AI at Scale: HPC Primitives in the Cloud-Native Era

Ricardo Rocha
CERN IT, CNCF TOC, CNCF End User TAB

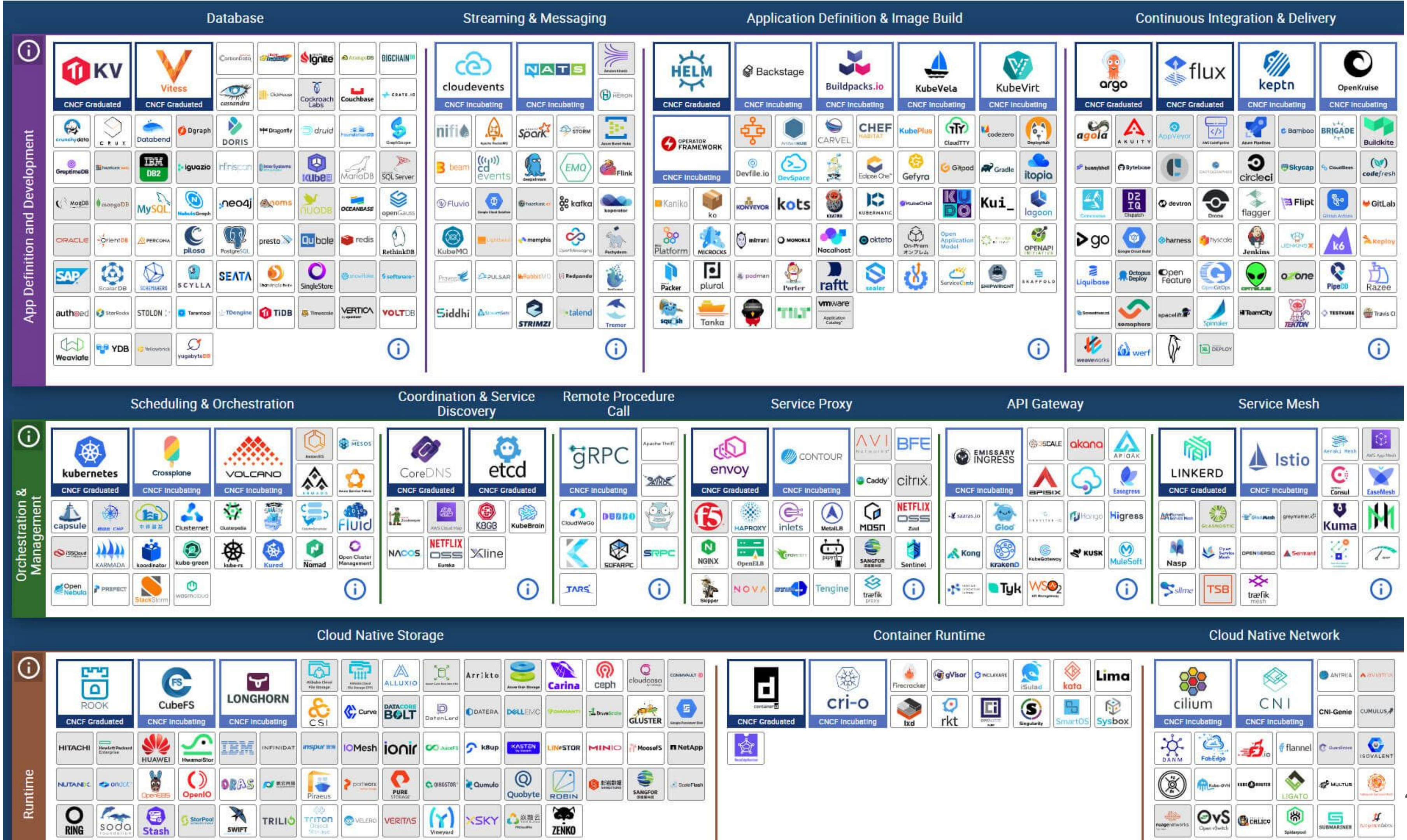
1. Mission of the Cloud Native Computing Foundation.

