
KubeCon
CloudNativeCon
North America 2019

Mitigating Noisy Neighbours
Advanced Container Resource Management

Alexander Kanevskiy, Intel
2019-11-20, v0.9

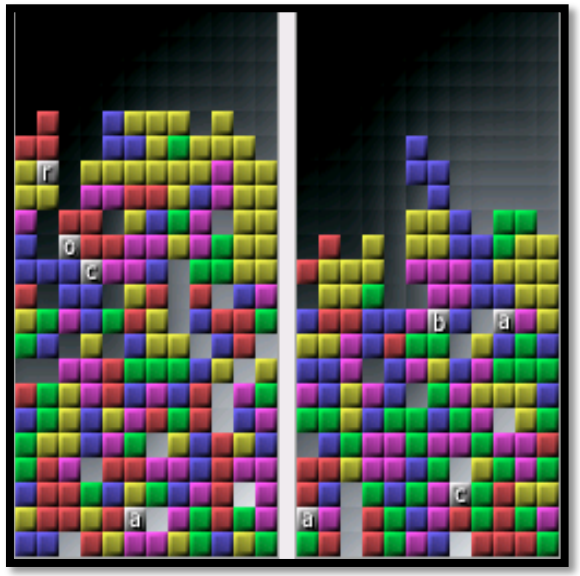
Foreword

- The real-life problem
  - … however, sometimes neither properly detected nor mitigated
- ”Silver bullet” does not exist
- Out of scope
  - Cluster level mitigations
  - Horizontal scaling
  - Dedicated nodes
  - …

* I Love Owls community

The “Noisy neighbour problem”

- In scope
  - Node hardware resources
    - CPU
    - Caches
    - Memory
    - Storage
    - Devices
  - Container runtimes
    - CRI-O*
    - containerd*
    - OCI runtimes: runc*, …



