INFINITIX

VST ECS

AI-Stack Introduction

Dec 2025

Thailand Representative

- **Somrath Tojumroen (ART)**
- **Mobile : 0888999692**

We are

© INFINITIX 2024 All rights reserved.

Company Overview

Information Software Services and Management Consulting

Computer and Office Equipment Retail

Capital
1M USD

Established
February 2003

No. of Emps
50

Headquarters: Taiwan (Taipei, Kaohsiung)
Subsidiaries: Japan, China, Hong Kong, Singapore

© INFINITIX 2024 All rights reserved.

About INFINITIX

To infinity and beyond AI.

INFINITIX took the lead in 2017 by venturing into the field of AI GPU resource scheduling and AI infrastructure management platforms, achieving many successful projects. Amid GPU shortages and rising prices, we helped companies increase GPU ROI by 10 times. Staying ahead of AI trends, INFINIT became a global partner of NVIDIA's Inception Program in 2019. **We are also the only company in Taiwan's AI software space to earn the rare NVIDIA Solution Advisor certification.**

Our core product, AI-Stack, is a leading **AI infrastructure management platform**, essential for businesses adopting AI services. It integrates third-generation GPU partitioning, GPU aggregation, multi-node computing, a new user interface, multi-cloud management, open-source deep learning tools, and environment deployment features, helping enterprises tackle the challenges brought by fast-paced AI advancements.

© INFINITIX 2024 All rights reserved.

Shareholders and Accelerators

台達 DELTA

仁寶電腦 COMPAL

ZERONE 零壹科技

ITIC Industrial Technology Investment Corporation

SparkLabs TAIPEI

StarFab accelerator

© INFINITIX 2024 All rights reserved.

Distributors and Partners

Americas

VOCOM

Europe / China

HeTone

Japan

MACNICA

kiwi technology Inc.

Toontec Solutions

Taiwan

ZERONE 零壹科技

Stark Technology Inc. 敦陽科技股份有限公司

Enlight 翔耀實業股份有限公司

UNIFOSA

MACNICA

APEX FONG YI TECHNOLOGY CO., LTD. 峰儀科技股份有限公司

敦新科技 DAWNING TECHNOLOGY INC.

豐康科技 www.fongcon.com.tw

tsti 大世科

無敵 BESTA

Southeast Asia

FLEXINFRA

VST ECS

INFINITIX

Server Partner

NVIDIA, AMD, DELL EMC, SUPERMICRO, Hewlett Packard Enterprise, PEGATRON, GIGABYTE, COMPAL, MSI, MiTAC

Storage Partner

MINIO, DELL EMC, GRAID

Edge Partner

ADLINK LEADING EDGE COMPUTING, MACNICA, aetina

AI Application Partner

PEGATRON 和碩聯合科技, Sightour, PROFET AI Realize AI within one week!!, POWER ARENA, is-land, MUEN, 科智企業, LEDA TECHNOLOGY

AI-Stack Introduction

1 Deploy Ai 10X Faster

2 Innovative GPU Solution for Manufacturing

3 Ai-Stack Admin demo

© INFINITIX 2024 All rights reserved.

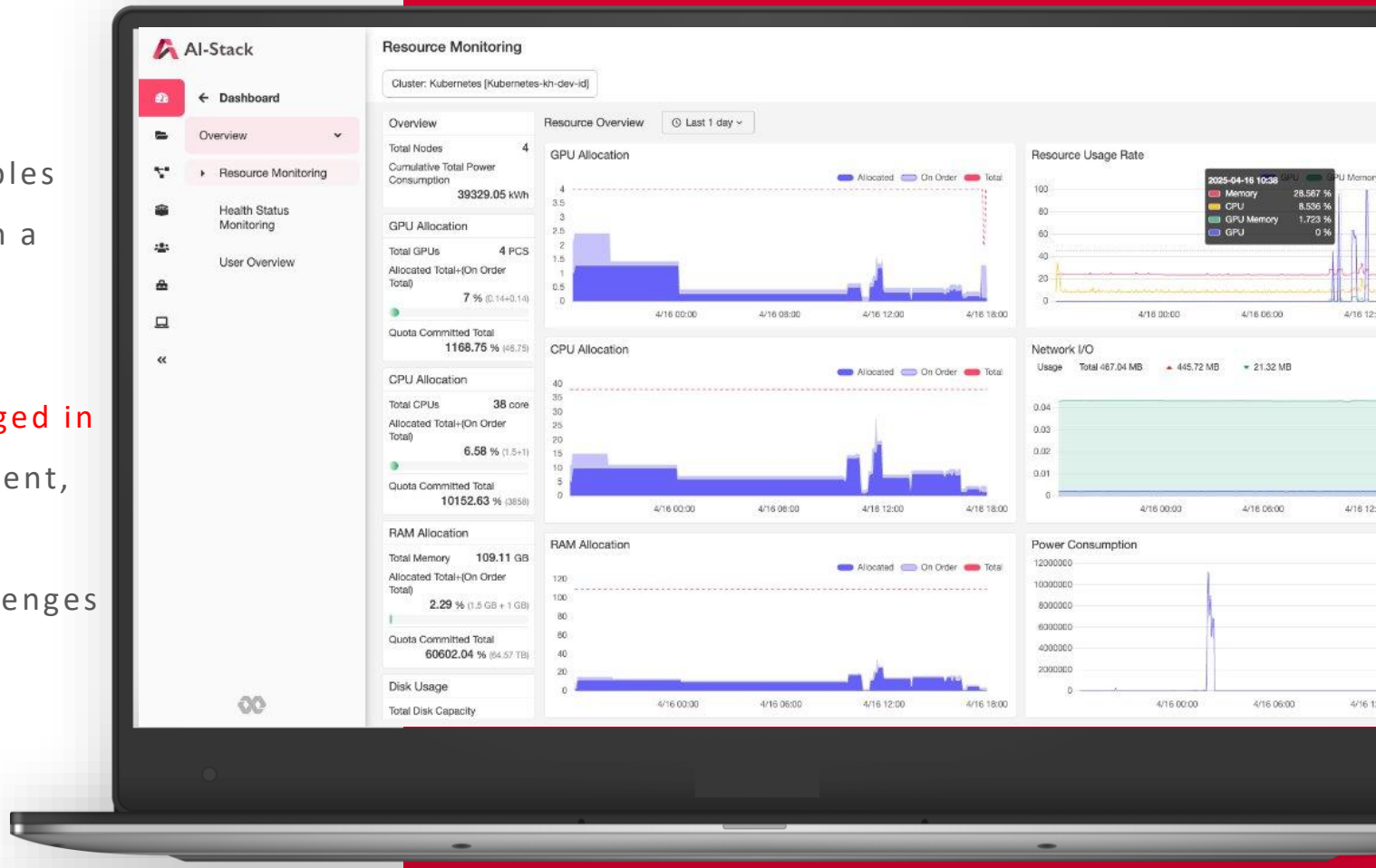
AI-Stack

AI Infrastructure and GPU Resource Scheduling Platform

AI-Stack is an **AI GPU resource scheduling and AI infrastructure management platform** that offers enterprises flexible and scalable solutions. It enables the rapid implementation of various AI services on a single platform while addressing development and deployment challenges.

AI-Stack is an essential core for enterprises engaged in AI development, integrating multi-cloud management, open-source deep learning frameworks, and infrastructure. It helps businesses tackle the challenges posed by advancing AI technology and optimize resources amid rising GPU costs, earning NVIDIA - Solution Advisor certification.

NVIDIA Partner



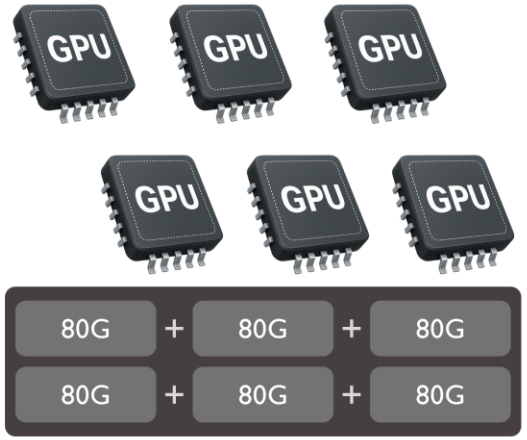
INFINITIX

AI Evolution: From Foundation to Autonomy

Foundational Stage
Training & Fine-Tune

1

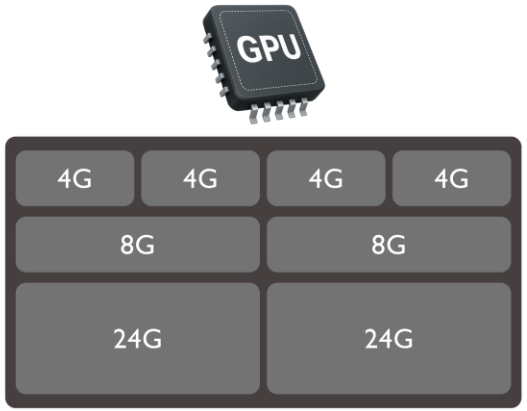
- Requiring significant GPU resources
- Utilizing GPU aggregation and elastic training for scalability



Application Stage
Inference & Predict

2

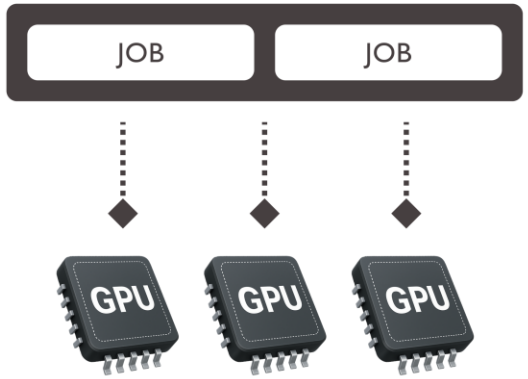
- Executing with minimal GPU resource
- Utilizing GPU partitioning to optimize resource allocation for lightweight model inference



