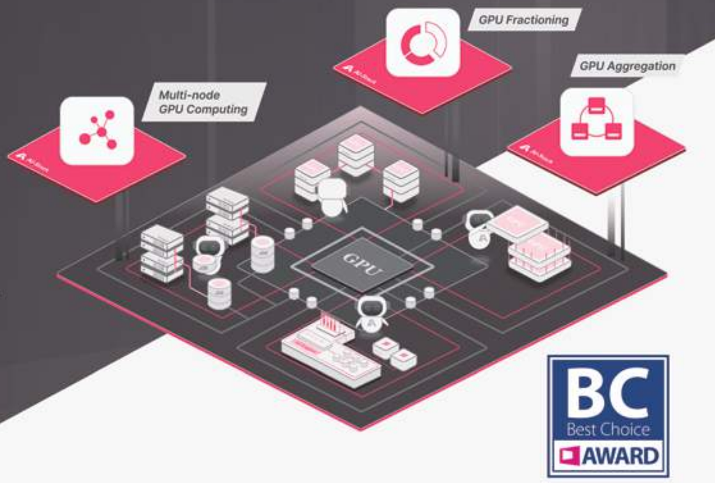
AI-Stack

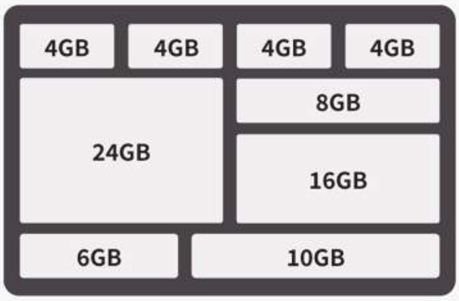
AI-Stack is Revolutionizing AI Infrastructure, Ushering in an Infinite Digital Era.

INFINITIX's AI-Stack offers comprehensive AI infrastructure management solutions, integrating GPU partitioning(NVIDIA/AMD), GPU aggregation, cross-node computing, intuitive user interface, MLOps workflow, open-source deep learning tools, and containerized environment deployment features. It is an essential key tool for enterprises when adopting AI services.



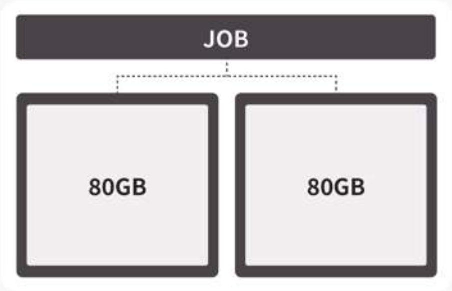
BC Best Choice AWARD

The 3 Core Technologies of AI-Stack for GPU Resource Management



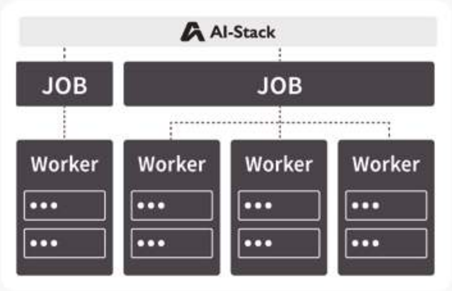
GPU Partitioning

Divides a single GPU into multiple independent units, accommodating training needs at various scales. This boosts the utilization rate of GPU to over 90%, significantly reducing the waste of computational resources.



GPU Aggregation

Aggregates the computing power of multiple GPUs to fulfill the training requirements of large AI/ML models. This greatly accelerates the training speed and enhances development efficiency.



Cross-Node Computing

Distributes training tasks across multiple nodes using distributed training techniques, enabling parallel processing of massive datasets and effectively reducing model training time.

Key Benefits

90% ↑

GPU Utilization

GPU Partitioning Increases Utilization from 30 % to 90 %

10x ↑

Workload Execution

10x Faster Workload Execution - Scaled Efficiency for Multiple Users and Tasks.

1 min ↓

DevEnv Setup

Dev Env Setup in 1 Minute - Cut from 2 Weeks to 60 Seconds.

10x ↑

Enhanced ROI

10x ROI Boost - Maximize Your AI Investment.

