Cloud Native AI Infra

Keith Chan
VP of The Linux Foundation APAC

2025

Email: kchan@apac.linux.com

Key indicators of how cloud native principles shape the systems required for AI development.

OPENAI

Scaling Kubernetes to 7,500 nodes

HUGGING FACE

Hugging Face Collaborates with Microsoft to launch Hugging Face Model Catalog on Azure

Cloud Native Artificial Intelligence is an evolving extension of Cloud Native.

Cloud Native Artificial Intelligence (CNAI) refers to approaches and patterns for building and deploying AI applications and workloads using the principles of Cloud Native. Enabling repeatable and scalable AI-focused workflows allows AI practitioners to focus on their domain.

The Cloud Native AI Ecosystem

Cloud Native AI Ecosystem

The cloud native AI ecosystem integrates AI and cloud native technologies to create an efficient, flexible, and scalable AI development and deployment environment.

The core competence

Combining containers, microservices, CI/CD, and DevOps to enhance AI development efficiency, deployment speed, and operational capabilities.

The Future

The cloud native AI ecosystem will continue to evolve, promoting the wider application of AI technology in various industries.

The core competence of Cloud Native AI

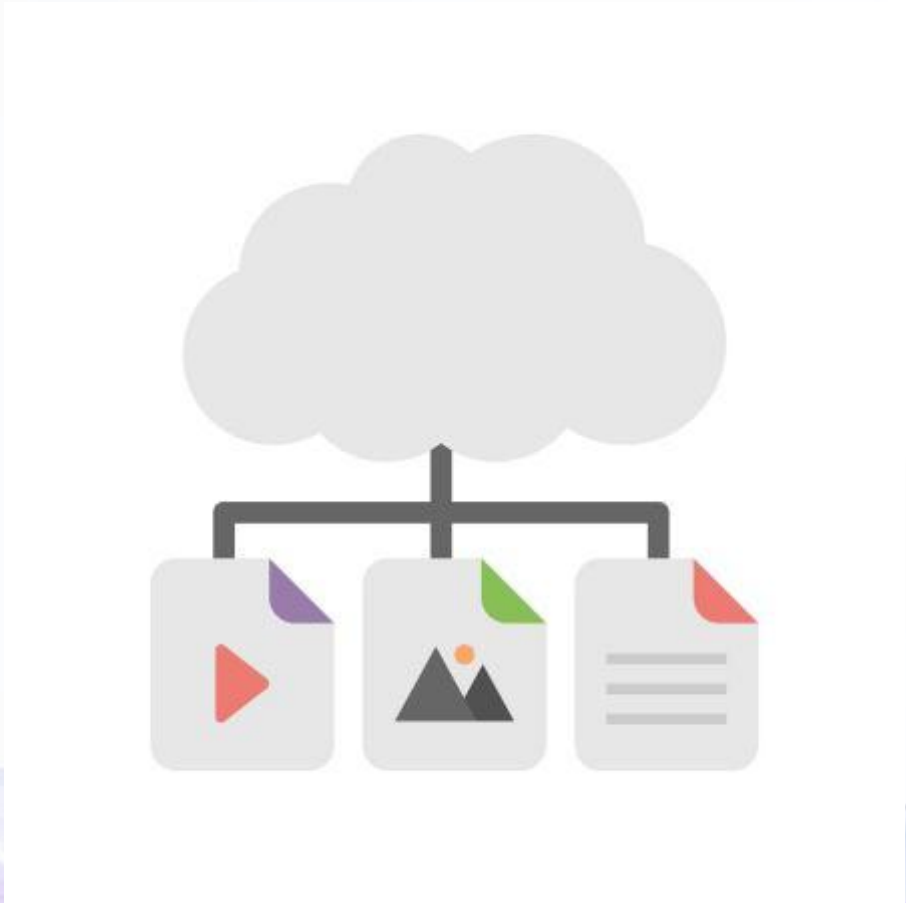
Elasticity & Scalability

Automatic resource allocation

The cloud native AI ecosystem supports automatic resource allocation based on AI workload requirements, achieving dynamic optimization and efficient utilization of resources.

Flexible expansion and contraction

The system can flexibly increase or decrease computing resources according to demand, ensuring sufficient computing power during peak periods and effectively saving costs during low periods.



Accelerate development cycles

Continuous Integration/Continuous Delivery

Combining CI/CD and containerization technology, AI model development, testing, and deployment are automated and standardized to accelerate iteration cycles.

Quickly respond to market demands

Through continuous integration and delivery, the team can quickly respond to changes, flexibly adjust model functionality and performance, and meet new market demands.

