
KubeCon
CloudNativeCon
North America 2023


KubeCon | CloudNativeCon

North America 2023

Lifting the Hood - to Take a Look at the Kubernetes Resource Management Evolution

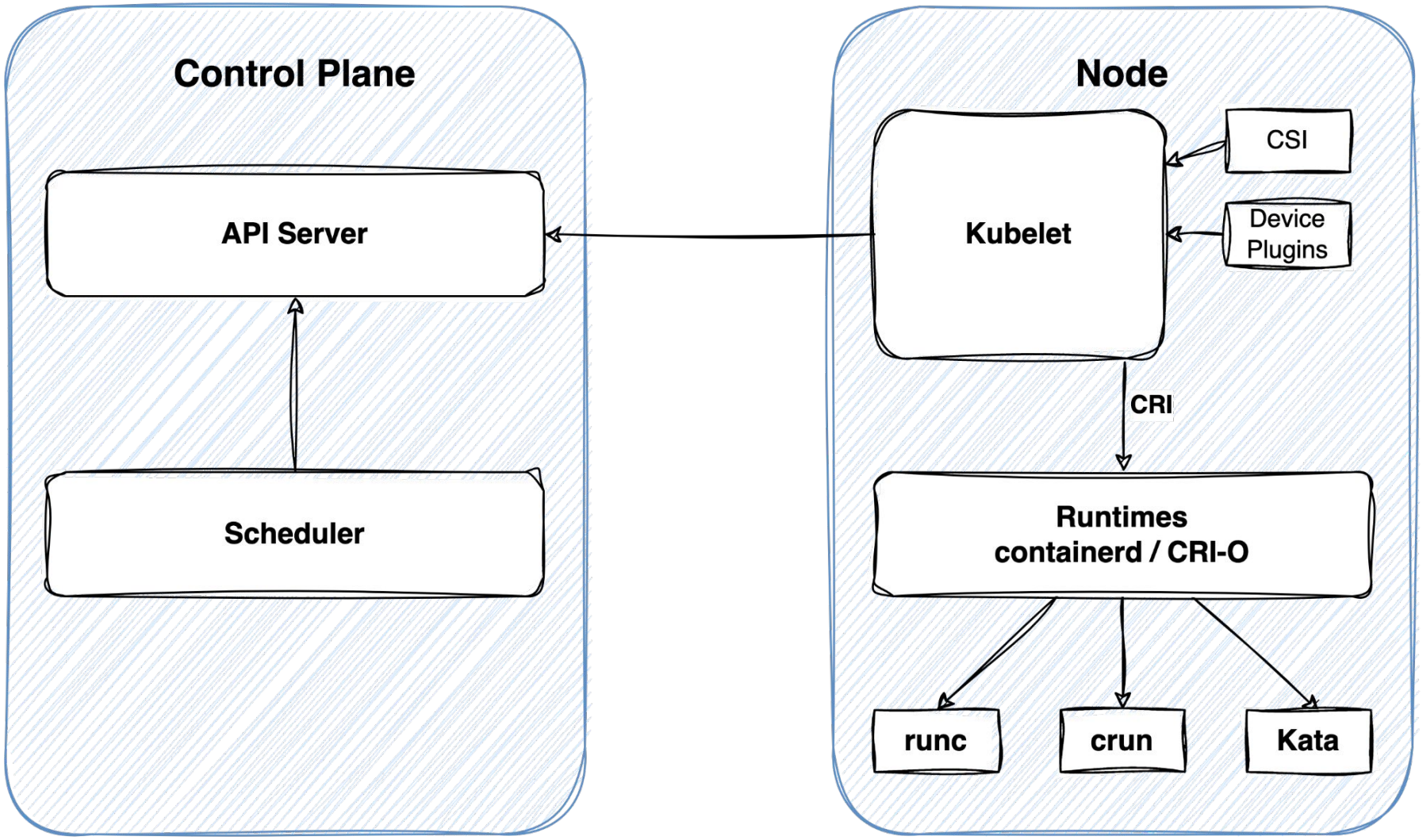
Mike Brown, IBM
Alexander Kanevskiy, Intel

Who we are?

