KubeCon

CloudNativeCon

THE LINUX FOUNDATION

OPEN SOURCE SUMMIT

AI_dev

Open Source GenAI & ML Summit

China 2024

KubeCon

CloudNativeCon

THE LINUX FOUNDATION

OPEN SOURCE SUMMIT

AI_dev

Open Source GenAI & ML Summit

China 2024

Kubespray Unleashed Navigating Bare Metal Services in Kubernetes for LLM and RAG

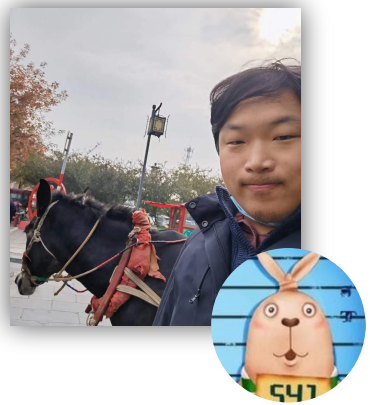
Alan Leung | Equinix

Kay Yan | DaoCloud

About Us

Alan Leung

Digital Technical Specialist
Equinix, Hong Kong



Kay Yan

Principal Engineer
DaoCloud, China
Github ID: yankay
Kubespray, Containerd/Nerdctl Maintainer

Deploy Kubernetes >500 various environment have been delivered in since 2016

1 What is Kubespray?

2 Features & Best Practices

3 RAG

4 QA

KUBESPRAY

What is Kubespray

Kubespray is an open-source project that provides tooling to deploy a Kubernetes cluster using **Ansible**.
It is designed to be a flexible and powerful way to manage **Kubernetes installations**, allowing you to deploy your clusters on various platforms, including **bare-metal environments** and most cloud providers. Kubespray is particularly noted for its ability to handle multi-cloud and on-premise deployments effectively.

kubernetes

ANSIBLE

When Using Kubespray

KubeCon CloudNativeCon China 2024 | THE LINUX FOUNDATION OPEN SOURCE SUMMIT | AI_dev

BootStrap
K8S SIG

kubeadm

Provisioner
K8S SIG

kops

Bare Metal
Production
Easy

Public Cloud

Amazon EKS

KUBERNETES ENGINE

Azure Kubernetes Service (AKS)

Developer
K8S SIG

minikube

kind

Commercial

RED HAT® OPENSHIFT Container Platform

VMware Tanzu

DaoCloud Enterprise

Main Features

- **Flexible** Can be used with **Cloud Environments (`terraform`) or Bare Metal (`ansible`).**
- **Highly availability** cluster(s) e.g control plane and etcd.
- **Configuration Options** (various aspects choice of CRI, CNI ..etc).
- Supports most popular **Linux distributions.**
- **Continuous integration tests.**

Options you can choose!

Cloud	Supported Linux	CRI	CNI	CSI	Others
Equinix	Flatcar Container Linux	Containerd	Calico	cephfs-provisioner	Coredns
Huawei Cloud	Ubuntu 20.04, 22.04, 24.04	Docker *	Cilium	rbd-provisioner	Metallb
AWS	RHEL/Oracle 7, 8, 9	CRI-O	cni-plugins (MacVlan…)	aws-ebs-csi-plugin	Ingress-nginx
Google Cloud	Alma/Rocky 8, 9	Crun	Multus	azure-csi-plugin	Kube-vip
Upcloud	Fedora 37, 38; CoreOS	Gvisor	Flannel	cinder-csi-plugin	Cert-manager
VMWare vSphere	Debian 10,11,12	Kata	Cannel	gcp-pd-csi-plugin	ArgoCD
Openstack	OpenSUSE Leap 15.x/Tumbleweed	Youki	Weave	local-path-provisioner	Registry
Hetzner	Amazon Linux 2		Kube-OVN	local-volume-provisioner	Helm
Nif cloud	Kylin V10; UOS Linux; openEuler		Customize		Node-Feature-Discovery

Life Cycle of cluster operations

- Support full lifecycle of cluster operations
  - New cluster
  - Upgrade cluster or control plane component
  - Scaling a cluster
  - Node management e.g remove/add nodes
  - Reset cluster
  - Configuration management
- Backup and restore
  - etcd snapshots taken during upgrade.

Release Cycle in Kubespray

Kubernetes Support matrix

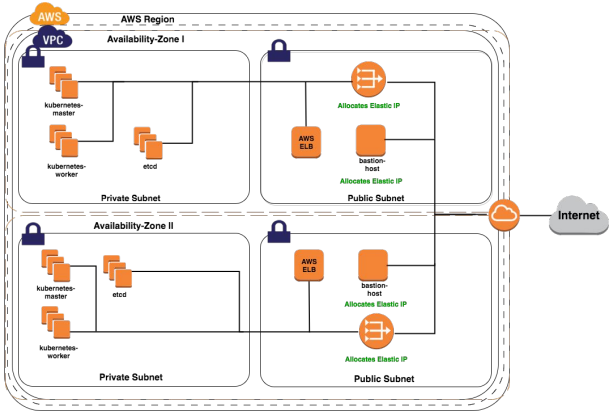
Kubespray	Kube Version (N-2)
master-branch	1.28 ~ **1.30**
2.25.0	1.27 ~ **1.29**
2.24.2	1.26 ~ **1.28**
2.23.3	1.25 ~ **1.27**