Creative Stage
Gen AI

3

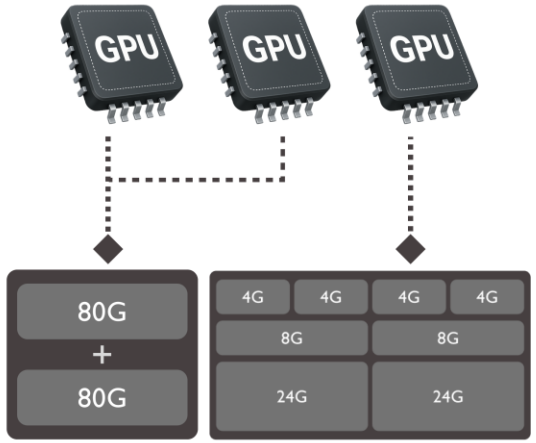
- Utilizing Normalization GPU Resource
- GPU Partitioning for balanced performance



Autonomous Stage
Agent AI

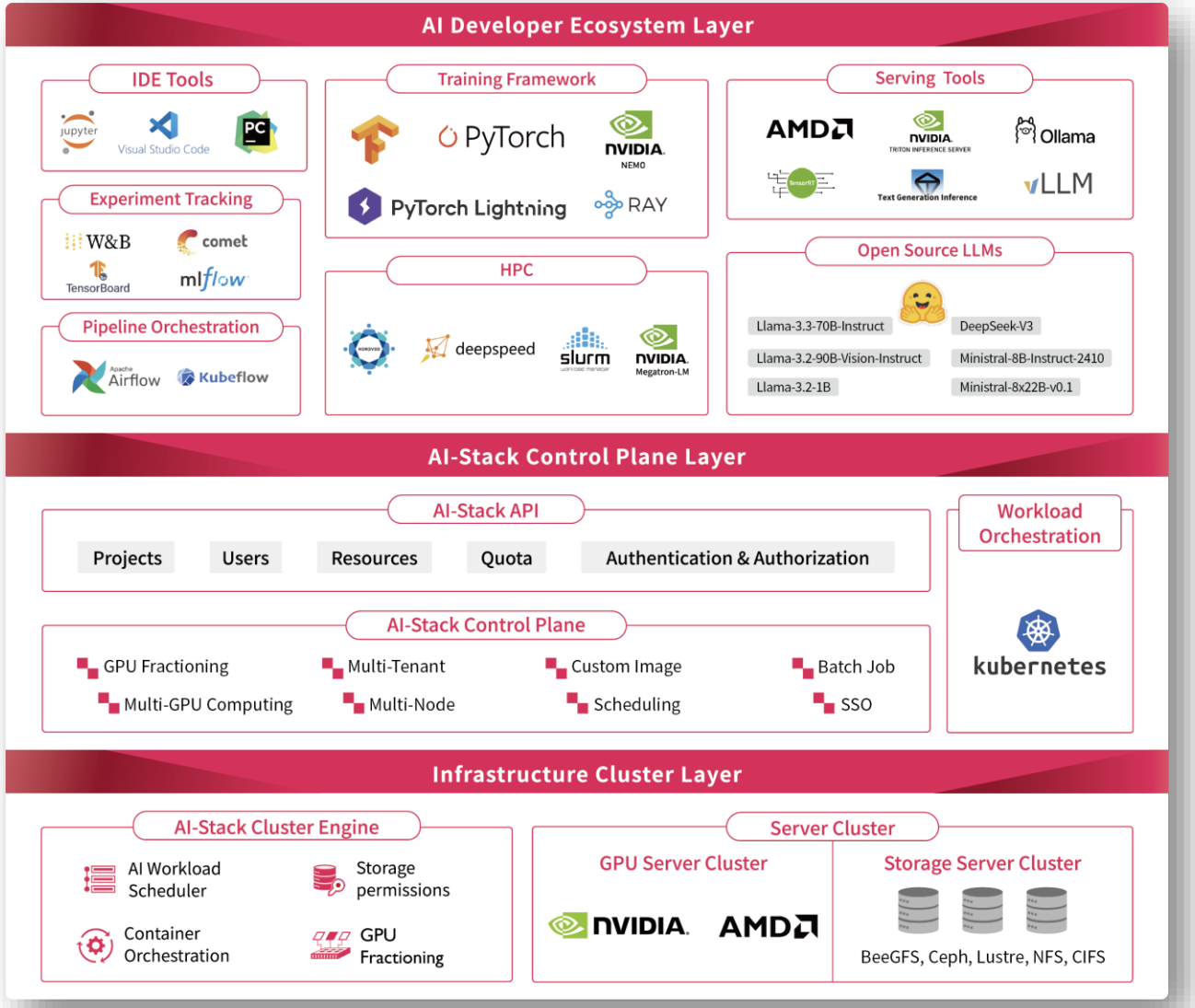
4

- Employing GPU Hybrid Allocation：GPU Partitioning and Aggregation
- With optimal GPU Scheduling Strategy



© INFINITIX 2024 All rights reserved.

AI-Stack Layered Architecture



User

Data Scientists and AI Researchers

- ✓ Focus on research and model training rather than DevOps
- ✓ Start AI container deployment in just 1 minute

Administrators

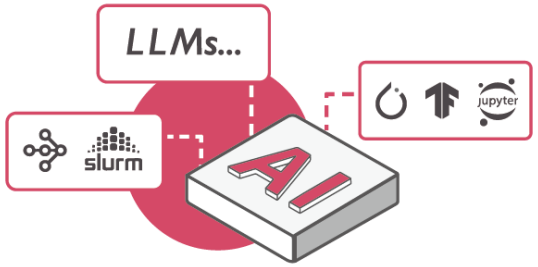
IT Administrators

- ✓ Enterprise-level security protection, efficient resource utilization
- ✓ Easy monitoring and comprehensive resource management

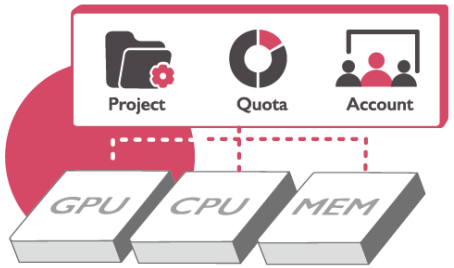
© INFINITIX 2024 All rights reserved.

AI-Stack provides the core infrastructure for AI development

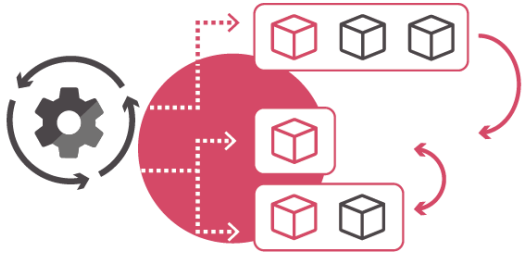
GPU Resource Management + MLOps



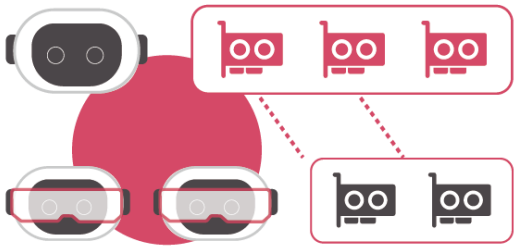
Seamless connect with commonly used AI development tools



Resource isolation, authority management and quota limitations



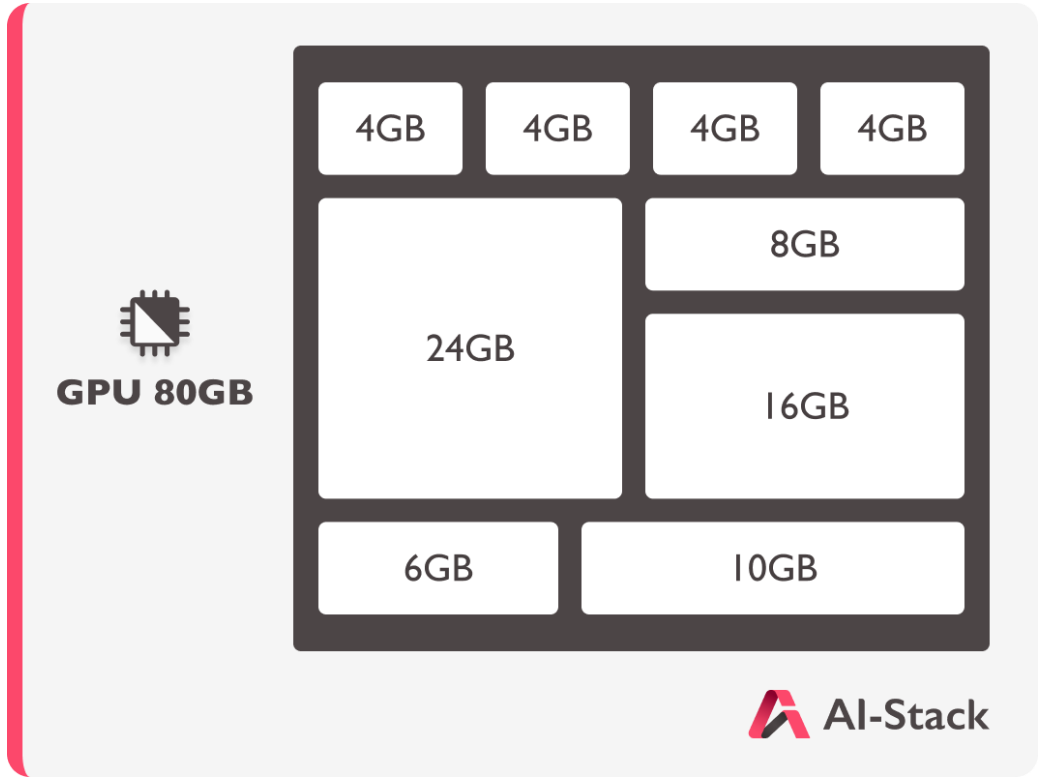
Automated execution, with options for single or batch tasks

