Running a High-Performance DPDK-Based Router on Kubernetes

DPDK Summit 2026, Stockholm

Andrea Panattoni
Red Hat

DPDK Red Hat

Telco Network Team @ Red Hat

Contributed to

- OVN-Kubernetes
- SR-IOV Network Operator
- SR-IOV Network Device Plugin
- Multus CNI Plugins

apanatto@redhat.com
github.com/zeeke
linkedin.com/in/andreapanattoni

Agenda

- DPDK on Containers
- Kubernetes resource allocation
- Operators
- OpenPERouter

The Problem

Core tension: Kubernetes treats devices as fungible. DPDK needs pinned, hardware-aware allocation.

Hugepages

EAL needs continuous memory via mmap

Isolated CPUs

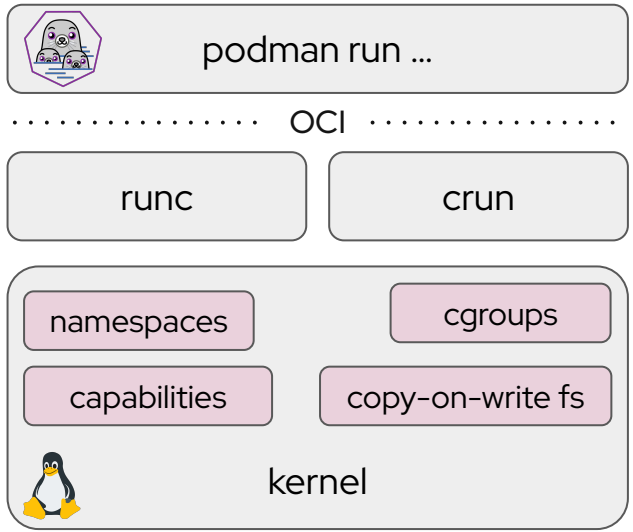
Poll-mode drivers can't share cores

Direct NIC access

Bypass kernel network stack

Containers

- Linux isolated process
  - namespaces (network, PID, user)
  - copy-on-write filesystems (e.g. OverlayFS)
    - pivot_root
  - cgroups
  - capabilities (CAP_NET_ADMIN, CAP_NET_RAW, ...)
- Open Container Initiative spec
  - github.com/opencontainers/runtime-spec
  - Implementations (runc, crun)



Grout

DPDK based software router using rte_graph.

- IPv4 forwarding
- IPv6 forwarding
- IPv6 router advertisements
- Multiple VRF domains
- VLAN sub interfaces
- L2 bridging
- VXLAN tunnels
- Bond/LACP interfaces

github.com/DPDK/grout

```
podman run --rm --name grout
    --device /dev/net/tun
    --device /dev/vfio
    -v /dev/hugepages:/dev/hugepages
    --cgroup-conf hugetlb.1GB.max=4G
    --cpuset-cpus 10-14
    --cap-add NET_ADMIN,IPC_LOCK
    quay.io/grout/grout:0.15
```

```
podman exec -it grout \
     grcli interface add port p0 devargs 0000:48:11.1 rxqs 1 qsize 2048
```

```
podman run --rm --name grout
    --device /dev/net/tun
    --device /dev/vfio
    -v /dev/hugepages:/dev/hugepages
    --cgroup-conf hugetlb.1GB.max=4G
    --cpuset-cpus 10-14
    --cap-add NET_ADMIN,IPC_LOCK
    quay.io/grout/grout:0.15
```

```
podman exec -it grout \
     grcli interface add port p0 devargs 0000:48:11.1 rxqs 1 qsize 2048
```

```
podman run --rm --name grout
    --device /dev/net/tun
    --device /dev/vfio
    -v /dev/hugepages:/dev/hugepages
    --cgroup-conf hugetlb.1GB.max=4G
    --cpuset-cpus 10-14
    --cap-add NET_ADMIN,IPC_LOCK
    quay.io/grout/grout:0.15
```

```
podman exec -it grout \
     grcli interface add port p0 devargs 0000:48:11.1 rxqs 1 qsize 2048
```

```
podman run --rm --name grout
    --device /dev/net/tun
    --device /dev/vfio
    -v /dev/hugepages:/dev/hugepages
    --cgroup-conf hugetlb.1GB.max=4G
    --cpuset-cpus 10-14
    --cap-add NET_ADMIN,IPC_LOCK
    quay.io/grout/grout:0.15
```