Release Cycle: A few weeks releases after the Kubernetes Release

Cloud Environment Deployment

Step1: Create VM/Net with Terraform

```
# Edit terraform config
# Apply with terraform
terraform init
terraform plan -no-color
terraform apply -no-color
```

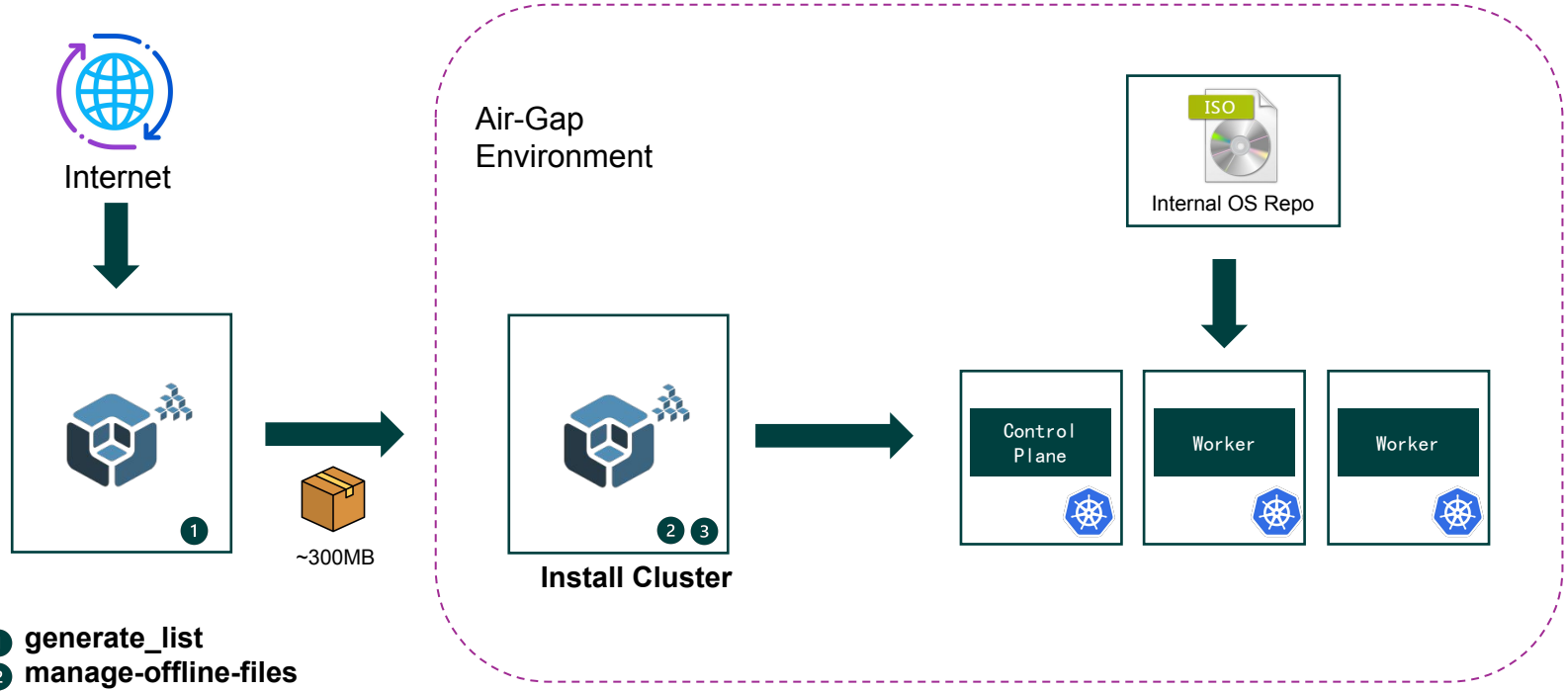


Step2: Install k8s with Ansible

```
# Edit ansible config
# Apply with ansible
ansible-playbook ... cluster.yml
```

AWS
EQUINIX
Google Cloud
EXOSCALE
NIFCLOUD
openstack.
HETZNER
vmware
UpCloud

Air-Gap Environment Deployment



1. **generate_list**
2. **manage-offline-files**
3. **manage-offline-container-images**

Docker-CE is a Special Case!

