Node Resource Management
The Big Picture

Sascha Grunert & Swati Sehgal, Red Hat
David Porter, Google
Evan Lezar, NVIDIA
Alexander Kanevskiy, Intel

Who we are?

KubeCon CloudNativeCon Europe 2023

David Porter
Senior Software Engineer
Google

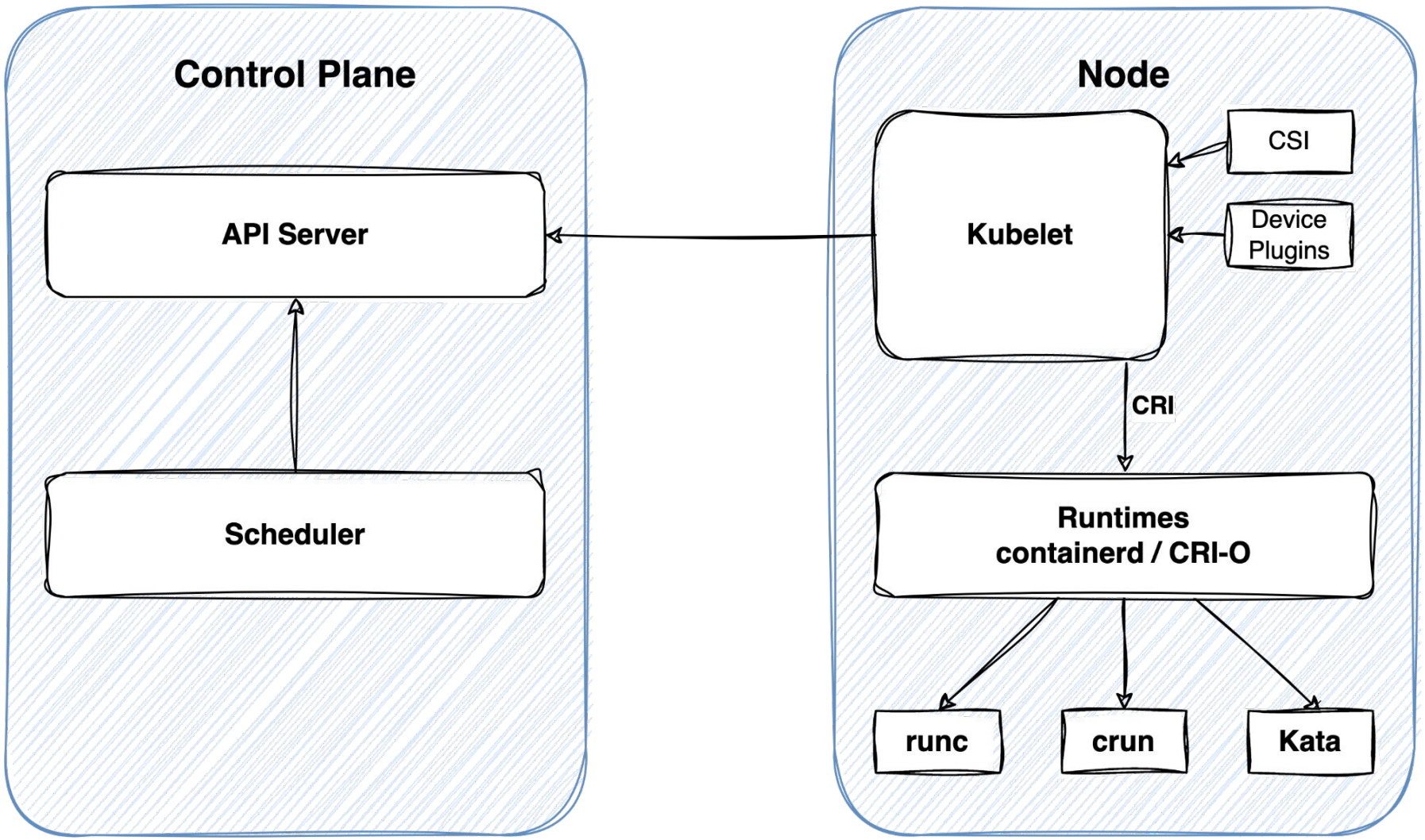
Swati Sehgal
Principal Software Engineer
Red Hat

