CERN openlab

AI and Kubernetes @ CERN - Challenges

Antonio Nappi

Ricardo Rocha

CERN openlab

Speakers

- Platform engineer, responsible for hosting infrastructure of CERN Java applications
- Part of CTO openlab team

- Lead, Platforms Infrastructure
- CNCF Technical Oversight Committee (TOC) + Technical Advisory Board (TAB)

What is Kubernetes and CNCF

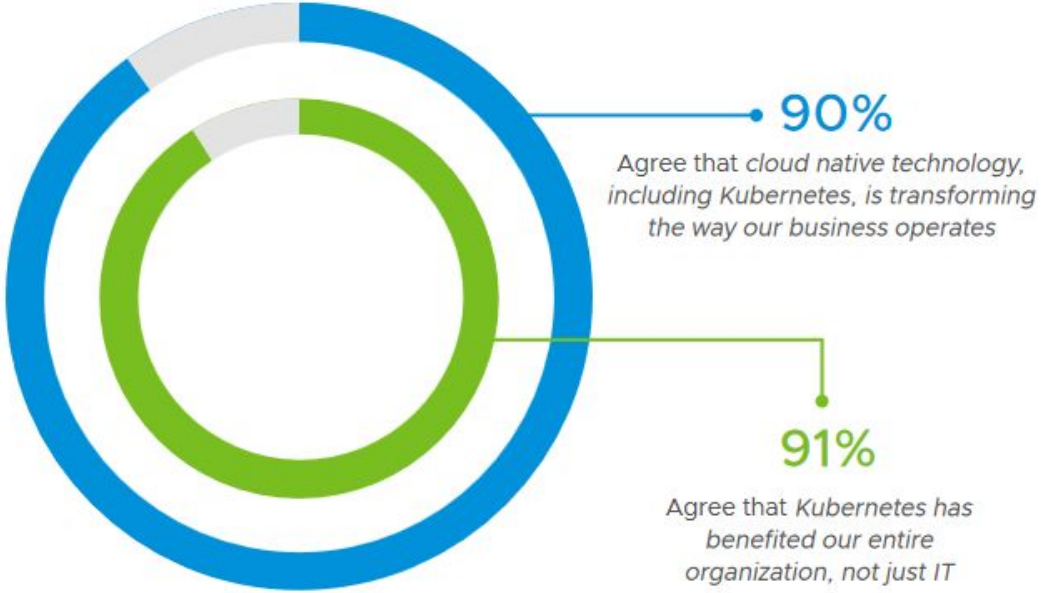
Kubernetes: *is an open-source orchestration system for automating deployment, scaling, and management of containerized applications*

Cloud Native Computing Foundation (CNCF): *is the open source, vendor-neutral hub of cloud native computing, hosting projects like Kubernetes and Prometheus to make cloud native universal and sustainable.*