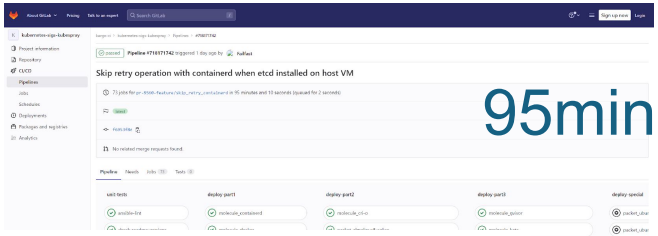
1. Not Support for Kubernetes in the Future.
2. Cray Dependency on Recent Version

Awesome CI Test

40+ test cases support:
- 13 operating systems
- 8 network plugins
- 30+ environments

CI strategies:
- All-in-one
- Separate roles
- HA
- Upgrade

CI Tips:
- Multi stage test cases
- Only run specific when coding
- Full automated



containerd

OS / CNI	calico	cilium	custom_cni	flannel	kube-ovn	kube-router	macvlan	weave
almalinux8	✓	✗	✗	✗	✓	✗	✗	✗
amazon	✓	✗	✗	✗	✗	✗	✗	✗
centos7	✓	✗	✗	✓	✗	✓	✗	✓
debian11	✓	✗	✓	✗	✗	✗	✓	✗
debian12	✓	✓	✓	✗	✗	✗	✗	✗
fedora37	✓	✗	✗	✗	✗	✓	✗	✗
fedora38	✗	✗	✗	✗	✓	✗	✗	✗
opensuse	✗	✗	✗	✗	✗	✗	✗	✗
rockylinux8	✓	✗	✗	✗	✗	✗	✗	✗
rockylinux9	✓	✓	✗	✗	✗	✗	✗	✗
ubuntu20	✓	✓	✗	✓	✗	✓	✗	✓
ubuntu22	✓	✗	✗	✗	✗	✗	✗	✗
ubuntu24	✓	✗	✗	✗	✗	✗	✗	✗

Best Practice: NTP & Sysctl

NTP

Time synchronization is very important for the Etcd and kubernetes.

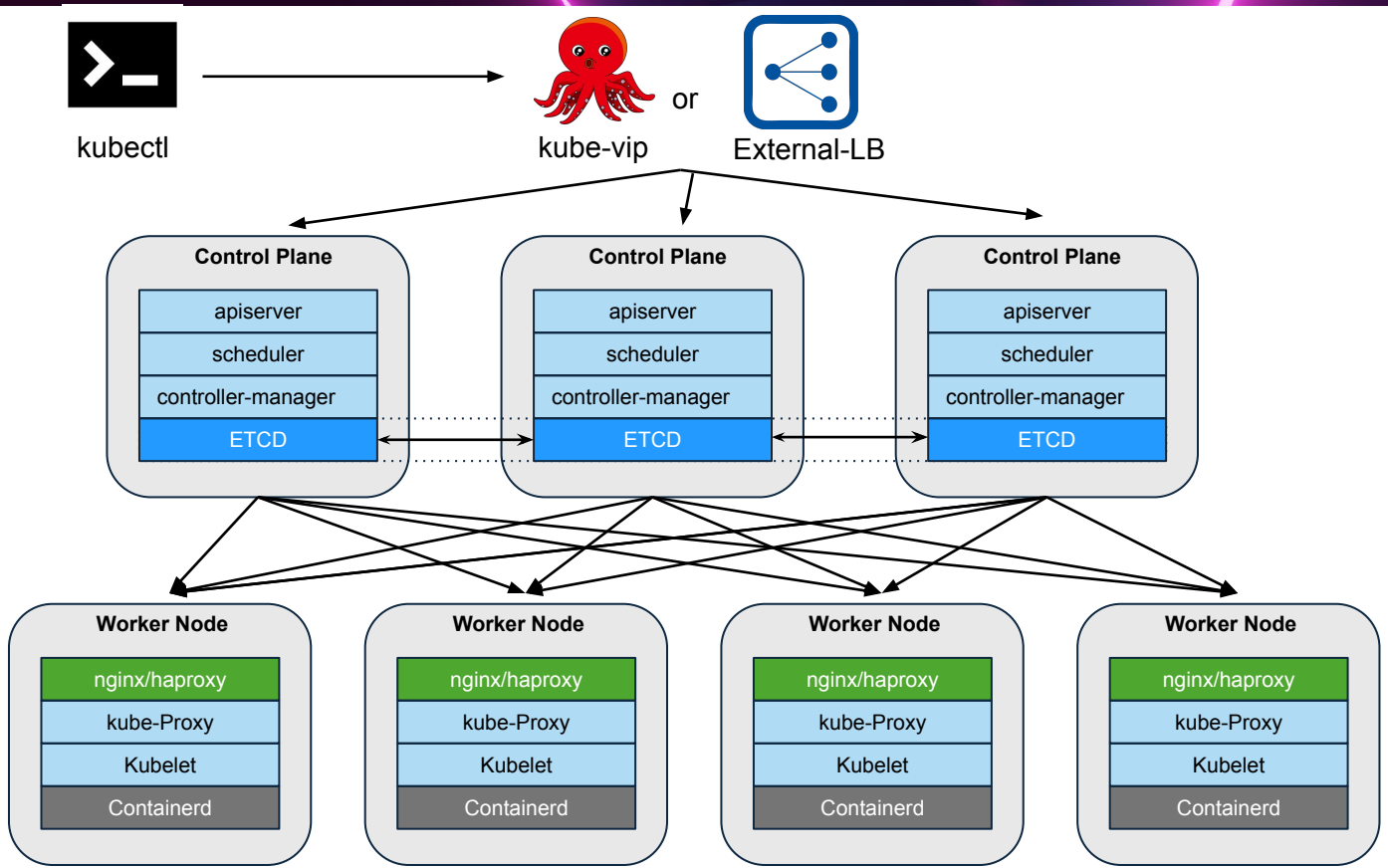
```
ntp_enabled: true
ntp_timezone: Asia/Shanghai
ntp_manage_config: true
ntp_force_sync_immediately: true
```