Mike Brown
Software Architect
IBM
containerd maintainer, OCI, …

Alexander Kanevskiy
Principal Engineer, Cloud Software
Intel
TAG-Runtimes, COD WG

The days of not so long past…

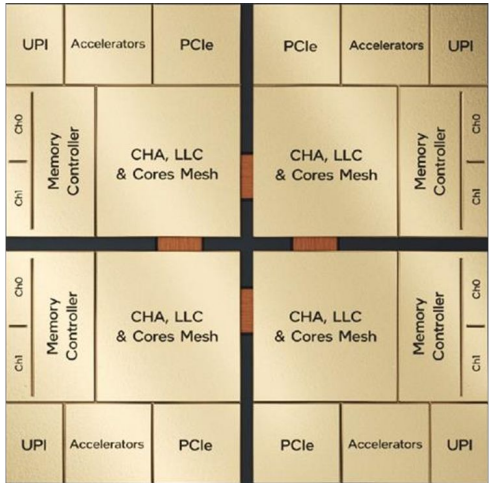


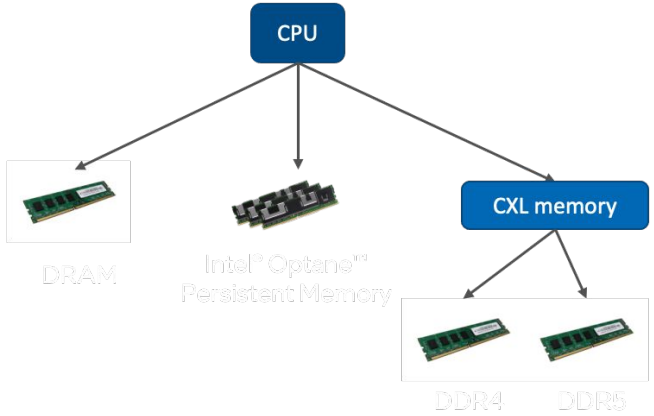
The world is changing…

- Software and workload driven changes
  - Raise of AI
    - … demand on accelerator devices
    - … new instruction sets for CPUs
    - … more CPU-intensive applications
    - … more memory needed for large models
    - … more control over “noisy neighbours” required
  - Confidential Computing
    - … stricter requirements to every element in the stack
    - … different lifecycle of the artifacts depending on runtime class implementation

The world is changing…

- Hardware driven changes
  - CPU is not anymore homogenous
    - Clusters of cores
      - logical groups or physical tiles
    - In-package interconnect between physical groups
    - Not all cores similarly sharing caches
    - Performance cores
      - 2 threads, “classical big cores”, higher frequencies
    - … and Efficient cores
      - 1 thread, L2 cache shared by group of 4 cores
    - Embedded accelerators
  - Memory is now heterogeneous dynamic resource
    - Multiple modes of operation for memory controllers
      - SNC(Sub-NUMA) / NPS(Nodes Per Socket)
    - More types of memory: DDR, HBM, PMEM, CXL.mem
    - Dynamic memory hot-plug for CXL.mem