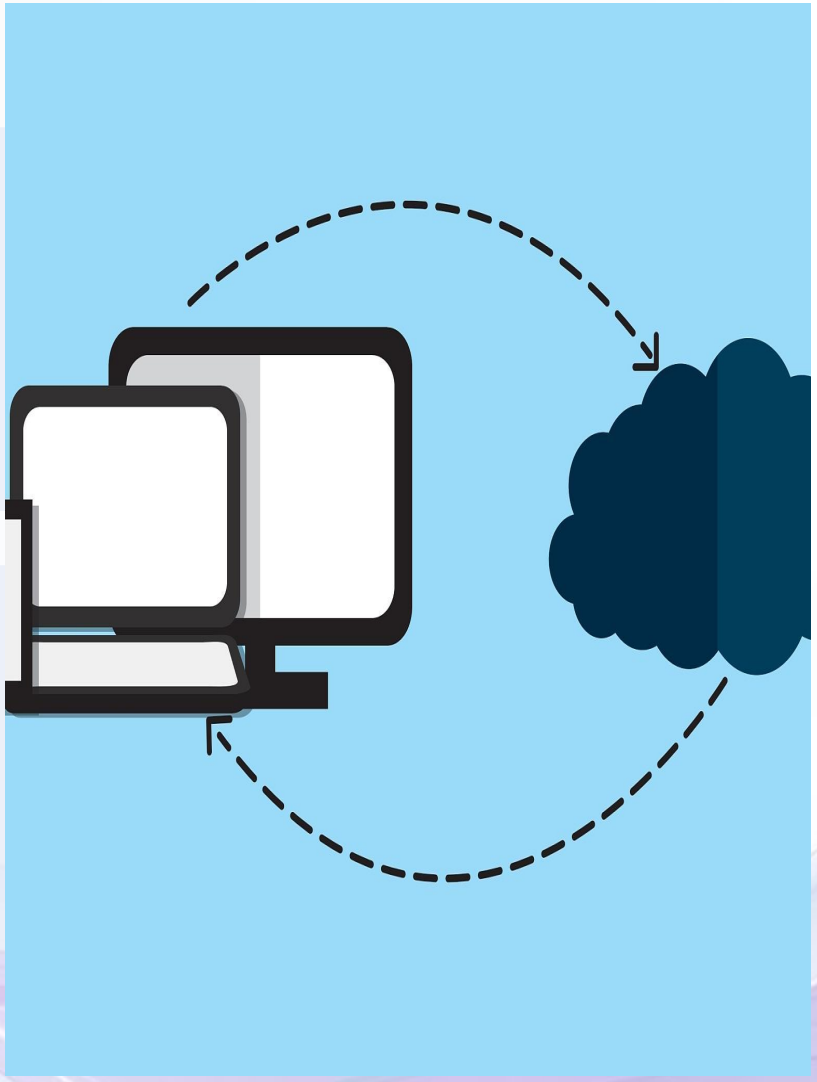
Cross-platform / Multi-Cloud compatibility

Containerized cross cloud migration

Containerization technology endows AI applications with the ability to seamlessly traverse different cloud platforms and local environments, enabling flexible migration and deployment.

Code as a Service Advantage

Cloud native architecture ensures that applications are tightly bound to their operating environment, reducing risks caused by environmental differences and improving deployment and operational efficiency.



Cost Optimization

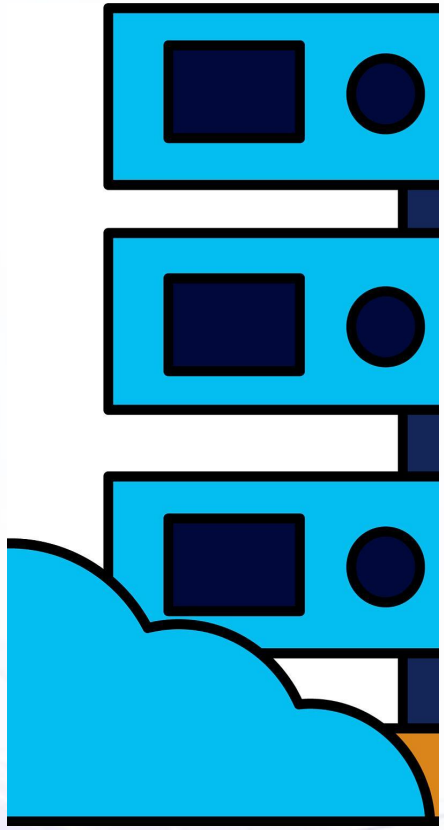
Resource allocation optimization

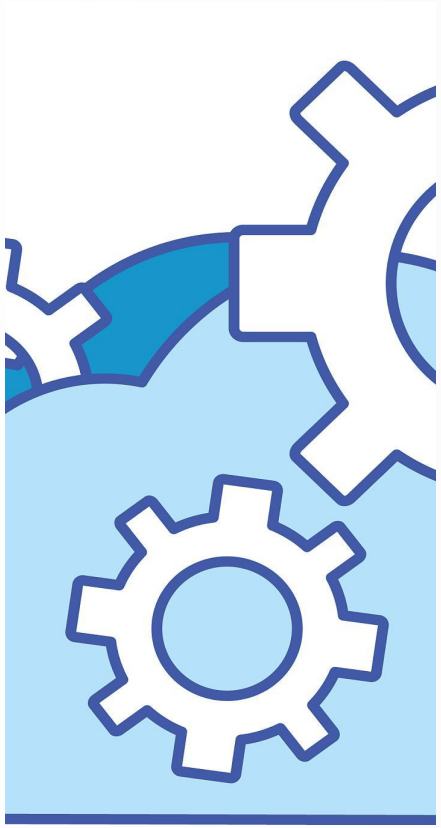
The cloud native AI ecosystem automatically allocates resources on demand, significantly reducing AI application development costs and improving resource utilization efficiency.