Sysctl

- Required **kernel variables** has been configured by default.
- Support customize the additional variables for tunning

```
additional_sysctl:
 - { name: kernel.pid_max, value: 4194304 }
 - { name: net.netfilter.nf_conntrack_max, value: 1048576 }
 - { name: fs.inotify.max_user_watches, value: 65536 }
 - { name: fs.inotify.max_user_instances, value: 8192 }
```

Best Practice: High Availability



Kubeadm Cert
Auto Renew Every
Month

Best Practice: Ansible Collection

Kubespray can work well with Ansible Collection.

1.Add Kubespray to your requirements.yml file

```
collections:
- name: https://github.com/kubernetes-sigs/kubespray
  type: git
  version: master # use the appropriate tag or branch for the version you need
```

2.Install your collection

```
ansible-galaxy install -r requirements.yml
```

3.Create a playbook to install your Kubernetes cluster

```
- name: Install Kubernetes
  ansible.builtin.import_playbook: kubernetes_sigs.kubespray.cluster
```

4.Update INVENTORY and PLAYBOOK so that they point to your inventory file and the playbook you created above, and then install Kubespray

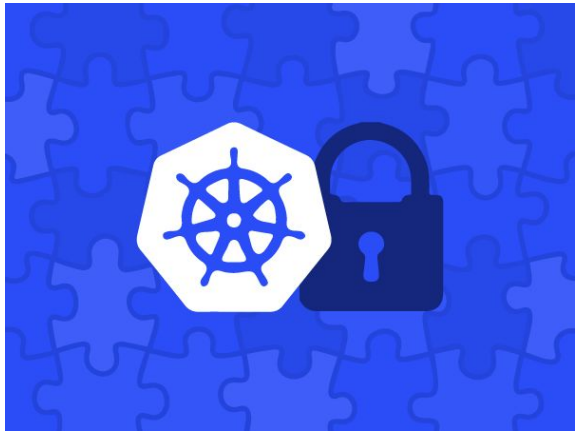
```
ansible-playbook -i INVENTORY --become --become-user=root PLAYBOOK
```

Best Practice: Cluster Hardening

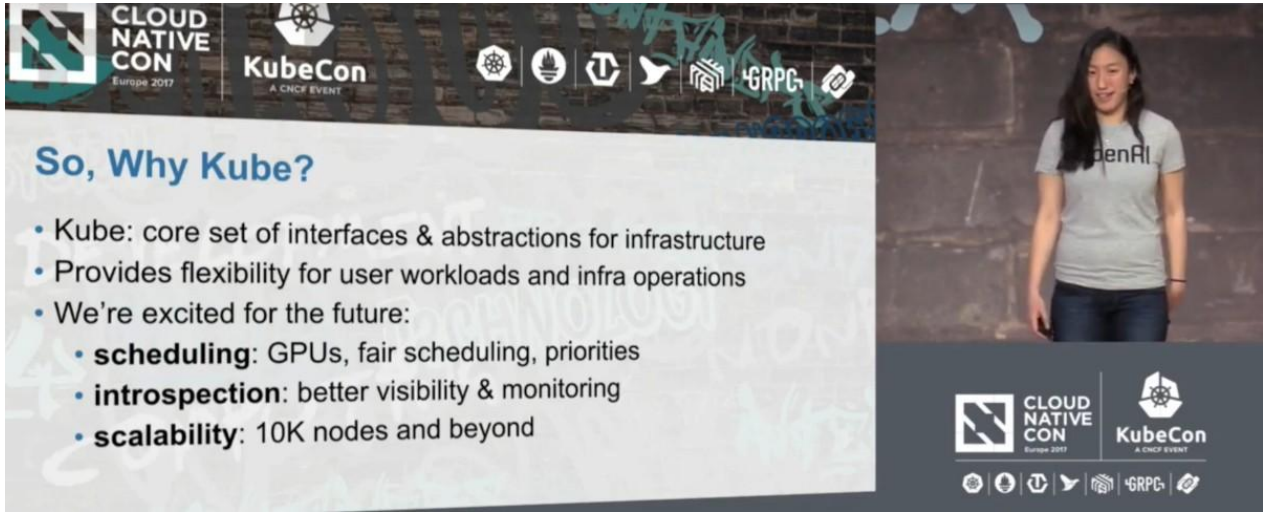
- Full support with config
- Cluster hardening guide