Kubernetes history

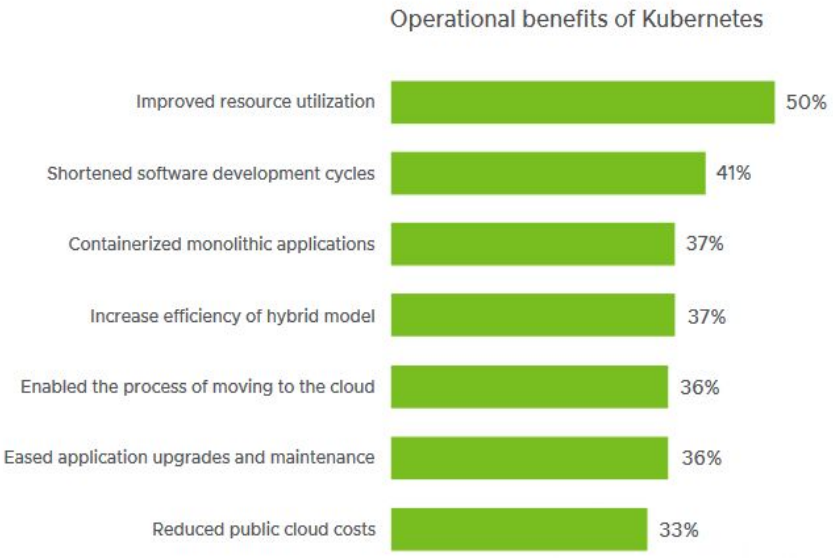
Kubernetes turns 10 in 2024

A decade of Kubernetes: impact



State of Kubernetes 2023, VMWare Tanzu

A decade of Kubernetes: benefits



State of Kubernetes 2023, VMWare Tanzu

Kubernetes @ CERN

Clusters	Nodes
528	3101
Cores	**RAM**
17117	40.5 TB

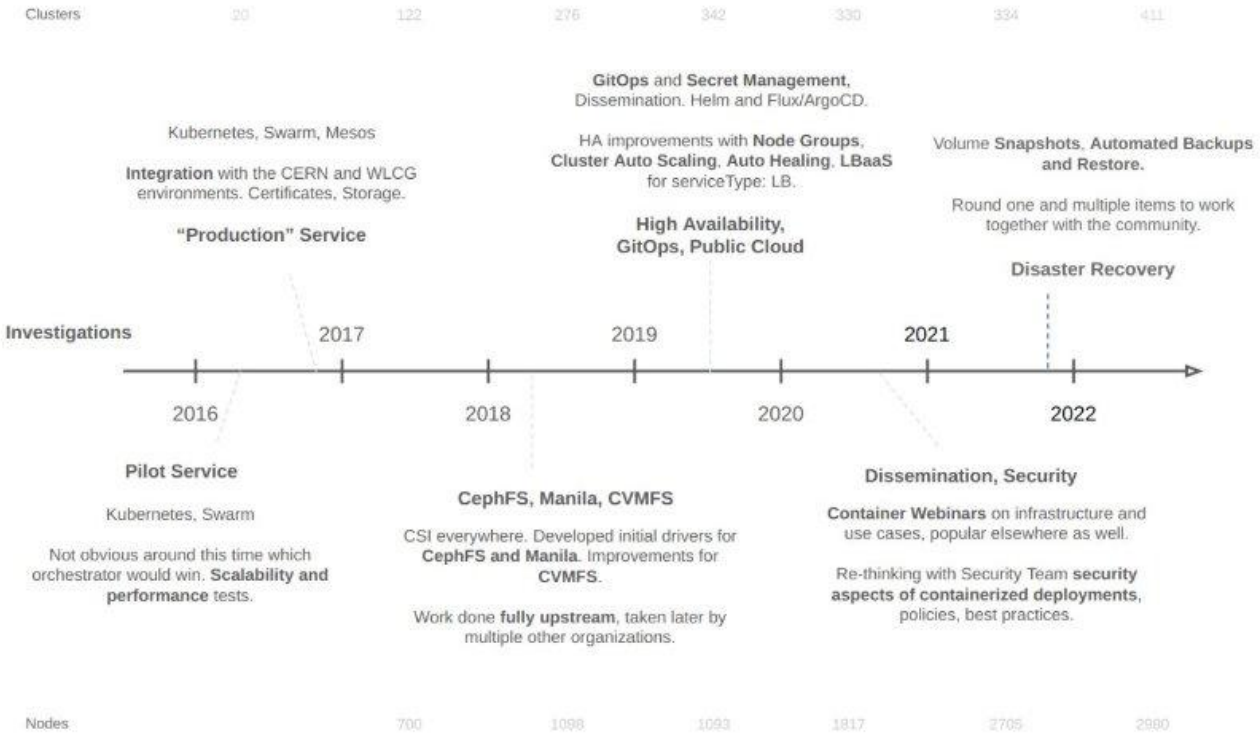
Service launched in 2016

CERN is a CNCF End User since 2020

Large and growing number of services on the platform

- EDH, Phonebook, EDMS, SSO infrastructure
- CERN Library, InspireHEP, HEPData
- GitLab + GitLab CI
- SWAN, Kubeflow/ML, REANA (Reproducible Analysis)
- CMSWeb, Rucio
- ATS / Accelerator Controls (ongoing work)
- And many more