The Foundation's mission is to make cloud native computing ubiquitous. The CNCF Cloud Native Definition v1.0 says:

Cloud native technologies empower organizations to build and run scalable applications in modern, dynamic environments such as public, private, and hybrid clouds. Containers, service meshes, microservices, immutable infrastructure, and declarative APIs exemplify this approach.

These techniques enable loosely coupled systems that are resilient, manageable, and observable. Combined with robust automation, they allow engineers to make high-impact changes frequently and predictably with minimal toil.

The Cloud Native Computing Foundation seeks to drive adoption of this paradigm by fostering and sustaining an ecosystem of open source, vendor-neutral projects. We democratize state-of-the-art patterns to make these innovations accessible for everyone.


Database
Streaming & Messaging
Application Definition & Image Build
Continuous Integration & Delivery
App Definition and Development
KV
CNCF Graduated
Vitess
CNCF Graduated
cloudevents
CNCF Incubating
NATS
CNCF Incubating
HELM
CNCF Graduated
Backstage
CNCF Incubating
Buildpacks.io
CNCF Incubating
KubeVela
CNCF Incubating
KubeVirt
CNCF Incubating
argo
CNCF Graduated
flux
CNCF Graduated
keptn
CNCF Incubating
OpenKruise
CNCF Incubating
OPERATOR FRAMEWORK
CNCF Incubating
Scheduling & Orchestration
Coordination & Service Discovery
Remote Procedure Call
Service Proxy
API Gateway
Service Mesh
Orchestration & Management
kubernetes
CNCF Graduated
Crossplane
CNCF Incubating
VOLCANO
CNCF Incubating
CoreDNS
CNCF Graduated
etcd
CNCF Graduated
gRPC
CNCF Incubating
envoy
CNCF Graduated
CONTOUR
CNCF Incubating
EMISSARY INGRESS
CNCF Incubating
LINKERD
CNCF Graduated
Istio
CNCF Incubating
Cloud Native Storage
Container Runtime
Cloud Native Network
Runtime
ROOK
CNCF Graduated
CubeFS
CNCF Incubating
LONGHORN
CNCF Incubating
containerd
CNCF Graduated
cri-o
CNCF Incubating
cilium
CNCF Incubating
CNI
CNCF Incubating

4

Immediate popularity and traction

For traditional, stateless services (micro services)

Later on also for stateful services (databases, in memory, …)

Jobs existed as a concept but not as we would expect them

Immediate popularity and traction

For traditional

Later on also

Jobs existed a

```
apiVersion: batch/v1
kind: Job
metadata:
  name: job-backoff-limit-per-index-example
spec:
  completions: 10
  parallelism: 3
  completionMode: Indexed  # required for the feature
  backoffLimitPerIndex: 1  # maximal number of failures per index
  maxFailedIndexes: 5      # maximal number of failed indexes before terminating the Job execution
  template:
    spec:
      restartPolicy: Never # required for the feature
      containers:
      - name: example
        image: python
        command:           # The jobs fails as there is at least one failed index
                           # (all even indexes fail in here), yet all indexes
                           # are executed as maxFailedIndexes is not exceeded.
        - python3
        - -c
        - |
          import os, sys
          print("Hello world")
          if int(os.environ.get("JOB_COMPLETION_INDEX")) % 2 == 0:
            sys.exit(1)
```