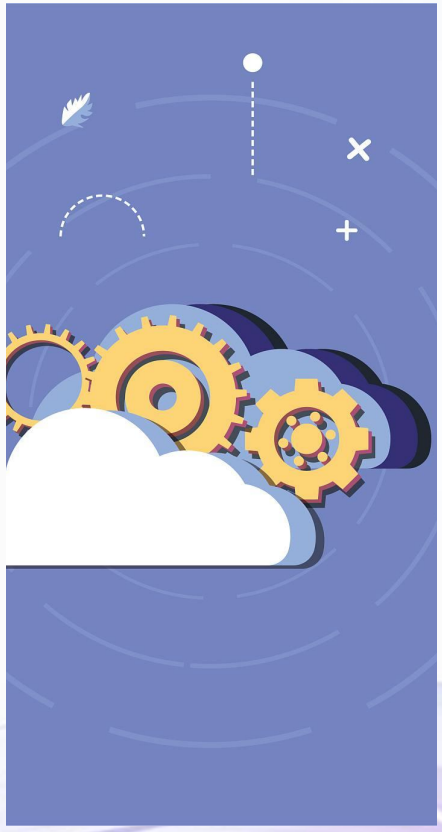
Reduced operation and maintenance costs

Automated operation and maintenance simplifies the operation process, reduces manual intervention, and further lowers the operation and maintenance costs of AI applications.

High availability and fault tolerance







Microservice architecture enhances availability

The cloud native microservice architecture achieves service splitting, increases the availability of AI applications, and avoids single points of failure.

Container orchestration enhances fault tolerance

Through container orchestration technology, AI applications can easily achieve fault isolation and automatic recovery, enhancing fault tolerance.

The Application of Cloud Native AI

Intelligent Recommendation System

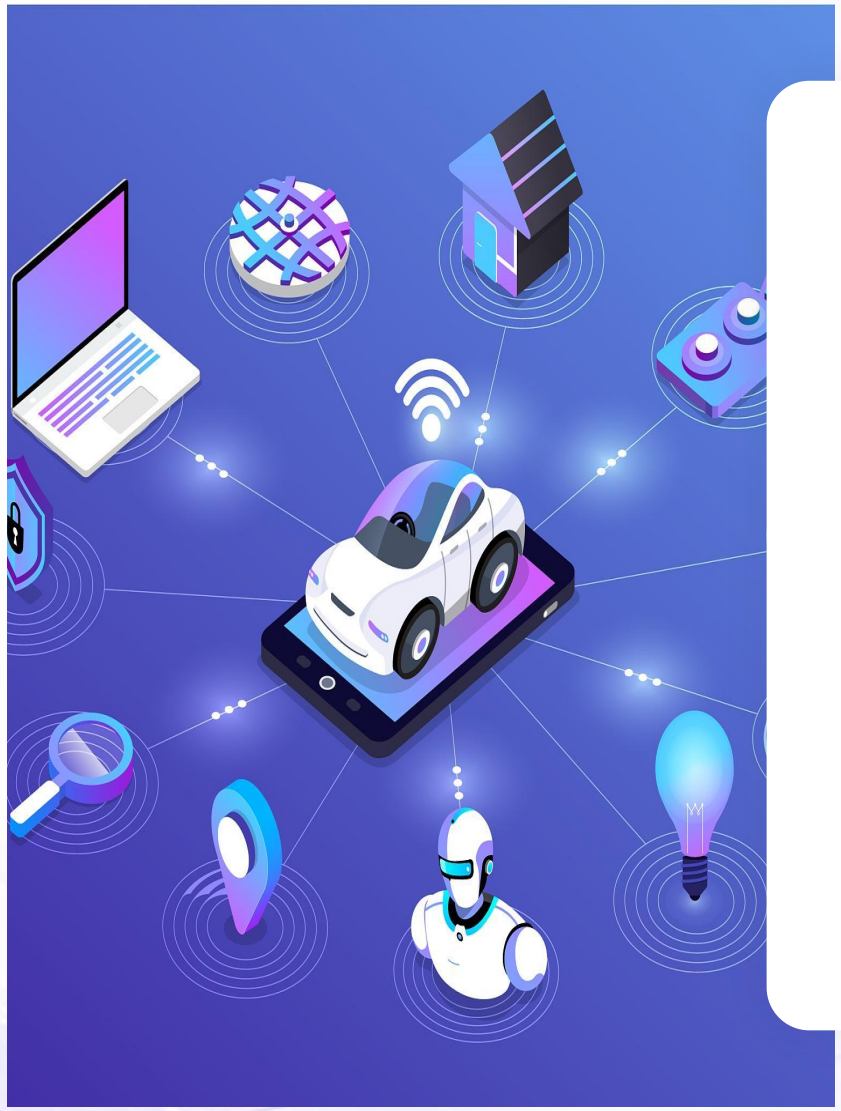
Cloud native elastic recommendation

By utilizing the elastic scalability of cloud native, intelligent recommendation systems can respond to user needs in real-time and provide personalized recommendation services.

Real time recommendation engine

Based on cloud native scalability, recommendation engines can quickly process massive amounts of data, achieve real-time recommendations, and enhance user experience.

Autonomous driving



Edge Cloud Collaboration

Combining edge computing and cloud native technology, real-time data processing and model updating are realized to improve the timeliness and accuracy of automatic driving decisions.

Intelligent Connected System

Through the elastic scalability of cloud native technology, it supports real-time perception, decision-making, and control of vehicles in complex environments, leading the future of intelligent connected vehicles.

Financial risk control

Risk monitoring

Through cloud native high availability and elastic scalability, real-time risk monitoring is supported to ensure that financial institutions can quickly identify and respond to potential risks.

