Sharding baremetal Kubernetes clusters with Kamaji and Kubevirt

LENNART JERN

10.12.2025

What is the problem with baremetal?

- Large Nodes in terms of CPU and RAM
  - Kubernetes designed for max 110 Pods per Node
  - Can you make use of the resources?
  - What about isolation of workloads from each other?
- Expensive if under utilized
  - Can you use the resources and stay within 110 Pods/Node?
  - Are you going to run workload on the CP Nodes?
  - 3 CP Nodes?

Can we get the best of both worlds?

- Hardware access and performance for the workload that needs it
- Isolation and multi-tenancy without wasting resources
- Both baremetal and virtual Nodes in the same cluster?
- Light-weight control planes but heavy-duty worker Nodes?

Let's look at three projects that can help us achieve this!

Kubevirt

- CNCF Incubating project
- Virtual machine management on Kubernetes
- Essentially, run your VMs as Pods in a Kubernetes cluster
- Makes it possible to run VM-based workloads along modern containerized applications

Cluster API infrastructure provider!

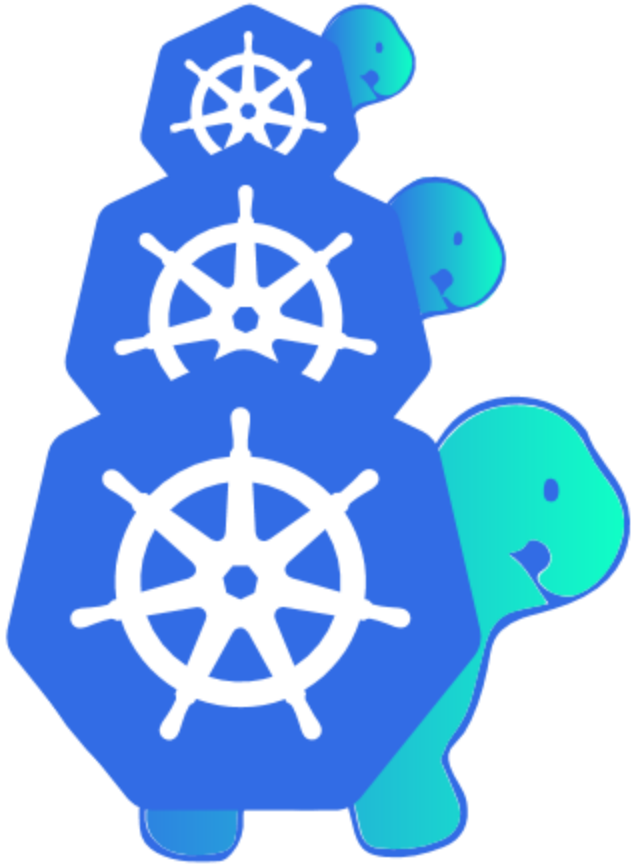
Metal3.io

- CNCF incubating project
- Baremetal management on Kubernetes
- Provision servers with disk images and start-up scripts

Cluster API infrastructure provider!

Kamaji

- Open source project by Clastix
- Run Kubernetes control plane components as Pods
- Allows light-weight but highly available CPs to be managed efficiently in a central cluster
- No dedicated control plane Nodes needed
- Multi-tenant ETCD can be shared by multiple control planes