Immediate popularity and traction

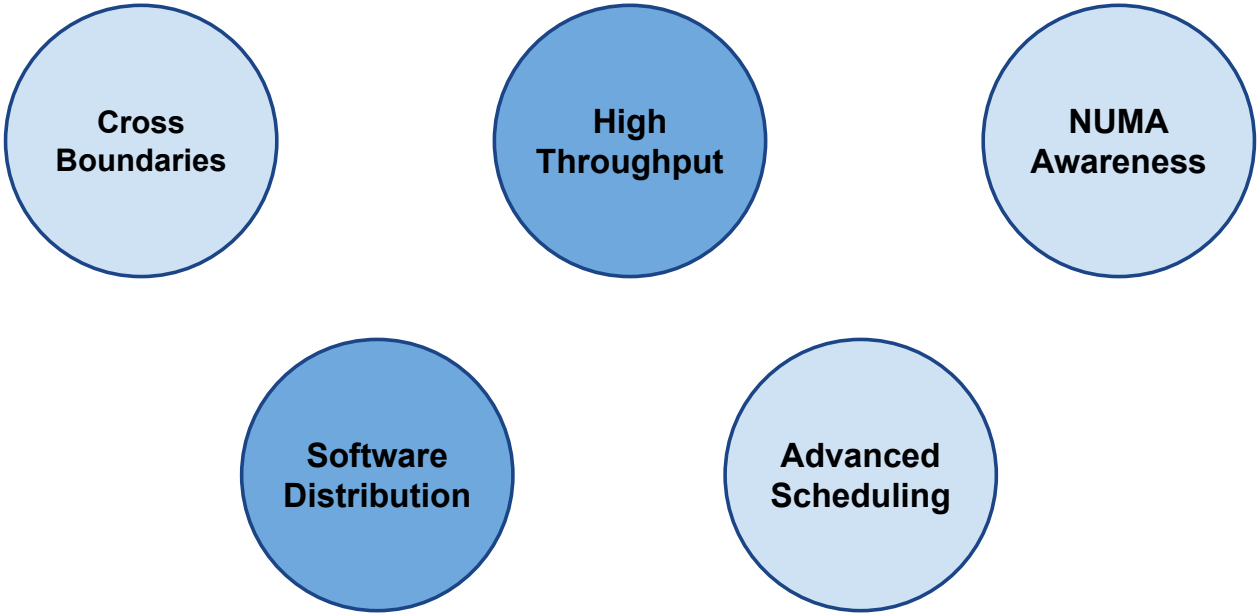
For traditional, stateless services (micro services)

Later on also for stateful services (databases, memory, …)

Jobs existed as a concept but not as we would expect them

And lacking concepts for queues, quotas, gang scheduling, fair sharing, …

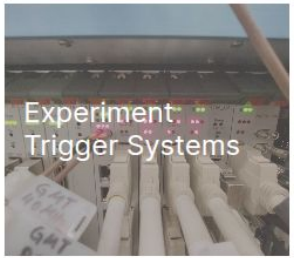
Meet GenAI!


Cross Boundaries
High Throughput
NUMA Awareness
Software Distribution
Advanced Scheduling

Where are we today?

Next Generation Triggers

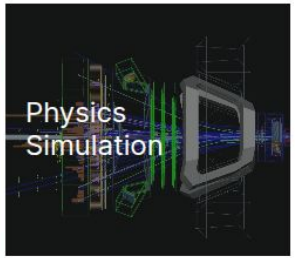
Our research



ATLAS and CMS experts in NextGen develop new workflows and data processing techniques to increase the sensitivity of future particle physics triggers

NextGen develops and benchmarks software and techniques to exploit accelerated computing architectures keeping in mind cost and energy efficiency

NextGen investigates the use of AI technologies to improve the experiments' physics impact leveraging massive data throughput pipelines and large-scale models