cri-o

containerd

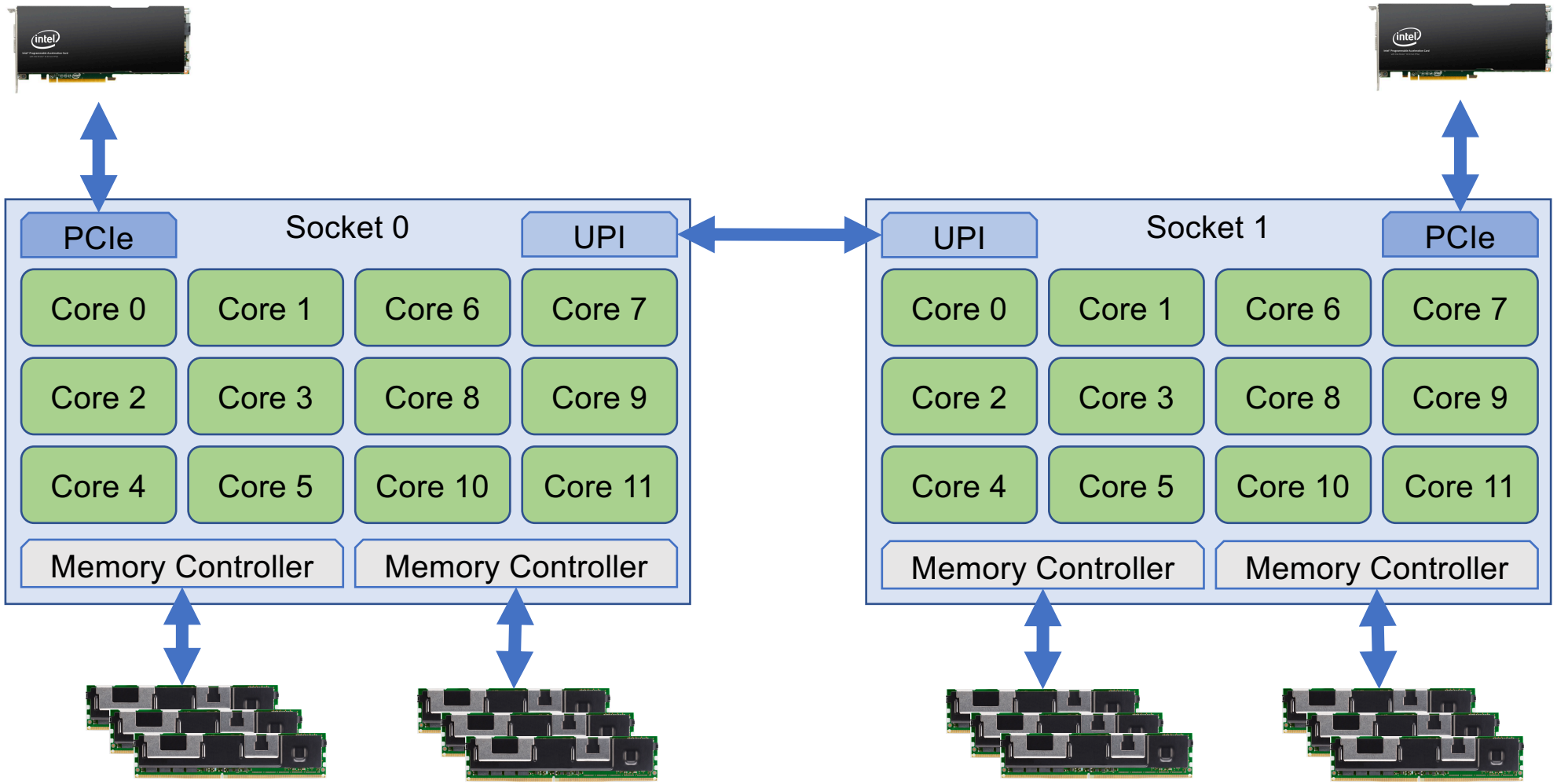
OPEN CONTAINER INITIATIVE

RUNC

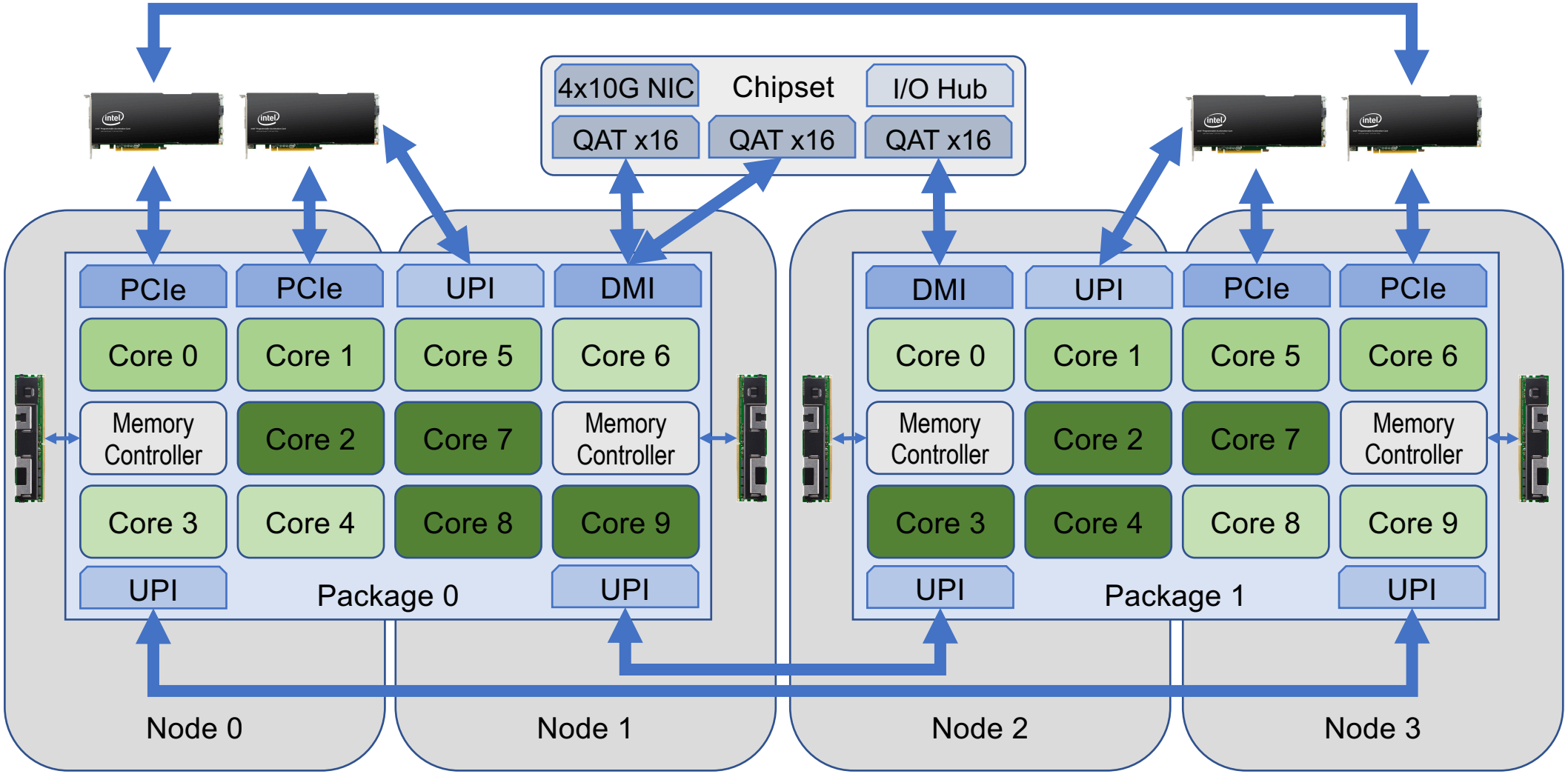
* Other names and brands may be claimed as the property of others.