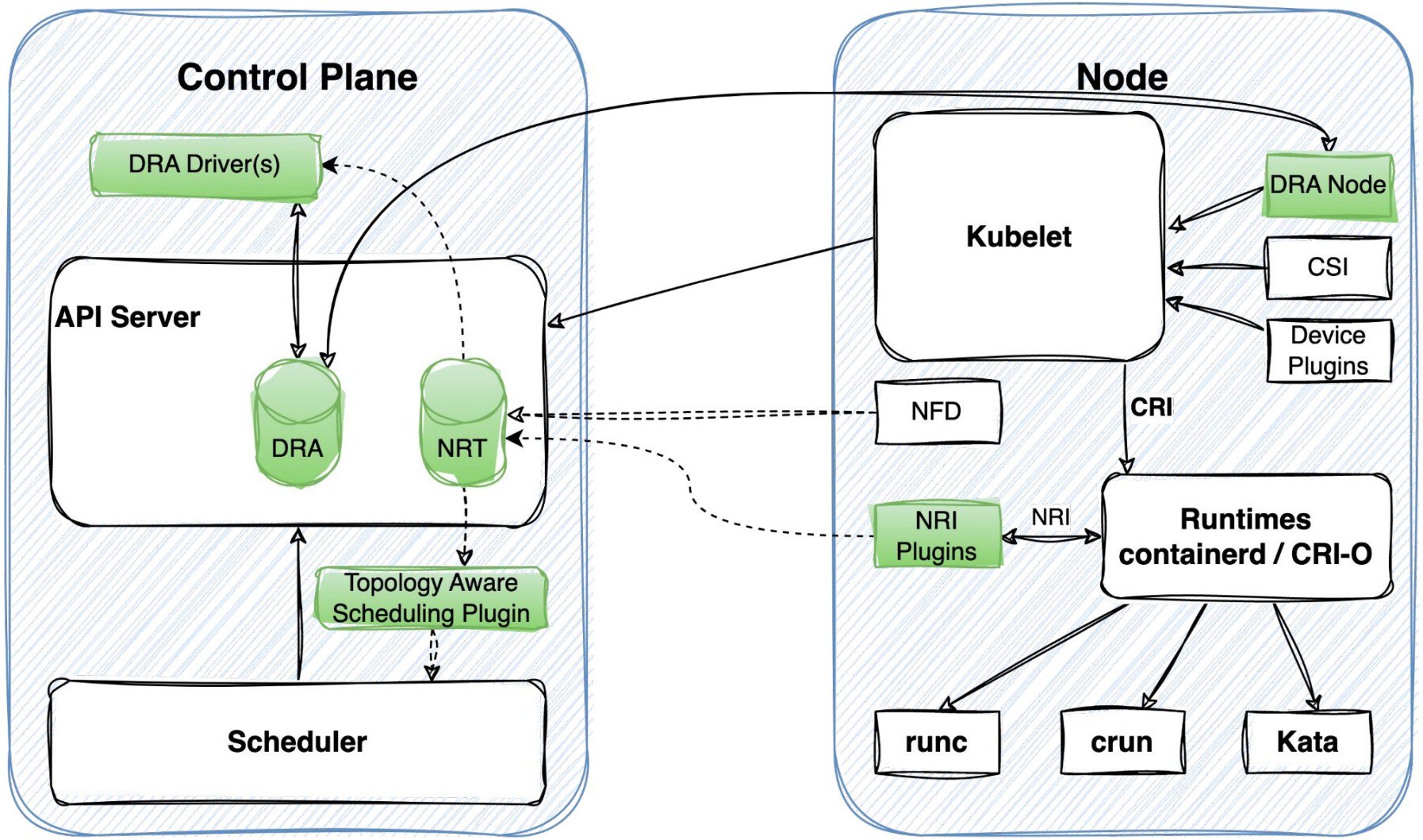


How to deal with new complex world?

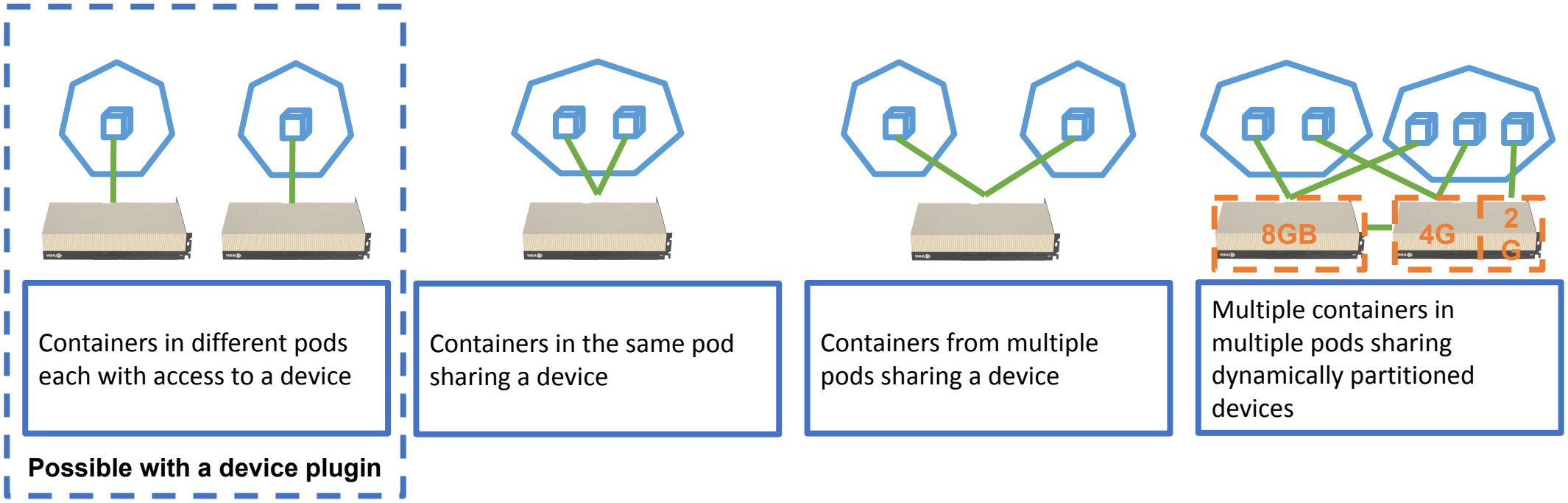
- Some of the problems are easy
  - e.g. Node Feature Discovery
  - Find the best node in the cluster to run your optimized workload
    - Discover your hardware
    - CPU instruction sets
      - AMX, AVX, AESNI, …
    - Devices and accelerators
      - Embedded, PCIe, CXL, USB, …
    - Network
    - Storage
    - System
    - Custom
      - Bring your own data sources
    - workload can use scheduler primitives to expose requirements on those properties