Kubernetes @ CERN



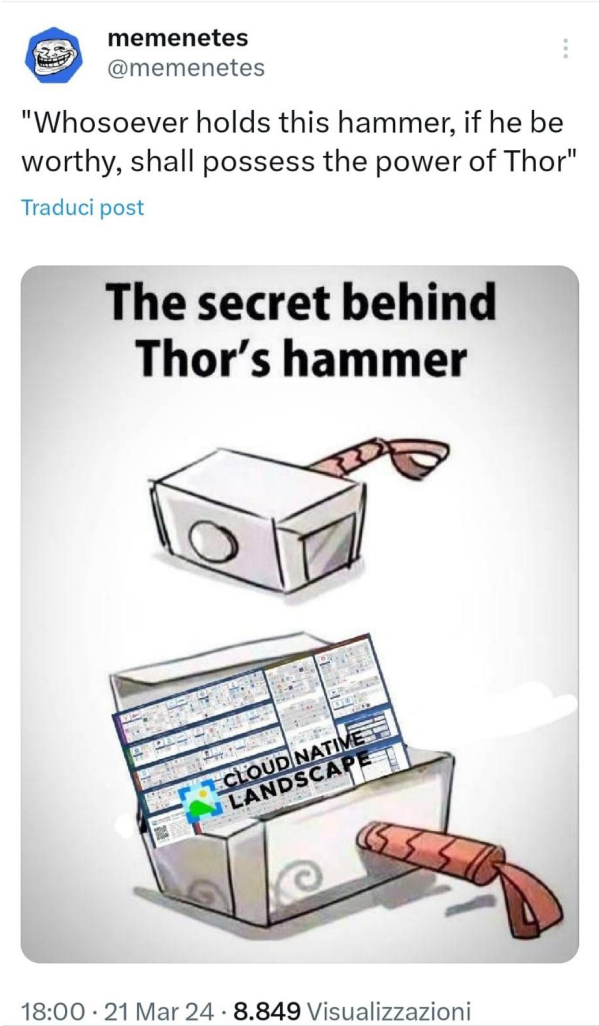
CNCF Top End User award

Challenges

Kubernetes has a steep learning curve

- Starting now is ***almost difficult**** as starting in 2016
- **CNCF landscape** is large and hard to navigate
- Multiple tools doing similar things. Need for a better way to identify best tool for each use case

* Kubernetes is a much more stable and mature product than 2016



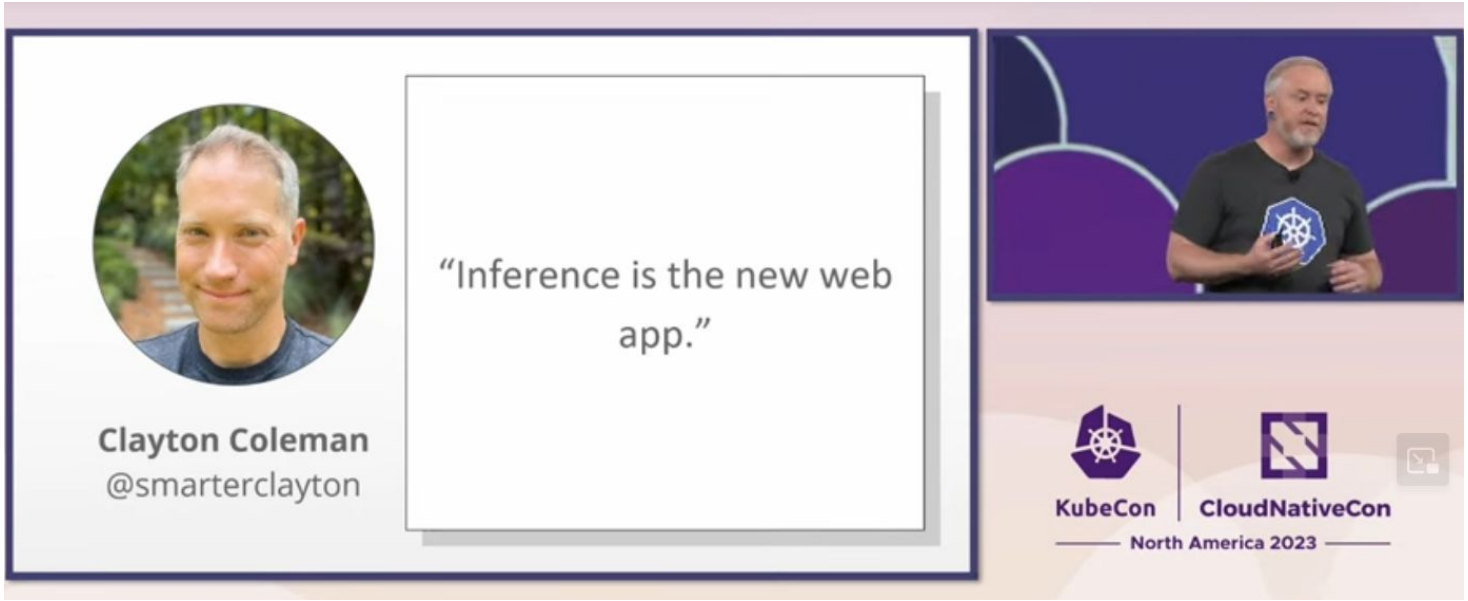
Challenges

Born for stateless web application but running anything

Stateful workloads are possible but not easy as stateless

AI at door, the challenge is to understand how Kubernetes can contribute

What in next decade ?