Hardware resources

System devices topology



System topology in real world



Resources in Kubernetes*

* Other names and brands may be claimed as the property of others.

Resources in Kubernetes*

Container level

- spec.containers[].resources
  - requests and limits
    - cpu
    - memory
- Extended resources
  - Arbitrary advertised by node capacity
  - Device Plugin managed resources
  - requests = limits

Pod level

- QoS
  - Best Effort, Burstable, Guaranteed
- Metadata:
  - spec.metadata.labels
  - spec.metadata.annotations

```
apiVersion: v1
kind: Pod
metadata:
  annotations:
    kubernetes.io/ingress-bandwidth: 1M
    kubernetes.io/egress-bandwidth: 1M
    seccomp.security.alpha.kubernetes.io/pod: xyz
```

Challenges: blkio

- More complex resource
  - Weight does not have capacity
  - Weight can be per device
  - Throttling is per device
- Cluster level policies
  - Classes?
- Node level
  - Mapping classes to actual per device parameters

```
"blockIO": {
    "weight": 10,
    "weightDevice": [
        { "major": 8, "minor": 0, "weight": 500 },
        { "major": 8, "minor": 16, "weight": 400 }
    ],
    "throttleReadBpsDevice": [
        { "major": 8, "minor": 0, "rate": 600 }
    ],
    "throttleWriteIOPSDevice": [
        { "major": 8, "minor": 16, "rate": 300 }
    ]
}
```