Sascha Grunert
Senior Software Engineer
Red Hat

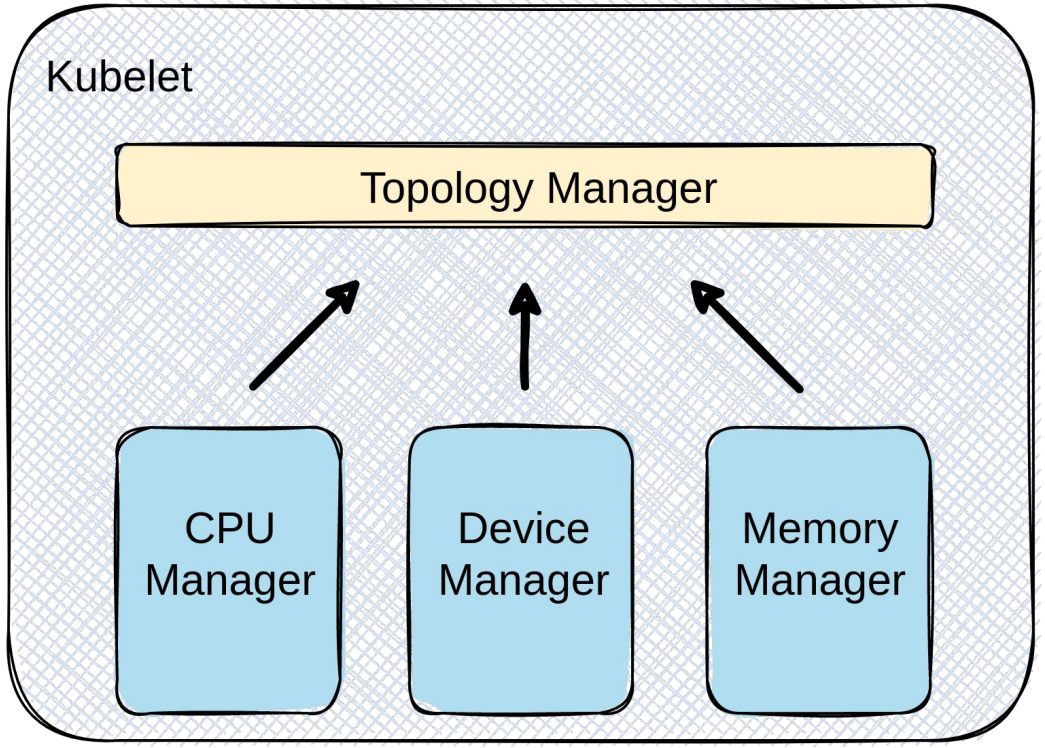
Alexander Kanevskiy
Principal Engineer, Cloud Software
Intel

Evan Lezar
Senior Systems Software Engineer
NVIDIA

The days of not so long past…



Kubelet - Resource Managers



Resource Manager	Alpha	Beta	GA
Device Manager	v1.8	v1.10	v1.26
CPU Manager	v1.8	v1.10	v1.26
Topology Manager	v1.16	v1.18	v1.27
Memory Manager	v1.21	v1.22	