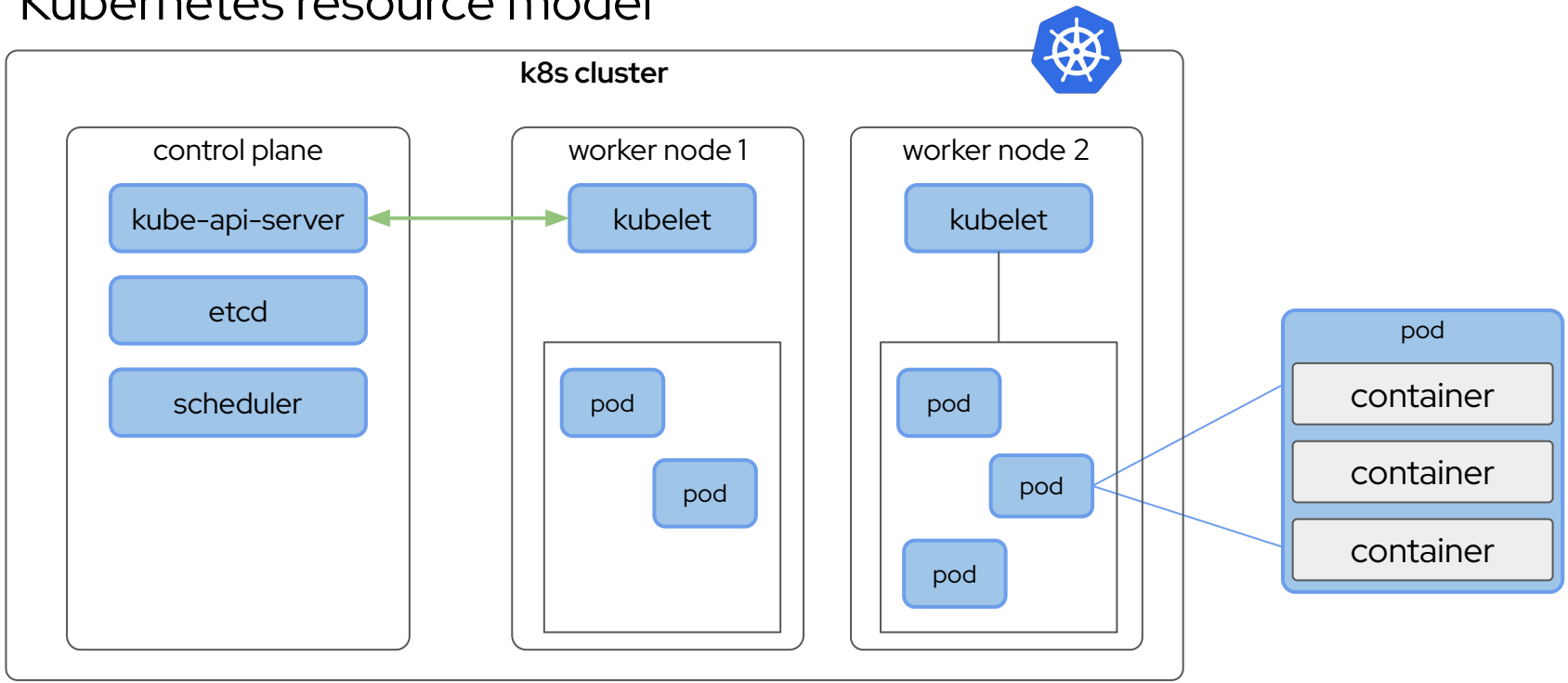
```
podman exec -it grout \
     grcli interface add port p0 devargs 0000:48:11.1 rxqs 1 qsize 2048
```

Bifurcated Driver

```
podman run --rm --name grout
    -v /dev/hugepages:/dev/hugepages
    --device /dev/net/tun
    --device /dev/infiniband
    --cap-add NET_ADMIN,IPC_LOCK,SYS_RESOURCE,NET_RAW
    --cpuset-cpus 10-14
    --cgroup-conf hugetlb.1GB.max=4G
    quay.io/grout/grout:0.15
```

```
ip link set ens7f0v12 netns `podman inspect --format '{{.State.Pid}}' grout`
podman exec -it grout \
     grcli interface add port p3 devargs 0000:c0:01.6 rxqs 1 qsize 2048
```