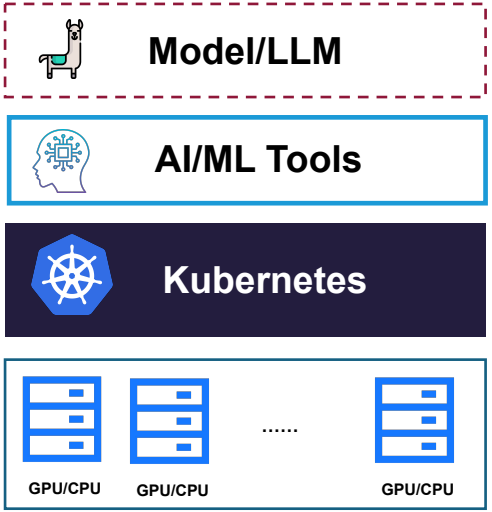
```
## kube-apiserver
authorization_modes: ['Node', 'RBAC']
kube_apiserver_request_timeout: 120s
kube_apiserver_service_account_lookup: true

# enable kubernetes audit
kubernetes_audit: true
audit_log_path: "/var/log/kube-apiserver-log.json"
audit_log_maxage: 30
audit_log_maxbackups: 10
audit_log_maxsize: 100
......
```

certified kubernetes

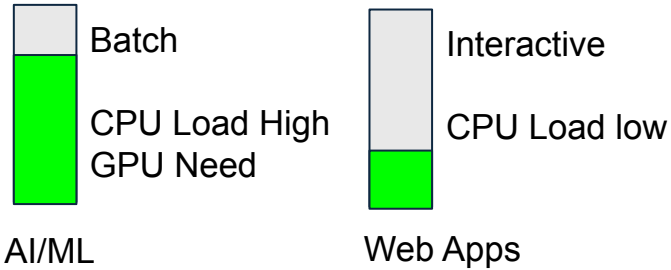


AI/ML Workload for Kubernetes





For Example:
Kubernetes that powers OpenAI infrastructure



AI-Optimized k8s Cluster

KubeCon | CloudNativeCon | THE LINUX FOUNDATION OPEN SOURCE SUMMIT | AI_dev
China 2024

Models

Model Develop | Model Training | Model Serving | LLM

AI Tools

AI/ML Framework
PyTorch | Tensorflow | Hugging face

AI/ML Operators
KubeRay | KubeFlow

Kubernetes

Scheduler for AI/Batch Job
Scheduler Plugin | Volcano

Queue
Kueue

Nvidia GPU Integration
NFD | DM | RDMA | MIG | vGPU | Kubevirt | DRA

Kubernetes

Docker/Containerd/CRI-O Support | Nvidia GPU Driver

Infrastructure

Bare Metal
CPU: Intel, ARM, AMD
NVIDIA GPU: 4090, A100, H100, T4

Cloud
Terraform → GPU instance
AWS, EQUINIX, Google Cloud

developing

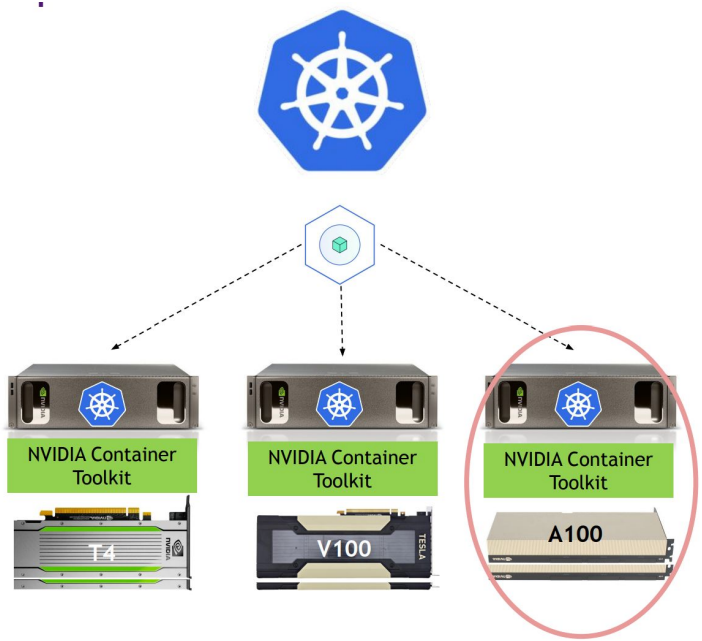
GPU Basic Usage


```yaml
apiVersion: v1
kind: Pod
metadata:
  name: gpu-example
spec:
  containers:
  - name: gpu-example
    image: nvidia/cuda
    resources:
      limits:
        nvidia.com/gpu: 2
  nodeSelector:
    nvidia.com/gpu.product: A100-PCIE-40GB
```

CUDA Image

Schedule by Device Plugin

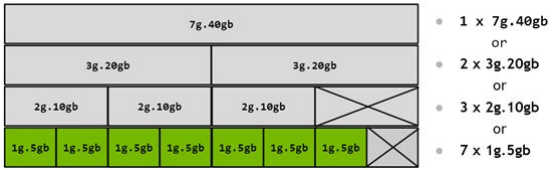
Label by gpu feature discovery



Ref to: Unlocking the Full Potential of GPUs for AI Workloads on Kubernetes - Kevin Klues, NVIDIA