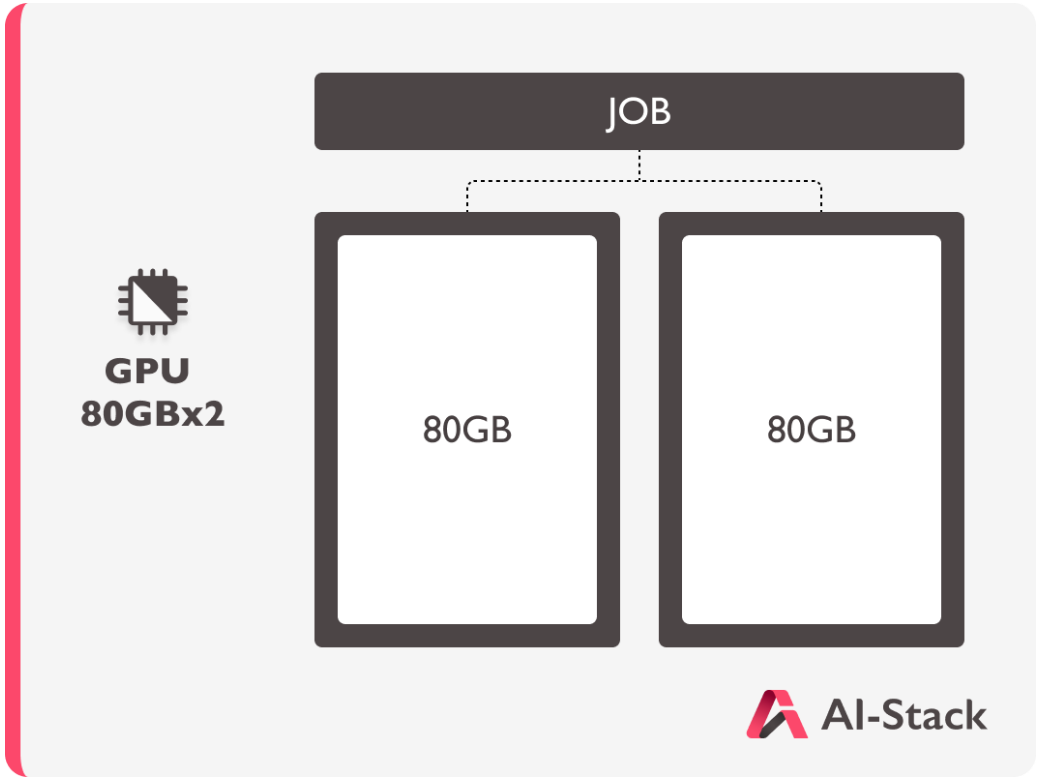


Help teams manage GPU resources more precisely and efficiently

© INFINITIX 2024 All rights reserved.

AI-Stack: GPU Partitioning and Aggregation Technology for Maximum Efficiency





Utilization rate ↑

One GPU → Multi GPU Fractions

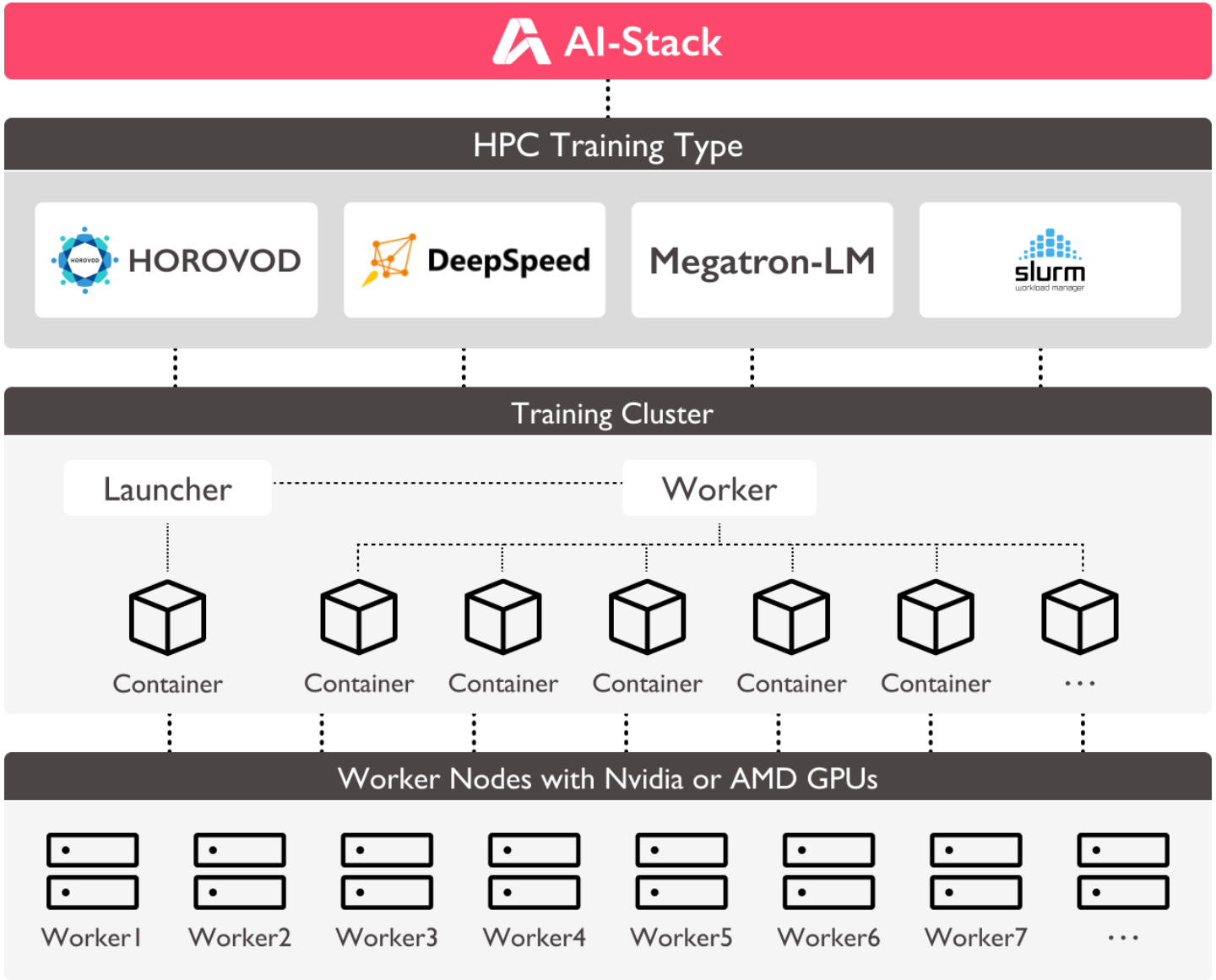
Computing performance ↑

Multi GPUs → One GPU Aggregation

© INFINITIX 2024 All rights reserved.

AI-Stack High-performance computing, HPC

The platform can **allocate training tasks across multiple nodes** based on demand and leverage **distributed training** technology to organize multiple containers into training groups. This enables **parallel and distributed processing** of massive datasets, effectively reducing model training time, enhancing computational efficiency, and optimizing resource utilization.



© INFINITIX 2024 All rights reserved.