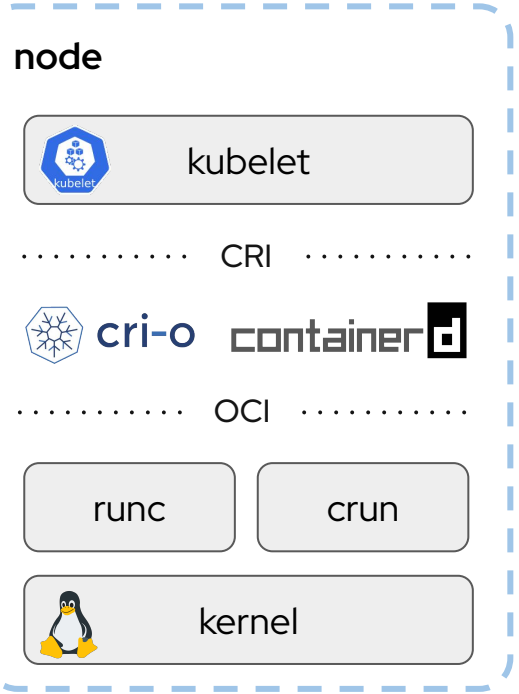
Kubernetes resource model



```yaml
apiVersion: v1
kind: Pod
metadata:
  name: hello-world
spec:
  containers:
  - name: c1
    image: busybox
    resources:
      requests:
        memory: "64Mi"
        cpu: "250m"
      limits:
        memory: "128Mi"
        cpu: "500m"
```

pod

yaml



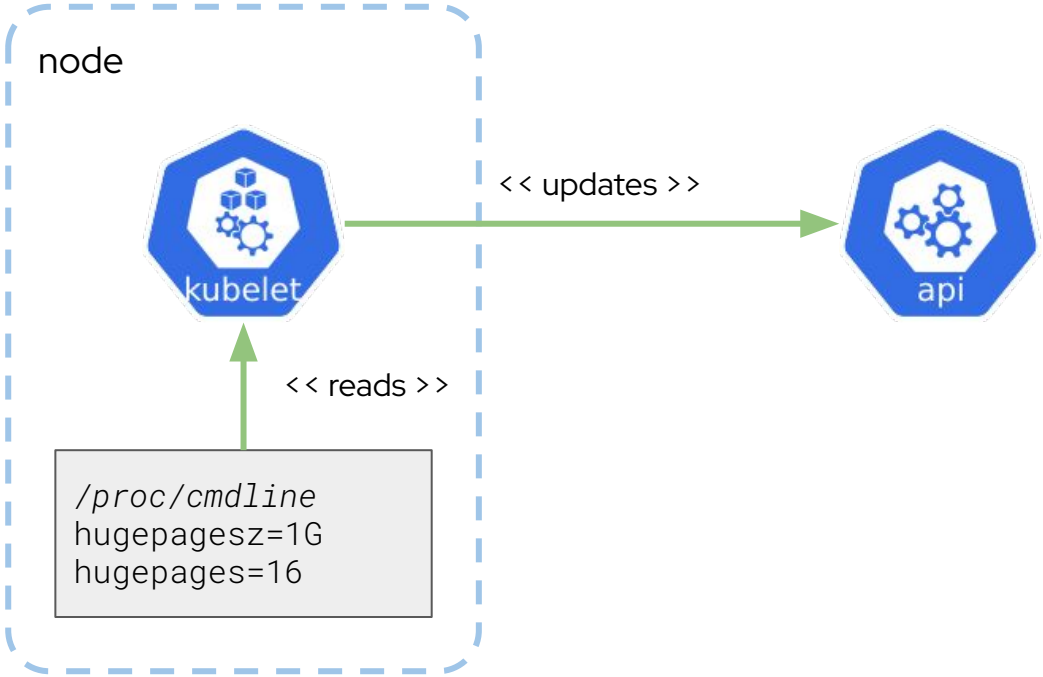
pull/unpack images
pod sandbox lifecycle
networking setup
container lifecycle mgmt