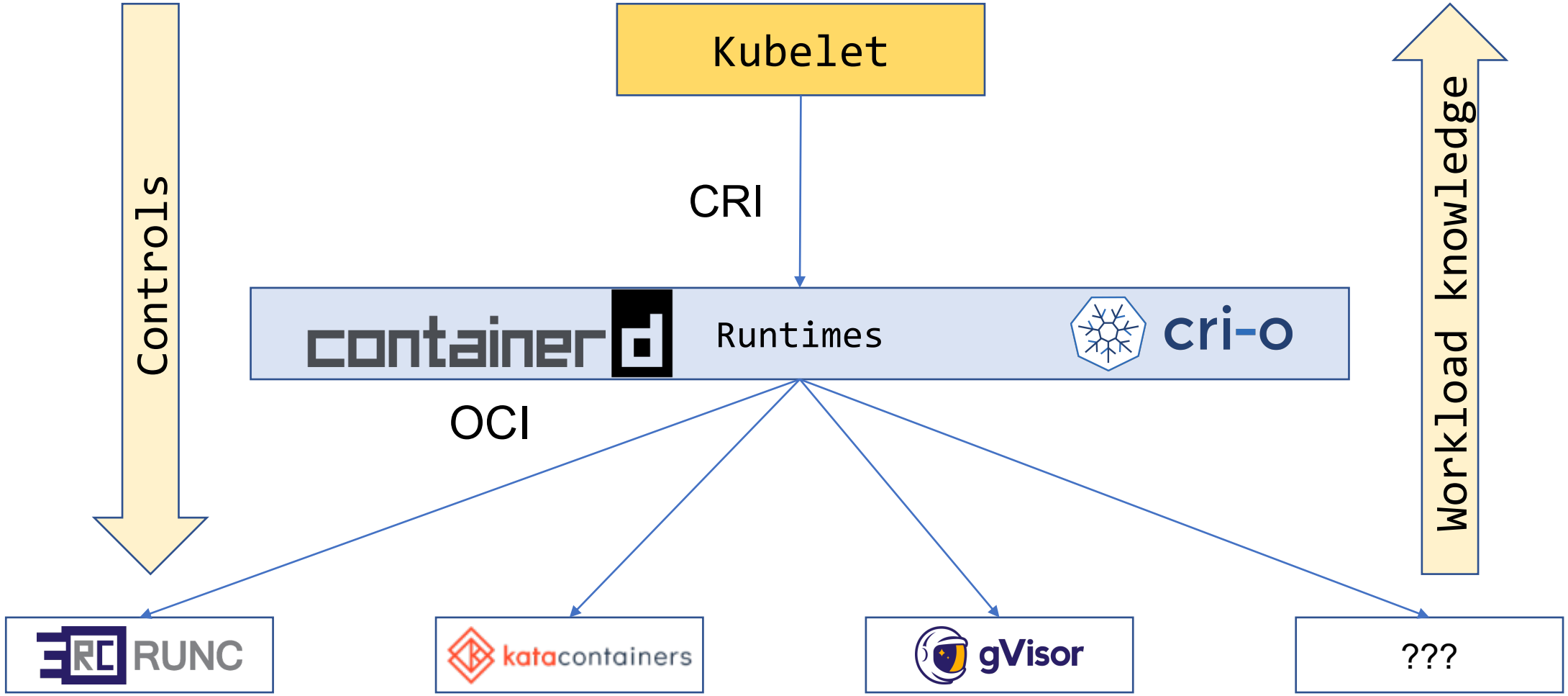
Challenges: resctrl

- Cache and Memory
  - Allocation and monitoring
  - Limited amount of classes
  - Exclusive cache lanes
  - Node hardware specific

```
"intelRdt": {
    "closID": "guaranteed_group",
    "l3CacheSchema": "L3:0=7f0;1=1f",
    "memBwSchema": "MB:0=20;1=70"
}
```

Resource controls

Runtime interfaces



Kubelet to runtimes: CRI

- Available:
  - CPU CFS parameters:
    - period, quota, shares
  - Memory
    - Limit
    - OOM Score
  - cpuset
    - cpus
    - mems

- What is lost:
  - CPU requests and limits
  - Memory requests
  - Extended resources
  - cpuset.mems not used
  - HugePages

Controls only on OCI* level

- runc*
  - blkio: weight
  - CPU real-time period
  - Kernel memory
  - Memory reservation
  - L3 cache schema
  - Memory Bandwidth schema

- OCI spec
  - blkio: IOPS / bps throttling
  - HugePages
  - Intel® RDT class

  - Hooks

* Other names and brands may be claimed as the property of others.

OCI* Hooks

* Other names and brands may be claimed as the property of others.

OCI* hooks configuration

- Executed by runtime
  - e.g. runc*
- Granularity: container
- Receive information
  - Container config (bundle)
  - Container annotations
- Can modify cgroups
- Can’t modify config.json
- More hooks: PR#1008

```
"hooks": {
    "prestart": [
        {
            "path": "/usr/bin/fix-mounts",
            "args": ["fix-mounts", "arg1", "arg2"],
            "env":  [ "key1=value1"]
        },
    ],
    "poststart": [
        {
            "path": "/usr/bin/notify-start",
            "timeout": 5
        }
    ],
    "poststop": [
        {
            "path": "/usr/sbin/cleanup.sh",
            "args": ["cleanup.sh", "-f"]
        }
    ]
}
```

* Other names and brands may be claimed as the property of others.

CRI-O* and OCI* hooks

/etc/crio/crio.conf

- Hooks are disabled by default
  - Comment out directive
    hooks_dir = []
- Default search paths
  - /etc/containers/oci/hooks.d/
  - /usr/share/containers/oci/hooks.d/
- Works only in CRI-O* so far
  - Containerd* hooks: PR#1248

/etc/containers/oci/hooks.d/hook.json

```
{
  "version": "1.0.0",
  "hook": {
    "path": "/opt/demo/hook"
  },
  "when": {
    "always": true
  },
  "stages": ["prestart"]
}
```

* Other names and brands may be claimed as the property of others.