INFINITIX

NVIDIA Partner

INFINITIX provides advanced GPU resources and its AI infrastructure management platform, AI-Stack, helping enterprises efficiently adopt AI. In 2021, we became an NVIDIA Certified Solution Advisor. In 2024, we received the 'AMD GPU Ecosystem Builder Award,' becoming a key strategic partner globally.

LinkedIn

infinitix.ai

Facebook

AI-Stack Application Industries

AI Data Center

Semi-conductor

Manufacturing

Finance Service

Academic

Energy

Government

Transportation

Administrators | IT Manager

- Enterprise-grade security protection and efficient resource utilization.
- Project permission management and quota limitations.
- Professional dashboard features for real-time monitoring of GPU resource usage.

Users | Data Scientist and AI Researcher

- Start AI container deployment in just 1 minute.
- Focus on research, model development, and model training.
- Automate scheduled training tasks.
- Quickly deploy inference models.

GPU Resource Management + MLOps

Support full range of NVIDIA and AMD GPUs.

More precise and effective management of GPU resources.

Seamless connection with commonly used AI development tools.

Comprehensive dashboard for real-time monitoring of GPU computing resources.

Rapid model inference deployment services.

Automated execution with options for single or batch tasks.

User-friendly interface that's easy to get started with.

Fast containerized environment deployment and MLOps tools.

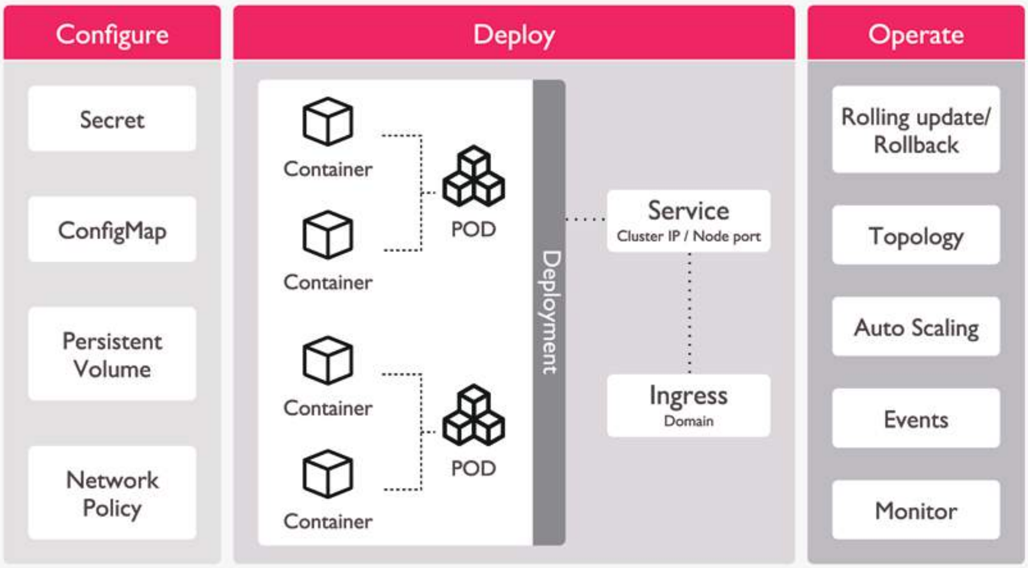
Resource isolation, permission management, and quota limitations.

Rapid Container Service(RCS)

Built on a Kubernetes architecture, RCS is designed for AI inference and application services, enabling enterprises to rapidly deploy, manage, and scale AI services.

Key Advantages of RCS:

- Rapid Deployment
- Real-Time Monitoring
- High Scalability
- Efficient Version Management



Server Partner & Distributors

AMD

NVIDIA

intel

MACNICA

ZERONE

敦新科技 DAWNING TECHNOLOGY INC.

Hardware Partners

DELL Technologies

Dell Technologies integrates talent, technology, data, and collaboration to help businesses enhance customer and employee experiences. With smart solutions, it enables secure, agile innovation and accelerates competitiveness for future growth.

Hewlett Packard Enterprise

Hewlett Packard Enterprise is a global edge-to-cloud, platform-as-a-service company that delivers a unified experience across all clouds and edges through a platform. It helps customers develop new business models and accelerate the transformation of ideas into business value.

msi

MSI is a global leader in gaming, content creation, business & productivity, and AIoT solutions. MSI develops its server products in-house, emphasizing design and manufacturing, reflecting the company's commitment to meeting customer needs and market demands.