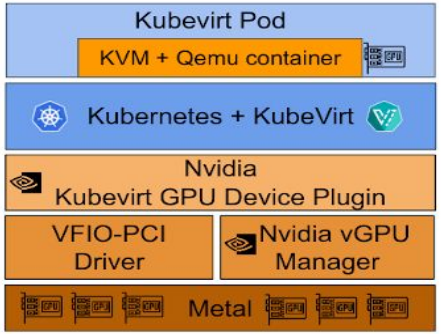
GPU Advanced Usage

MIG



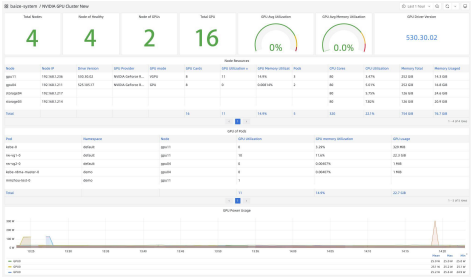

```yaml
apiVersion: v1
kind: Pod
metadata:
  name: gpu-example
spec:
  containers:
  - name: gpu-example
    image: nvidia/cuda
    resources:
      limits:
        nvidia.com/mig-1g.5gb.shared: 1
        nvidia.com/mig-1g.5gb: 1
```

KubeVirt




```yaml
apiVersion: kubevirt.io/v1alpha3
kind: VirtualMachineInstance
metadata:
  name: vmi-gpu
spec:
  domain:
    devices:
      gpus:
      - deviceName:
nvidia.com/GP102GL_Tesla_P40
        name: gpu1
...
```

DCGM exporter



Grafana Prometheus NVIDIA DCGM

ref: https://docs.nvidia.com/datacenter/tesla/mig-user-guide/index.html

Discuss: Helm vs Ansible Template

Q: Using helm template or ansible template in Kubespray ？

Benefit

- Maintenance by the original project itself

Challenge

- Almost all features is implement by ansible, Migrate means break change

Benefit

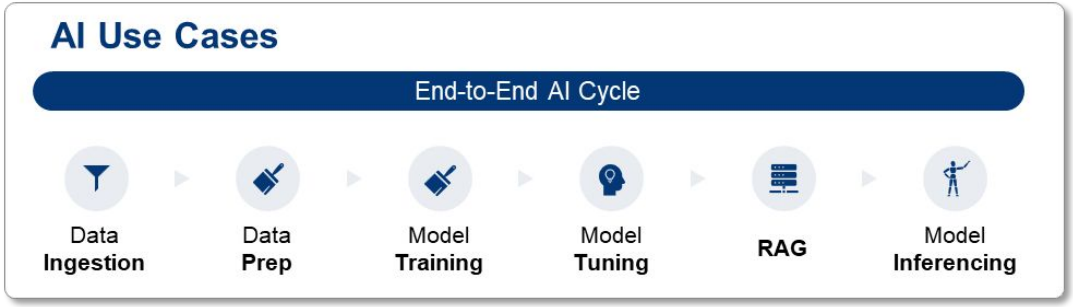
- All Templates is include by Kubespray, make it’s easy to customized.
- Ready for air-gap environment