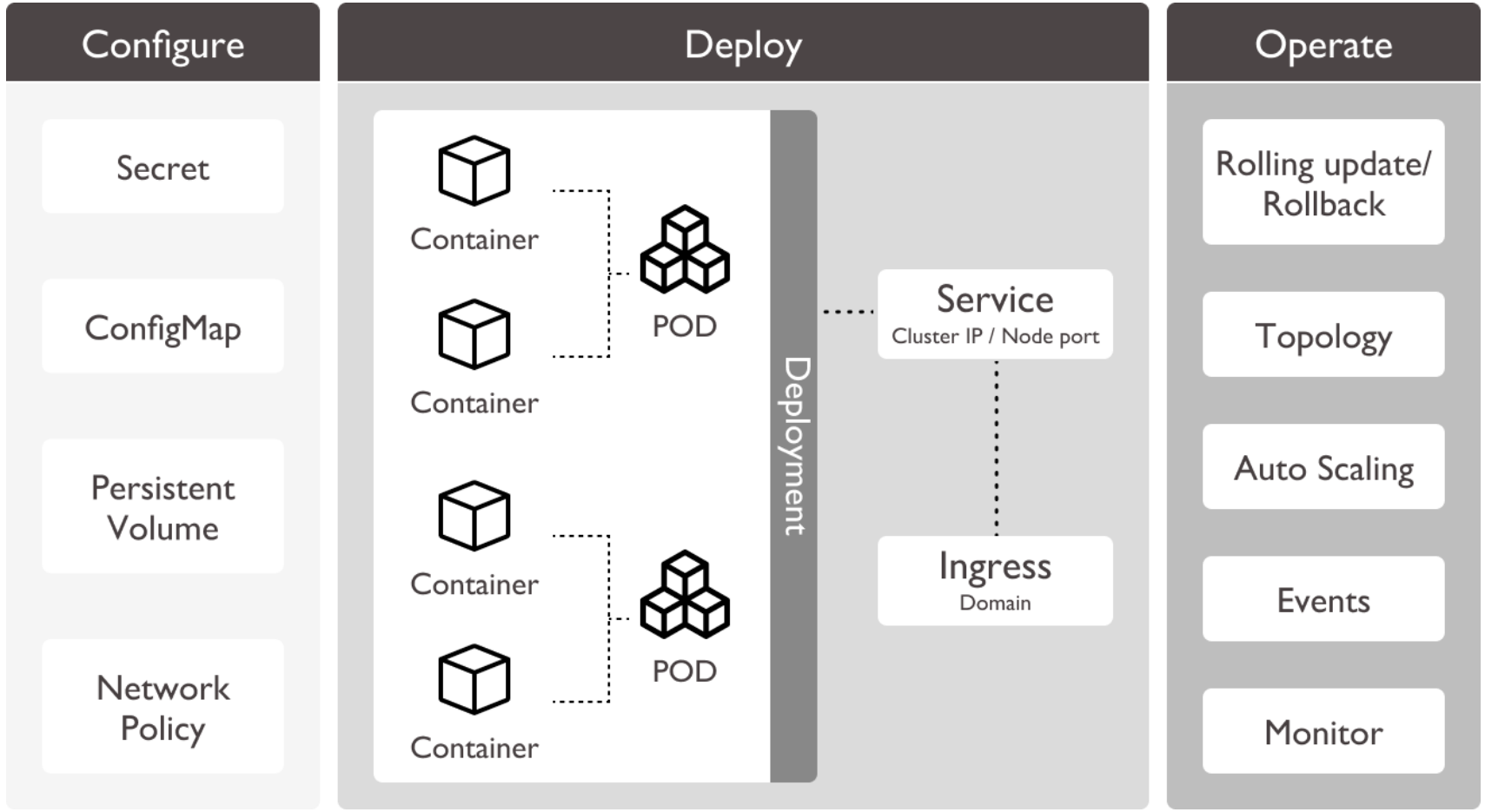
AI-Stack RCS

Rapid Container Service

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

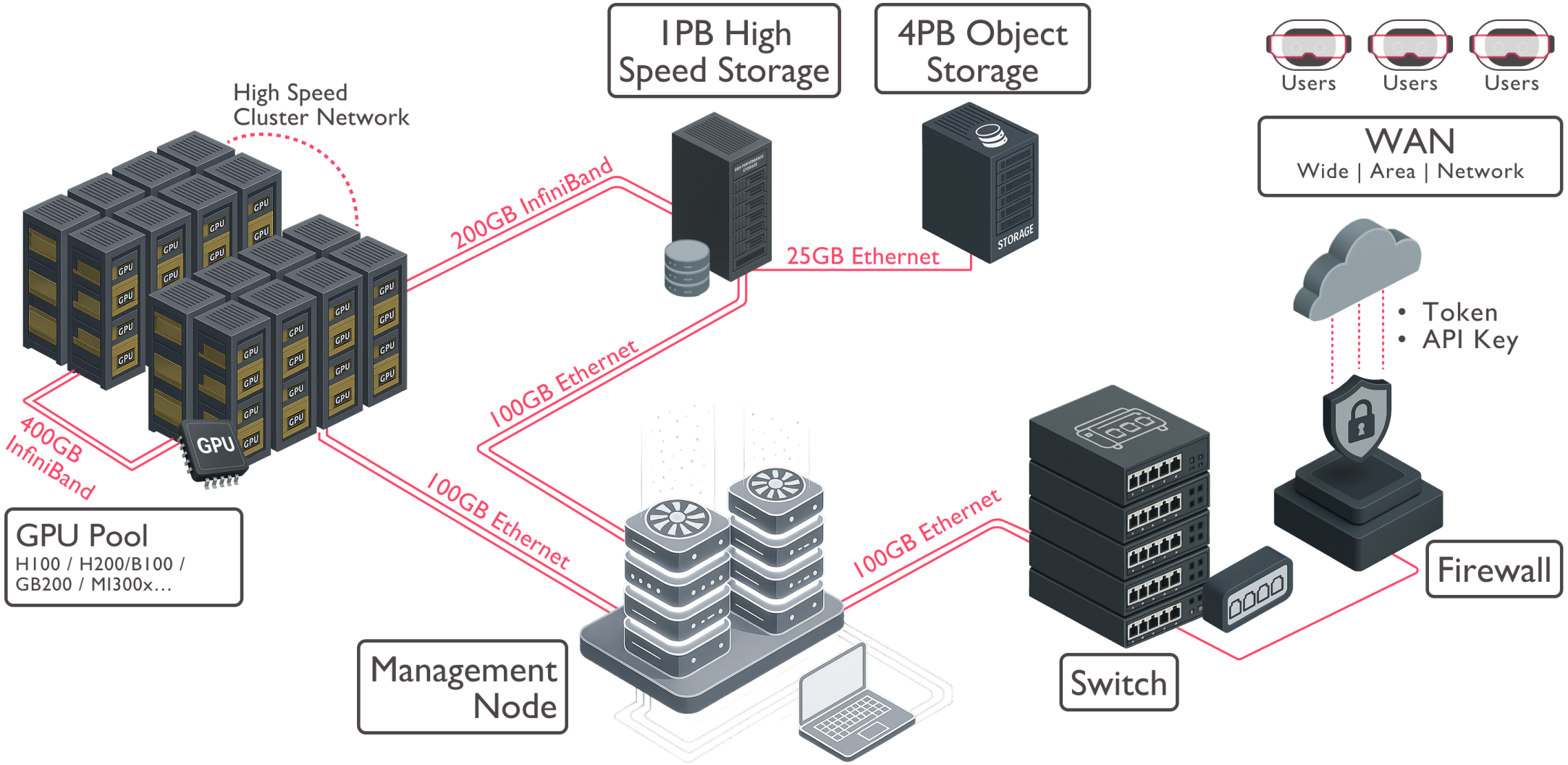
Key Advantages of RCS:

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



© INFINITIX 2024 All rights reserved.

High-Performance Computing Architecture



© INFINITIX 2024 All rights reserved.

AI-Stack Data Center

Comprehensive Computing Center Solution

AI-Stack Data Center

Developer

- API Catalog
- Resource Management (Images, Models)

Administrator

- Billing, Cost Analysis
- Access Management
- Account Management

Operator

- Resource Allocation
- Dashboard
- Health Monitoring, Alert system
- Power Consumption and Carbon Emissions Monitoring
- Usage Data Analysis
- High Availability, Disaster Recovery

© INFINITIX 2024 All rights reserved.

AI-Stack enhances GPU efficiency when helping enterprises implement AI

90% ↑

GPU Utilization

10x ↑

Workload execution

1min ↓

Development Environment Setup

10x ↑

Enhanced ROI

GPU Utilization

GPU Partitioning Increases Utilization 30% → 90%

Workload execution

Multiple users and tasks increase efficiency by 10x

DevEnv Setup

Reduces setup time from 2 weeks → 1 min

Enhanced ROI

Boosts return on investment by 10x

© INFINITIX 2024 All rights reserved.

AI-Stack Application Industries

Semi-conductor

UMC
Himax
fitipower
SiS
REALTEK

Manufacturing

中鋼公司
台塑企業 FORMOSA PLASTICS GROUP
DELTA

Finance Service

永豐金控 SinoPac Holdings
iPASS 一卡通

Academic

中央研究院 ACADEMIA SINICA
NATIONAL CHENG KUNG UNIVERSITY 1931
1896
TAIPEI TECH
I-SHOU UNIVERSITY

Energy

台灣電力公司

Government

NARLabs 國家實驗研究院
國家高速網路與計算中心
National Center for High-performance Computing
moda 數位發展部 Ministry of Digital Affairs
交通部中央氣象局 CENTRAL WEATHER BUREAU

Medical

Transportation

Taoyuan Airport 桃園機場
Metro Taipei
metro Taoyuan

Over Sea

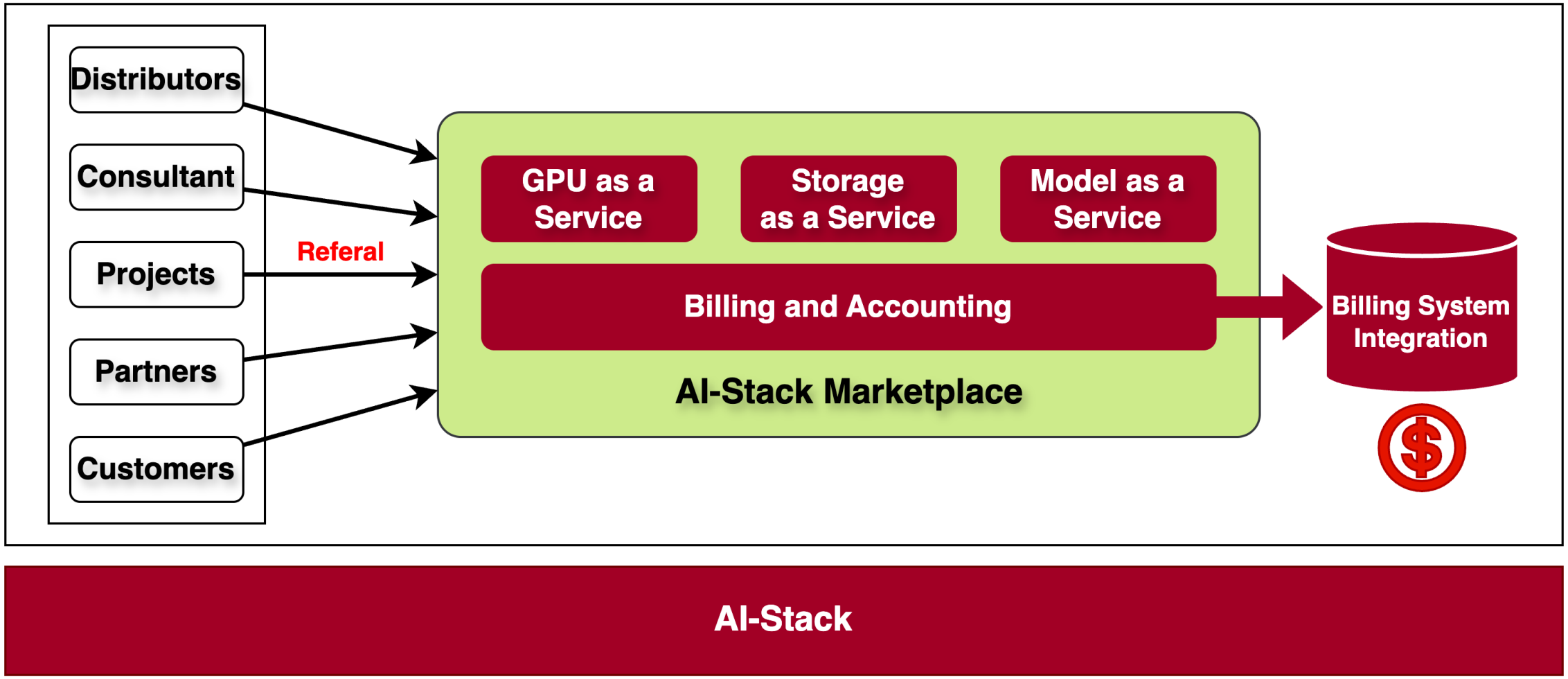
TEL
HITACHI Inspire the Next
MITSUBISHI
UNION TOOL
maxis

© INFINITIX 2024 All rights reserved.

Cooperate Model

© INFINITIX 2024 All rights reserved.