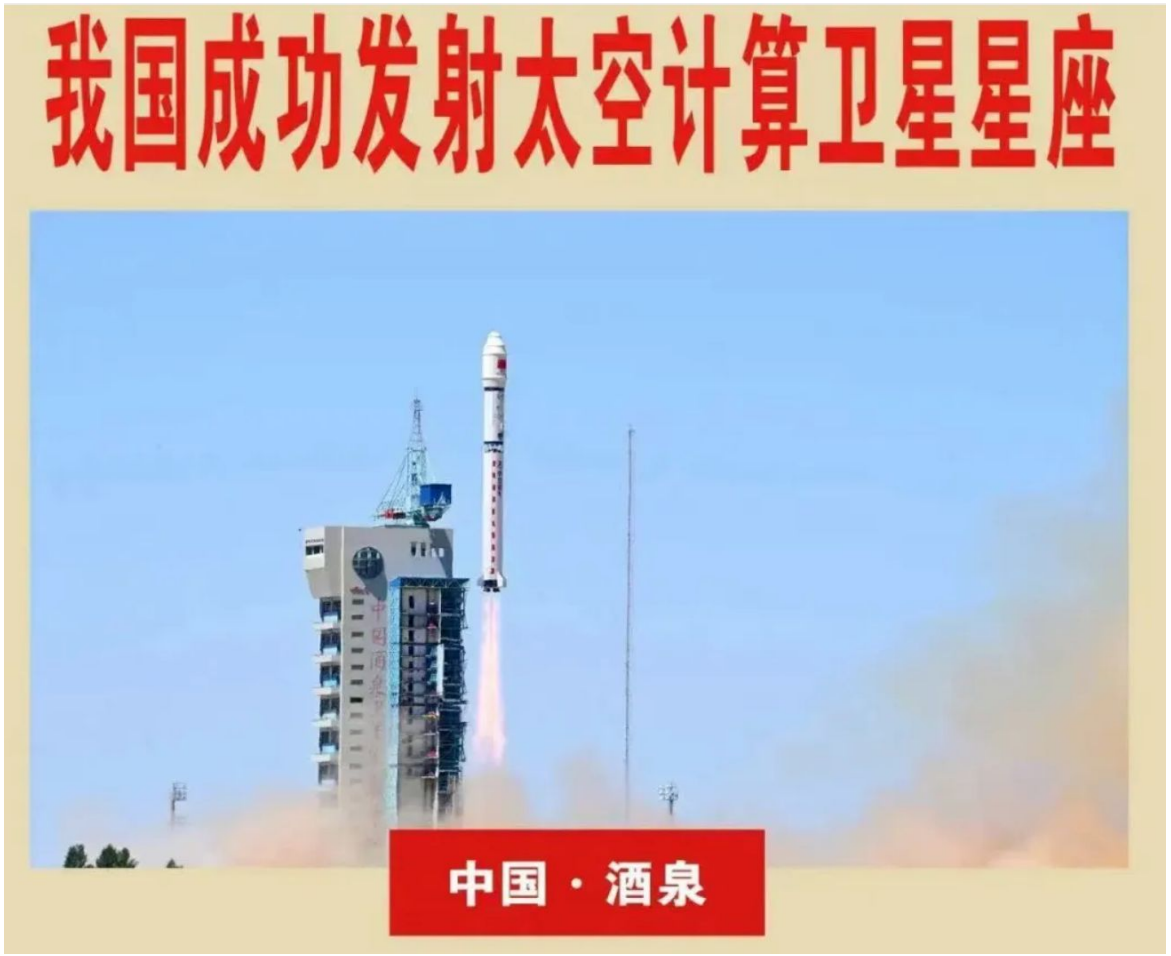
Intelligent risk control

The cloud native AI ecosystem provides powerful technical support for financial risk control, automates risk assessment, and ensures the security of financial transactions.

In China: Very Interesting Case Stu

CubeFS
Cloud Native Storage

- 2025 May 14th, China has launched its space computing satellite to build up the space-based intelligent computing infrastructure
- CubeFS was chosen in this space-based distributed operating system to manage data distributed across numerous satellites because:
  - High chance of single node failure because of radiation, need to ensure high availability and strong consistency of data and quickly recover.
  - Massive small file processing capability



In China: Very Interesting Case Stud

KARMADA
Scheduling & Orchestration

- Rednote leverages Karmada to build its multi cloud IT infrastructure:
  - creating a unified platform entrance for applications,
  - addressing the challenges of cluster and resource management in the rapidly developing business process
  - KARMADA End User Community

In China: Very Interesting Case Stu

- SF Express recently released the "Effective GPU Technology White Paper", which elaborates on its engineering practice in the field of GPU resource pooling and scheduling. This technology architecture deeply integrates HAMi's core capabilities in efficient scheduling of diverse heterogeneous GPUs, unified management, and observability.

In China: Very Interesting Case Studies

- Bytedance AIBrix open source project: the first enterprise level inference system project based on Kubernetes

- It collaborates deeply with inference engines such as vLLM to continuously optimize inference efficiency, and integrates multiple cutting-edge research results to promote large-scale model inference towards a more efficient and practical production stage.

Cloud Native 2024: Approaching a Decade of Code, Cloud, and Change

Cloud native adoption reaches a new high of 89%, and it's equally popular in companies of all sizes.

91% of organizations use containers for production, but their deployment is challenged primarily by **cultural changes in the development team.**

It's a Kubernetes world today: 93% of companies use it in production, are piloting it or are actively evaluating it.

Kubernetes dominates CNCF's top graduated products, with ties to five of the top products: Helm, etcd, CoreDNS, Cert Manager and Argo.

Incubating projects tell a technologically diverse story, from workload management to security and dev/user experience.

Survey respondents reported **significant progress in making their software more secure** as compared to where they were in 2023.

Use of CI/CD in production grew 31% from 2023 to 2024.

Organizations are releasing code faster than ever before: 29% are pushing it out the door multiple times a day, up from 23% in 2023.

