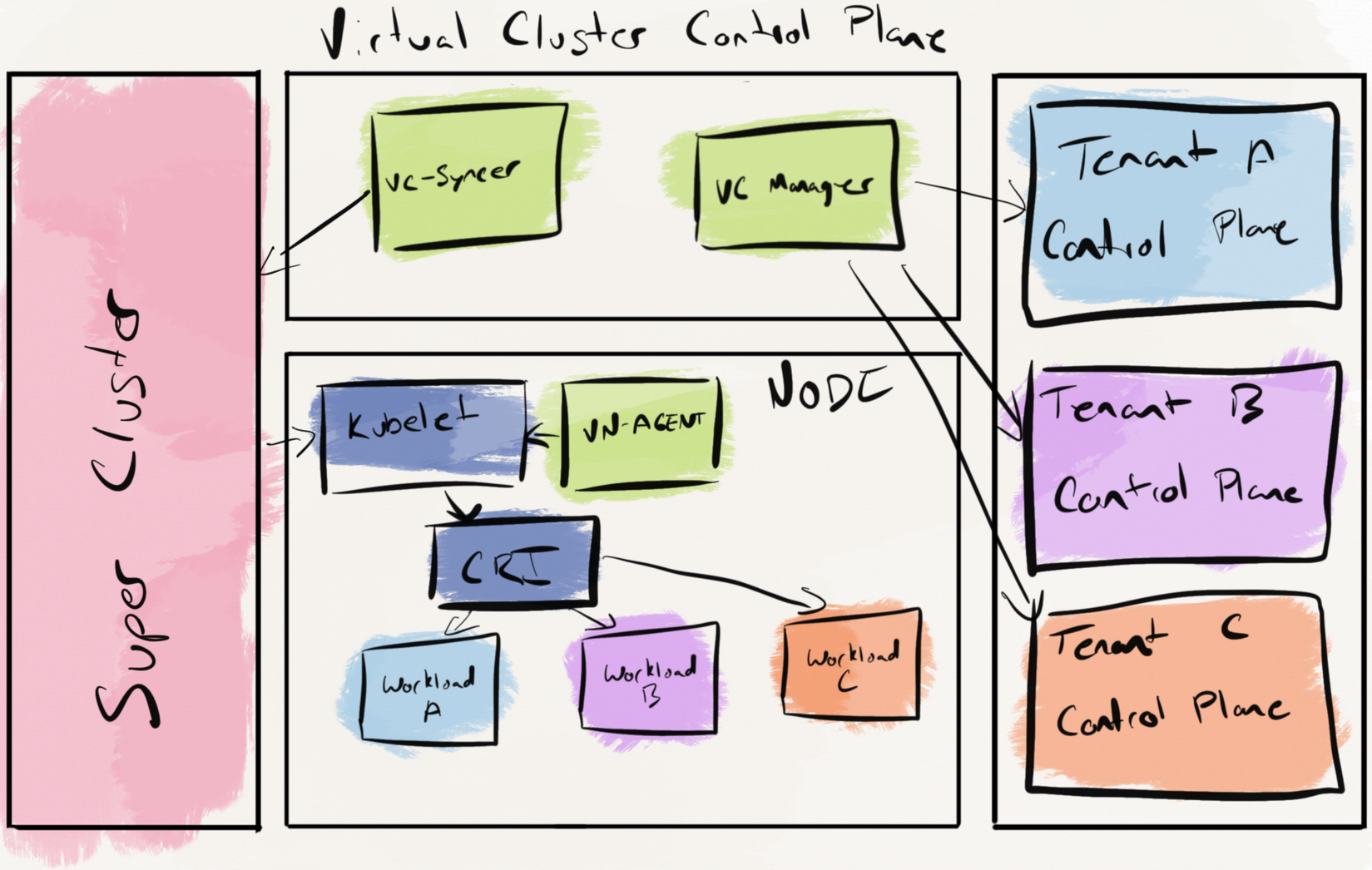
Sandboxed Runtimes

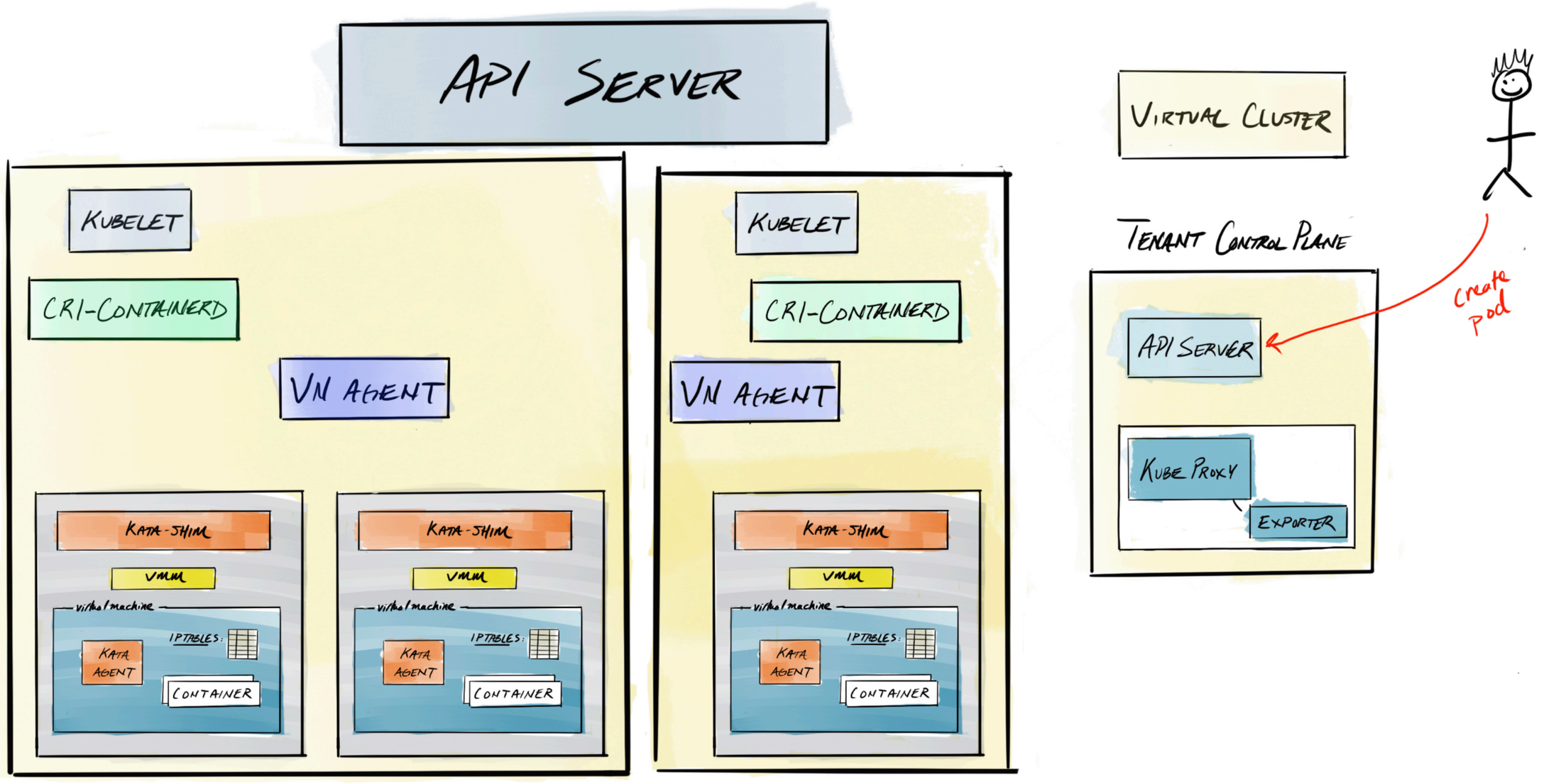
Sandboxed Runtimes

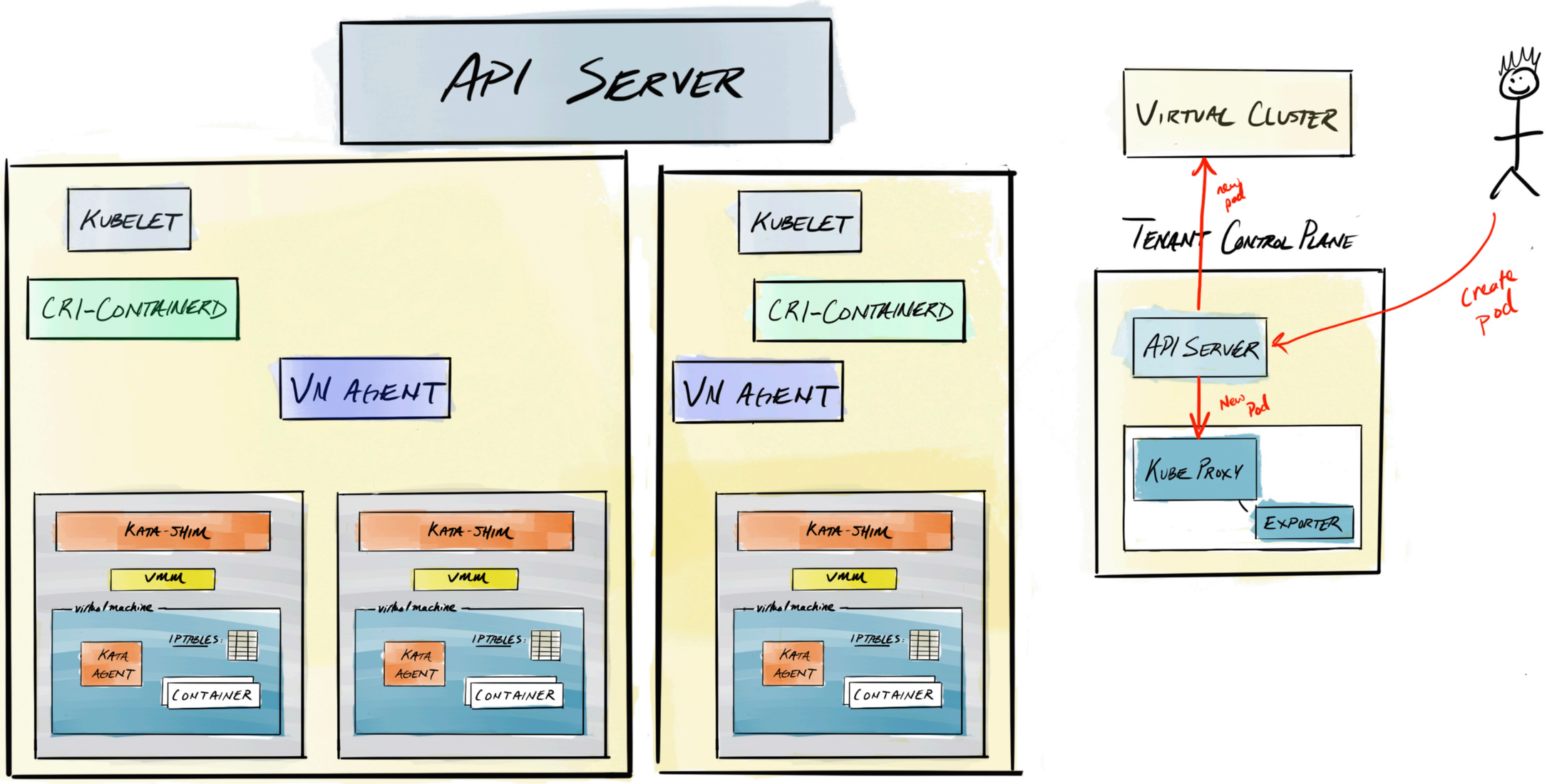

namespaces
WORKLOAD A
WORKLOAD B
• CAPABILITIES
• SECCOMP, SELINUX
HOST LINUX KERNEL
CPU
MEM
NETWORK
STORAGE
Virtual Machine
WORKLOAD A
LINUX KERNEL
Virtual Machine
WORKLOAD B
LINUX KERNEL
HARDWARE VIRTUALIZATION
HARDWARE VIRTUALIZATION
HOST LINUX KERNEL