Container Runtime Interface (CRI) - (api.proto)

```yaml
apiVersion: v1
kind: Pod
metadata:
  annotations:
    k8s.v1.cni.cncf.io/networks: [{"name": "net-a"}]
  name: grout
spec:
  containers:
  - name: grout
    image: quay.io/grout/grout:v0.15
    resources:
      requests:
        cpu: "4"
        memory: 2Gi
        hugepages-1Gi: 2Gi
        sriov/nic1: “1”
      limits:
        cpu: "4"
        memory: 2Gi
        hugepages-1Gi: 2Gi
        sriov/nic1: “1”
    volumeMounts:
      - name: hugepages
        mountPath: /dev/hugepages
  volumes:
    - name: hugepages
      emptyDir:
        medium: HugePages
```

yaml

Hugepages

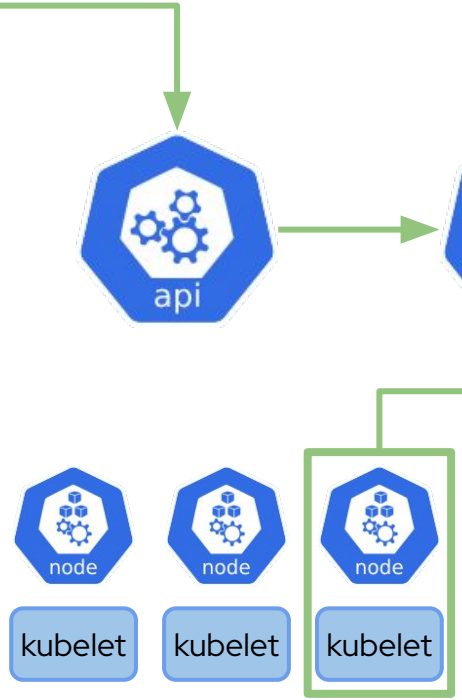



```
apiVersion: v1
kind: Node
status:
  capacity:
    hugepages-1Gi: 16Gi
  allocatable:
    hugepages-1Gi: 16Gi
```

K8s docs - Manage HugePages

```yaml
kind: Pod
...
    resources:
      requests:
        cpu: "4"
        memory: 2Gi
        hugepages-1Gi: 2Gi
        sriov/nic1: “1”
      limits:
        cpu: "4"
        memory: 2Gi
        hugepages-1Gi: 2Gi
        sriov/nic1: “1”
    volumeMounts:
      - name: hugepages
        mountPath: /dev/hugepages
  volumes:
    - name: hugepages
      emptyDir:
        medium: HugePages
```

yaml




```yaml
apiVersion: v1
kind: Node
status:
  capacity:
    hugepages-1Gi: 16Gi
  allocatable:
    hugepages-1Gi: 16Gi
```

```
kind: Pod
...
    resources:
      requests:
        cpu: "4"
        memory: 2Gi
        hugepages-1Gi: 2Gi
        sriov/nic1: “1”
      limits:
        cpu: "4"
        memory: 2Gi
        hugepages-1Gi: 2Gi
        sriov/nic1: “1”
    volumeMounts:
      - name: hugepages
        mountPath: /dev/hugepages
  volumes:
    - name: hugepages
      emptyDir:
        medium: HugePages
```

CRI

OCI

```
“linux”: {
  “resources”: {
    "hugepageLimits": [
      {
        "limit": 2147483648,
        "pageSize": "1GB"
      }
    ],
“mounts”: [
  {
    "destination": "/dev/hugepages",
```