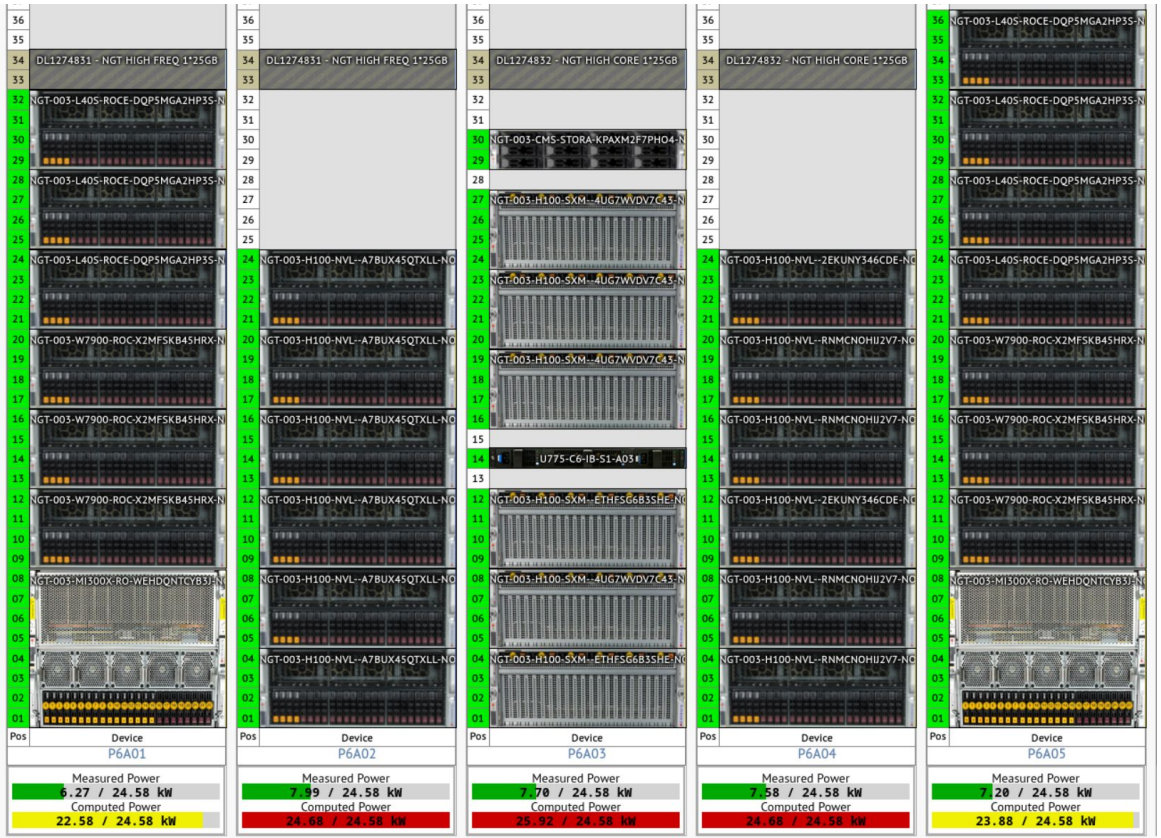


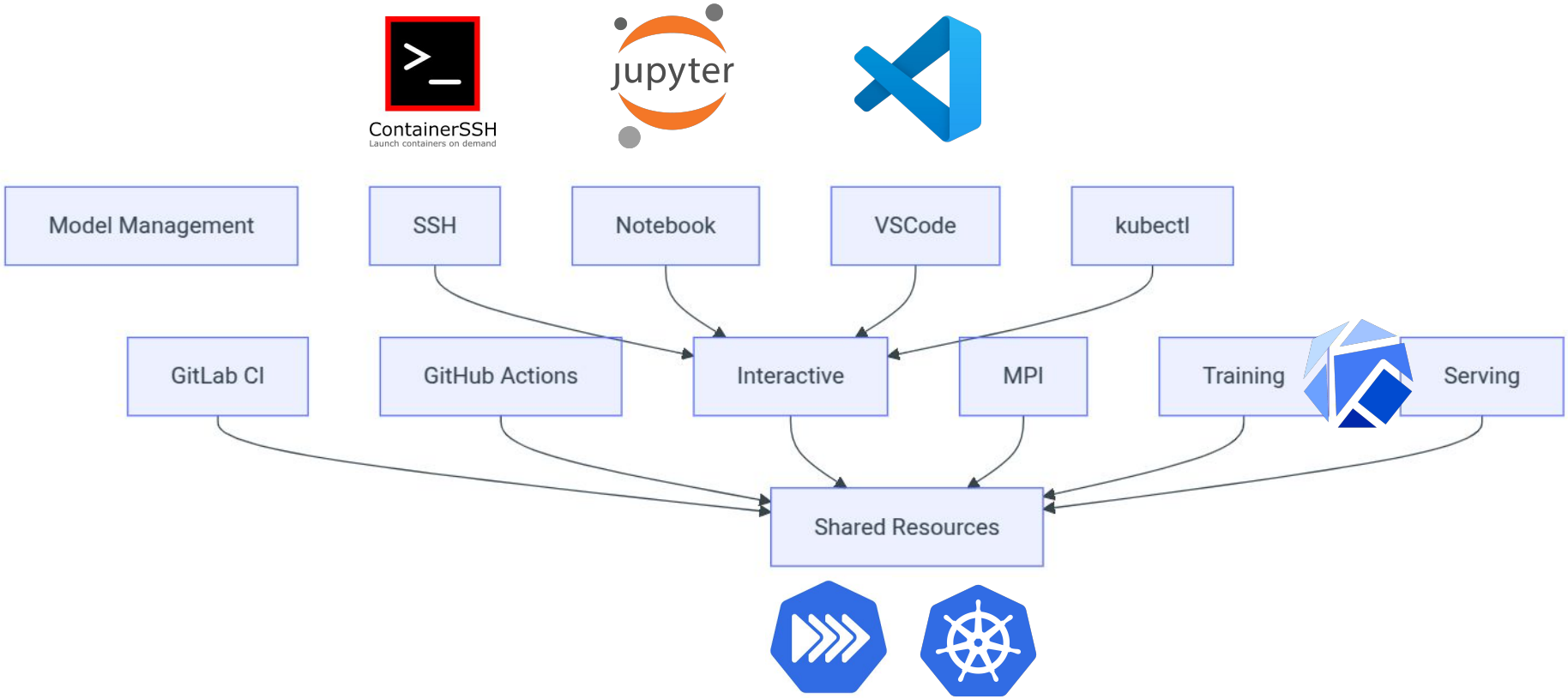
Theoretical physicists and software engineers in NextGen look at new event generators techniques to improve simulation and detection of exotic signatures