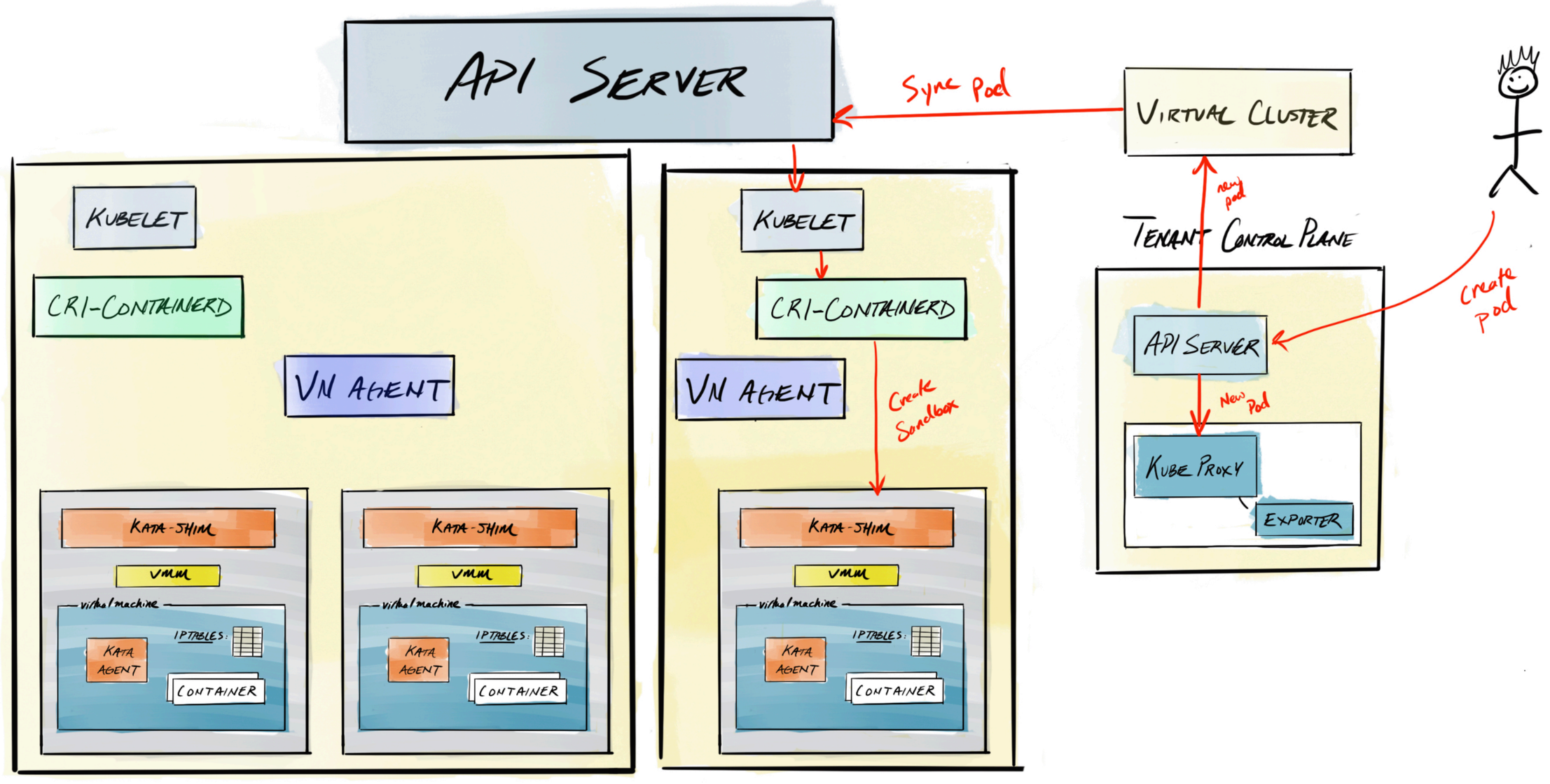

virtual machine
KUBELET
CONTAINER Workload
CRIO / CONTAINERD
KATA-AGENT
KATA SHIM
GUEST KERNEL
virtio net
virtio vsock
VIRTUAL MACHINE MONITOR
tc-mirror
netns
TAP
rx
tx
tx
rx
veth
Linux Bridge
NIC