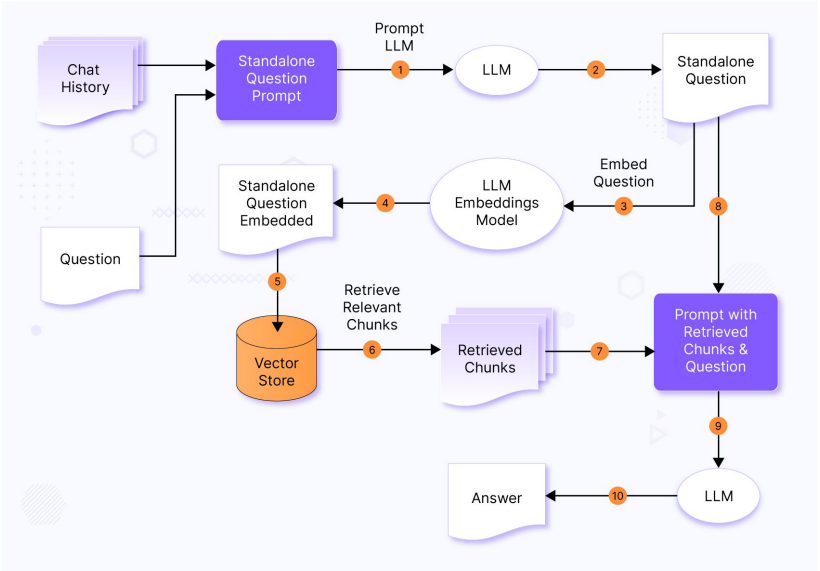
Challenge

- huge maintenance load

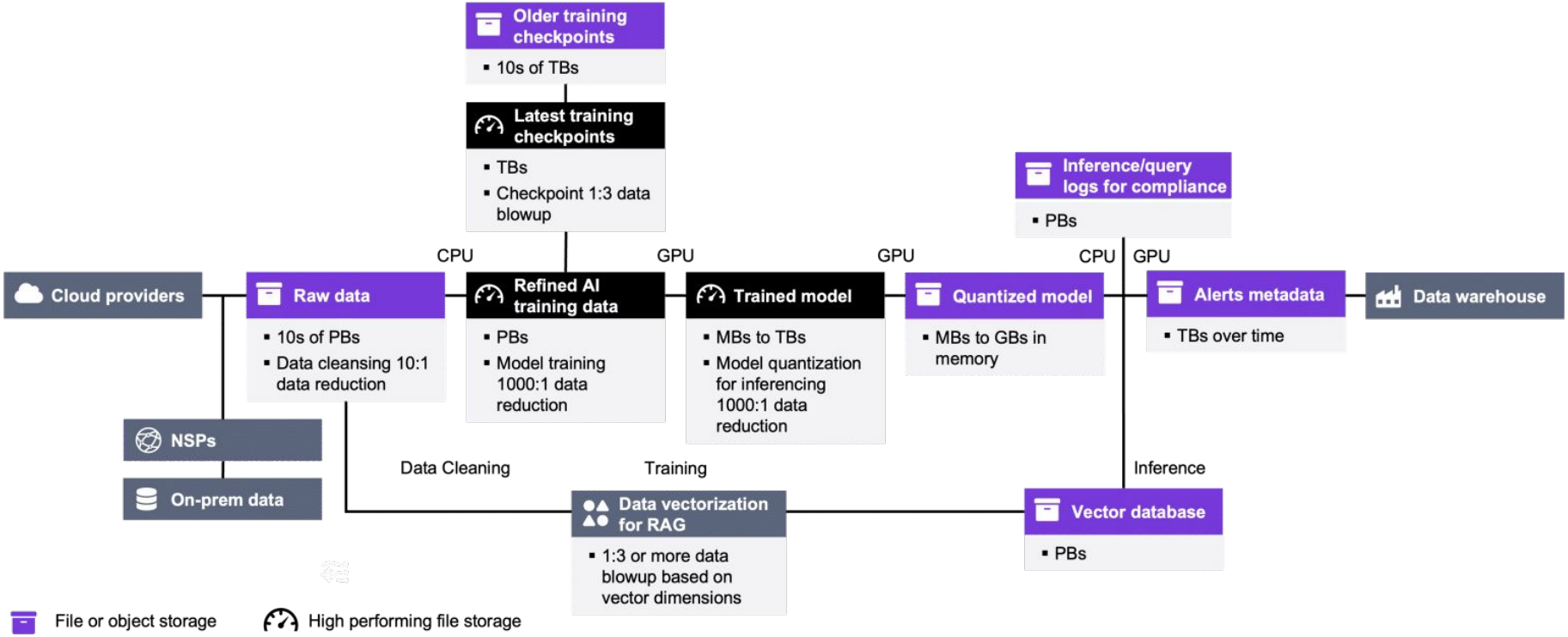
A: The Kubespray should the helm and ansible template both. Ansible is better for System Components, and Helm is better for Apps.

LLM and RAG in AI Cycle





Data types and storage of RAG in the AI pipeline



Vector database for RAG inferencing

Size: Petabytes (depending on the size of document stores, number of dimensions indexed)

Type of storage: High-performance file and capacity-based object storage

Considerations: Retrieval-Augmented Generation (RAG) is an AI technique used to make large language model (LLM) results more accurate by providing more contextual information in the input prompts. Vector databases are a crucial component of RAG inferencing. Companies encode documents and store them in vector databases. Moving from raw data to a vector database can result in a 1:3 (or greater) data blowup because the number of dimensions determine the number of indices and the size of storage.

Think about your storage strategy for RAG optimization

1. **Storage Selection**: Choosing the right storage for each AI phase is a crucial, yet often overlooked, part of AI preparation.
2. **Holistic Approach**: When deciding on the location for your AI infrastructure, consider the entire AI pipeline and the unique storage requirements of each phase.
3. **Data Management Strategy**: The success with AI is significantly influenced by your overall data management strategy.
4. **Optimum Execution Location**: The second part of the guide will discuss the best locations for executing different AI pipeline phases, whether in the cloud, a cloud adjacent location like Equinix, or on-premises. It will also cover key AI-related storage innovations, including improved GPU storage access, consolidation of various storage types to reduce costs, and storage fabrics for data movement between distributed AI sites.

Benefits of Bare Metal Kubernetes for LLM and RAG Workloads

1. **Resource efficiency**
   Unlock your hardware's full potential with bare metal Kubernetes. Direct access to CPU, RAM, and storage eliminates virtualization overhead, maximizing performance and efficiency.

2. **Direct access**
   Bare metal Kubernetes directly accesses physical hardware, ideal for low-latency tasks needing fast responses.

3. **Optimize costs**
   Bare metal Kubernetes offers full server control, eliminating virtualization "black box."