COMPAL

Compal Server specializes in HPC, delivering enterprise-grade solutions that balance performance, efficiency, and sustainability. From AI data centers to cloud platforms and high-density computing, Compal offers flexible, reliable infrastructure.

MiTAC

MiTAC Computing

MiTAC Computing Technology Corp. delivers energy-efficient server solutions backed by industry expertise dating back to the 1990s. The company provides tailored platforms for data centers, HPC, and AI applications, ensuring performance and scalability.

Graid Technology Inc.

Graid Technology's innovative GPU RAID solution is tailor-made for NVMe SSDs, breaking through the limitations of traditional RAID solutions. It has been dedicated to building end-to-end big data solutions from edge to cloud, ushering in a new era of storage revolution.

HeTone Group

HeTone focuses on next-generation liquid cooling technologies - direct-to-chip(DTC)and immersion cooling - designed for HPC. The solutions reduce energy use and cost while enhancing performance, ideal for AI, cloud, and big data applications.

鼎新數智

Data Systems Consulting Co., Ltd offers data-driven AI solutions through its PaaS platform, METIS, combining industry know-how with generative AI to enhance analytics and decision-making. Supports both cloud and on-premises deployment.

SIGHTIFY

Sightify is a software company dedicated to providing no-code AI solutions that help businesses automate workflows while ensuring data security. Its core product, AI Agents, is an on-premise GenAI platform for managing company knowledge and generating insights or reports, deployable via private cloud, on-premises servers, and embedded systems.