Custom Runtimes

Runtime Classes

Runtime Class definition

```
apiVersion: node.k8s.io/v1beta1
kind: RuntimeClass
metadata:
  name: blkio
handler: blkio
```

Pod Runtime Class usage

```
apiVersion: v1
kind: Pod
metadata:
  name: mypod
spec:
  runtimeClassName: blkio
  # ...
```

Runtime Class handlers

CRI-O*
/etc/crio/crio.conf

```
[crio.runtime.runtimes.blkio]
runtime_path = "/opt/demo/runc.blkio"
```

containerd*
/etc/containerd/config.toml

```
[plugins.cri.containerd.runtimes.blkio]
    runtime_type = "io.containerd.runc.v1"
    pod_annotations = ["*"]
    container_annotations = ["*"]

[plugins.cri.containerd.runtimes.blkio.options]
    BinaryName = "/opt/demo/runc.blkio"
```

runc* wrapper


```
#!/bin/bash
# WARNING: demo only, contains bugs
if [ "$1" == "start" ]; then
    if [ -n "$2" ]; then
        BUNDLE=`/usr/bin/runc state $2 2>/dev/null | jq .bundle -r`
        if [ -n "$BUNDLE" -a -f "$BUNDLE/config.json" ]; then
            CGROUP=`jq .linux.cgroupsPath $BUNDLE/config.json -r`
            if [[ "$CGROUP" == *burstable* ]]; then
                W=50
            elif [[ "$CGROUP" == *besteffort* ]]; then
                W=10
            fi
            if [ -n "$W" ]; then /usr/bin/runc update --blkio-weight $W $2 ; fi
        fi
    fi
fi
exec /usr/bin/runc "$@"
```

* Other names and brands may be claimed as the property of others.