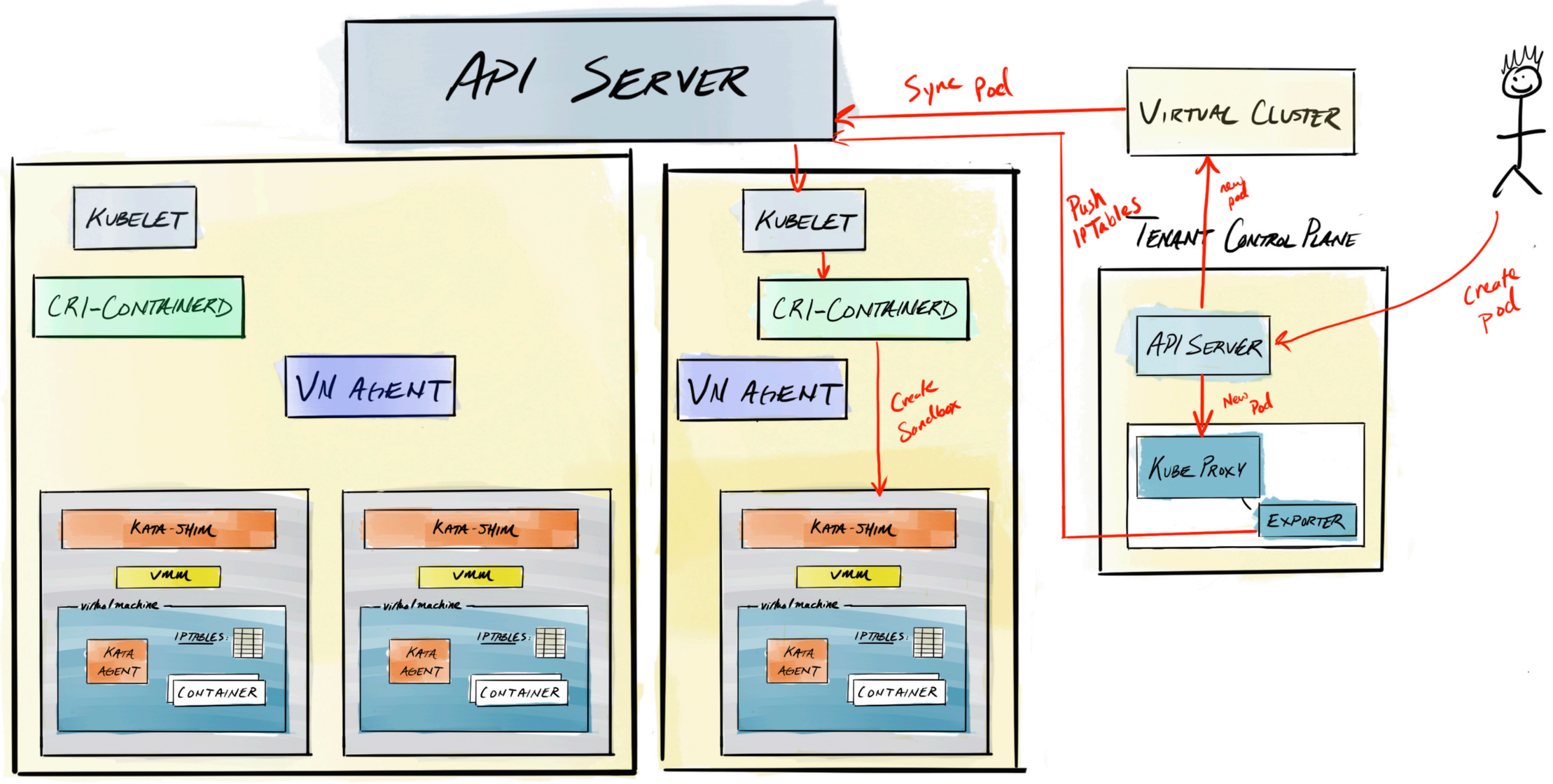
VirtualCluster Service Networking

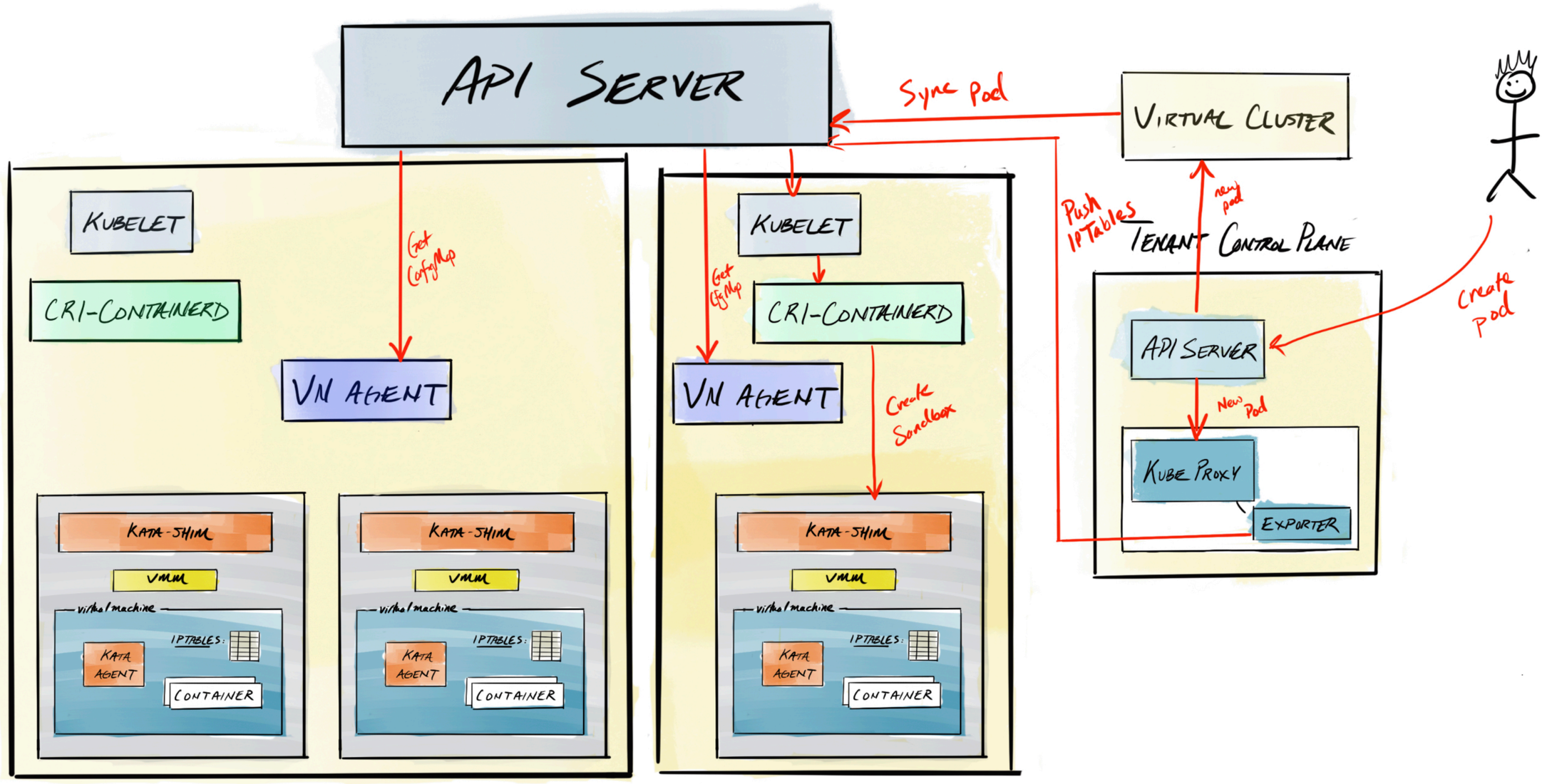

Virtual Cluster Control Plane
Super Cluster
vc-syncer
VC Manager
Tenant A Control Plane
Tenant B Control Plane
Tenant C Control Plane
NODE
Kubelet
VN-AGENT
CRI
Workload A
Workload B
Workload C