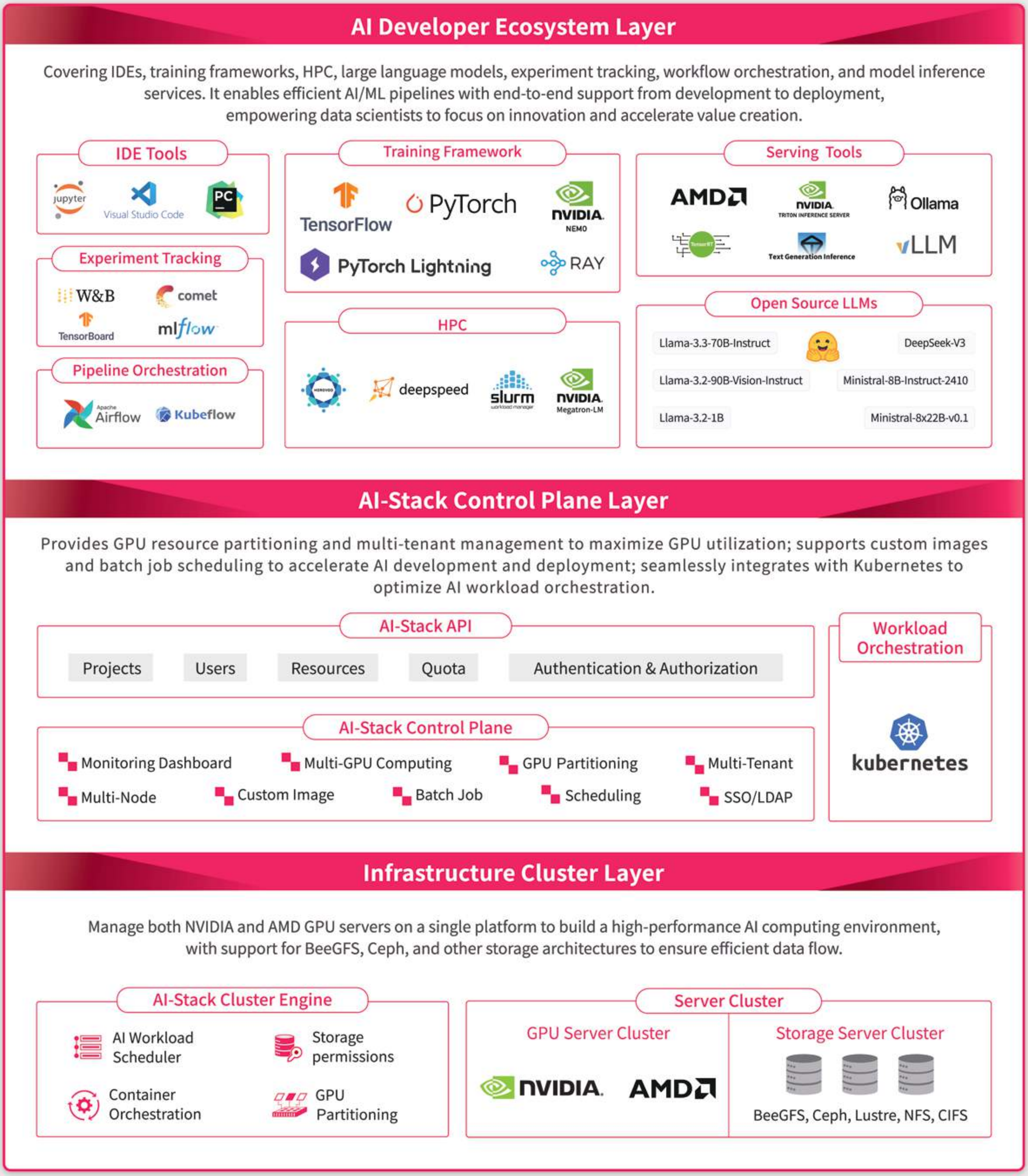
morpho

Morpho, a Japanese company, has over 20 years of experience in R&D and product development in image processing and AI. Its image enhancement, intelligent detection, and recognition are widely adopted by leading global smartphone brands, laptop manufacturers, and automotive companies.

Fulfilling all MLOps Requirements from Development to Management

AI-Stack is built for enterprises, providing comprehensive AI infrastructure solutions, covering everything from hardware management to model deployment. It maximizes GPU utilization, simplifies the AI development process, lowers the barriers to AI adoption, and accelerates digital transformation.

AI-Stack Architecture



AI Developer Ecosystem Layer

Covering IDEs, training frameworks, HPC, large language models, experiment tracking, workflow orchestration, and model inference services. It enables efficient AI/ML pipelines with end-to-end support from development to deployment, empowering data scientists to focus on innovation and accelerate value creation.

IDE Tools

jupyter, Visual Studio Code, PC

Experiment Tracking

W&B, comet, TensorBoard, mlflow

Pipeline Orchestration

Apache Airflow, Kubeflow

Training Framework

TensorFlow, PyTorch, NVIDIA NEMO, PyTorch Lightning, RAY

HPC

HOROVOD, deepspeed, slurm workload manager, NVIDIA Megatron-LM

Serving Tools

AMD, NVIDIA TRITON INFERENCE SERVER, Ollama, TensorRT, Text Generation Inference, vLLM

Open Source LLMs

Llama-3.3-70B-Instruct, DeepSeek-V3, Llama-3.2-90B-Vision-Instruct, Ministral-8B-Instruct-2410, Llama-3.2-1B, Ministral-8x22B-v0.1

AI-Stack Control Plane Layer

Provides GPU resource partitioning and multi-tenant management to maximize GPU utilization; supports custom images and batch job scheduling to accelerate AI development and deployment; seamlessly integrates with Kubernetes to optimize AI workload orchestration.

AI-Stack API

Projects, Users, Resources, Quota, Authentication & Authorization

AI-Stack Control Plane

Monitoring Dashboard, Multi-GPU Computing, GPU Partitioning, Multi-Tenant, Multi-Node, Custom Image, Batch Job, Scheduling, SSO/LDAP

Workload Orchestration

kubernetes

Infrastructure Cluster Layer

Manage both NVIDIA and AMD GPU servers on a single platform to build a high-performance AI computing environment, with support for BeeGFS, Ceph, and other storage architectures to ensure efficient data flow.