77% of respondents said some, much, or nearly all of their deployment practices and tools **adhere to GitOps principles.**

CLOUD NATIVE COMPUTING FOUNDATION

200+ Projects

728 CNCF Members

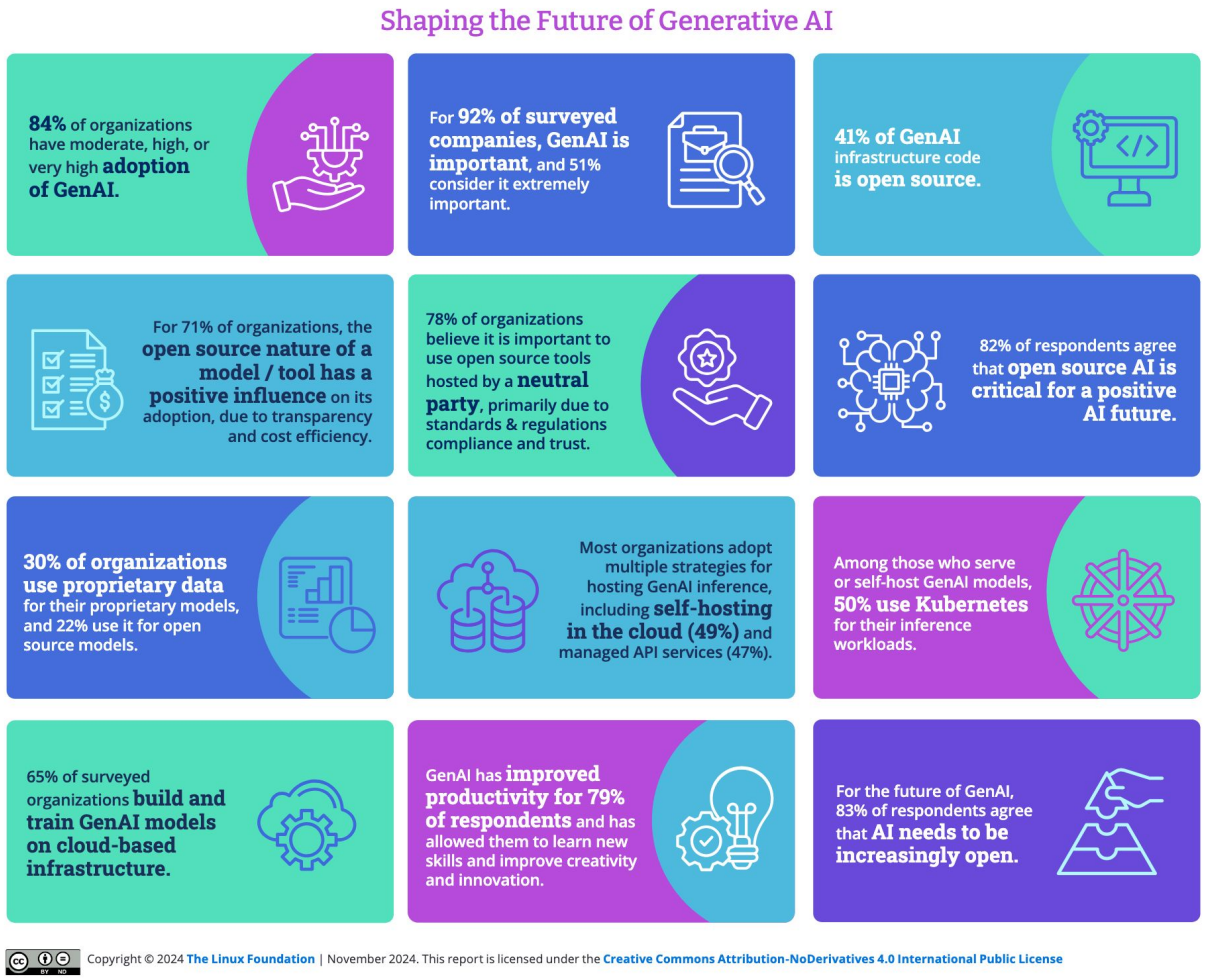
270,000+ Contributors

100,000+ community.cncf.io Group Members

The Impact of Open Source Innovation: Shaping the Future of Generative Artificial Intelligence

We know that GenAI is changing the industry at an unprecedented speed. As this technology enters the mainstream, organizations unanimously agree that the future of AI must be open. In fact, 82% of organizations believe that open source AI is crucial for ensuring a positive future for AI, and 83% of organizations agree that AI needs to become increasingly open to promote trust, collaboration, and innovation.

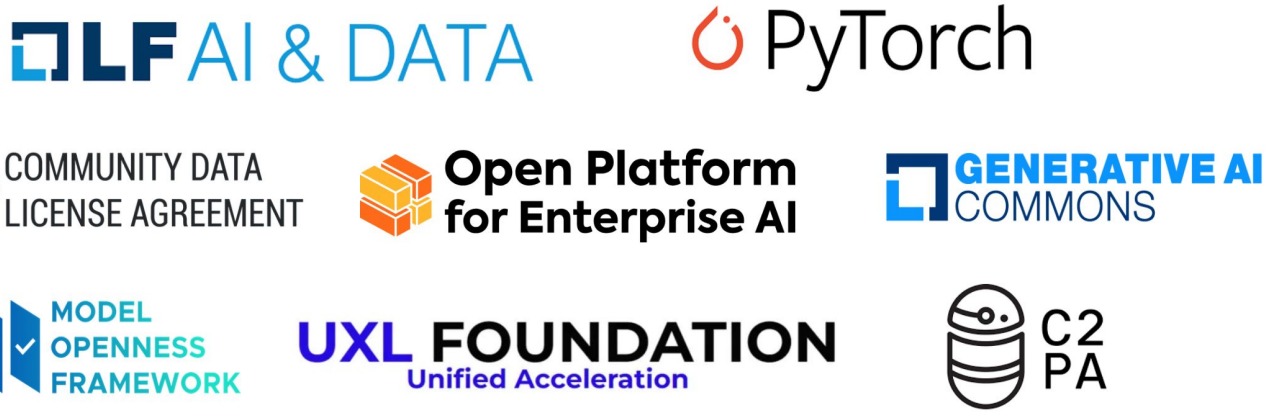
https://www.linuxfoundation.org/hubfs/LF%20Research/lfr_genai24_111924.pdf?hsLang=en