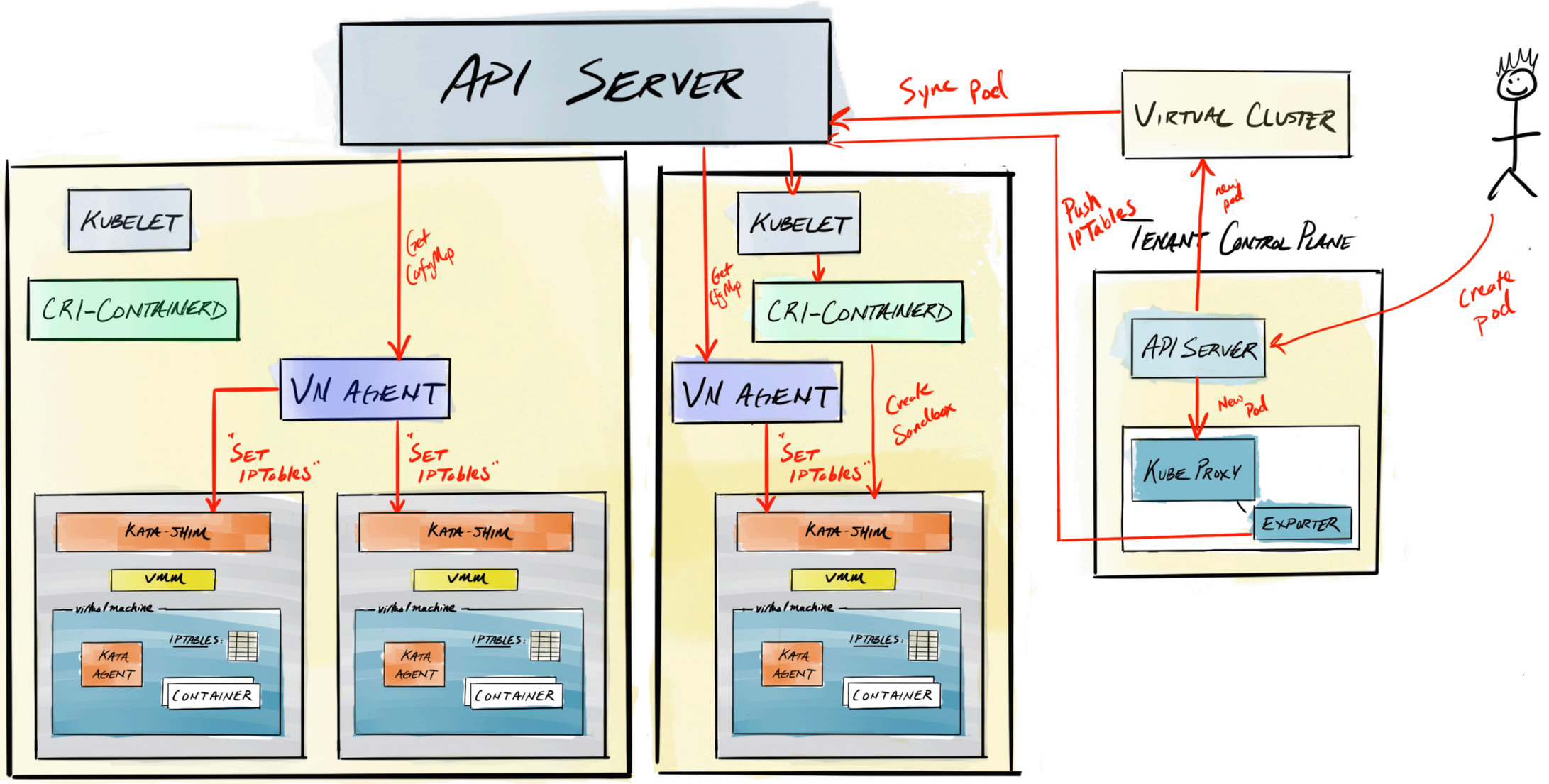

API SERVER
KUBELET
CRI-CONTAINERD
VN AGENT
KATA-SHIM
VMM
virtual machine
KATA AGENT
IPTABLES
CONTAINER
KUBELET
CRI-CONTAINERD
VN AGENT
KATA-SHIM
VMM
virtual machine
KATA AGENT
IPTABLES
CONTAINER
KATA-SHIM
VMM
virtual machine
KATA AGENT
IPTABLES
CONTAINER
VIRTUAL CLUSTER
TENANT CONTROL PLANE
API SERVER
KUBE PROXY
EXPORTER
create pod