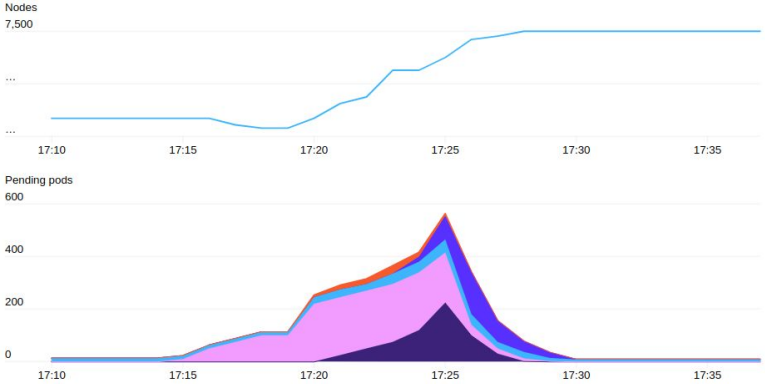


https://openai.com/research/scaling-kubernetes-to-7500-nodes

January 25, 2021 Compute, Software engineering, Conclusion

Research

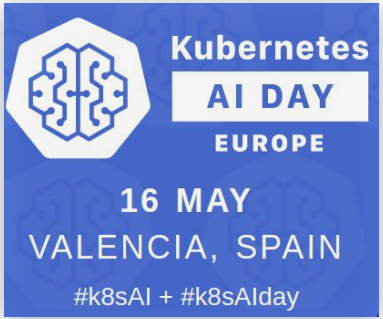
Scaling Kubernetes to 7,500 nodes



A large machine learning job spans many nodes and runs most efficiently when it has access to all of the hardware resources on each node. This allows GPUs to cross-communicate directly using NVLink, or GPUs to directly communicate with the NIC using GPUDirect. So for many of our workloads, a single pod occupies the entire node. Any NUMA, CPU, or PCIE resource contention aren't factors for scheduling. Bin-packing or fragmentation is not a common problem. Our current clusters have full bisection bandwidth, so we also don't make any rack or network topology considerations. All of this means that, while we have many nodes, there's relatively low strain on the scheduler.

We've scaled Kubernetes clusters to 7,500 nodes, producing a scalable infrastructure for large models like GPT-3, CLIP, and DALL·E, but also for rapid small-scale iterative research such as Scaling Laws for Neural Language Models.

2022



2023

https://events.linuxfoundation.org/archive/**2023**/kubecon-cloudnativecon-europe/program/schedule/

No results found for **llm**

Try searching again or browsing the types and venues below.