AI-Stack Cluster Engine

AI Workload Scheduler, Storage permissions, Container Orchestration, GPU Partitioning

Server Cluster

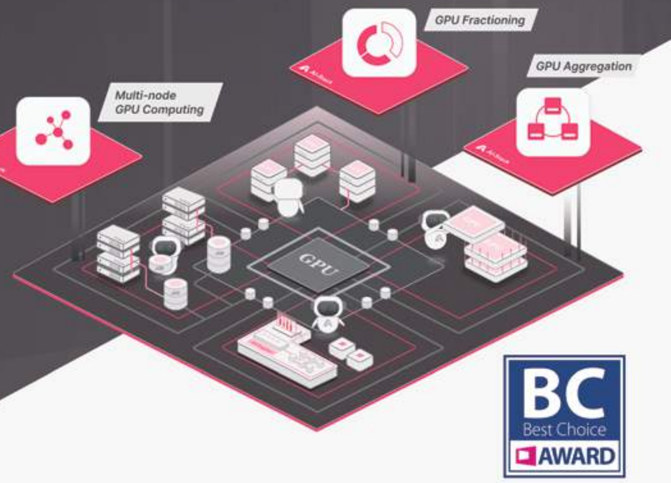
GPU Server Cluster: NVIDIA, AMD

Storage Server Cluster: BeeGFS, Ceph, Lustre, NFS, CIFS

AI-Stack

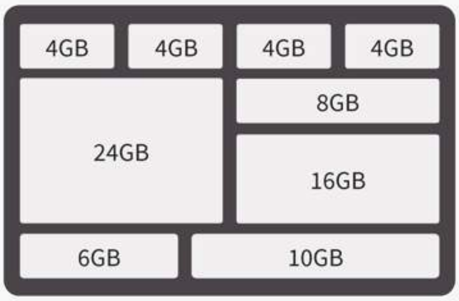
引領AI基礎設施新革命，驅動數位無限新紀元

數位無限的AI-Stack提供一站式的AI基礎設施管理解決方案，整合GPU切割技術(NVIDIA/AMD)、GPU多片聚合、跨節點運算、直覺的使用者介面、容器化與MLOps流程、開源深度學習工具、環境部署功能，是企業導入AI服務時必備的關鍵工具。



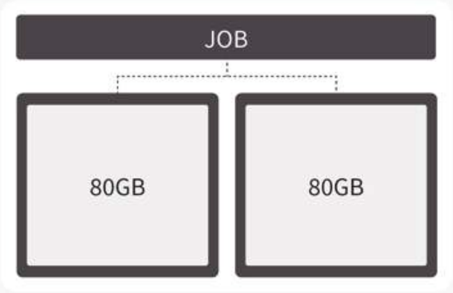
BC Best Choice AWARD

三大GPU算力管理層核心技術



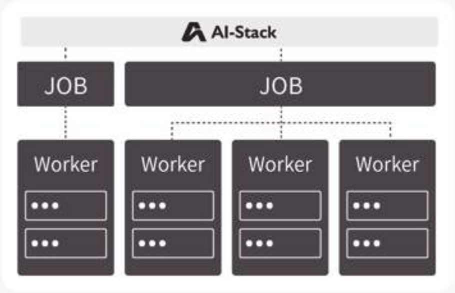
GPU單片切割技術

將GPU切分為多個區塊，滿足不同大小模型的訓練需求。提高GPU使用率至90%以上，大幅減少運算資源的浪費。



GPU 多片聚合技術

整合多張GPU的算力來應對大型AI/ML模型的訓練需求。顯著加快大型模型訓練速度，提高開發效率。



跨節點運算技術

將訓練任務分配至多個節點運算，利用分散式訓練技術，平行分散處理巨量數據，有效縮減模型訓練時間。

關鍵效益

90% ↑

GPU利用率

切割技術讓使用率
30% → 90%

10倍 ↑

工作負載運行

多人多項任務讓效率
↑ 10倍

1分鐘 ↓

開發環境建置

縮短開發建置時間
2週 → 1分鐘

10倍 ↑

提升投資效益

增加投資效益
↑ 10倍

INFINITIX

NVIDIA Partner

數位無限提供最先進的GPU資源和AI基礎設施管理平台AI-Stack，協助企業高效導入AI。2021年成為NVIDIA認證解決方案顧問 (Solution Advisor)。2024年獲「AMD GPU生態建設夥伴獎」成為全球關鍵影響力策略合作夥伴。