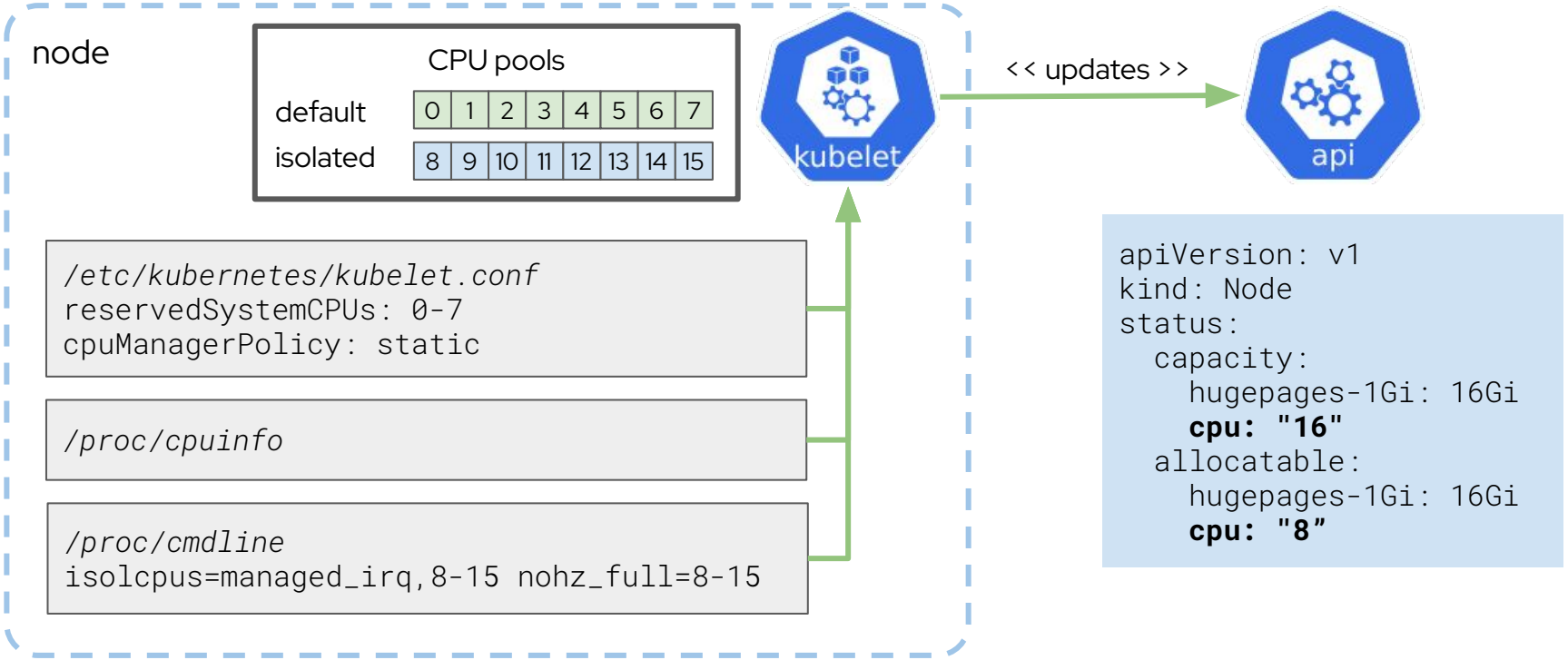
```
# grep -i huge /proc/mounts
nodev /dev/hugepages hugetlbfs
rw,seclabel,relatime,pagesize=1024M 0 0

# cat
/sys/fs/cgroup/kubepods.slice/<pod_id>/hugetlb.1GB.max
2147483648
```