Search

18 APRIL 2023 | 13:30-17:00

RAI, AMSTERDAM, THE NETHERLANDS

#k8sBatch #k8sHPC

Hall 7 | Room A

13:29 CEST	Kubernetes Batch + HPC Day Hosted by CNCF - Half Day Event | SOLD OUT HALL 7, ROOM A | GROUND FLOOR | EUROPE COMPLEX
13:30 CEST	Welcome + Opening Remarks, Program Committee Member - Ricardo Rocha, CERN HALL 7, ROOM A | GROUND FLOOR | EUROPE COMPLEX
13:45 CEST	Sharing is Caring – Fractional GPU Allocations With MetaGPU Device Plugin - Dmitry Kartsev, cnvrg.io HALL 7, ROOM A | GROUND FLOOR | EUROPE COMPLEX
14:15 CEST	Building a Batch System for the Cloud with Kueue - Aldo Culquicondor, Google & Kante Yin, DaoCloud HALL 7, ROOM A | GROUND FLOOR | EUROPE COMPLEX
14:45 CEST	Making the Most Out of Your Hardware Accelerators in a Kubernetes Cluster - Rishit Dagli, University of Toronto & Shivay Lamba HALL 7, ROOM A | GROUND FLOOR | EUROPE COMPLEX
15:20 CEST	Kubernetes Batch Processing at Scale - A Scheduling Perspective - Lim Haw Jia & Fan Deliang, Bytedance HALL 7, ROOM A | GROUND FLOOR | EUROPE COMPLEX
15:50 CEST	SLA Aware Batch Scheduling in Apache YuniKorn with Multi-Tenant Preemption - Sunil Govindan & Craig Condit, Cloudera HALL 7, ROOM A | GROUND FLOOR | EUROPE COMPLEX
16:20 CEST	⚡ Lightning Talk: How to Bring Data Locality to I/O-Intensive Workloads on Kubernetes - Shouwei Chen, Alluxio HALL 7, ROOM A | GROUND FLOOR | EUROPE COMPLEX
16:30 CEST	⚡ Lightning Talk: Orchestrating Kubernetes Clusters on HPC Infrastructure - Elia Oggian, Swiss National Supercomputing Centre HALL 7, ROOM A | GROUND FLOOR | EUROPE COMPLEX
16:40 CEST	⚡ Lightning Talk: Does Cloud Elasticity Pay off for HPC Workloads? - Joris Cramwinckel, Ortec Finance HALL 7, ROOM A | GROUND FLOOR | EUROPE COMPLEX
16:45 CEST	Closing Remarks, Program Committee Member - Aldo Culquicondor, Google HALL 7, ROOM A | GROUND FLOOR | EUROPE COMPLEX

2024 https://events.linuxfoundation.org/kubecon-cloudnativecon-europe/

llm | Search | Clear

Tuesday, March 19

- 16:10 CET — WasmEdge, portable and lightweight runtime for AI/LLM workloads | Project Lightning Talk
- 18:05 CET — ⚡ Lightning Talk: Locking the Monster: Strategies to Isolate Resource Big Eaters - Peter Pan, DaoCloud (Description: llm)

Wednesday, March 20

- 09:40 CET — Sponsored Keynote: Build an Open Source Platform for AI/ML - Jorge Palma, Principal PM Lead, Microsoft (Description: llm)
- 10:00 CET — Keynote: The Cloud Native News Show: AI Breakthroughs Revealed - Nikhita Raghunath & Rajas Kakodar, VMware by Broadcom; Patrick Ohly & Cathy Zhang, Intel (Description: llm)
- 11:15 CET — AI HUB | 🤖 Welcome + Keynote: Platform Engineering Foundations for AI Innovation (Description: llm)
- Gen AI at the Edge: How Cloud Native Technologies Enable the Next Wave of Intelligent Applications - Kevin Wang, Huawei; Tina Tsou, LF Edge; Yin Ding, Google; Hongbing Zhang, DaoCloud (Description: llm)
- 11:55 CET — AI HUB | 🤖 Unconference Pitches + Talk Selections (Description: llm)
- 12:10 CET — Accelerating Kubernetes Data Intensive APPs with Cloud Native Local Storage - Simon YN Zhao & Zhou Mingming, DaoCloud (Description: llm)
- Future of Intelligent Cluster Ops: LLM-Azing Kubernetes Controllers - Rajas Kakodkar, VMware & Amine Hilaly, AWS
- 12:25 CET — AI HUB | 🤖 Unconference Session: Managing and Running LLMs & Embedding Models in the Cloud
- AI HUB | 🤖 Unconference Session: My Models are Centralized But My Data Is Not, How Do I Move my Models To by Data in a Cloud Native Friendly Way? (Description: llm)
- 14:30 CET — Navigating the Processing Unit Landscape in Kubernetes for AI Use Cases - Mofi Rahman & Kaslin Fields, Google & Rob Koch, Slalom (Description: llm)

- AI HUB | 🤖 AI Q&A Panel (Description: llm)
- Cloud-Native LLM Deployments Made Easy Using LangChain - Ezequiel Lanza & Arun Gupta, Intel
- Self-Hosted LLMs on Kubernetes: A Practical Guide - Hema Veeradhi & Aakanksha Du Red Hat
- Tutorial: Cloud Native Sustainable LLM Inference in Action - Chen Wang, Eun Kyung Le Wen, IBM; Huamin Chen, Red Hat; Cathy Zhang, Intel
- 15:05 CET — AI HUB | 🤖 Demos (Description: llm)
- 15:25 CET — Strategies for Efficient LLM Deployments in Any Cluster - Angel M De Miguel Meana, VMware & Francisco Cabrera, Microsoft
- Building AI-Ready Platforms - Symphony for Developer and Platform Engineer - Thom Vitale, Systematic & Lize Raes, LangChain4j (Description: llm)