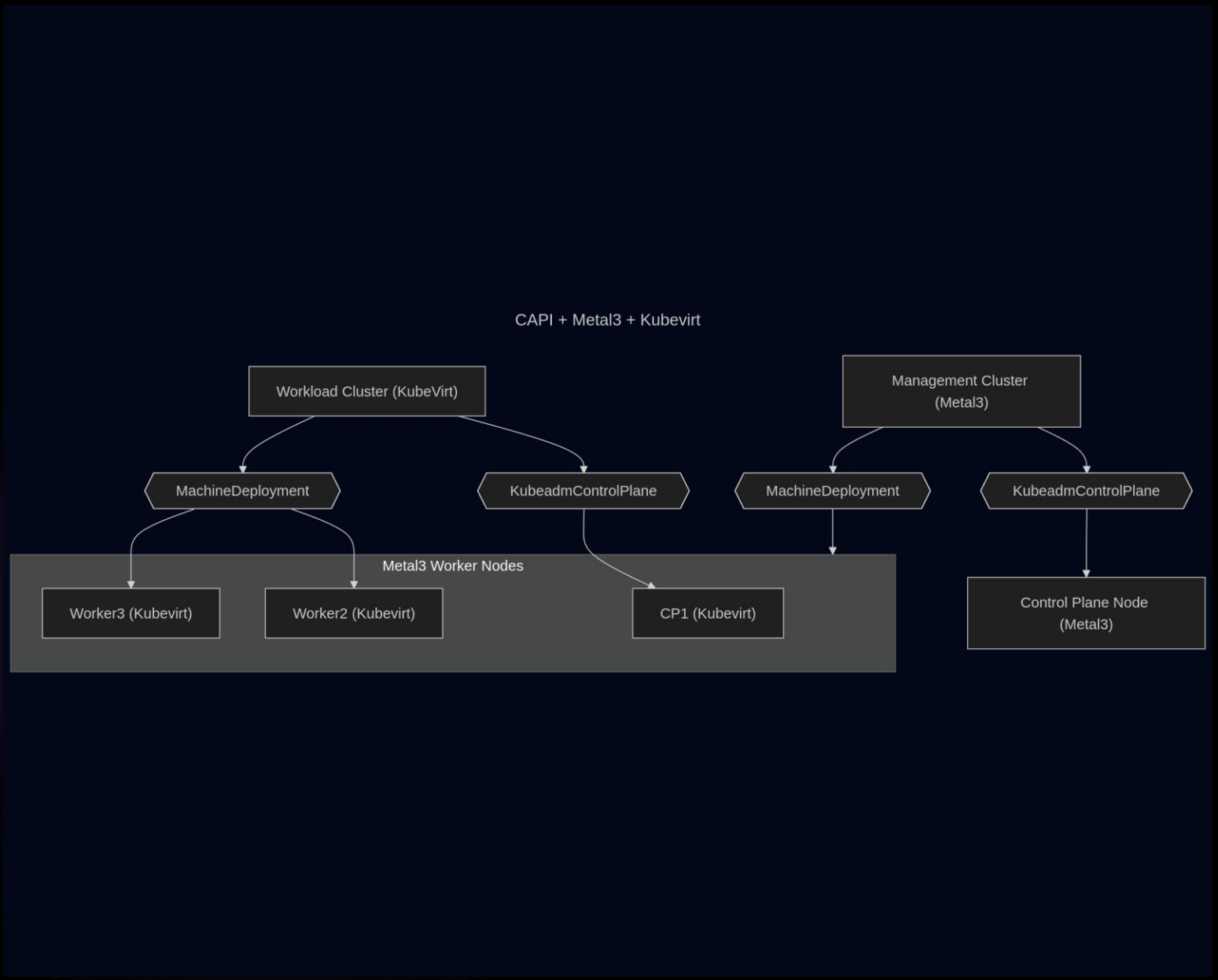


"Turtles all the way down..."