NextGen works with the High-Energy Physics community and computer science experts to provide new generations of scientists with the right skills for tomorrow's challenges

https://nextgentriggers.web.cern.ch/

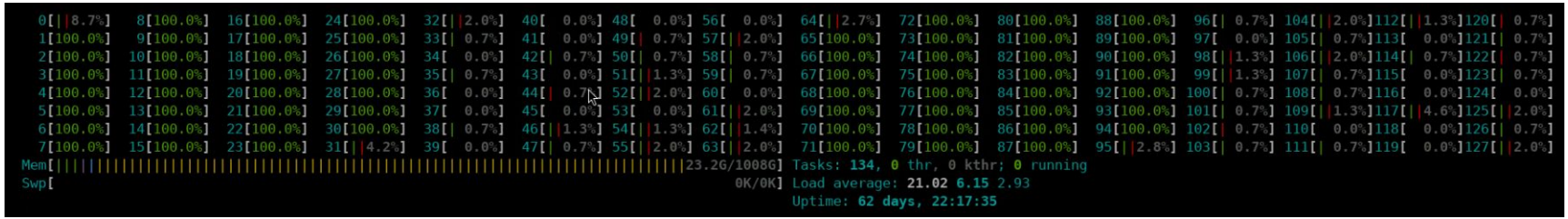
Description	Nodes	Resources	Cores	RAM	Network
Nvidia H100 188GB NVL NVLink	12	8x GPUs	192	1.6TB	100G
Nvidia H100 80GB SXM w/IB	6	4x GPUs	64	1.6TB	100G + 4x400G IB
Nvidia L40S 48GB w/ RoCEv2	7	4x GPUs	128	768GB	200G
AMD MI300X HBM3 192GB w/RoCEv2	2	8x GPUs	128	768GB	200G
AMD Radeon PRO W7900 48GB w/RoCEv2	6	4x GPUs	128	768GB	200G




ContainerSSH
Launch containers on demand
jupyter
Model Management
SSH
Notebook
VSCode
kubectl
GitLab CI
GitHub Actions
Interactive
MPI
Training
Serving
Shared Resources