- However, most of other problems require new solutions
  - e.g. Accelerator is not anymore “1 GPU”, rather “1 GPU with 32Gb of memory, time shared between 4 workloads”

Kubernetes Resources world nowadays



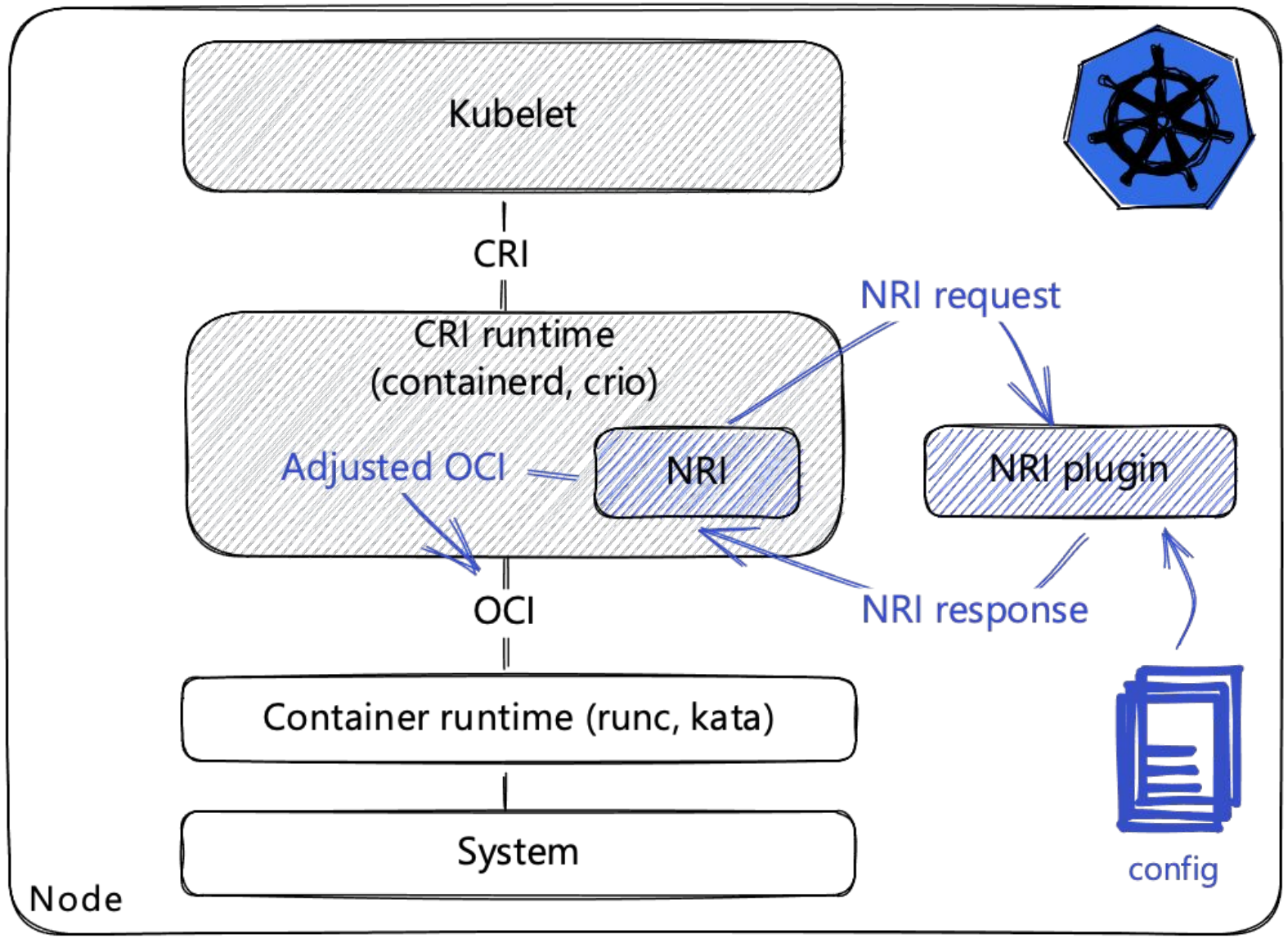
Dynamic Resource Allocation



- Enables a more flexible allocation and usage of accelerator resources in Kubernetes
  - high level abstraction for declaring accelerator request requirements
  - relies on Container Device Interface in runtime to convert opaque allocation ID to actual device properties
    - CDI new home under CNCF-Tags: https://github.com/cncf-tags/container-device-interface
- Kubernetes 1.27+, containerd 1.7+ or CRI-O 1.26+

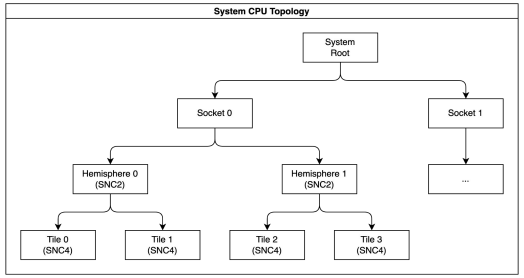
NRI: Node Resource Interface

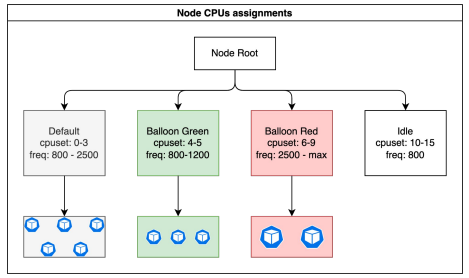
- What
  - Common plugin interface to enhance container runtimes
  - Originally, by Apple as subproject of containerd
  - Improved version adopted by CRI-O
  - New use cases from Google
- Usage
  - Flexible resource management
  - Hook injection
  - Logging and debugging
  - Security introspection and enforcement
  - Prototyping Runtimes extensions
- Availability
  - Containerd 1.7+
  - CRI-O 1.26+
    - OpenShift 4.13+
- Maturity: experimental feature



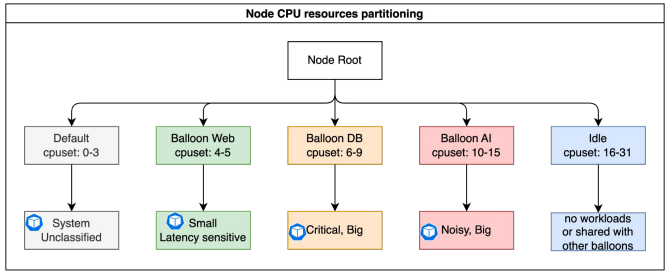
Currently available NRI Plugins

- Reference plugins in NRI repository
  - https://github.com/containerd/nri
  - NEW in v0.4.0: ulimit-adjuster plugin
  - WIP: integration with Pod network lifecycle: NRI PR#57