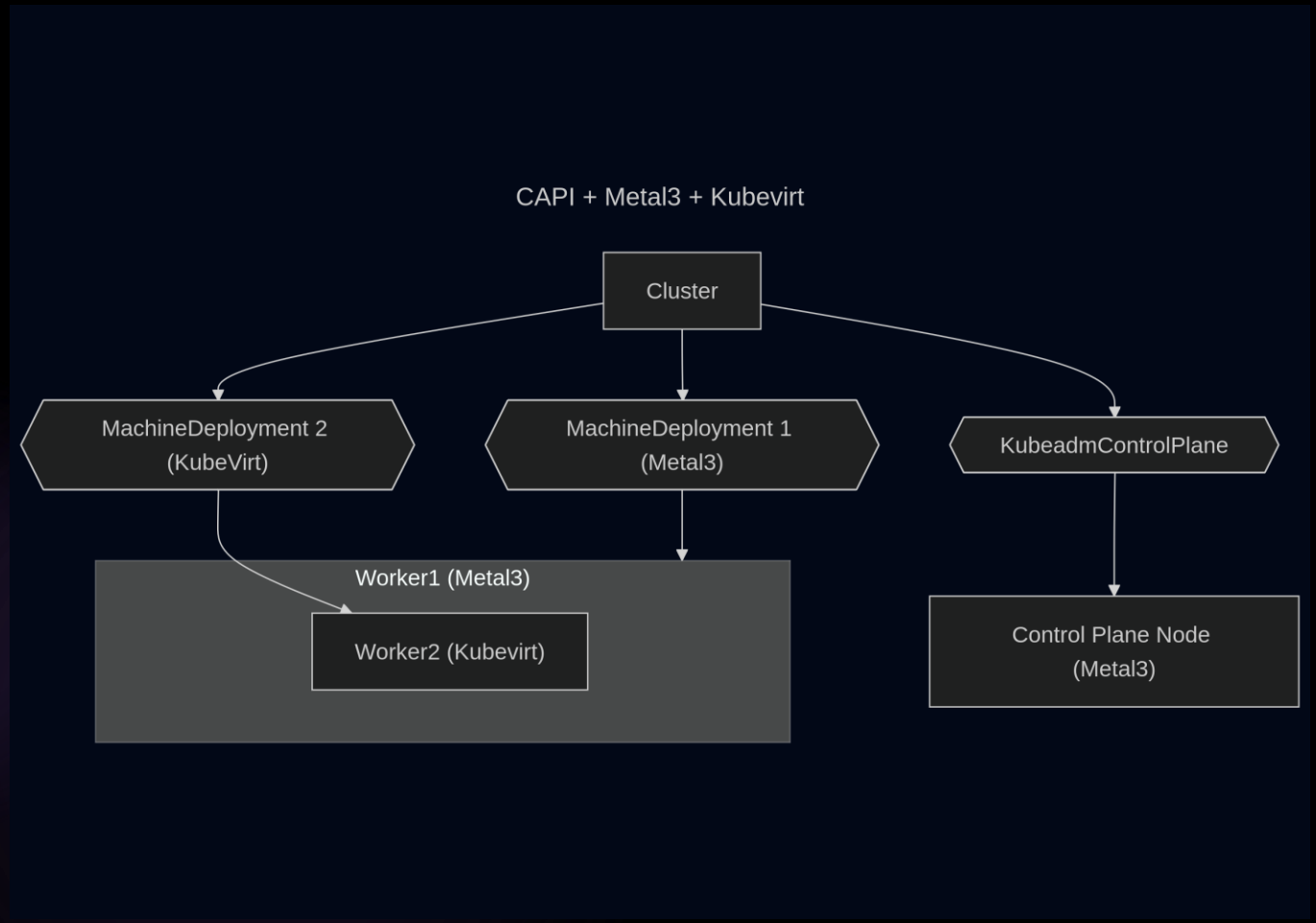
Architectures

- Management cluster: baremetal using Metal3
- Virtualized workload cluster: Kubevirt only – running inside the management cluster
- Hack – mixed cluster: Join Kubevirt nodes to Metal3 cluster
  - Requires an "extra" InfraCluster for Kubevirt with manually added owner reference
- Kamaji-Kubevirt cluster: CP hosted in management cluster, Nodes from Kubevirt
- Kamaji-Metal3 cluster: CP hosted in management cluster, Nodes from Metal3