AI/HPC workloads on Kubernetes

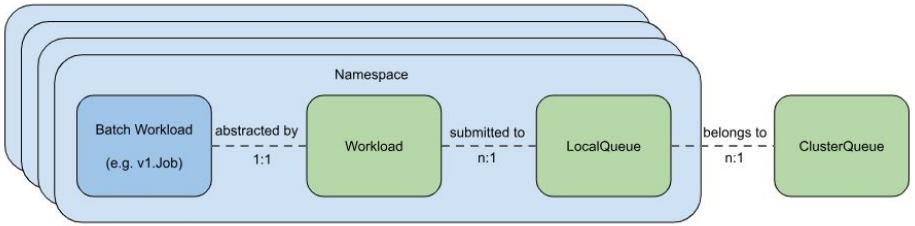
CPU Pinning, NUMA awareness

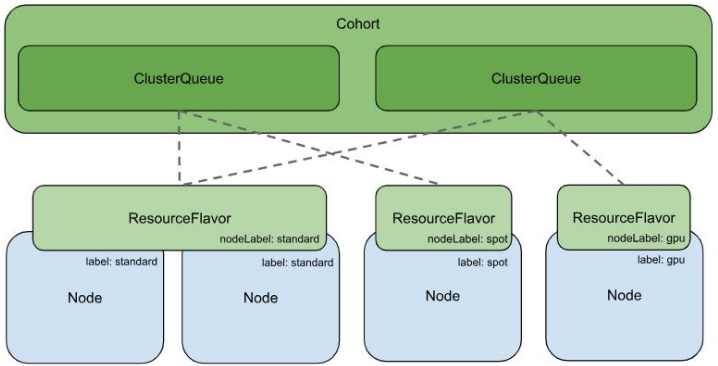


AI/HPC workloads on Kubernetes

CPU Pinning, NUMA awareness

Kueue as a built-in component for advanced scheduling primitives



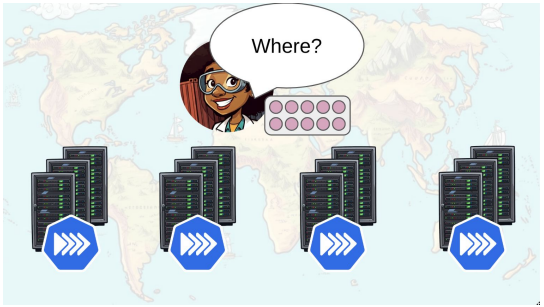


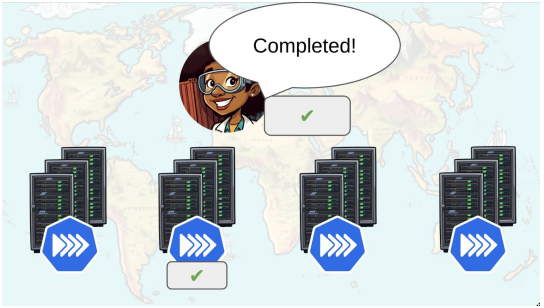
AI/HPC workloads on Kubernetes

CPU Pinning, NUMA awareness

Kueue as a built-in component for advanced scheduling primitives

Support for multi-cluster, multiple administrative domains





AI/HPC workloads on Kubernetes

CPU Pinning, NUMA awareness

Kueue as a built-in component for advanced scheduling primitives

Support for multi-cluster, multiple administrative domains

Support for Infiniband, RoCEv2

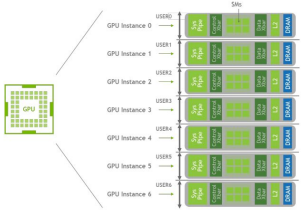
What’s coming next?

Dynamic Resource Allocation

Previous to DRA resources are made available in a static way

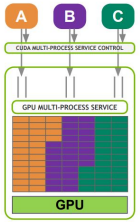
Strong demand for a more dynamic resource allocation (partitioning, sharing)

Multi-Instance GPU (MIG)



Physical partitioning
Full isolation between workloads

Multi-Process Service (MPS)



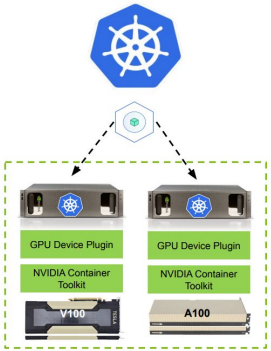
Logical partitioning
Limited isolation between workloads[1]