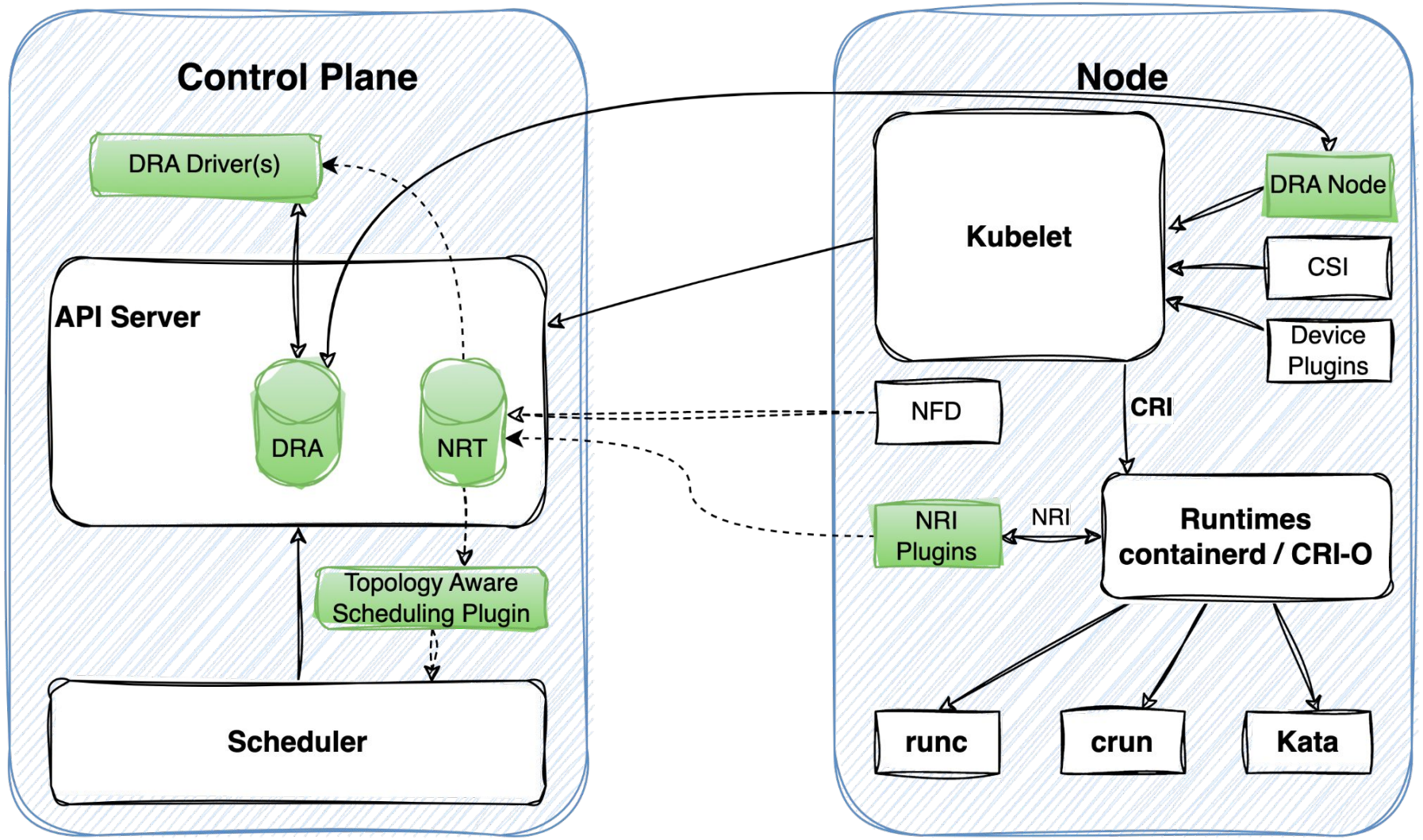
Blog Posts:

[Kubernetes Topology Manager Moves to Beta - Align Up!]

[Kubernetes v1.26: CPUManager goes GA]

[Kubernetes 1.26: Device Manager graduates to GA]

Kubernetes Resources world nowadays



Dynamic Resource Allocation

- Resource counts replaced with the notion of **`ResourceClaims`** that separate resource declaration from resource consumption

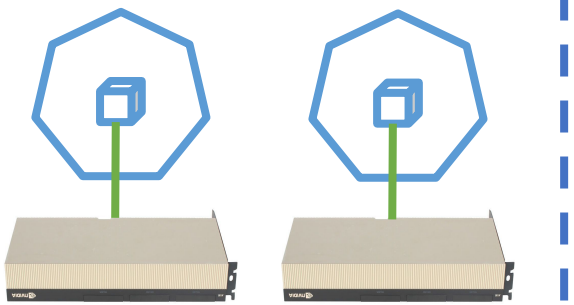
```
apiVersion: v1
kind: Pod
metadata:
  name: gpu-example
spec:
  containers:
  - name: ctr
    image: nvidia/cuda
    command: ["nvidia-smi", "-L"]
    resources:
      limits:
        nvidia.com/gpu: 2
```

```
apiVersion: v1
kind: Pod
metadata:
  name: gpu-example
spec:
  containers:
  - name: ctr
    image: nvidia/cuda
    command: ["nvidia-smi" "-L"]
    resources:
      claims:
      - gpu0
      - gpu1
  resourceClaims:
  - name: gpu0
    claim:
      template:
        spec:
          resourceClassName: gpu.nvidia.com
  - name: gpu1
    claim:
      template:
        spec:
          resourceClassName: gpu.nvidia.com
```