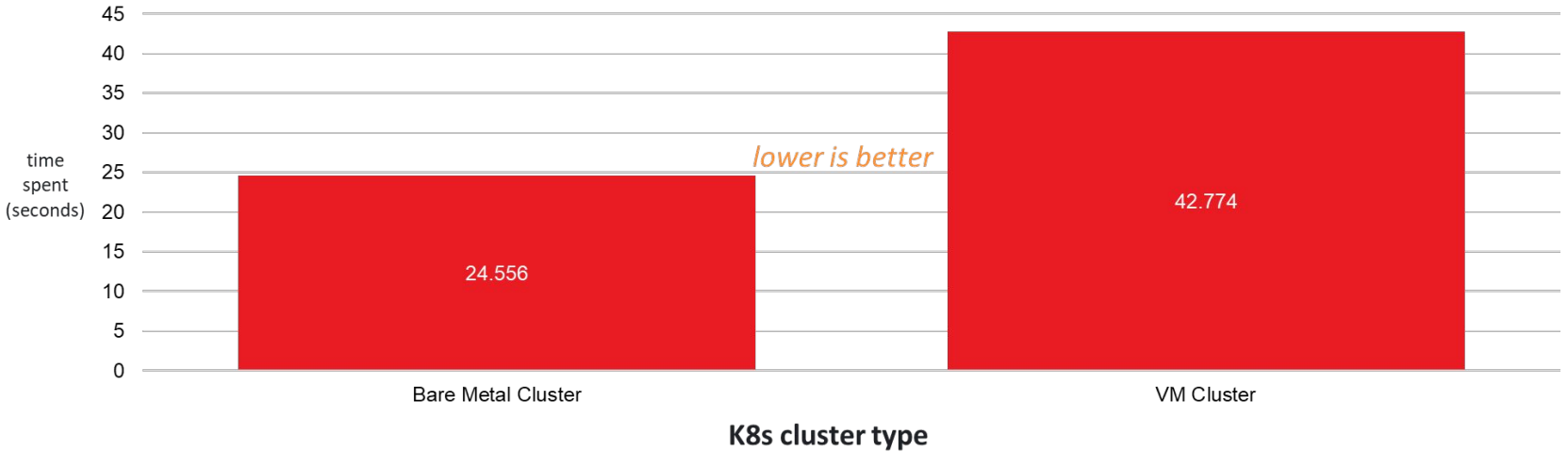
4. **Physical isolation**
   Bare metal Kubernetes offers audit-friendly isolation, deploying apps on a single-tenant physical server with no external interactions.

Processor Performance Test

CPU Benchmark

Average time spent for 10 retries on the calculation of pi with an accuracy of 10,000 decimal places

- **Bare metal worker node:** 1x Intel Xeon E-2378G 8C/16T 3.2 GHz / 64 GB / Ubuntu 22.04
- **Virtual machine worker node:** 16 vCPU / 64 GiB Memory / Ubuntu 22.04



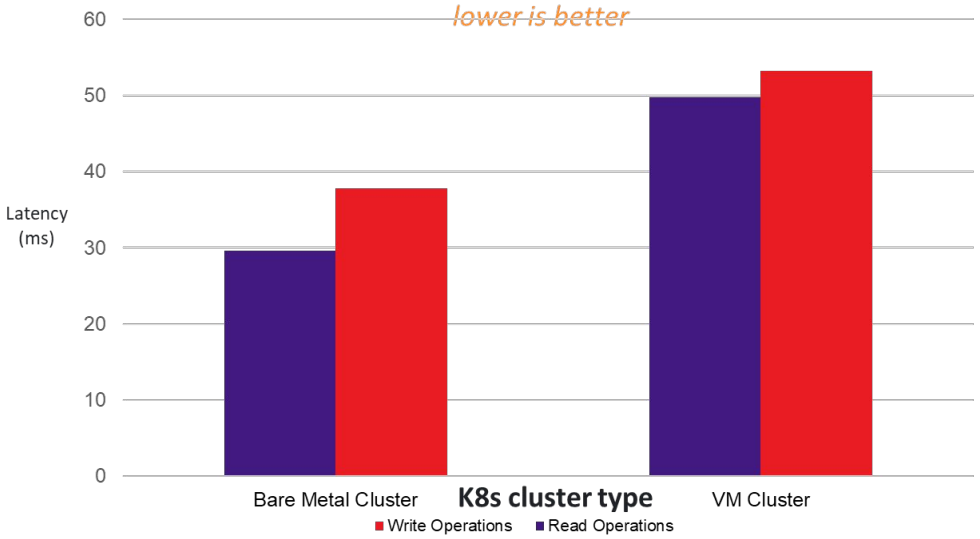
GitHub - alexdedyura/cpu-benchmark: CPU benchmark by calculating Pi, powered by Python3

Memory Performance Test

RAM Latency

Use sysbench and transfer 6400 GB of data through RAM and measure the latency in ms

- **Bare metal worker node:** 1x Intel Xeon E-2378 8C/16T 3.2 GHz / 64 GB / Ubuntu 22.04
- **Virtual machine worker node:** 16 vCPU / 64 GiB Memory / Ubuntu 22.04



Sysbench: https://github.com/akopytov/sysbench

Thanks to Community

a Pure Bazaar Global Community

Developers World Wide

1443 Developers

>50 in a release

7765 Commits

9 years

From Oct 2015

Tico88612

VannTen

ErikJiang

New Maintainers last year

MrFreezeex

Ant31

Honored Returnee

Need your help!

Thank You!

Q & A

slack #kubespray for general questions & support
#kubespray-dev for Kubespray development

GitHub kubernetes-sigs/kubespray