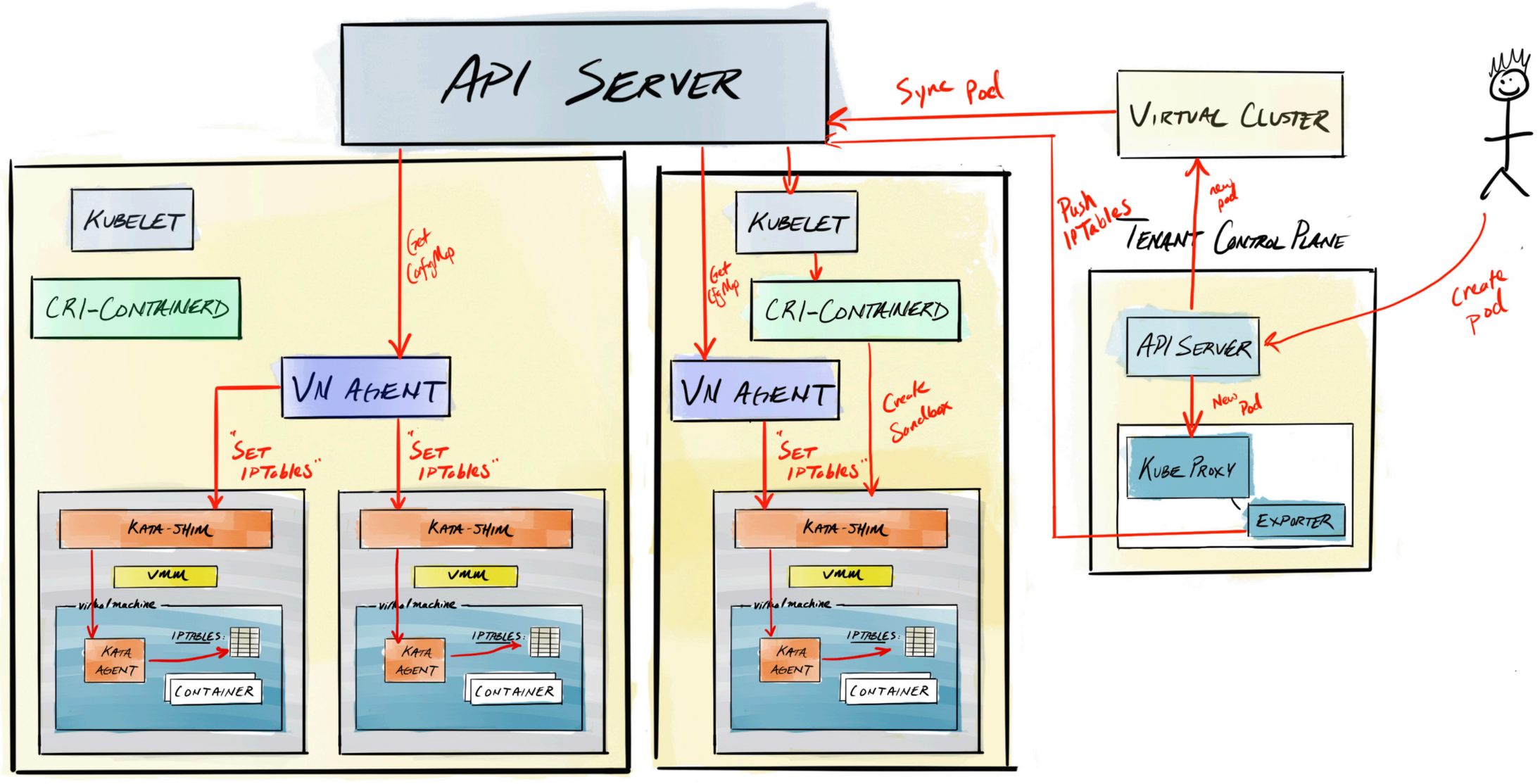

API SERVER
KUBELET
CRI-CONTAINERD
VN AGENT
KATA-SHIM
VMM
virtual machine
IPTABLES:
KATA AGENT
CONTAINER
KUBELET
CRI-CONTAINERD
VN AGENT
KATA-SHIM
VMM
virtual machine
IPTABLES:
KATA AGENT
CONTAINER
KATA-SHIM
VMM
virtual machine
IPTABLES:
KATA AGENT
CONTAINER
VIRTUAL CLUSTER
new pod
TENANT CONTROL PLANE
API SERVER
New Pod
KUBE PROXY
EXPORTER
create Pod

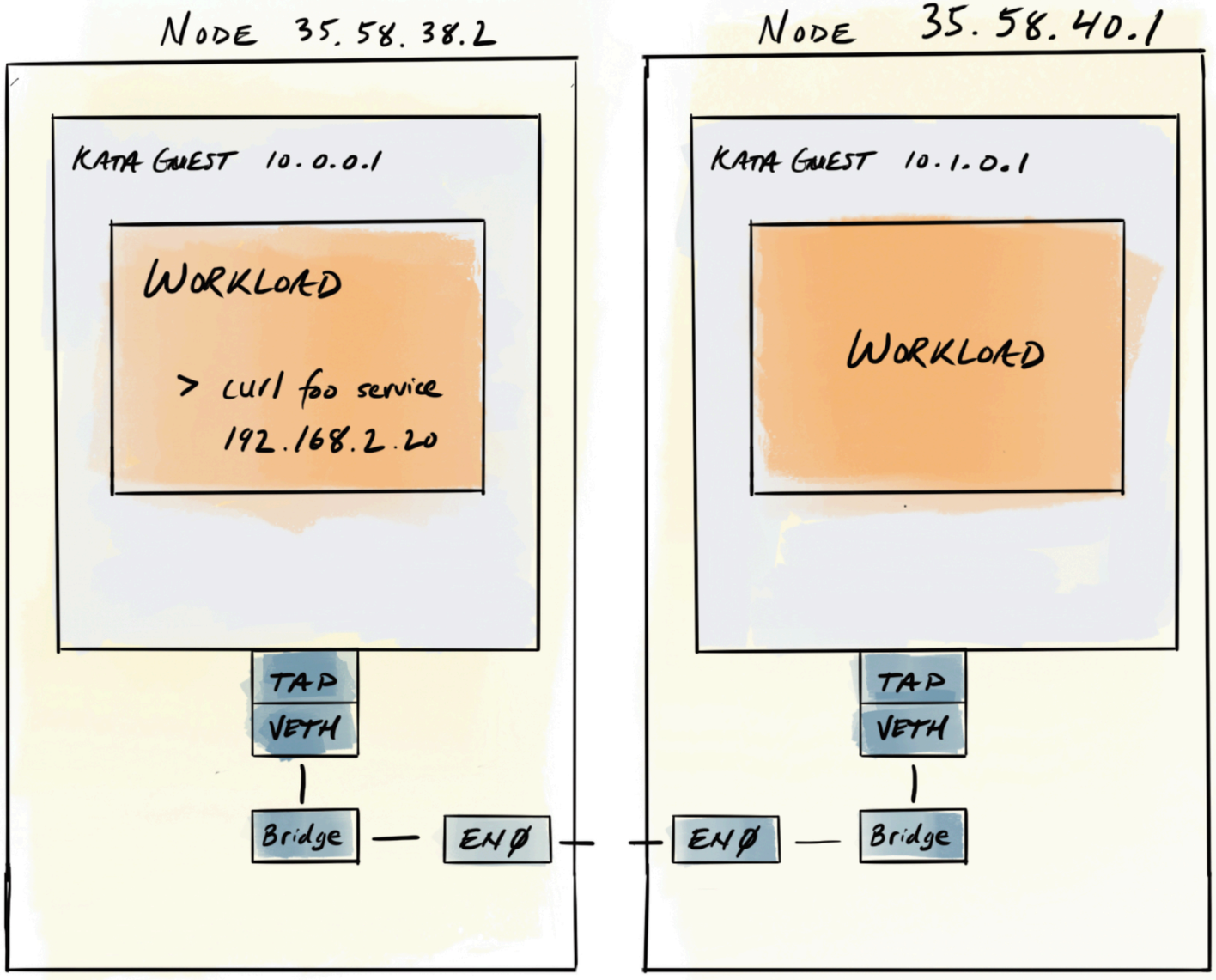

API SERVER
Sync Pod
VIRTUAL CLUSTER
KUBELET
CRI-CONTAINERD
VN AGENT
KATA-SHIM
VMM
virtual machine
IPTABLES
KATA AGENT
CONTAINER
KUBELET
CRI-CONTAINERD
VN AGENT
Create Sandbox
KATA-SHIM
VMM
virtual machine
IPTABLES
KATA AGENT
CONTAINER
new pod
TENANT CONTROL PLANE
API SERVER
New Pod
KUBE PROXY
EXPORTER
create pod