Open Container Initiative / Config Linux # Hugepages

DPDK

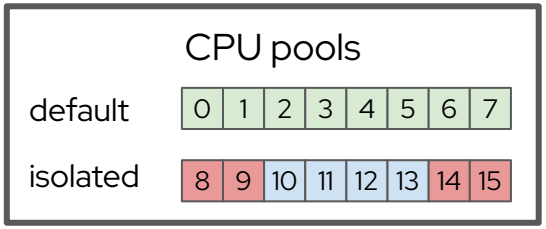
Red Hat

CPU Isolation



https://kubernetes.io/docs/tasks/administer-cluster/cpu-management-policies/

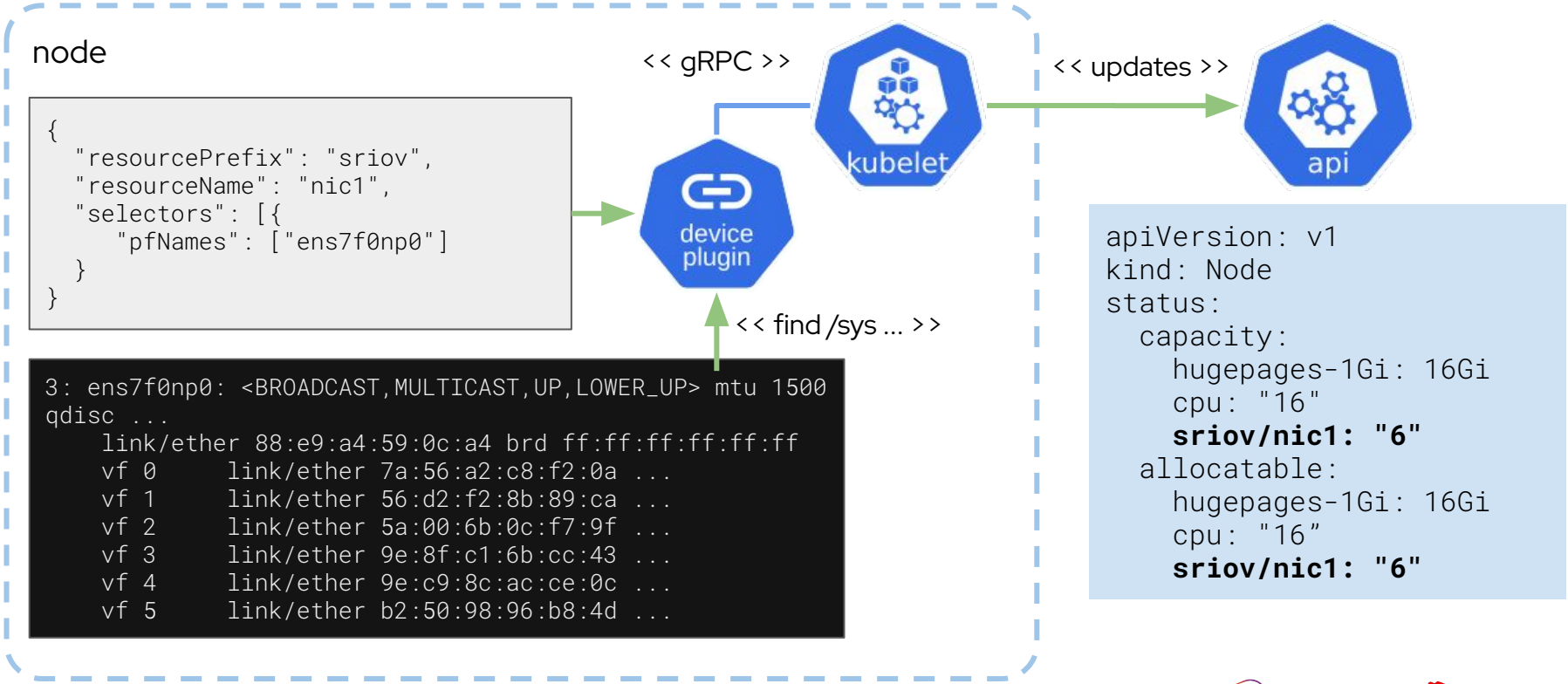
```
apiVersion: v1
kind: Pod
spec:
  containers:
    ...
    resources:
        requests:
          cpu: "4"
          memory: 2Gi
          hugepages-1Gi: 2Gi
        limits:
          cpu: "4"
          memory: 2Gi
          hugepages-1Gi: 2Gi

status:
  qosClass: Guaranteed
```




```
"linux": {
  "resources": {
    "cpusetCpus": "8-9,14-15",
```

```
# cat /sys/fs/cgroup/kubepods.slice/<pod_id>/cpuset.cpus.effective
8-9,14-15

# taskset -c -p 1
pid 1's current affinity list: 8,9,14,15
```

Direct NIC Access



github.com/k8snetworkplumbingwg/**sriov-network-device-plugin**

```
apiVersion: v1
kind: Pod
metadata:
  annotations:
    k8s.v1.cni.cncf.io/networks:
      [{"name": "net-a"}]
spec:
  containers:
    ...
    resources:
        requests:
          cpu: "4"
          memory: 2Gi
          hugepages-1Gi: 2Gi
          sriov/nic1: “1”
        limits:
          cpu: "4"
          memory: 2Gi
          hugepages-1Gi: 2Gi
          sriov/nic1: “1”

status:
  qosClass: Guaranteed
```

```
apiVersion: k8s.cni.cncf.io/v1
kind: NetworkAttachmentDefinition
metadata:
  annotations:
    k8s.v1.cni.cncf.io/resourceName: sriov/nic1
  name: net-a
spec:
  config: |-
    {
        "cniVersion": "1.0.0",
        "name": "net-a",
        "type": "sriov",
    }
```