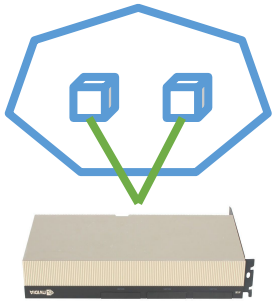
Indicates which resource driver is responsible for the claim

See the talk today at 11:55 in room G104-G105:
Device Plugins 2.0: How to Build a Driver for Dynamic Resource Allocation - Kevin Klues, NVIDIA & Alexey Fomenko, Intel

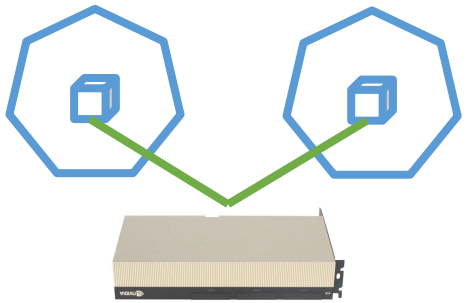
Dynamic Resource Allocation



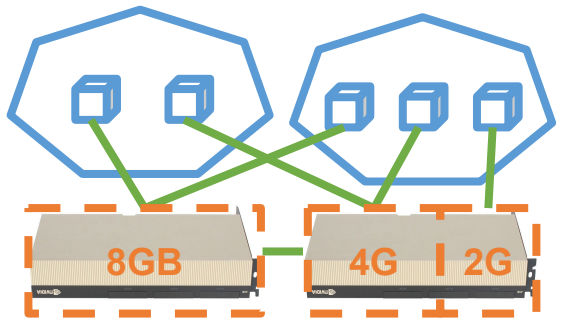
Containers in different pods each with access to a device