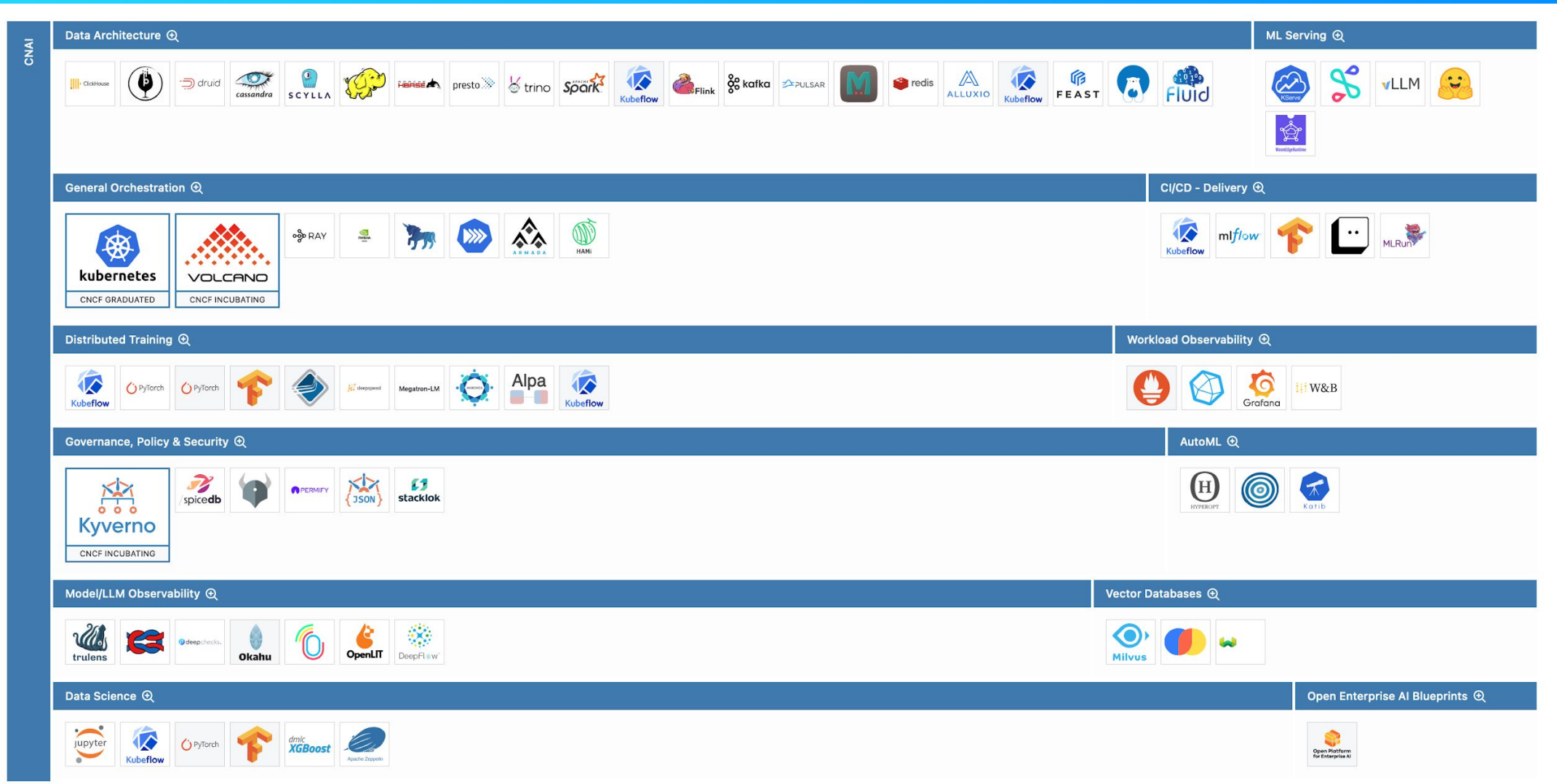
Open Source AI: Open Source is the Core of Artificial Intelligence Innovation

Open source is the core of artificial intelligence innovation:

Artificial intelligence (AI) is rapidly shifting from early adoption to mainstream, and the Linux Foundation plays a key role in guiding this transformation by fostering a strong open-source AI ecosystem. The foundation's efforts cover core technologies, legal and IP frameworks, innovation sandboxes, and a series of newly launched conferences that bring together leading ideas in the fields of AI and open source on one platform. These measures aim to ensure that the future of AI remains based on open source principles.



Multiple Linux Foundation (LF) organizations, including CNCF, LF AI&Data, and AI Alliance, provide a CNAI Landscape that allows you to see established and developing projects by functional area.



Complexity

Challenges in cloud native AI

The combination of cloud native and AI technology has potential, but it also brings significant increases in system complexity, thereby increasing the difficulty of technology implementation and operation, requiring higher technical capabilities and experience.