KubeCon | CloudNativeCon

North America 2019

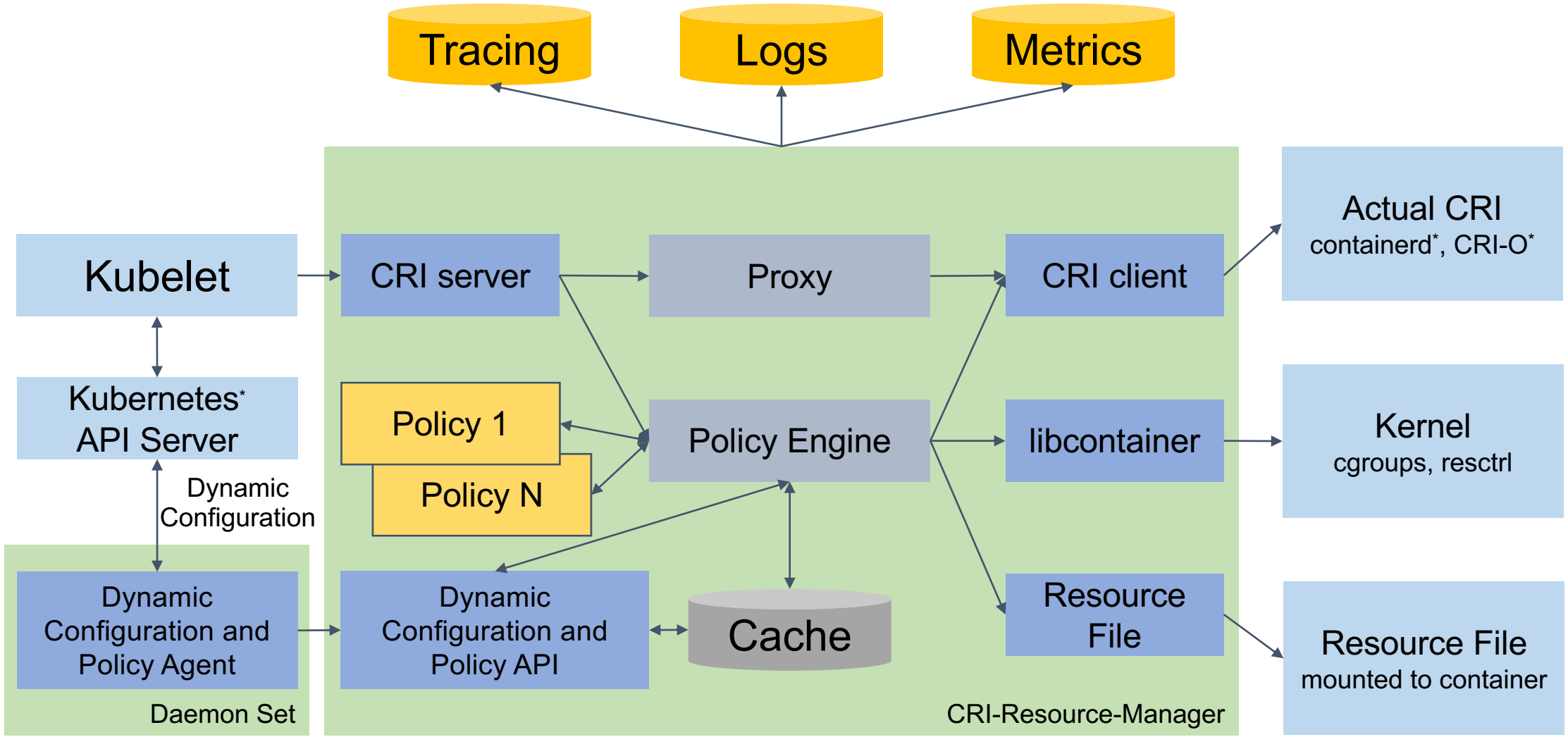
CRI Resource Manager

https://bit.ly/cri-r-m

CRI Resource Manager

- What?
  - Basically it is a Container Runtime Interface proxy
- How?
  - Applies (hardware) resource policies to containers by
    - modifying proxied container requests, or
    - generating container update requests, or
    - triggering extra policy-specific actions during request processing
    - can interact directly with kernel interfaces
- Why?
  - Started as internal debug and tracing tool
  - Instrumentation of CRI interface
  - Enables easy prototyping of features before upstreaming

CRI Resource Manager



* Other names and brands may be claimed as the property of others.

CRI Resource Manager: now

- Policies:
  - Static
    - Same as Kubelet's CPU manager, with support of isolcpus
  - Static+
    - As above, with support of mixed shared + exclusive CPUs
    - Downwards API exposed to container
  - Topology-aware
    - Multilayered topological set of pools for shared, exclusive and isolated CPUs
    - CPU and memory alignment based on devices and storage volumes hints
    - Containers affinity/anti-affinity
- Intel® RDT: L3 Cache and Memory Bandwidth allocation
- Dynamic configuration API
  - Global, groups and individual node configs

CRI Resource Manager: WIP

- Block I/O classification and tuning
- Better monitoring of resources usage
  - Block I/O usage
  - NUMA memory consumption stats
  - L3 Cache monitoring
  - Memory Bandwidth monitoring
  - …
- Dynamic rebalancing
- External Policy APIs

Demos
CRI Resource Manager

KubeCon | CloudNativeCon

North America 2019

Demo: Static policy

https://bit.ly/cri-r-m-s-demo

Demo: Static Plus policy

https://bit.ly/cri-r-m-sp-demo

Demo: Topology Aware policy

https://bit.ly/cri-r-m-t-demo

Key takeaways

- Hardware
  - Not all “CPUs” reported by the OS are equal
  - The “C” in “NUMA” stands for “CPU”
  - Even if your environment is virtualized, keep in mind underlying hardware
  - … we live in the world where assumptions about hardware are changing frequently and drastically
- Kubernetes* resources
  - Not everything can be easily represented as simple countable object
  - Time to think about user experience for other types of resources?
- Do your own experiments
  - CRI Resource Manager can give you hand for your custom resource policies
  - … and share ideas and results of your experiments with the community

Thank you!

GitHub*: @kad
Kubernetes* Slack*: @akanevskiy


KubeCon
CloudNativeCon
North America 2019

Legal notices and disclaimers

- Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation.
- Performance varies depending on system configuration.
- No computer system can be absolutely secure.
- Check with your system manufacturer or retailer or learn more at www.intel.com.
- Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
- *Other names and brands may be claimed as the property of others.
-