Container Network Interface
github.com/k8snetworkplumbingwg/**multus-cni**
github.com/k8snetworkplumbingwg/**sriov-cni**

```
apiVersion: v1
kind: Pod
metadata:
  annotation
    k8s.v1.cni.cncf.io/networks:
      [{"name": "net-a"}]
spec:
  containers:
    ...
    resources:
        requests:
          cpu: "4"
          memory: 2Gi
          hugepages-1Gi: 2Gi
          sriov/nic1: “1”
        limits:
          cpu: "4"
          memory: 2Gi
          hugepages-1Gi: 2Gi
          sriov/nic1: “1”

status:
  qosClass: Guaranteed
```

```
apiVersion: k8s.cni.cncf.io/v1
kind: NetworkAttachmentDefinition
metadata:
  annotations:
    k8s.v1.cni.cncf.io/resourceName: sriov/nic1
  name: net-a
spec:
  config: |-
    {
        "cniVersion": "1.0.0",
        "name": "net-a",
        "type": "sriov",
    }
```

```
# env | grep PCI
PCIDEVICE_SRIOV_NIC1=0000:c0:01.1

# ip link show dev net1
184: net1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500
qdisc mq state UP mode DEFAULT group default qlen 1000
    link/ether 32:b6:3f:b3:36:ce brd ff:ff:ff:ff:ff:ff

# ethtool -i net1 | grep bus
bus-info: 0000:c0:01.1
```

Operators

- Extends K8s cluster functionalities
- Defines custom object types

k8snetworkplumbingwg/**sriov-network-operator**

openshift/**cluster-node-tuning-operator**

openshift/**machine-config-operator**

SR-IOV Network Operator

- NIC Configuration (VFs)
- Device Plugin Deployment
- Device Plugin Configuration
- NetworkAttachmentDefinition
- Kernel Hugepages
- Kernel CPU isolation
- Kubelet Configuration
- Pod Deployment

SR-IOV Network Operator

- NIC Configuration (VFs)
- Device Plugin Deployment
- Device Plugin Configuration
- NetworkAttachmentDefinition
- Kernel Hugepages
- Kernel CPU isolation
- Kubelet Configuration
- Pod Deployment

```
apiVersion: sriovnetwork.openshift.io/v1
kind: SriovNetworkNodePolicy
metadata:
  name: policy-nic1
  namespace: openshift-sriov-network-operator
spec:
  deviceType: netdevice
  nicSelector:
    pfNames:
      - "ens7f0np0"
  numVfs: 16
  resourceName: nic1
  nodeSelector:
    node-role.kubernetes.io/worker: ""
```

openshift_container_platform/4.20/html/hardware_networks/configuring-sriov-device

- NIC Configuration (VFs)
- Device Plugin Deployment
- Device Plugin Configuration
- NetworkAttachmentDefinition
- Kernel Hugepages
- Kernel CPU isolation
- Kubelet Configuration
- Pod Deployment

```
apiVersion: sriovnetwork.openshift.io/v1
kind: SriovNetworkNodePolicy
...
```

```
apiVersion: sriovnetwork.openshift.io/v1
kind: SriovNetwork
metadata:
  name: net-a
spec:
  resourceName: nic1
  spoofChk: "off"
  trust: "on"
```

```
apiVersion: k8s.cni.cncf.io/v1
kind: NetworkAttachmentDefinition
metadata:
  annotations:
    k8s.v1.cni.cncf.io/resourceName:
      sriov/nic1
  name: net-a
spec:
  ...
```

openshift_container_platform/4.20/html/hardware_networks/configuring-sriov-net-attach