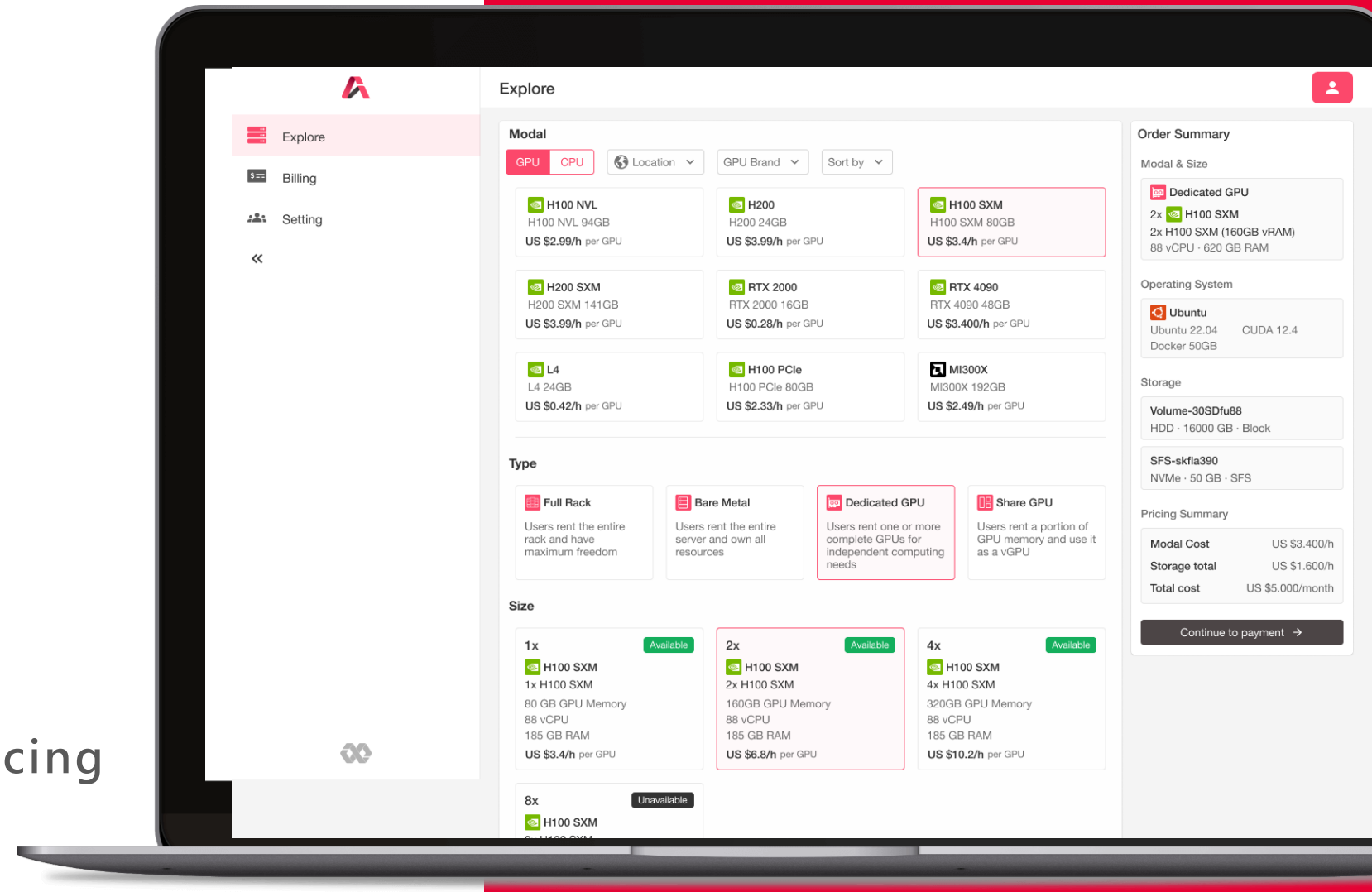
AI-Stack - The best cooperative partner of AIDC



© INFINITIX 2024 All rights reserved.

Tenant

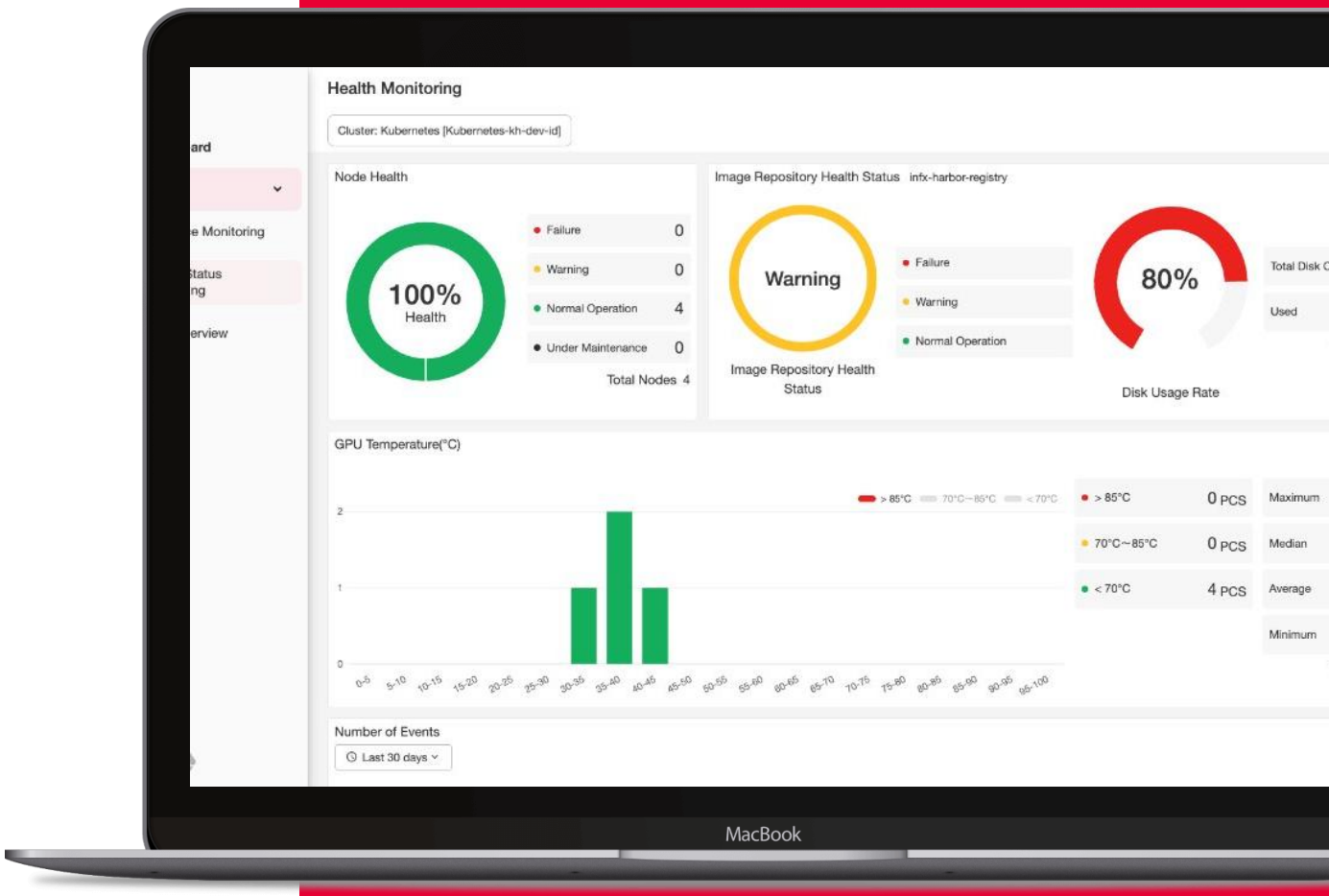
- Full Rack
- Bare Metal
- Dedicated GPUs
- Shared GPUs
- Storage Rental
- Professional Software Services
- Flexible Pricing, Spot Pricing



INFINITIX

Operators

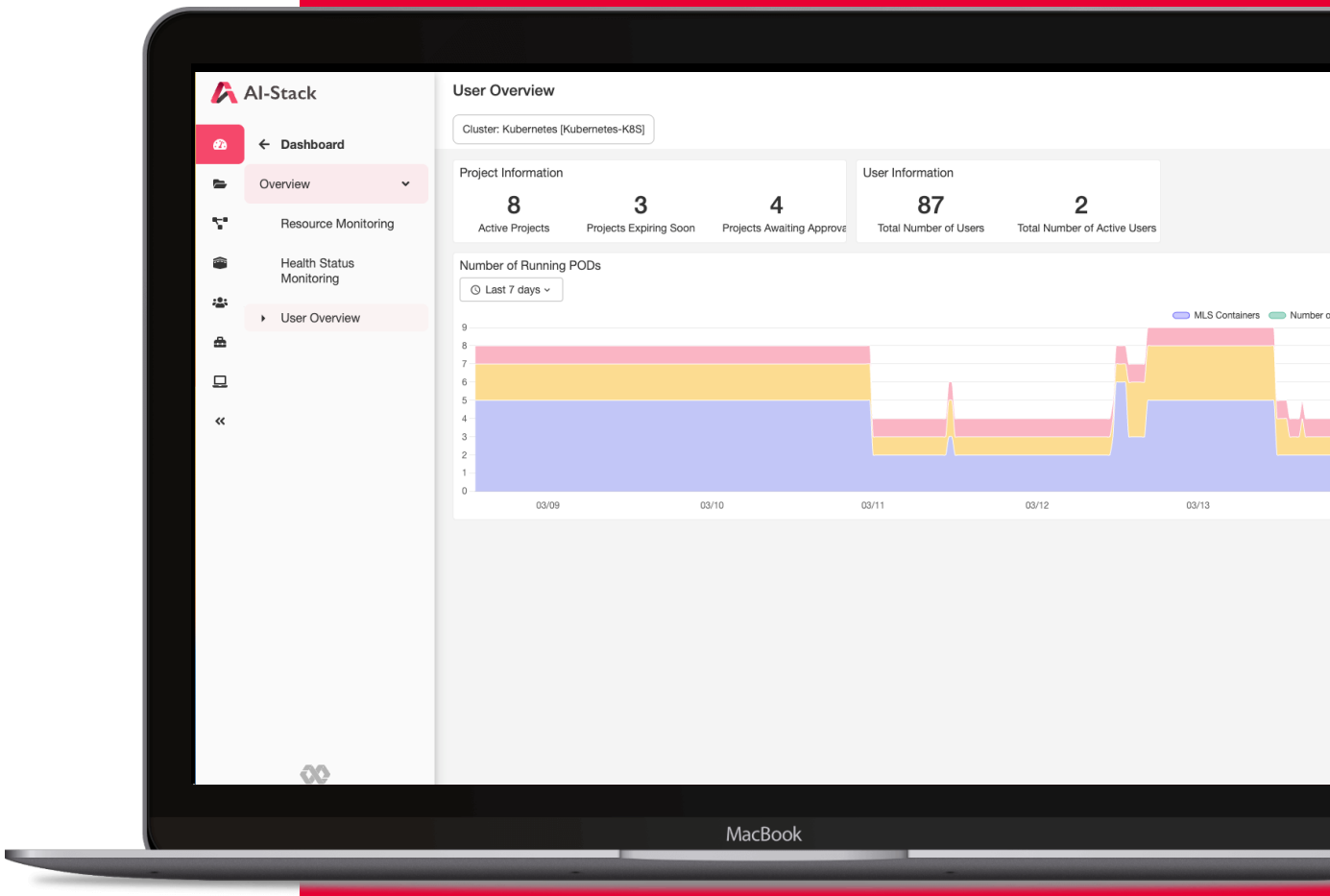
- Setting resource specifications
- Creating an AIDC Management Platform (Multiple)
- Publish Service to Marketplace