Thursday, March 21

- 11:00 CET — Unleashing the Power of DRA (Dynamic Resource Allocation) for Just-in-Time GPU Sli Abhishek Malvankar & Olivier Tardieu, IBM (Description: llm)
- 14:30 CET — Intelligent Observability: The Foundation for Operating Smarter in the Age of AI - Aloli Sharma, Apple (Description: llm)
- 16:30 CET — From Insanity to Ingenuity: Seven Practical Tips for Navigating the AI Storm in DBaaS Evolution - Lisa-Marie Namphy, Independent; Joseph Sandoval, Adobe; Eddie Wassef, Vonage; Bart Farrell, Learnk8s; Monica Sarbu, Xata.io (Description: llm)
- Confidential Containers for GPU Compute: Incorporating LLMs in a Lift-and-Shift Strategy for AI - Zvonko Kaiser, NVIDIA

Friday, March 22

- 11:00 CET — How to Stabilize a GenAI-First, Modern Data LakeHouse: Provision 20,000 Ephemeral D Lakes/Year - Shirley Yang, LinkedIn (Description: llm)
- Mastering GPU Management in Kubernetes Using the Operator Pattern - Shiva Krishna & Kevin Klues, NVIDIA (Description: llm)
- Kubernetes MLSec: Securing AI in Space - Francesco Beltramini & James Callaghan, ControlPlane (Description: llm)
- 11:55 CET — Cloud Native Batch Computing with Volcano: Updates and Future - William Wang, Huaw Mengxuan Li, 4paradigm (Description: llm)
- Precision Matters: Scheduling GPU Workloads on Kubernetes - Amit Kumar & Gaurav Ku Uber (Description: llm)
- 14:00 CET — Create Cloud Native Agents and Extensions for LLMs - Xiaowei Hu, Second State
- Prompt: Help Me Debug a Cluster! - Anusha Ragunathan & Lili Wan, Intuit Inc (Descriptio llm)
- 16:00 CET — Production-Ready AI Platform on Kubernetes - Yuan Tang, Red Hat (Description: llm)

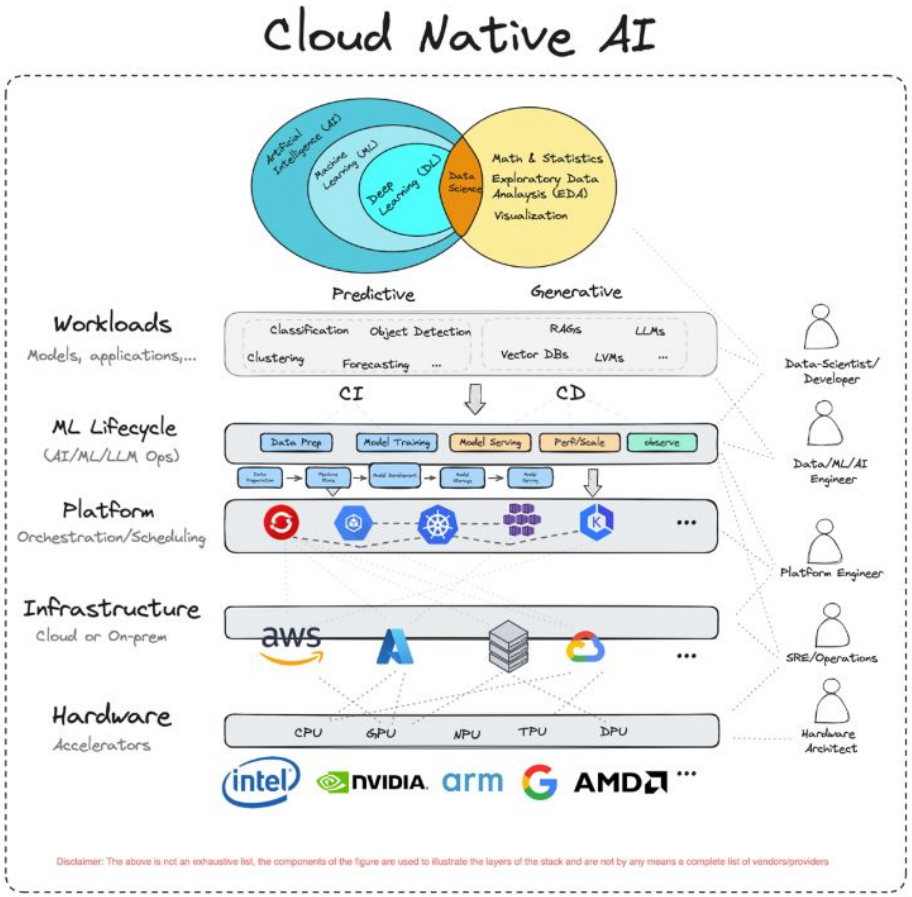
2024 https://events.linuxfoundation.org/kubecon-cloudnativecon-europe/