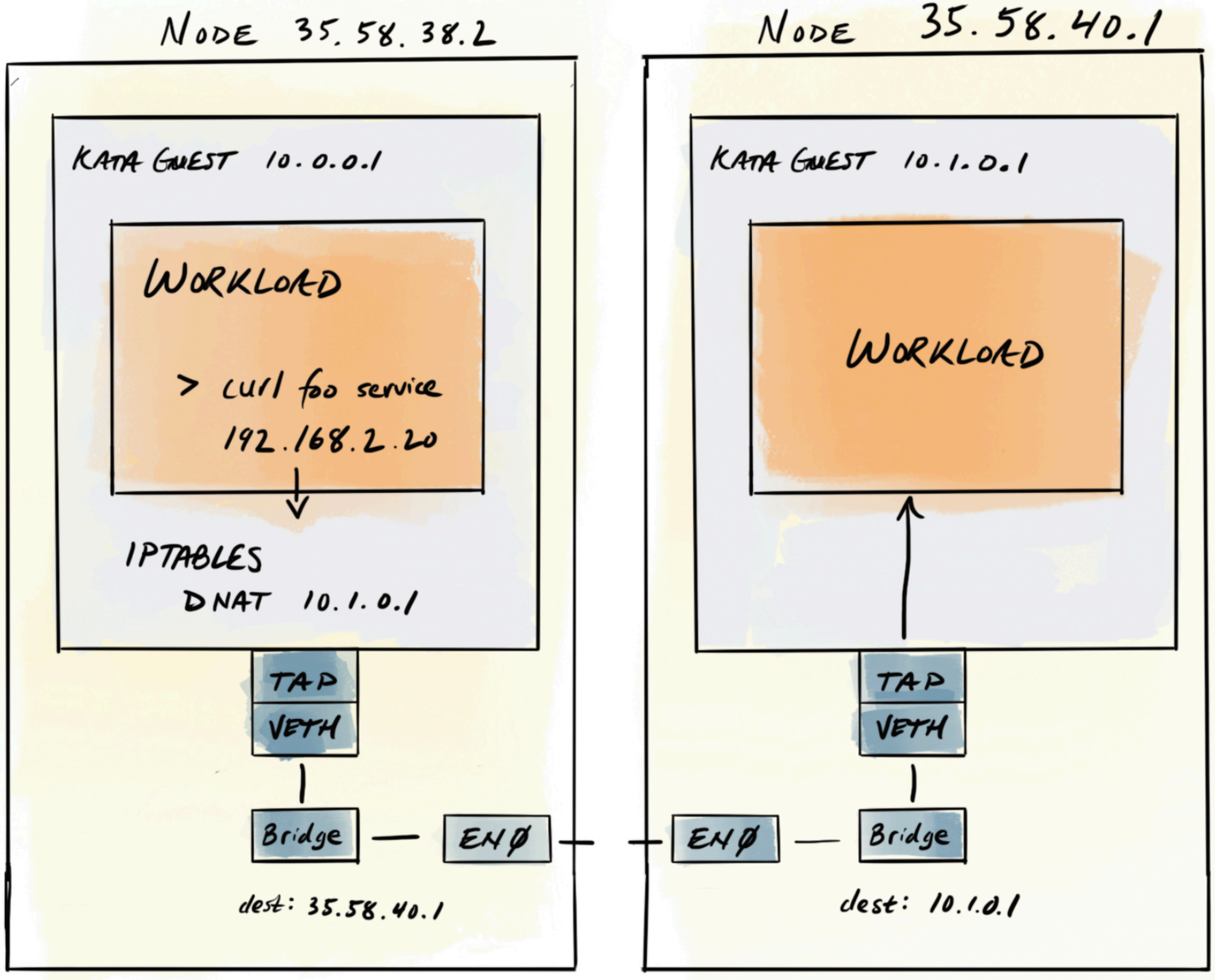

API SERVER
Sync Pod
VIRTUAL CLUSTER
KUBELET
CRI-CONTAINERD
VN AGENT
KATA-SHIM
VMM
virtual machine
KATA AGENT
IPTABLES
CONTAINER
KUBELET
CRI-CONTAINERD
VN AGENT
Create Sandbox
KATA-SHIM
VMM
virtual machine
KATA AGENT
IPTABLES
CONTAINER
Push IPTables
new pod
TENANT CONTROL PLANE
API SERVER
New Pod
KUBE PROXY
EXPORTER
create pod


API SERVER
Sync Pod
VIRTUAL CLUSTER
KUBELET
CRI-CONTAINERD
Get ConfigMap
VN AGENT
KUBELET
CRI-CONTAINERD
Get CfgMap
VN AGENT
Create Sandbox
Push IPTables
new pod
TENANT CONTROL PLANE
API SERVER
New Pod
KUBE PROXY
EXPORTER
Create Pod
KATA-SHIM
VMM
virtual machine
IPTABLES:
KATA AGENT
CONTAINER
KATA-SHIM
VMM
virtual machine
IPTABLES:
KATA AGENT
CONTAINER
KATA-SHIM
VMM
virtual machine
IPTABLES:
KATA AGENT
CONTAINER