LinkedIn　infinitix.ai　Facebook

AI-Stack 應用場域

AI資料中心

半導體業

製造業

金融業

學術單位

能源產業

政府部門

交通產業

管理者| IT管理員

- 企業級安全防護，資源有效運用
- 專案權限管理及配額限制
- 專業儀錶板功能，即時監測GPU資源

使用者| 資料科學家與AI研究員

- 只花 1 分鐘即刻開始AI容器佈署
- 專注於研究模型開發和訓練
- 自動執行排程訓練
- 快速部署推論模型

GPU資源管理 + MLOps

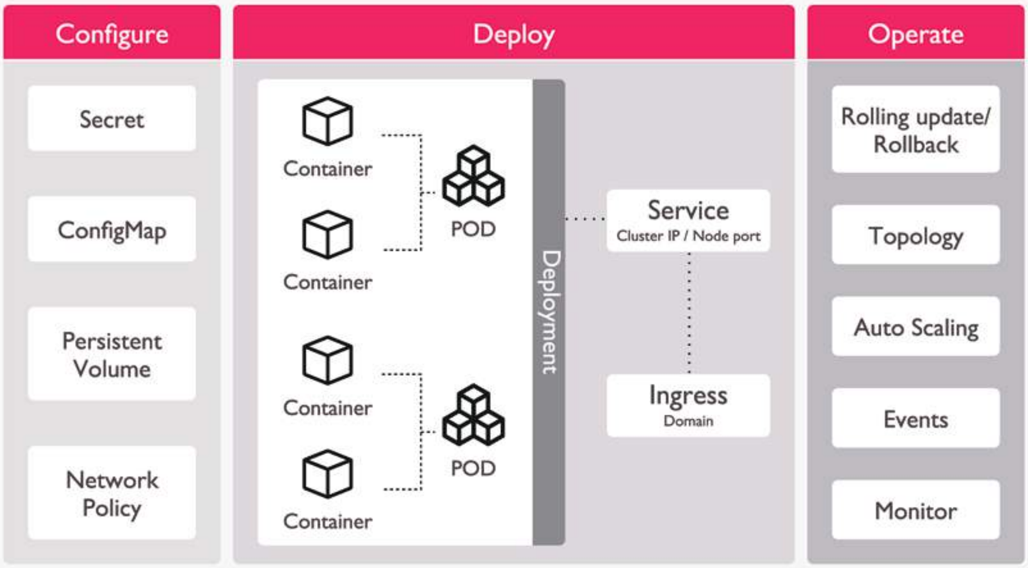
- 支援NVIDIA和AMD全系列GPU
- 更精準、有效管理 GPU資源
- 無縫對接常用 AI 開發工具
- 完整儀錶板，即時監測 GPU算力資源使用情況
- 模型推論快速部署服務
- 自動執行，預設單次或批次任務
- 直覺的使用者介面，輕鬆上手
- 快速容器化環境部署功能與MLOps管理工具
- 資源隔離、權限管理、配額限制

快速部署服務 Rapid Container Service(RCS)

基於Kubernetes架構，專為模型推論服務與AI應用服務設計，幫助企業快速部署AI各類服務，管理及擴展AI應用。

RCS功能優勢：

- 快速部署
- 即時監控
- 高彈性擴展
- 高效版本管理



伺服器合作夥伴 & 經銷代理

AMD

NVIDIA

intel

MACNICA

ZERONE

敦新科技
DAWNING TECHNOLOGY INC.

硬體合作夥伴

DELL Technologies

戴爾科技集團匯聚專才、技術、數據與協作機制，透過創新方法協助企業提升客戶與員工體驗。憑藉全方位智慧解決方案，企業可實現簡單、安全、敏捷的創新，加速競爭力提升，掌握未來發展先機。

Hewlett Packard Enterprise

慧與科技是一家全球性的邊緣到雲端、平台即服務公司，以全方位的技術平台，在所有的雲端和邊緣提供一致的體驗，協助客戶開發新的業務模式。藉由雲端運算、資料中心與工作場所解決方案，協助客戶更快將理念轉化為商業價值。

msi

MSI是全球領先的遊戲、內容創作、商業與生產力以及人工智慧物聯網解決方案提供者。MSI的伺服器產品完全為內部開發，注重設計和製造，體現了公司對滿足客戶需求和適應市場需求的承諾。

COMPAL

仁寶深耕高效能運算領域，打造兼具效能、效率與永續性的企業級解決方案。無論是AI資料中心、雲端平台、高密度運算環境，仁寶伺服器都能提供靈活、高效、可靠的基礎架構，協助企業快速應對未來運算需求，掌握數位轉型契機。

MiTAC MiTAC Computing

神雲科技股份有限公司為神達控股集團旗下子公司，憑藉自1990年代以來的深厚產業經驗，提供多元、節能高效的伺服器解決方案。神雲科技為超大規模數據中心、HPC及 AI應用，提供量身打造的解決方案，確保最佳效能與高擴展性。

Graid Technology Inc.