...

```
apiVersion: v1
kind: Pod
metadata:
  name: gpu-example
spec:
  containers:
    - name: gpu-example
      image: nvidia/cuda
      resources:
        limits:
          nvidia.com/gpu: 1
```

App

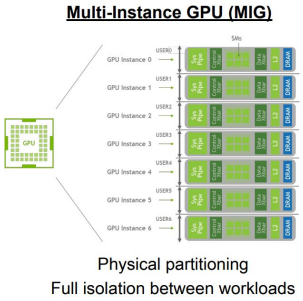
GPU



Dynamic Resource Allocation

Resources are made available in a static way

Strong demand for a more dynamic resource allocation (partitioning, sharing)

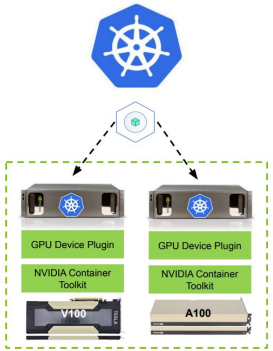



```yaml
apiVersion: v1
kind: Pod
metadata:
  name: gpu-example
spec:
  containers:
  - name: ctr
    image: nvidia/cuda
    resources:
      limits:
        nvidia.com/gpu: 2
```

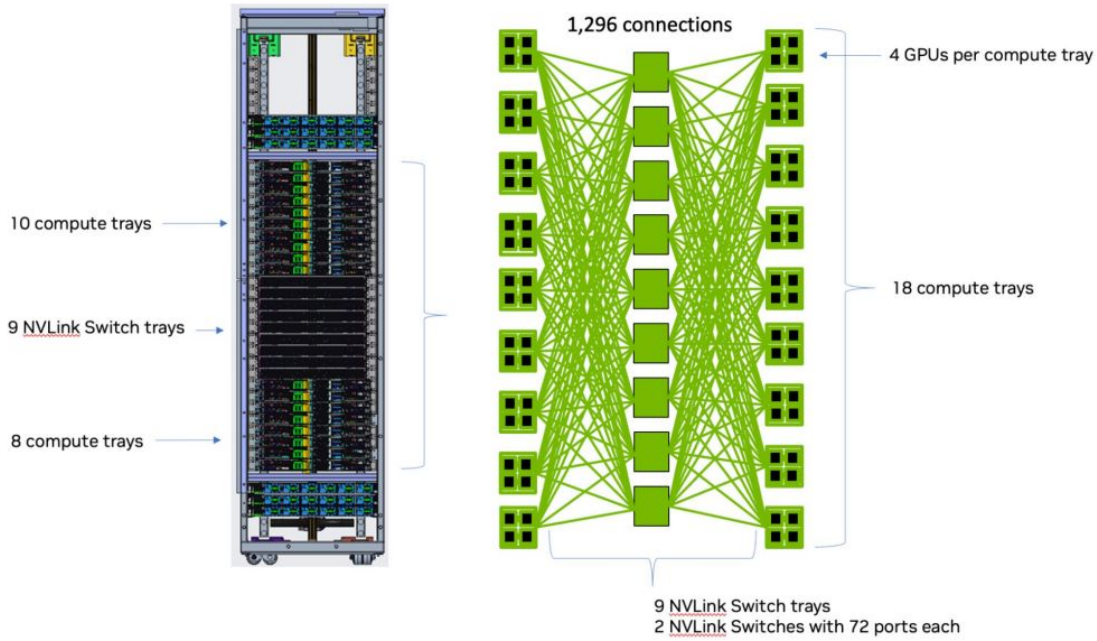

```yaml
apiVersion: resource.k8s.io/v1alpha3
kind: ResourceClaimTemplate
metadata:
  name: unique-gpu
spec:
  spec:
    devices:
      requests:
      - name: gpu
        deviceClassName: gpu.nvidia.com
---
apiVersion: v1
kind: Pod
metadata:
  name: gpu-example
spec:
  containers:
  - name: ctr
    resources:
      claims:
      - name: gpu0
      - name: gpu1
  resourceClaims:
  - name: gpu0
    resourceClaimTemplateName: unique-gpu
  - name: gpu1
    resourceClaimTemplateName: unique-gpu
```