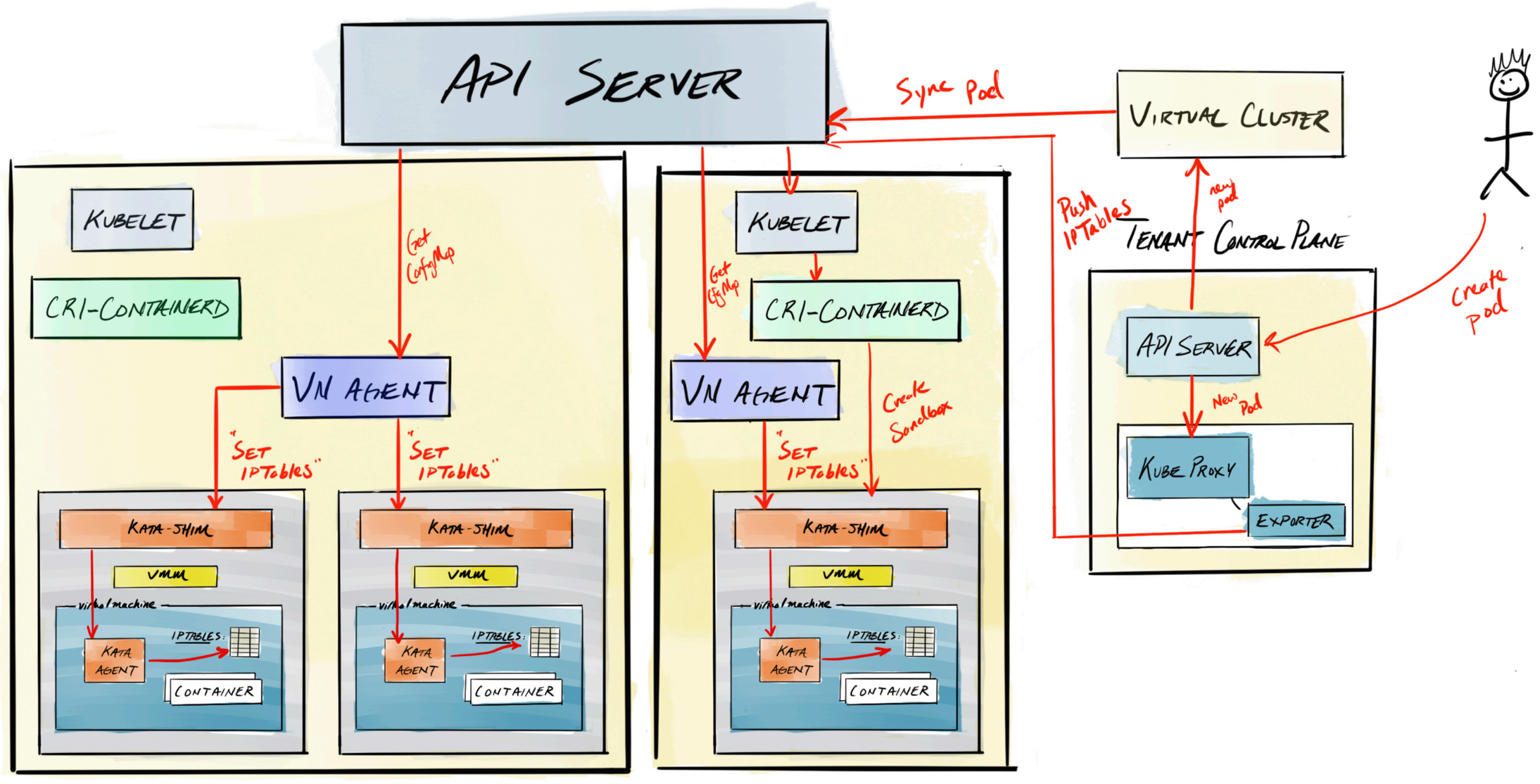

API SERVER
Sync Pod
VIRTUAL CLUSTER
KUBELET
CRI-CONTAINERD
Get CofigMap
VN AGENT
"SET IPTABLES"
"SET IPTABLES"
KATA-SHIM
VMM
virtual machine
IPTABLES:
KATA AGENT
CONTAINER
KUBELET
Get CfgMap
CRI-CONTAINERD
VN AGENT
Create Sandbox
"SET IPTABLES"
KATA-SHIM
VMM
virtual machine
IPTABLES:
KATA AGENT
CONTAINER
Push IPTables
new pod
TENANT CONTROL PLANE
API SERVER
New Pod
KUBE PROXY
EXPORTER
create Pod


API SERVER
Sync Pod
VIRTUAL CLUSTER
KUBELET
CRI-CONTAINERD
Get CofigMap
VN AGENT
"SET IPTABLES"
"SET IPTABLES"
KATA-SHIM
VMM
virtual machine
IPTABLES:
KATA AGENT
CONTAINER
KUBELET
Get CfgMap
CRI-CONTAINERD
VN AGENT
Create Sandbox
"SET IPTABLES"
KATA-SHIM
VMM
virtual machine
IPTABLES:
KATA AGENT
CONTAINER
Push IPTables
new pod
TENANT CONTROL PLANE
API SERVER
New Pod
KUBE PROXY
EXPORTER
create Pod


NODE 35.58.38.2
KATA GUEST 10.0.0.1
WORKLOAD
> curl foo service
192.168.2.20
TAP
VETH
Bridge
ENØ
NODE 35.58.40.1
KATA GUEST 10.1.0.1
WORKLOAD
ENØ
TAP
VETH
Bridge


NODE 35.58.38.2
KATA GUEST 10.0.0.1
WORKLOAD
> curl foo service
192.168.2.20
IPTABLES
DNAT 10.1.0.1
TAP
VETH
Bridge
ENØ
dest: 35.58.40.1
NODE 35.58.40.1
KATA GUEST 10.1.0.1
WORKLOAD
TAP
VETH
ENØ
Bridge
dest: 10.1.0.1

Demo


API SERVER
Sync Pod
VIRTUAL CLUSTER
KUBELET
CRI-CONTAINERD
Get CofigMap
VN AGENT
"SET IPTABLES"
"SET IPTABLES"
KATA-SHIM
VMM
virtual machine
IPTABLES:
KATA AGENT
CONTAINER
KUBELET
Get CfgMap
CRI-CONTAINERD
VN AGENT
Create Sandbox
"SET IPTABLES"
KATA-SHIM
VMM
virtual machine
IPTABLES:
KATA AGENT
CONTAINER
Push IPTables
new pod
TENANT CONTROL PLANE
API SERVER
New Pod
KUBE PROXY
EXPORTER
create Pod

Future

Integration with KPNG

“New Kubernetes NG Proxy” - https://sigs.k8s.io/kpng

Separate VPCs/Overlays Per VirtualCluster.

Feedback

Thanks!