圖睿科技創新GPU RAID技術，專為 NVMe SSD量身打造，突破傳統磁碟陣列限制，釋放 CPU資源，帶來前所未有的計算效能！致力打造大數據邊緣到雲端的全方位解決方案，開啟新一代儲存革命！

HeTone Group

合通企業專注於新一代液冷技術，包含直接接觸式(DTC)與浸沒式冷卻，專為高效能運算設計。解決方案可有效降低能耗與成本，提升運算效能，特別適用於AI、雲端與大數據等領域。

鼎新數智

鼎新數智是領先的數據和智能方案提供商。創新研發數智驅動PaaS平台METIS，結合各行業知識經驗與生成式AI技術，以「數據自決」與「智能生成」驅動數據分析、預測與決策優化，企業可依需求彈性選擇雲端或地端AI部署方案。

SIGHTIFY

視旅科技是一家致力於提供無需程式能力即可使用的AI解決方案的軟體公司，協助企業自動化工作流程，同時確保資料安全。主要產品包括 AI Agents，一個可管理企業知識並產生洞察或報告的地端版生成式 AI 平台。提供多元的部署選項，包含私有雲、本地伺服器與嵌入式系統等。

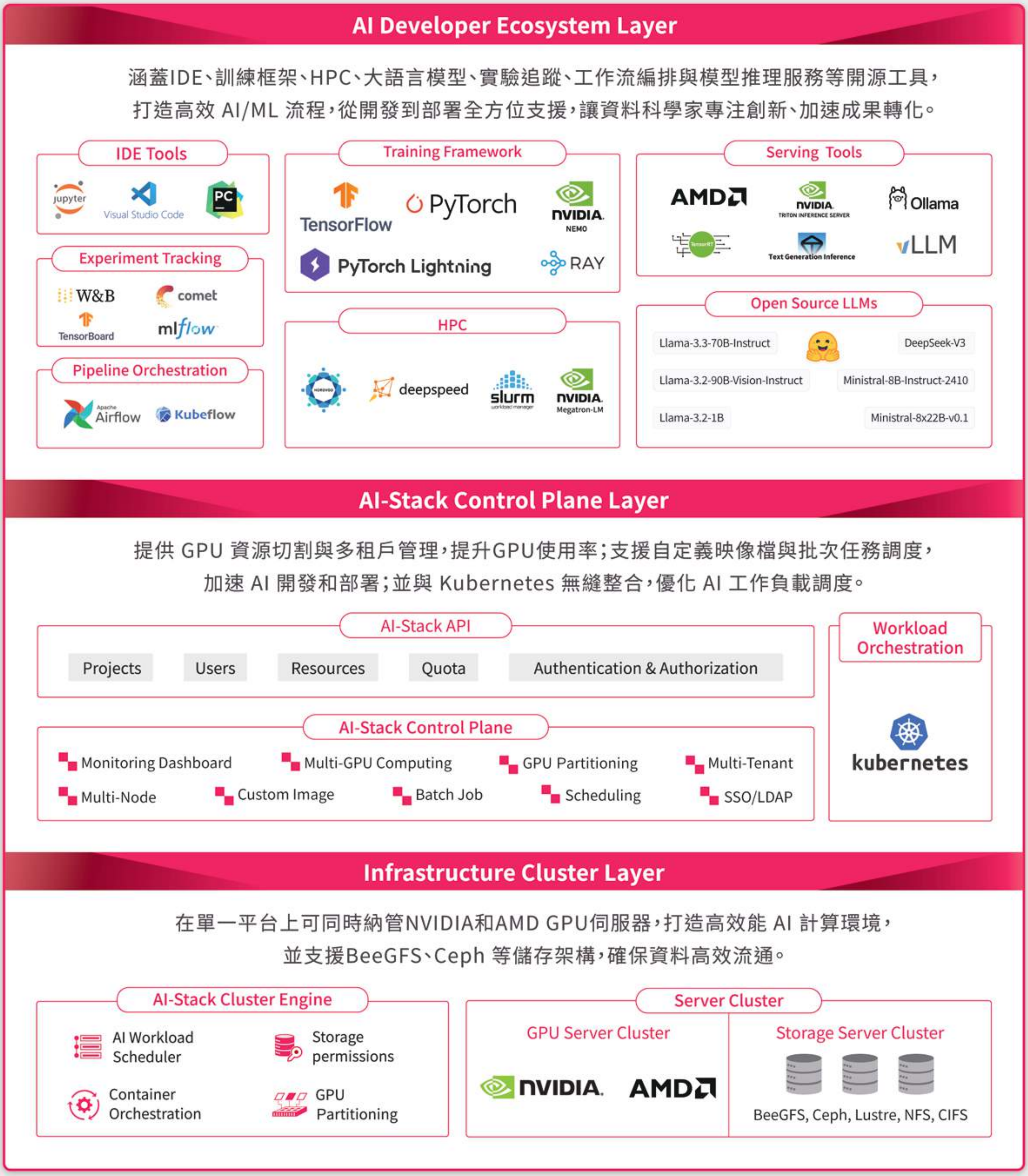
morpho

日商Morpho在影像處理與人工智慧領域擁有超過20年的研發與產品開發經驗。其影像優化、智能檢測與辨識等核心技術，已廣泛應用於多家國際知名的手機品牌、筆記型電腦製造商及汽車廠商。

全方位滿足開發到營運管理的各種MLOps需求

AI-Stack 專為企業打造，提供全方位的AI基礎設施解決方案，從硬體管理到模型部署一應俱全，提升GPU使用率，簡化AI開發流程，協助企業降低 AI導入門檻，加速數位轉型。

AI-Stack 架構



AI Developer Ecosystem Layer

涵蓋IDE、訓練框架、HPC、大語言模型、實驗追蹤、工作流編排與模型推理服務等開源工具，打造高效 AI/ML 流程，從開發到部署全方位支援，讓資料科學家專注創新、加速成果轉化。

IDE Tools

jupyter · Visual Studio Code · PC

Experiment Tracking

W&B · comet · TensorBoard · mlflow

Pipeline Orchestration

Apache Airflow · Kubeflow

Training Framework

TensorFlow · PyTorch · NVIDIA NEMO · PyTorch Lightning · RAY