Architecture 1 – Virtualized workload cluster

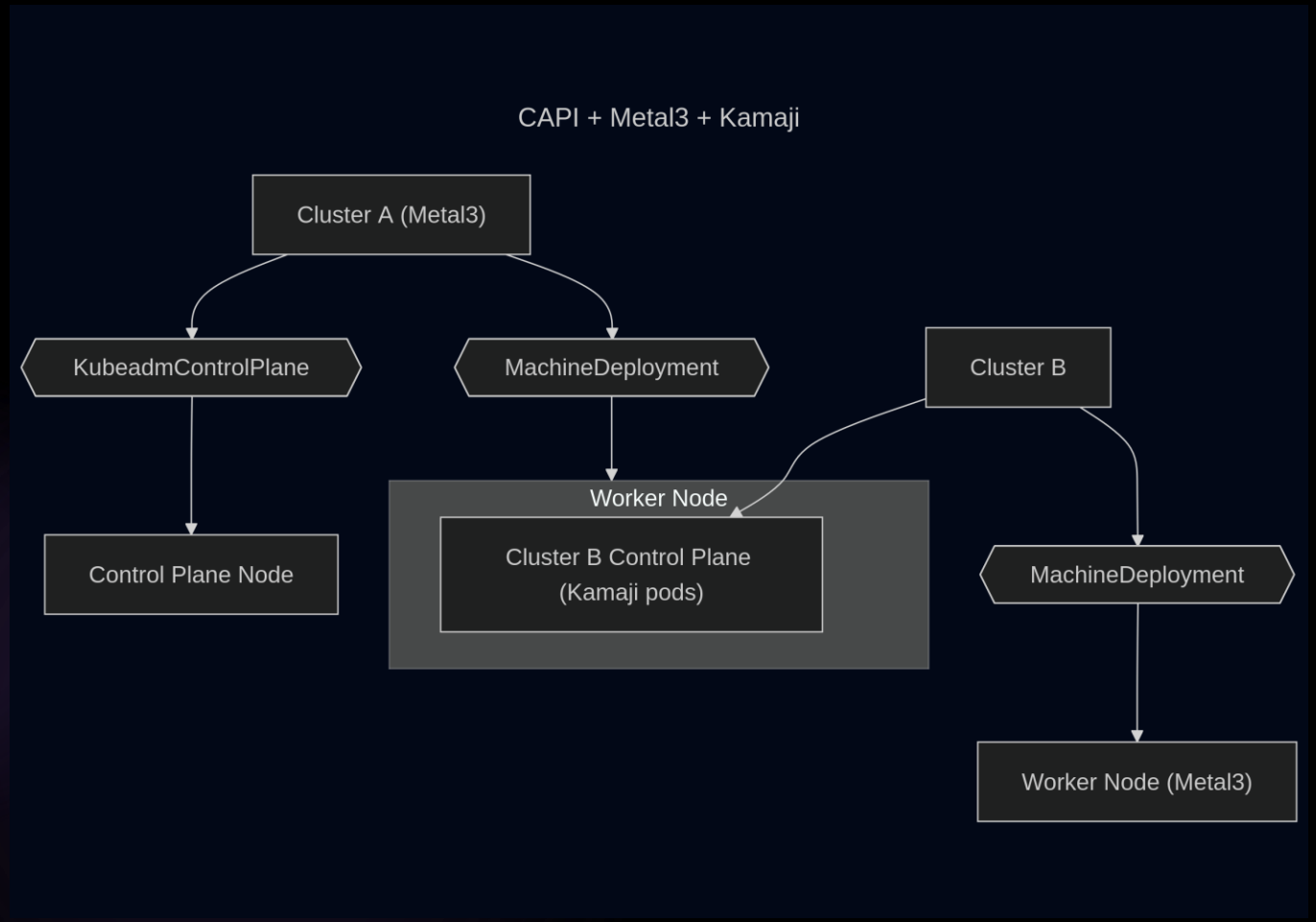


- Baremetal management cluster
- Kubevirt cluster
- Kubevirt nodes runs inside Metal3 workers



Architecture 2 – mixed cluster

- Baremetal management cluster
- Kubevirt MachineDeployment
- Kubevirt worker runs inside Metal3 worker



Architecture 3 – Kamaji-<Infra>

- Baremetal management cluster
- Workload cluster with hosted CP
- Metal3/Kubevirt MachineDeployment

Thank you!
Questions?

Links

- https://kubevirt.io/
- https://metal3.io/
- https://kamaji.clastix.io/
- https://cluster-api.sigs.k8s.io/
- Considerations for large clusters
- Demo with Kamaji + Metal3: https://asciinema.org/a/734801
- Demo with Kubevirt mixed cluster: https://asciinema.org/a/761208