Integration challenges need to be addressed

The complexity challenges of the cloud native AI ecosystem are mainly reflected in technology integration, system deployment, and continuous operation and maintenance. It is necessary to overcome the difficulties of technology integration and ensure the stable and efficient operation of the system.

Data Security and Privacy

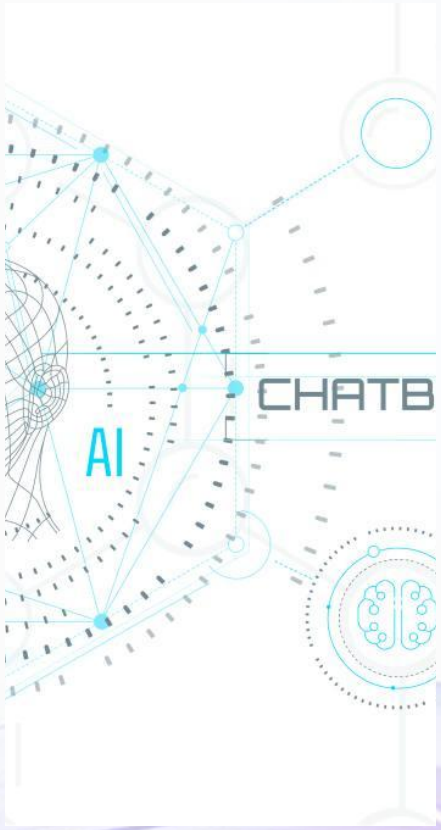
Data security and privacy concerns

In a distributed environment, data security and privacy protection have become the primary challenges, facing more complex and evolving threats.

Protective measures need to be strengthened

To address these challenges, it is necessary to strengthen security measures such as data encryption, access control, monitoring and auditing to ensure the security and privacy protection of data.

Resource scheduling optimization



Difficulty in scheduling heterogeneous resources

Efficient scheduling of heterogeneous computing resources such as GPUs and TPUs is still a problem that needs to be solved in the cloud native AI ecosystem, which puts high demands on technical implementation.

Optimizing scheduling is key

To maximize the utility of heterogeneous computing resources, it is necessary to continuously optimize scheduling algorithms, improve resource utilization efficiency, and ensure the smooth operation of AI applications.

Model Management and Governance

Model Management Challenge

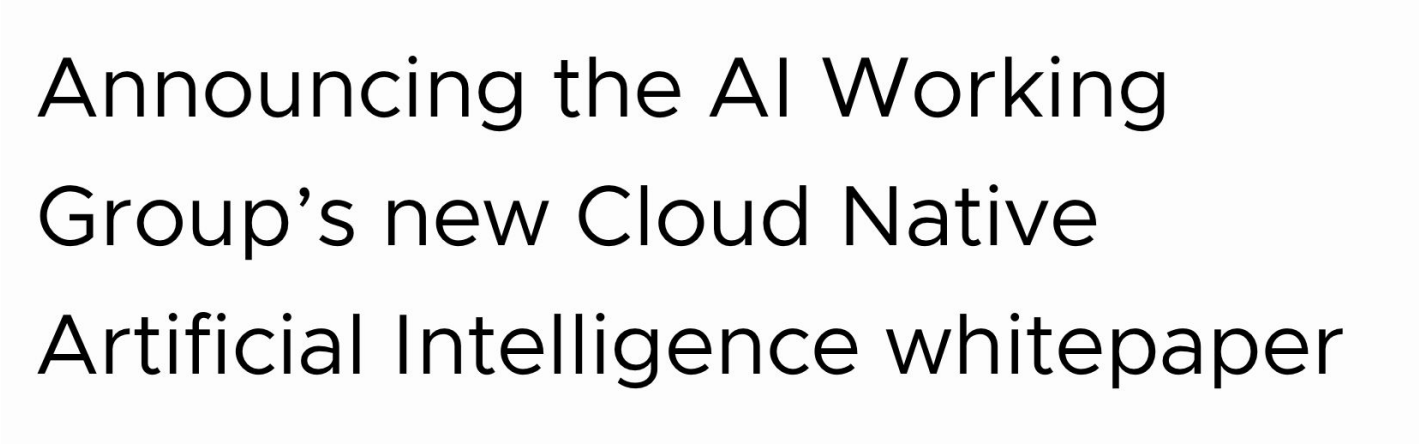
With the increasing number of AI models, how to effectively manage, monitor, and govern models has become a major challenge.

Model governance strategy

Develop a reasonable model management strategy, including model version control, lifecycle management, security auditing, etc.

1. CNCF Cloud Native AI Working Grou
2. Cloud Native AI Whitepaper

CNCF Wechat Account



https://github.com/cncf/tag-runtime/blob/main/wgs/cnai/charter/index.md

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

Equip yourself

The Linux Foundation Training and Certification

https://training.linuxfoundation.cn/certificates

官方证书

- 全部
- SkillCred技能证书
- DevOps
- IoT 及 嵌入式
- 开源最佳实践
- 微服务
- Node.JS
- 系统管理认证
- 区块链
- 网络
- 云技术及容器
- 人工智能

云技术及容器

CKA (Certified Kubernetes Administrator)

CKA(Certified Kubernetes Administrator) 认证考试可确保Kubernetes管理人员在从业时具备应有的技能、知识和能力。CKA现有以下两种考试方式可供选择： 英...

CKA | Kubernetes | 管理员认证

云技术及容器

CKS (Certified Kubernetes Security Specialist)

获得CKS认证的Kubernetes安全专家在构建、部署和运行时确保基于容器的应用程序和Kubernetes平台的安全及最佳实践。 CKS现有以下两种考试方式可供选择：...