HPC

HOROVOD · deepspeed · slurm workload manager · NVIDIA Megatron-LM

Serving Tools

AMD · NVIDIA TRITON INFERENCE SERVER · Ollama · TensorRT · Text Generation Inference · vLLM

Open Source LLMs

Llama-3.3-70B-Instruct		DeepSeek-V3
Llama-3.2-90B-Vision-Instruct		Ministral-8B-Instruct-2410
Llama-3.2-1B		Ministral-8x22B-v0.1

AI-Stack Control Plane Layer

提供 GPU 資源切割與多租戶管理，提升GPU使用率；支援自定義映像檔與批次任務調度，加速 AI 開發和部署；並與 Kubernetes 無縫整合，優化 AI 工作負載調度。

AI-Stack API

Projects · Users · Resources · Quota · Authentication & Authorization

AI-Stack Control Plane

- Monitoring Dashboard
- Multi-GPU Computing
- GPU Partitioning
- Multi-Tenant
- Multi-Node
- Custom Image
- Batch Job
- Scheduling
- SSO/LDAP

Workload Orchestration

kubernetes

Infrastructure Cluster Layer

在單一平台上可同時納管NVIDIA和AMD GPU伺服器，打造高效能 AI 計算環境，並支援BeeGFS、Ceph 等儲存架構，確保資料高效流通。

AI-Stack Cluster Engine

- AI Workload Scheduler
- Storage permissions
- Container Orchestration
- GPU Partitioning

Server Cluster

GPU Server Cluster: NVIDIA · AMD

Storage Server Cluster: BeeGFS, Ceph, Lustre, NFS, CIFS