Defines the "template" for creating a ResourceClaim

Associated with the DRA Driver and installed by the cluster admin



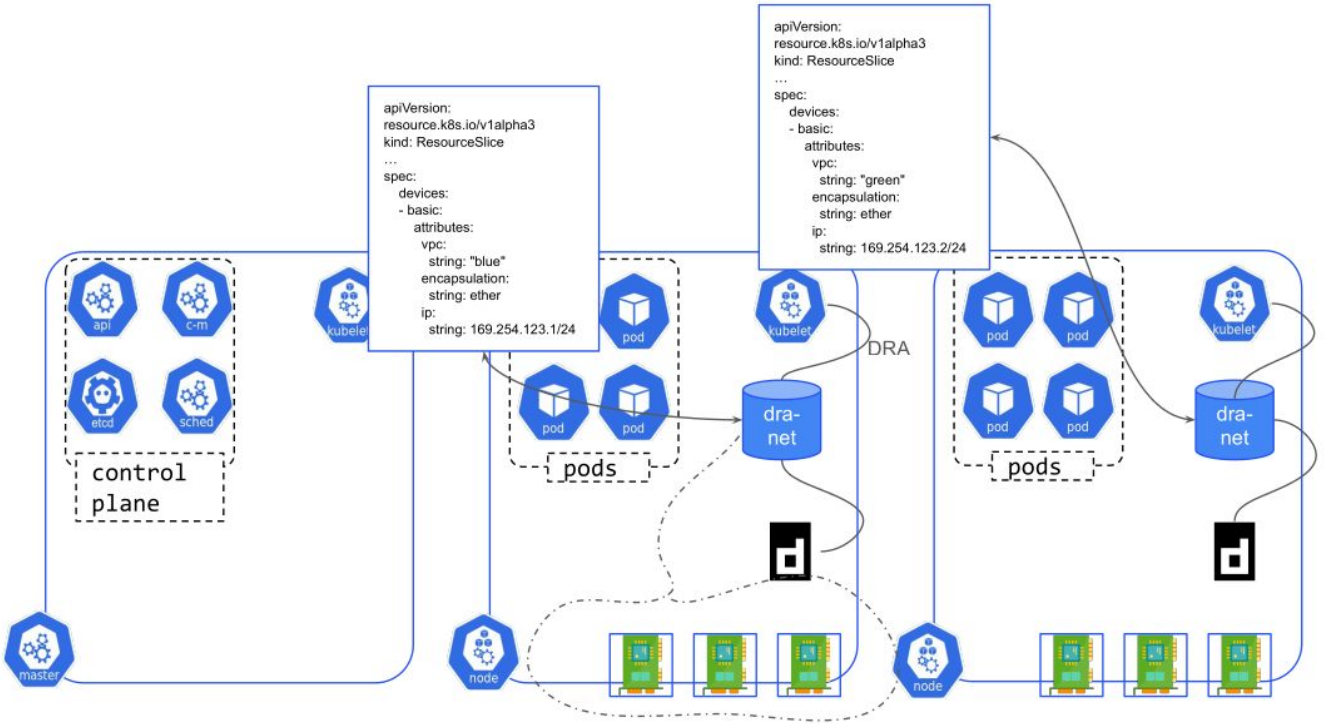
NVIDIA GB200 NVL72




```yaml
apiVersion: resource.nvidia.com/v1beta1
kind: ComputeDomain
metadata:
  name: compute-domain
spec:
  numNodes: 18
  channel:
    resourceClaimTemplate:
      name: compute-domain-channel
```

```yaml
apiVersion: apps/v1
kind: Deployment
metadata:
  name: test-workload
spec:
  replicas: 18
  template:
    spec:
      containers:
        resources:
          limits:
            nvidia.com/gpu: 4
          claims:
          - name: channel
      resourceClaims:
      - name: channel
        resourceClaimTemplateName: compute-domain-channel
```

Dynamic Resource Allocation… networking



Q & A