INFINITIX

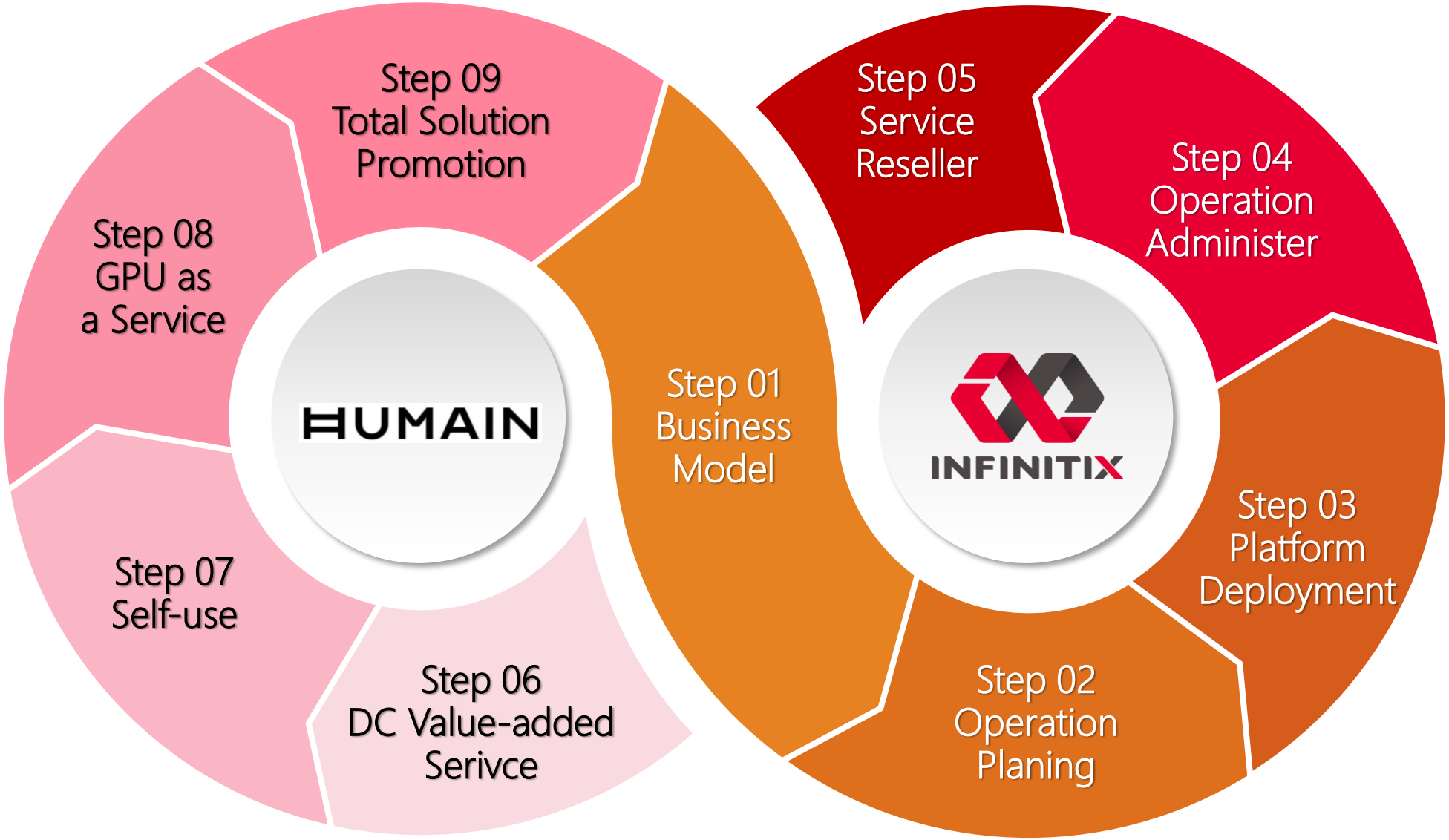
LLM Service

- Guardrail Service
- Quantitative Tools
- Performance Testing
- Automatic Image Packaging
- Rapid Deployment (NIM/HuggingFace)
- Proxy Server
- GPU Slicing for AI Models



INFINITIX

Cooperation Framework



© INFINITIX 2024 All rights reserved.

Summarize : AI-Stack

INFINITIX

ข้อสรุปความเหมือนและความแตกต่าง : Strong points
ความเหมือน:
ทั้งสองแพลตฟอร์มมี Core Competency เดียวกัน คือการใช้ Kubernetes เพื่อจัดการ, แบ่งส่วน (Fractional/Partitioning), จัดตารางงาน, และรวมทรัพยากร GPU เพื่อเพิ่มประสิทธิภาพการใช้งานในงาน AI/ML อย่างสูงสุด
ความแตกต่าง:
Run.ai: มุ่งเน้นการเป็น AI Resource Orchestration Layer ที่ทรงพลังและยืดหยุ่น โดยมีชื่อเสียงในระดับโลกและได้รับการรับรองจาก NVIDIA อย่างเป็นทางการ (ปัจจุบัน Run.ai ถูกควบรวมโดย NVIDIA)

Ai-Stack :
Infinitix Ai-Stack: เน้นการนำเสนอโซลูชันที่ เร็วและใช้งานง่าย กว่า (No-Code UI, 1-Minute Setup) มุ่งเน้นไปที่การลดภาระงานของทีม IT และเร่งให้ทีม R&D สามารถเริ่มงานได้ทันที เหมาะสำหรับองค์กรที่อาจต้องการความรวดเร็วในการเริ่มต้นและเน้นความครบถ้วนของ MLOps ในแพลตฟอร์มเดียว

Note: If you are considering using:หากคุณกำลังพิจารณาเลือกใช้:
Run.ai อาจเป็นตัวเลือกที่ดีกว่าหากคุณมี MLOps Team ที่เชี่ยวชาญ และต้องการ ความยืดหยุ่น ในการปรับแต่ง Workflow และมีคลัสเตอร์ GPU ขนาดใหญ่หลายโหนด
Infinitix Ai-Stack อาจเป็นตัวเลือกที่ดีกว่าหากคุณต้องการ โซลูชันแบบเบ็ดเสร็จ (Turn-Key Solution) ที่ ติดตั้งและเริ่มต้นใช้งานได้รวดเร็ว โดยเน้นที่ความง่ายในการบริหารจัดการทรัพยากร GPU ภายในองค์กร

© INFINITIX 2024 All rights reserved.

TOR : Thai Version

INFINITIX

ขอบเขตของงาน (TOR)

AI-Stack

AI STACK – ข้อกำหนดทางเทคนิค

เอกสารฉบับนี้ระบุข้อกำหนดด้านการทำงานและข้อกำหนดทางเทคนิคขั้นต่ำสำหรับการติดตั้งและใช้งาน AI STACK แพลตฟอร์มการจัดการทรัพยากร (Resource Management Platform) เพื่อบริหารจัดการทรัพยากร GPU/CPU/Memory, งานแบบ Containerized Workloads และบริการคอมพิวติ้งด้าน AI ในสภาพแวดล้อมที่มีความเสถียรสูง และสามารถทำงานได้ตลอดเวลาโดยที่ไม่เกิดการล่มหรือ Downtime. (High Availability)

1. การขออนุมัติและการสมัครโครงการ (Project Application and Approval)

- แพลตฟอร์มต้องสามารถรองรับการสมัคร ขออนุมัติ การเปิดใช้งาน และการยกเลิกโครงการ (Project) ได้
- เมื่อโครงการได้รับอนุมัติแล้ว ต้องสามารถกำหนดค่าพารามิเตอร์การใช้งาน GPU ได้
- การสมัครที่อยู่ในขอบเขตการกำหนด สามารถดำเนินการต่อได้โดยไม่ต้องผ่านการอนุมัติจากผู้ดูแลระบบ (Admin)

2. การจัดการโครงการและผู้ใช้งาน (Project and User Management)

- ผู้ดูแลระบบสามารถสร้าง แก้ไข และลบเนื้อหาโครงการและสมาชิกได้
- ต้องมีฟังก์ชันการลบโครงการแบบกลุ่ม (Batch Deletion)
- ผู้ดูแลระบบสามารถกำหนดข้อจำกัดการใช้ GPU บางรุ่นสำหรับโครงการ และผู้ใช้โครงการจะเห็นเฉพาะ GPU รุ่นที่ได้รับอนุญาตเท่านั้น
- สามารถรองรับผู้ดูแลระบบหลายระดับ
- สามารถการบันทึกประวัติการใช้งานของผู้ใช้และผู้ดูแลระบบในทุกการดำเนินการ
- ผู้ดูแลระบบสามารถปรับการจัดสรรทรัพยากร (GPU/CPU/RAM Quota) สำหรับแต่ละโครงการได้

3. การจัดการ Image และการปรับใช้ (Image and Deployment Management)

- รองรับการดึงอัตโนมัติ (Automatic Image Pulling) จาก NVIDIA GPU Cloud (NGC) เพื่อให้ผู้ดูแลสามารถสร้างแม่แบบบริการ (Service Templates) ได้
- รองรับการนำเข้า Image จาก NGC และ Docker Hub
- ผู้ดูแลระบบสามารถเลือกการเข้าถึง Container ผ่าน SSH ทั้งแบบรหัสผ่านและแบบ Key-based Authentication ได้
- รองรับการเชื่อมต่อบัญชีผู้ใช้กับ AD, LDAP หรือ SSO
- ต้องมี Dashboard แบบกราฟิก แสดงการจัดสรรทรัพยากร GPU ใน 2 โหมด: Aggregation Mode และ Shared Mode