Type: **Cloud Native AI Day** [Clear Filter]

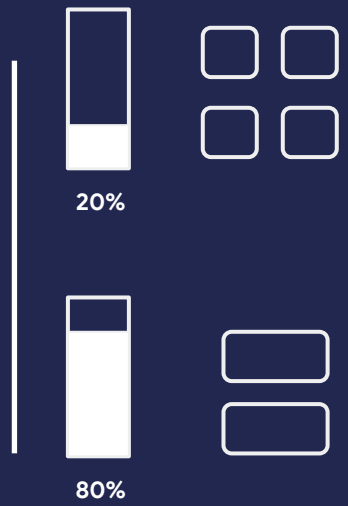
Tuesday, March 19

09:00 CET	Cloud Native AI Day	Welcome + Opening Remarks - Yuan Tang, Red Hat & Rajas Kakokdar, VMware
09:15 CET	Training and Optimisation of Large Transformer Models: An ATLAS and CERN Use Case - Ricardo Rocha, CERN & Maxence Draguet, University of Oxford - ATLAS
09:50 CET	Gen-AI at Scale: Simplifying Orchestration of Healthcare Applications Across Multi-Cluster Environme - Selvi Kadirvel, Elotl & Christopher Nuland, Red Hat
10:25 CET	⚡ Lightning Talk: Best Practices for LLM Serving with DRA - Chen Wang & Abhishek Malvankar, IBM
10:45 CET	Unleashing Kubernetes Intelligence - running k8sgpt utilizing your own fine-tuned LLM - Mario Fahlandt, Kubermatic
11:05 CET	Pods Everywhere! InterLink: A Virtual Kubelet Abstraction Streamlining HPC Resource Exploitation - Diego Ciangottini, INFN
11:50 CET	⚡ Lightning Talk: Cloud Native Networking for AI : Strengthen CNI for RDMA - Weizhou Lan & Junnan Shi, Daocloud
12:05 CET	Panel: Beyond the Clouds: Charting the Course for AI in the CloudNative World - Rajas Kakodkar, VMware; Ricardo Aravena, TruEra; Alolita Sharma, Apple; Madhuri Yechuri, Elotl; Cathy Zhang, Intel
13:30 CET	The Hitchhiker's Guide to Kubernetes Platforms: Don't Panic, Just Launch! - Alexa Griffith & Tessa Pham, Bloomberg
14:05 CET	Efficient Multi-Cluster GPU Workload Management with Karmada and Volcano - Kevin Wang, Huawei
14:40 CET	Make Descheduler Smarter and Safer: How We Apply Reinforcement Learning in Descheduling Strategies - Xuming Wang & Haosong Huang, Shopee
15:15 CET	Effortless Scalability: Orchestrating Large Language Model Inference with Kubernetes - Rohit Ghumare, devrelasservice.com & Joinal Ahmed, Navatech AI
15:50 CET	Resource-Aware Scheduling for Production GenAI with RAG running on Multicluster Cloud Kubernetes - Anne Holler, Elotl & Dave Southwell, Deft Computing
16:25 CET	Building Serverless AI Apps with Spin and WebAssembly - Matt Butcher & Radu Matei, Fermyon
17:00 CET	Scale Your Batch / Big Data / AI Workloads Beyond the Kubernetes Scheduler - Antonin Stefanutti & Anish Asthana, Red Hat
17:25 CET	Closing Remarks - Rajas Kakokdar, VMware