Cluster Node Tuning Operator

- NIC Configuration (VFs)
- Device Plugin Deployment
- Device Plugin Configuration
- NetworkAttachmentDefinition
- Kernel Hugepages
- Kernel CPU isolation
- Kubelet Configuration
- Pod Deployment

openshift/**cluster-node-tuning-operator**

- NIC Configuration (VFs)
- Device Plugin Deployment
- Device Plugin Configuration
- NetworkAttachmentDefinition
- Kernel Hugepages
- Kernel CPU isolation
- Kubelet Configuration
- Pod Deployment

cluster-node-tuning-operator

```yaml
apiVersion: performance.openshift.io/v2
kind: PerformanceProfile
metadata:
  name: high-performance
spec:
  cpu:
    reserved: "0-7"
    isolated: "8-15"
  hugepages:
    defaultHugepagesSize: 1G
    pages:
      - size: 1G
        count: 16
  nodeSelector:
    node-role.kubernetes.io/worker: ""
```

openshift_container_platform/4.20/html-single/node_apis/#performanceprofile-performance-openshift-io-v2

- NIC Configuration (VFs)
- Device Plugin Deployment
- Device Plugin Configuration
- NetworkAttachmentDefinition
- Kernel Hugepages
- Kernel CPU isolation
- Kubelet Configuration
- Pod Deployment

cluster-node-tuning-operator

machine-config-operator

```yaml
apiVersion: performance.openshift.io/v2
kind: PerformanceProfile
metadata:
  name: high-performance
spec:
  cpu:
    reserved: "0-7"
    isolated: "8-15"
  hugepages:
    defaultHugepagesSize: 1G
    pages:
      - size: 1G
        count: 16
  nodeSelector:
    node-role.kubernetes.io/worker: ""
```

```yaml
apiVersion: tuned.openshift.io/v1
kind: Tuned
```

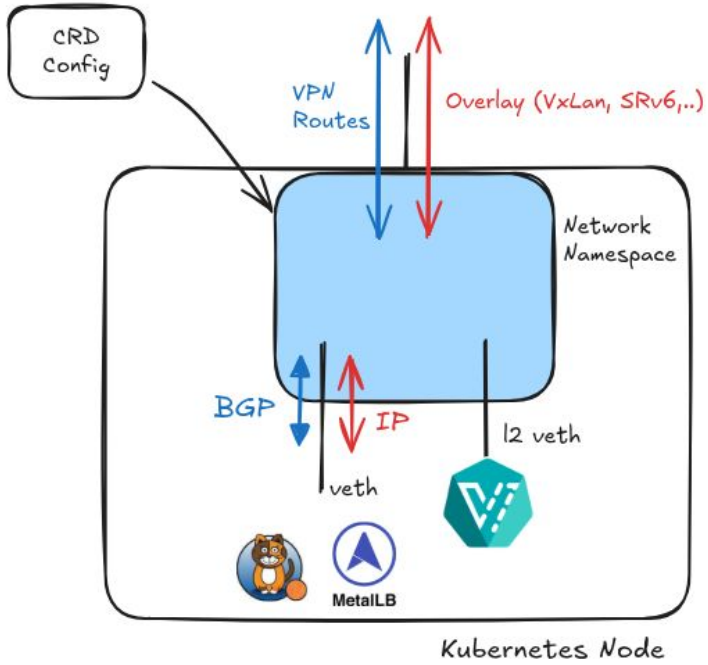
```yaml
apiVersion: machineconfiguration.openshift.io/v1
kind: KubeletConfig
```

```yaml
apiVersion: machineconfiguration.openshift.io/v1
kind: MachineConfig
```

https://tuned-project.org/

OpenPERouter

- Provider Edge router implementation
  - Moves top of the rack device configuration to k8s nodes
- Leverage FRRouting
  - BGP, EVPN
- Kernel based dataplane
  - VxLan, SRv6
  - Moving to Grout DPDK !



https://openperouter.github.io/

DPDK

Red Hat

Thank you!


DPDK/grout


openperouter


kubernetes


kubelet

Container Runtime Interface (CRI)


podman


cri-o


containerd

sriov-network-operator

sriov-network-device-plugin

sriov-cni

multus-cni


Open Container Initiative

opencontainers/runc

containers/crun


machine-config-operator

cluster-node-tuning-operator


TuneD

man7/namespaces

man7/cgroups

man7/capabilities

overlayfs

DPDK

Red Hat