4. การแบ่งพาร์ติชัน GPU และการจัดการต้นทุน (GPU Resource Partitioning and Cost Management)

- รองรับความสามารถ GPU Slicing เพื่อแบ่งหน่วยความจำของ GPU แต่ละรุ่น และกำหนดทรัพยากร CPU/Memory ให้ผู้ใช้เลือกได้
- รองรับการคำนวณค่าใช้ GPU รายชั่วโมง
- ผู้ดูแลระบบสามารถกำหนดโควตาการใช้งานรายโครงการได้ โดยเลือกเพิ่มโควตาหรือกำหนดให้ใช้งานไม่จำกัดได้
- เมื่อโครงการใดใช้เกินโควตาที่กำหนด ระบบต้องแจ้งเตือนในขั้นตอนการสร้าง Container

5. การตรวจสอบและการรายงาน (Monitoring and Reporting)

- ต้องมี Dashboard แบบรวมศูนย์ (Centralized Graphical Dashboard) เพื่อตรวจสอบการใช้งานฮาร์ดแวร์ของ GPU Servers (CPU, RAM, GPU, GPU RAM)
- รองรับการจัดตารางการใช้งานทรัพยากรแบบ Reservation-Based Scheduling เพื่อให้ผู้ใช้สามารถจอง GPU/CPU/DRAM ตามช่วงเวลาได้
- หากไม่มีทรัพยากรเพียงพอ ระบบต้องแจ้งผู้ใช้และแสดงรายการรอ (Waiting Queue) ให้ผู้ดูแลระบบเห็น

6. การสลับเวอร์ชัน CUDA (CUDA Version Switching)

- ต้องมีอินเทอร์เฟซแบบกราฟิก(GUI)สำหรับผู้ดูแลในการกำหนดและสลับเวอร์ชัน CUDA ในแต่ละ Node ได้

7. การจัดเก็บและความปลอดภัย (Storage and Security)

- รองรับการเชื่อมต่อกับ MinIO Object Storage
- ต้องมีการสแกน Container เพื่อหาช่องโหว่ด้านความปลอดภัยตามรอบเวลา
- รองรับการเชื่อมต่อกับ Label Studio สำหรับการทำ Data Labeling
- ต้องรองรับการจัดการทรัพยากร GPU ทั้ง NVIDIA และ AMD

8. ความพร้อมใช้งานสูง (High Availability – HA) และ RCS (Rapid Container Service)

- รองรับการออกแบบระบบ HA เพื่อลดความเสี่ยงจากการล้มเหลวของจุดใดจุดหนึ่ง (Single Point of Failure)
- มี RCS (Rapid Container Service) ทำให้ผู้ใช้สามารถปรับใช้งานโมเดล/อิมเมจได้โดยไม่ต้องใช้ GPU สำหรับงาน Inference

9. สภาพแวดล้อมการพัฒนา AI ที่ติดตั้งล่วงหน้า (Pre-installed AI Development Environment)

- แพลตฟอร์มมาพร้อมสภาพแวดล้อมการพัฒนา Deep Learning ที่สมบูรณ์ ได้แก่: Ubuntu 22.04 LTS หรือใหม่กว่า, NVIDIA Driver, NVIDIA CUDA 11.5 หรือใหม่กว่า, cuDNN, NVIDIA Docker, Jupyter
- และในระยะเวลารับประกัน สามารถได้รับการอัปเดตอย่างต่อเนื่องสำหรับ OS และซอฟต์แวร์ดังกล่าว

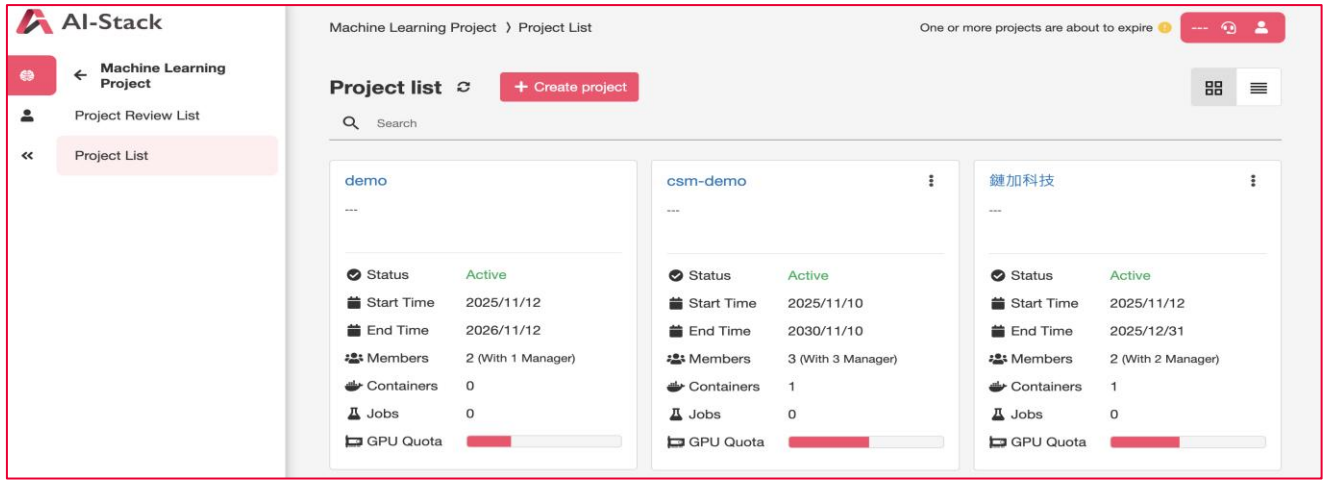
10. การฝึกอบรมแบบยืดหยุ่น (Elastic Training)

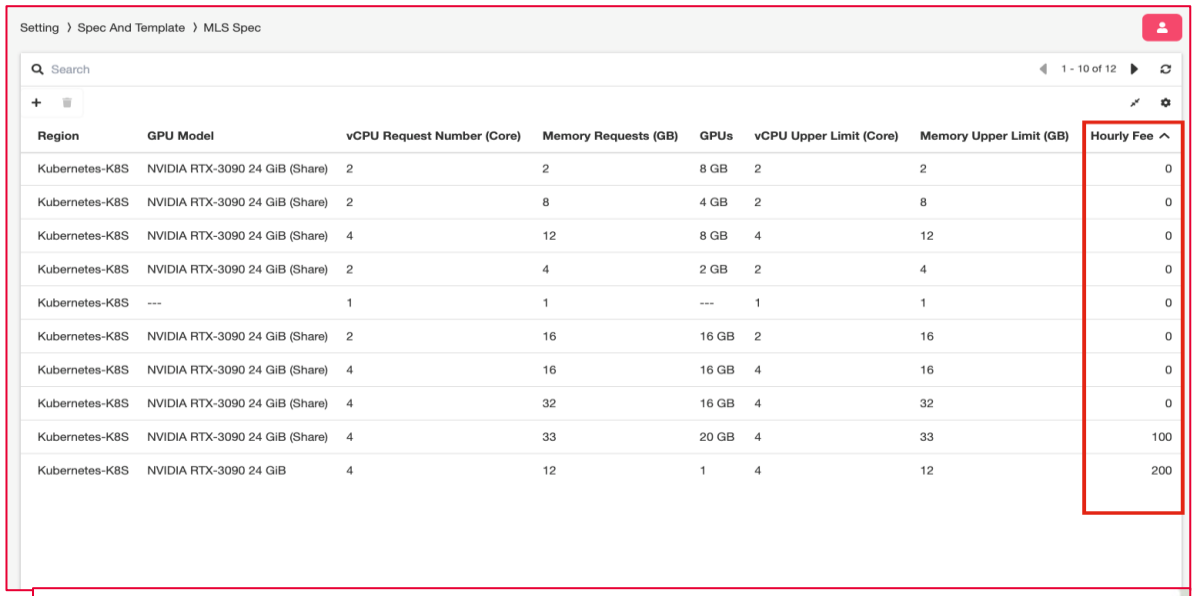
- รองรับ Elastic Training ที่สามารถปรับขยายหรือลดจำนวน Worker Containers ระหว่างการฝึกอบรมได้ตามทรัพยากรใน Cluster
- เพื่อเพิ่มประสิทธิภาพการใช้ทรัพยากรให้เหมาะสมที่สุด

11. การอนุญาตสิทธิการใช้งาน (Licensing)