Containers in the same pod sharing a device



Containers from multiple pods sharing a device

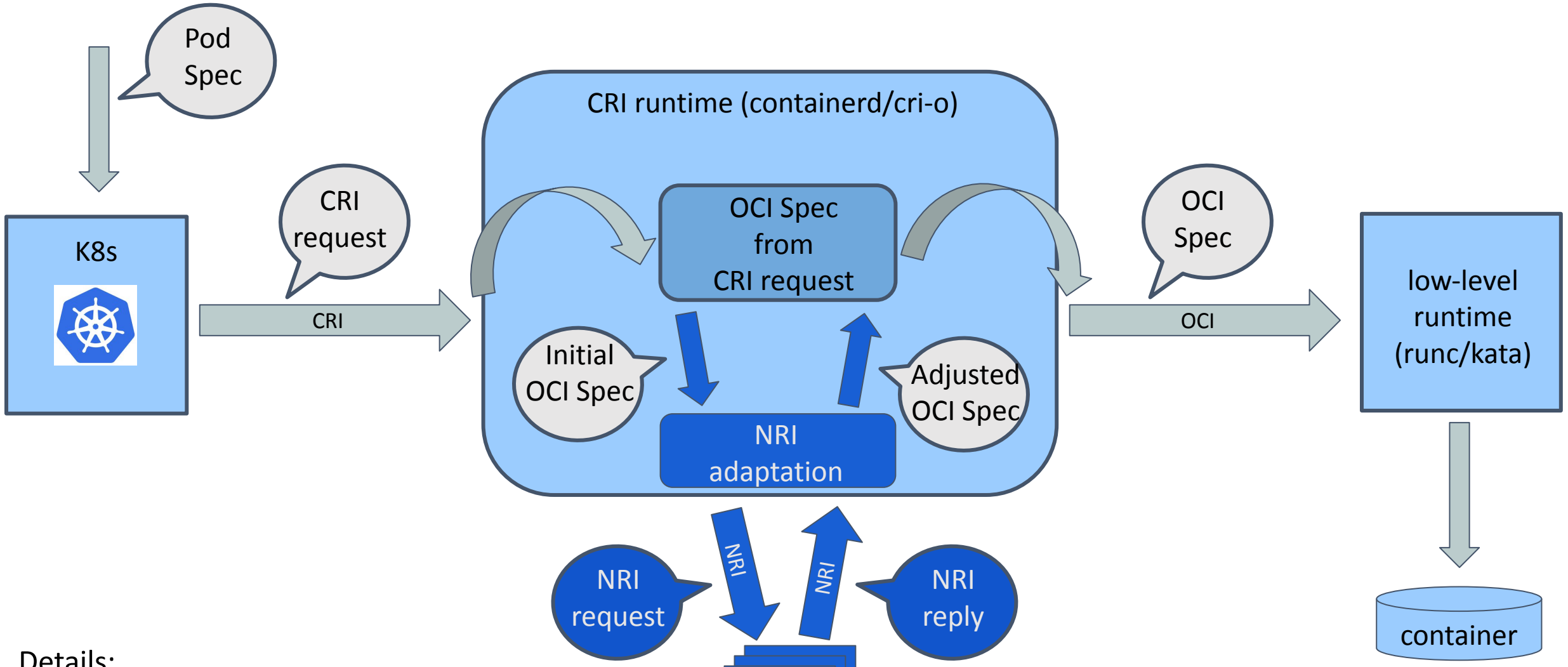


Multiple containers in multiple pods sharing dynamically partitioned devices

Possible with a device plugin

- Enables a more flexible allocation and usage of resources
- CDI specification is (typically) generated dynamically for a claim
- Relies on CDI support in CRI runtimes (cri-o, containerd)

NRI: containerd 1.7+ & CRI-O 1.26+



Details:

- Available in containerd 1.7 and CRI-O 1.26
- NRI KubeConNA'22 presentation

- https://github.com/containerd/nri/
- https://github.com/containers/nri-plugins/

NodeResourceTopology API

CRD based API to capture node resource topology information

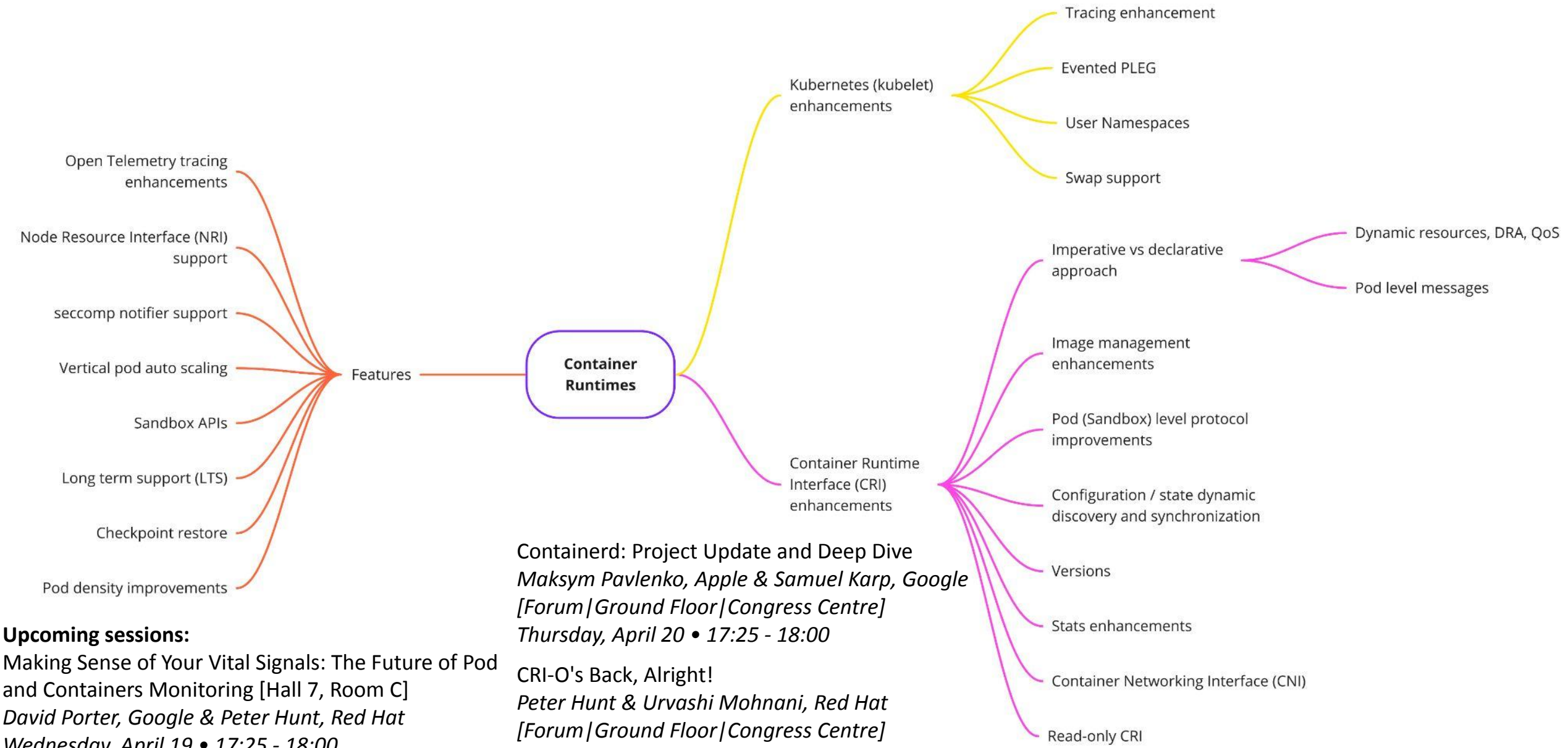
Use Cases:

1. **Topology-aware Scheduling**
   a. **NFD-topology updater** exposes resource topology information as CR instances created per node.
   b. **TAS Scheduler plugin** uses the per-node CRs to make topology aware scheduling decisions.
   c. For more details: Topology-aware Scheduling Kubecon EU 2022 Presentation
2. The resource policies implemented as **NRI plugins** are utilizing this API to share internal state assignments, which in turn can be used by other components like DRA controllers or scheduler plugins.

   Details: https://github.com/k8stopologyawareschedwg/noderesourcetopology-api

See the talk today at 15:25 in room E105-106:
WG Batch: What's New and What Is Next? - Swati Sehgal, Red Hat & Aldo Culquicondor, Google

Container Runtimes



Upcoming sessions:

Making Sense of Your Vital Signals: The Future of Pod and Containers Monitoring [Hall 7, Room C]
David Porter, Google & Peter Hunt, Red Hat
Wednesday, April 19 • 17:25 - 18:00

Containerd: Project Update and Deep Dive
Maksym Pavlenko, Apple & Samuel Karp, Google
[Forum|Ground Floor|Congress Centre]
Thursday, April 20 • 17:25 - 18:00

CRI-O's Back, Alright!
Peter Hunt & Urvashi Mohnani, Red Hat
[Forum|Ground Floor|Congress Centre]
Friday, April 21 • 14:55 - 15:30

cgroup v2 platform roadmap

- cgroup v2 is the next generation API (GA in k8s 1.25)
- Future roadmap
  - MemoryQoS (alpha 1.27)
  - PSI metrics for eviction
  - IO isolation
  - Swap
  - Network
  - What issues are you facing? We want to hear from you!

- Refer to *"cgroup v2 Is Coming Soon To a Cluster Near You" - Kubecon NA Detroit 2022 (David Porter, Mrunal Patel) for more details*

How to participate?

KubeCon CloudNativeCon Europe 2023

- CNCF
  - TAG-Runtime & Container Orchestrated Devices WG
    - Slacks:
      - #tag-runtime
      - #containerd
      - #crio
  - Projects
    - containerd
    - CRI-O
    - NRI & NRI Plugins
- Kubernetes
  - SIG-Node
    - Slack: #sig-node @ Kubernetes

KubeCon CloudNativeCon Europe 2023

KubeCon CloudNativeCon North America 2021

Please scan the QR Code above
to leave feedback on this session


KubeCon
CloudNativeCon
Europe 2023
TiKV