Kubernetes | K8s | 安全 | CKS

云技术及容器

CKAD (Certified Kubernetes Application Developer)

CKAD （Certified Kubernetes Application Developer）认证考试证明考生为Kubernetes设计、构建和部署云原生应用程序的能力。CKAD现有以下两种考试方式可...

Kubernetes认证 | K8s认证 | CKAD | 开发者认证

云技术及容器

KCNA (Kubernetes and Cloud Native Associate)

KCNA (Kubernetes and Cloud Native Associate) 展示考生在Kubernetes和更广泛的云原生生态系统中的基础知识和技能。KCNA现有以下两种考试方式可供选择...

KCNA | Kubernetes | CKA | CKAD | CKS | Cloud Native

云技术及容器

PCA (Prometheus Certified Associate)

PCA (Prometheus Certified Associate) 认证考试展示了工程师对可观察性的基本知识和使用Prometheus(开源系统监控和警报工具包)的技能。

Prometheus | Cloud Native | k8s

云技术及容器

ICA (Istio Certified Associate)

ICA (Istio Certified Associate)认证考试展示了对Istio原则、术语和最佳实践的深入理解，以便建立Istio。

Istio

云技术及容器

KCSA (Kubernetes and Cloud Native Security Associate)

KCSA (Kubernetes and Cloud Native Security Associate) 考试展示了考生对Kubernetes集群基线安全配置的能力，以合乎合规的要求，这包括加强安全控...

Kubernetes | Cloud Native | Security

云技术及容器

CCA（Cilium Certified Associate）

Cilium Certified Associate（CCA）认证考试证明用户具有使用Cilium连接、保护和观察Kubernetes集群所需的知识。

Cilium

云技术及容器

OTCA (OpenTelemetry Certified Associate)

随着云原生系统变得越来越复杂，对能够利用遥测数据的专业人员的需求正在迅速增长。开启您的新的职业道路，证明您在Opentelmetry的专业知识，包括跟踪，...

OpenTelemetry

云技术及容器

CBA (Certified Backstage Associate)

内部开发平台对于开发人员和工程人员的快速入职、跨团队协作和产品创新至关重要。获得CBA证书证明您有技能和思维模式与Backstage工作，以推进您的职业...

Backstage

云技术及容器

KCA (Kyverno Certified Associate)

作为管理和保护Kubernetes环境的专家，Kyverno专业知识表明您了解云管理和安全的高级方面，这是当今最受欢迎的技能组合之一。

Kyverno

OpenInfra Day Korea 2025
Exclusive LF Training and Certification Discount

Exclusive **40%** Discount Code:

OPENINFRADAYKR25

Valid till **Aug 31, 2025**

The voucher is applicable to: LF's E-learning, LF's Certification Exam Only , Standard E-Learning Course – Certification Exam bundles at LF Education Global Website: https://training.linuxfoundation.org/

LFAPAC Education and Certification
Community Partner Program

커뮤니티 파트너 프로그램에 지원하고
다양한 파트너 혜택을 만나보세요!

파트너 주요 혜택 (Benefits of Partnership)

- 파트너 전용 교육, 자격증 할인혜택
- 행사 기회 및 커뮤니티 행사 홍보 지원 및 발표자 추천
- 추천 성과에 따른 인센티브

파트너 지원하기
↓

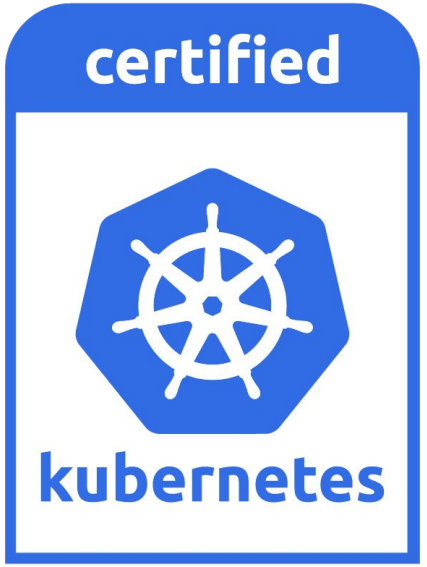
bitly

Certified Kubernetes Conformance

Software conformance ensures that every vendor's version of Kubernetes supports the required APIs, as do open source community versions.

For organizations using Kubernetes, conformance enables interoperability from one Kubernetes installation to the next. It allows them the flexibility to choose between vendors.

CNCF runs the Certified Kubernetes Conformance Program. The world's leading enterprise software vendors and cloud computing providers have Certified Kubernetes offerings.



100+ Certified Kubernetes Conformance Partners

Kubernetes Certified Service Provider

A pre-qualified tier of vetted service providers who have deep experience helping enterprises successfully adopt Kubernetes through support, consulting, professional services and/or training.

Benefits

- Placement on the first tab of https://kubernetes.io/partners/
- Recognized in the community as a leader and expert in helping businesses adopt Kubernetes.
- Increase awareness of your brand when end users are searching for consulting partners
- KCSPs are featured on https://kubernetes.io/partners/#kcsp, and https://www.cncf.io/certification/kcsp/, and https://landscape.cncf.io which in aggregate receive more than 25,000 page views per month on their listings of KCSPs.

Requirements

- Three or more engineers who pass the CKA exam
- A business model to support enterprise end users
- A Kubernetes professional services landing page on your website that details your training, consulting, implementation, and support service offerings.
- Be a CNCF and LF member

240+ Kubernetes Certified Service Providers

Thank you

Keith Chan
Email: kchan@apac.linux.com