https://www.cncf.io/reports/cloud-native-artificial-intelligence-whitepaper/



Sharing and Efficient Usage of GPU resources



Interactive, Inference
Spiky utilization, day vs night time, immediate access

Batch, Training
Sustained utilization, long running, queues and priorities

Resource Sharing, Partitioning

Scheduling, Priorities, Pre-emption

Online + Offline → **TIDAL CO-LOCATION**

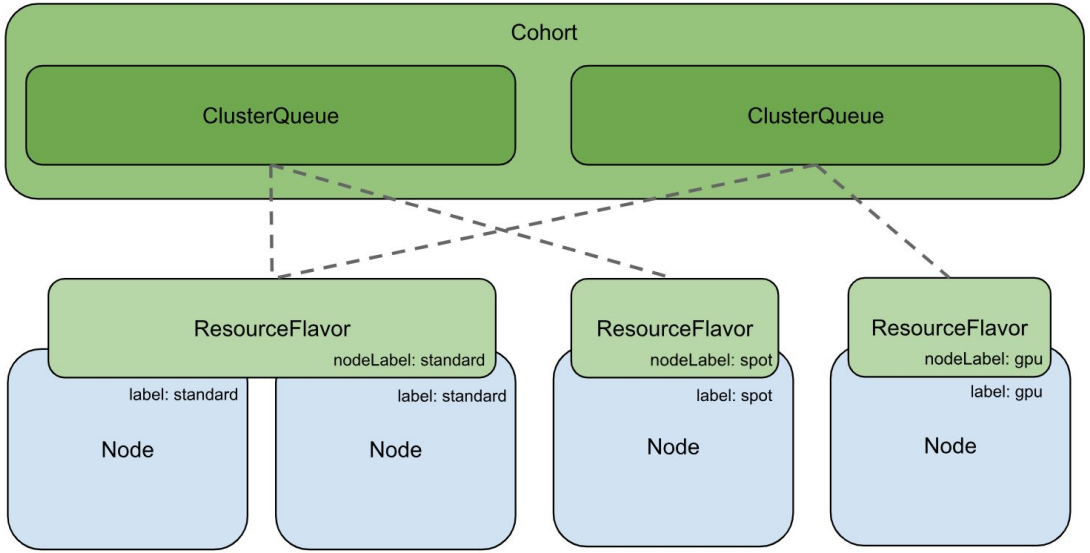
```yaml
apiVersion: v1
kind: Pod
metadata:
  name: gpu-pod
spec:
  restartPolicy: Never
  containers:
    - name: cuda-container
      image: nvcr.io/nvidia/k8s/cuda-sample:vectoradd-cuda10.2
      resources:
        limits:
          nvidia.com/gpu: 1 # requesting 1 GPU
```

```yaml
apiVersion: resource.k8s.io/v1alpha2
kind: ResourceClass
name: resource.example.com
driverName: resource-driver.example.com
---
apiVersion: cats.resource.example.com/v1
kind: ClaimParameters
name: large-black-cat-claim-parameters
spec:
  color: black
  size: large
---
apiVersion: resource.k8s.io/v1alpha2
kind: ResourceClaimTemplate
metadata:
  name: large-black-cat-claim-template
spec:
  spec:
    resourceClassName: resource.example.com
    parametersRef:
      apiGroup: cats.resource.example.com
      kind: ClaimParameters
      name: large-black-cat-claim-parameters
```

```yaml
apiVersion: v1
kind: Pod
metadata:
  name: pod-with-cats
spec:
  containers:
  - name: container0
    image: ubuntu:20.04
    command: ["sleep", "9999"]
    resources:
      claims:
      - name: cat-0
  - name: container1
    image: ubuntu:20.04
    command: ["sleep", "9999"]
    resources:
      claims:
      - name: cat-1
  resourceClaims:
  - name: cat-0
    source:
      resourceClaimTemplateName: large-black-cat-claim-template
  - name: cat-1
    source:
      resourceClaimTemplateName: large-black-cat-claim-template
```

https://kubernetes.io/docs/concepts/scheduling-eviction/dynamic-resource-allocation/

```yaml
apiVersion: batch/v1
kind: Job
metadata:
  name: sample-job
  labels:
    kueue.x-k8s.io/queue-name: user-queue
    kueue.x-k8s.io/priority-class: sample-priority
spec:
...
```



https://kueue.sigs.k8s.io/docs/overview/

KEP-693: MultiKueue #1380

Merged k8s-ci-robot merged 1 commit into kubernetes-sigs:main from mwielgus:mk-kep on Dec 28, 2023

GPU-free LLM Inference

Sustainable Compute
Better Ease of Use and Availability
Proven Performance w/ More Flexibility

Kubernetes + Small Parameter LLM
Robust Open Source Ecosystem
Pragmatic Choice

Cost Efficient
Up to 80+% lower cost per (Million) token
(to GPU)

Slide from Ampere
Kubecon Europe 2024

CERN openlab