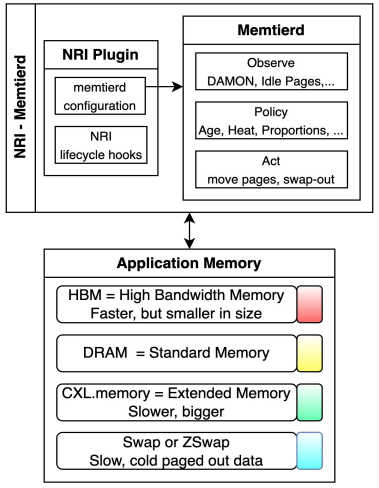




- Community maintained NRI plugins collection:
  - https://github.com/containers/nri-plugins
  - Reference resource policy plugins
    - Topology-Aware and Balloons
  - Memory Management plugins
    - memory-QoS, memtierd, SGX-EPC
      - Advancing Memory Management in Kubernetes: Next Steps with Memory QoS - Dixita Narang, Google & Antti Kervinen, Intel @ Tuesday 7, 2023, 2:30pm

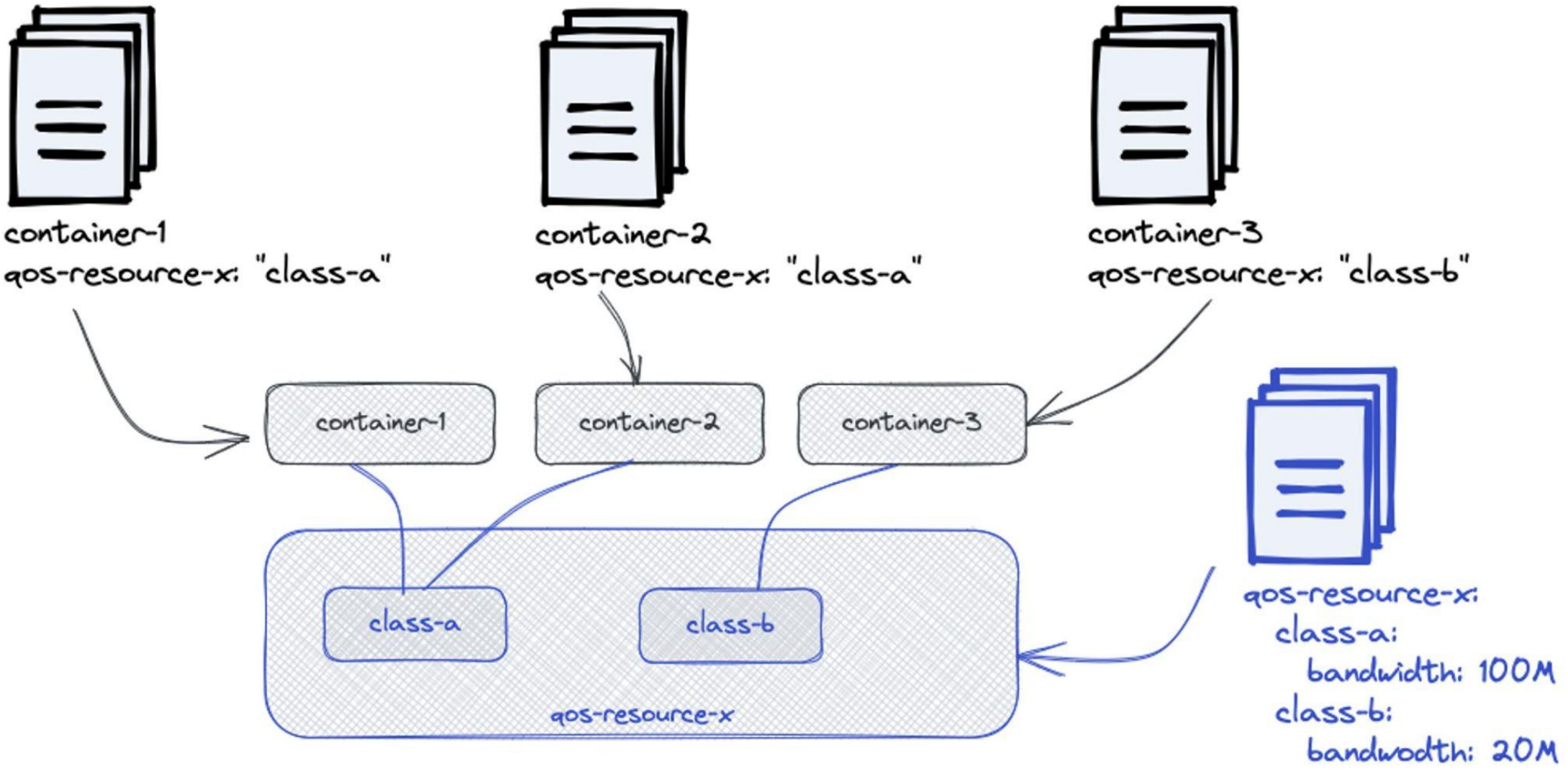


Memory: QoS, Swap and memory tiers

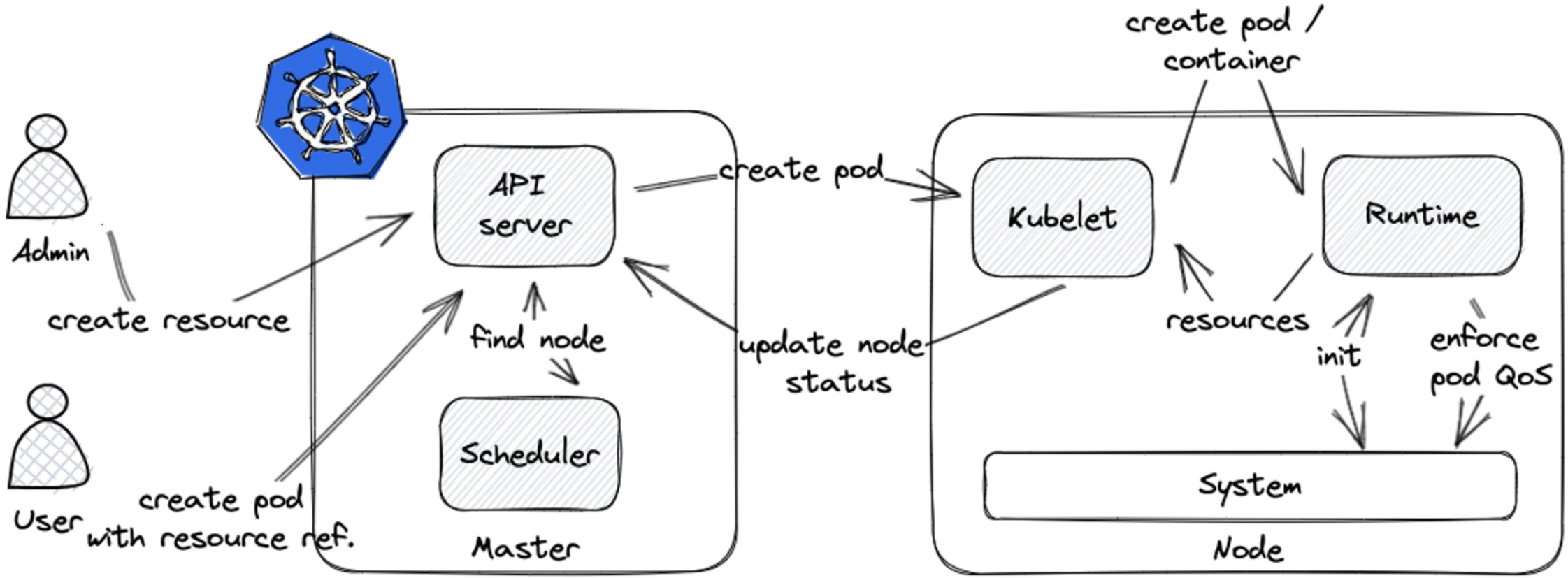


- Usage
  - Helm charts at ArtifactHUB.io:
    - https://artifacthub.io/packages/search?repo=nri-plugins
  - Cloud Providers
    - Helm charts validated on Google Cloud and Microsoft Azure
  - NRI interface can be easily enabled in popular Kubernetes deployment tools
    - kubespray (master, backported to v2.23.1)
    - kOps (upcoming releases)

KEP-3008: QoS class resources - Concepts



QoS class resources - Architecture



Join to the discussion in the [KEP-3008: QoS-class resources]

Get involved!

- 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

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


KubeCon
CloudNativeCon
North America 2023