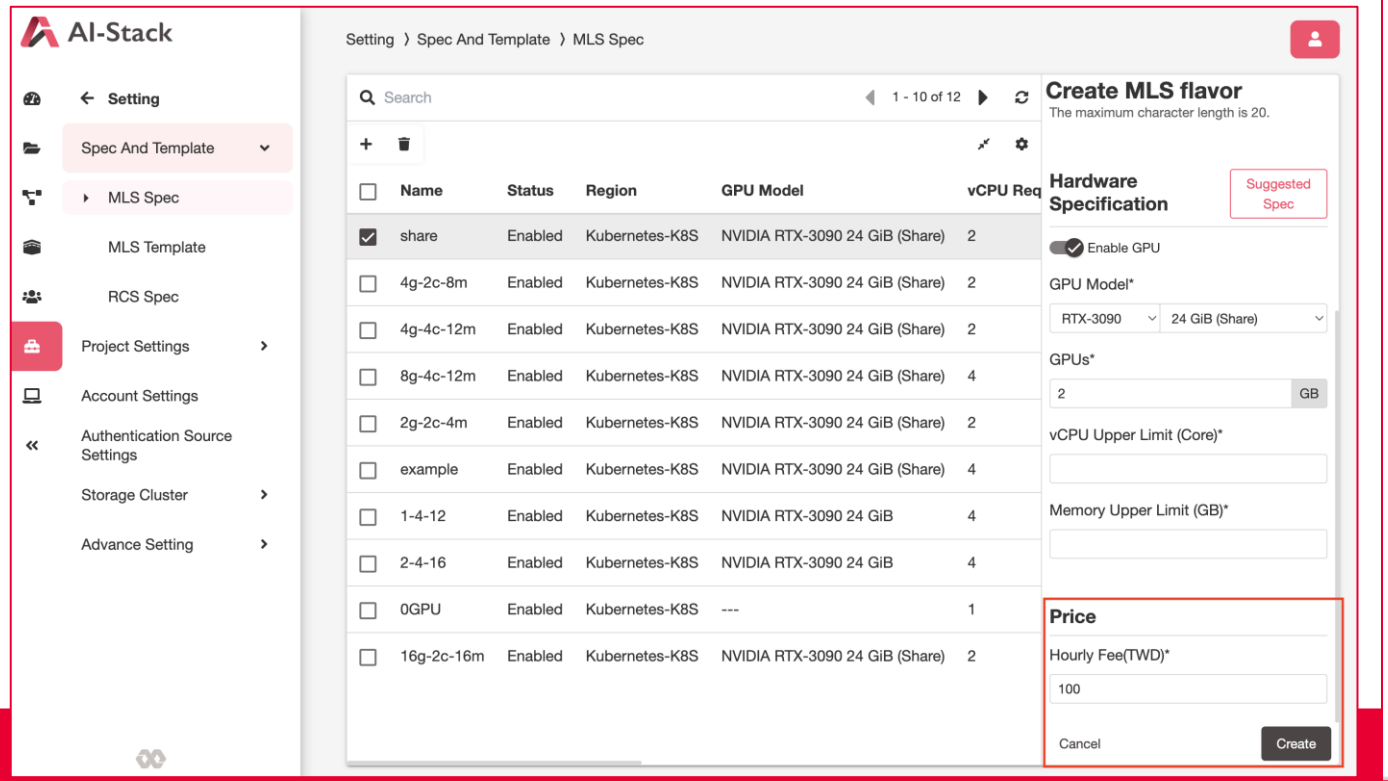
- ซอฟต์แวร์เชิงพาณิชย์ของแพลตฟอร์มต้องให้ภายใต้สิทธิการใช้งานถาวร (Perpetual License – Buy-out Model)
- สิทธิการใช้งานต้องครอบคลุมการเข้าถึงแพลตฟอร์มอย่างต่อเนื่องภายหลังครบกำหนดระยะเวลา 3 ปีแรก

Billing System





Region	GPU Model	vCPU Request Number (Core)	Memory Requests (GB)	GPUs	vCPU Upper Limit (Core)	Memory Upper Limit (GB)	Hourly Fee
Kubernetes-K8S	NVIDIA RTX-3090 24 GiB (Share)	2	2	8 GB	2	2	0
Kubernetes-K8S	NVIDIA RTX-3090 24 GiB (Share)	2	8	4 GB	2	8	0
Kubernetes-K8S	NVIDIA RTX-3090 24 GiB (Share)	4	12	8 GB	4	12	0
Kubernetes-K8S	NVIDIA RTX-3090 24 GiB (Share)	2	4	2 GB	2	4	0
Kubernetes-K8S	---	1	1	---	1	1	0
Kubernetes-K8S	NVIDIA RTX-3090 24 GiB (Share)	2	16	16 GB	2	16	0
Kubernetes-K8S	NVIDIA RTX-3090 24 GiB (Share)	4	16	16 GB	4	16	0
Kubernetes-K8S	NVIDIA RTX-3090 24 GiB (Share)	4	32	16 GB	4	32	0
Kubernetes-K8S	NVIDIA RTX-3090 24 GiB (Share)	4	33	20 GB	4	33	100
Kubernetes-K8S	NVIDIA RTX-3090 24 GiB	4	12	1	4	12	200



Billing report

Tenant: project1762740787795,project1763521744972,project1762916754929,admin,
Service type: MLS
Currency: TWD

Print date: 2025-11-20 15:57
Duration: 2025-10-01 00:00~2025-11-20 15:56

Tenant/Owner: csm-demo / infinitix

Order No.	Cloud provider	Region	Instance name	Service type	Spec	Costing type	Start time	End time	Hours of use	Rate	Total price
2025111000003	infinitiessoft-k8s	K8S	test	MLS	8 NVIDIA RTX-3090/2core/2GiB	Hour	2025-11-10 14:42	2025-11-11 16:36	26.0	0.00	0.00
2025111000004	infinitiessoft-k8s	K8S	tst111001	MLS	8 NVIDIA RTX-3090/2core/2GiB	Hour	2025-11-10 16:22	2025-11-14 15:45	95.5	0.00	0.00
2025111100001	infinitiessoft-k8s	K8S	nfs	MLS	8 NVIDIA RTX-3090/2core/2GiB	Hour	2025-11-11 14:30	2025-11-11 16:36	2.1	0.00	0.00
2025111100002	infinitiessoft-k8s	K8S	ubuntu24	MLS	16 NVIDIA RTX-3090/8core/32GiB	Hour	2025-11-11 18:00	2025-11-11 18:34	0.6	0.00	0.00
2025111100004	infinitiessoft-k8s	K8S	nvidia	MLS	16 NVIDIA RTX-3090/8core/16GiB	Hour	2025-11-11 19:02	2025-11-12 17:39	22.7	0.00	0.00
2025111100005	infinitiessoft-k8s	K8S	offline	MLS	16 NVIDIA RTX-3090/8core/16GiB	Hour	2025-11-11 19:37	2025-11-12 17:39	22.1	0.00	0.00
2025111200002	infinitiessoft-k8s	K8S	tt	MLS	1 NVIDIA RTX-3090/4core/12GiB	Hour	2025-11-12 14:03	2025-11-12 17:39	3.7	200.00	740.00
2025111400001	infinitiessoft-k8s	K8S	pytorch	MLS	8 NVIDIA RTX-3090/4core/12GiB	Hour	2025-11-14 10:51	2025-11-17 10:29	71.7	0.00	0.00
2025111700004	infinitiessoft-k8s	K8S	22222222	MLS	20 NVIDIA RTX-3090/4core/33GiB	Hour	2025-11-17 15:00	2025-11-17 15:47	0.8	100.00	80.00
2025111800006	infinitiessoft-k8s	K8S	iss111801	MLS	2 NVIDIA RTX-3090/2core/4GiB	Hour	2025-11-18 11:50	2025-11-20 15:00	51.2	0.00	0.00
2025111800007	infinitiessoft-k8s	K8S	4444444	MLS	2 NVIDIA RTX-3090/2core/4GiB	Hour	2025-11-18 14:00	2025-11-18 14:46	0.8	0.00	0.00
									Personal total		820.00

Tenant/Owner: csm-demo / CSM-Wenju

Order No.	Cloud provider	Region	Instance name	Service type	Spec	Costing type	Start time	End time	Hours of use	Rate	Total price
2025111200001	infinitiessoft-k8s	K8S	demo	MLS	2 NVIDIA RTX-3090/2core/4GiB	Hour	2025-11-12 10:55	2025-11-12 11:19	0.5	0.00	0.00
									Personal total		0.00
									Tenant total		820.00

The Comparison of AI-Stack

Comparision Ai-Stack & Run.Ai		
มิติการเปรียบเทียบ	Run.ai (NVIDIA Run:ai)	Infinitix Ai-Stack Software
วัตถุประสงค์หลัก	MLOps & GPU Orchestration (จัดการ, จัดตาราง, และเพิ่มประสิทธิภาพการใช้ GPU ในคลัสเตอร์)	AI Infrastructure & GPU Resource Scheduling (แพลตฟอร์มโครงสร้างพื้นฐาน AI ครบวงจร, การจัดการ GPU และ MLOps)
กลุ่มผู้ใช้งานหลัก	Data Scientists, ML Engineers, และ IT/MLOps Teams ในองค์กรขนาดใหญ่ที่มี GPU Cluster	ทีมพัฒนา AI, สถาบันวิจัย, และองค์กรที่ต้องการโซลูชัน AI แบบครบวงจรและบริหารจัดการ GPU ได้อย่างมีประสิทธิภาพ
รากฐานเทคโนโลยี	Kubernetes เป็นหลักในการทำ Orchestration และ MLOps	Kubernetes & Docker (สร้างสภาพแวดล้อมทำงานที่ใช้งานง่าย)
ความสามารถหลัก (GPU)	* Fractional GPU: แบ่ง GPU ส่วนย่อย (Memory & Compute)	* GPU Partitioning (MIG/AMD): การแบ่งส่วน GPU (เช่น NVIDIA MIG)
	* Dynamic Workload Scheduling: การจัดตารางงานแบบไดนามิกและจัดลำดับความสำคัญ	* Dynamic Resource Scheduling: การจัดสรรทรัพยากรแบบไดนามิก
	* GPU Pooling & Aggregation: การรวม GPU ข้ามโหนด (Cross-node computing)	* Cross-Node Computing: การรวม GPU ข้ามโหนด (Aggregation)
ความสามารถหลัก (MLOps)	* MLOps Tools Integration: ผสานรวมกับเครื่องมือ MLOps อื่นๆ	* End-to-End MLOps Workflow: รองรับตั้งแต่ Development ถึง Deployment
	* Centralized Management: การควบคุมและมองเห็นภาพรวมการใช้ทรัพยากรแบบรวมศูนย์	* Control Plane: หน้าจอควบคุมแบบรวมศูนย์, RBAC, Quota Management
	* Experiment Tracking & Artifact Management	* No-Code UI for Project Containers & Resource Allocation
ประสิทธิภาพ & ROI	เน้นการเพิ่ม Utilization ของ GPU (มักอ้างถึงการใช้งานที่สูงขึ้น) เพื่อเพิ่ม ROI	เน้นการเพิ่ม Utilization ของ GPU (อ้างถึง 5-10 เท่าของ ROI และเพิ่ม Utilization จาก 30% เป็น 90%)
	* ลดเวลาการรอของ Data Scientist	* Rapid Environment Setup: ตั้งค่าสภาพแวดล้อม AI ได้ใน 1 นาที (Automatic Provisioning)
การติดตั้ง & การใช้งาน	เป็นโซลูชันที่ติดตั้งบน Infrastructure ของลูกค้า (On-premise/Cloud) และมักต้องใช้ความเชี่ยวชาญด้าน Kubernetes/MLOps	เป็นโซลูชันที่ติดตั้งบน Hardware ของลูกค้า (เช่น NVIDIA DGX) และอ้างถึงการติดตั้งรวดเร็ว (3 ชั่วโมง) และมี UI แบบ No-Code สำหรับการจัดการบางส่วน
การจำหน่าย & ราคา	Custom Quote/Enterprise Pricing (ไม่มีราคาเผยแพร่ทั่วไป) มักขายเป็น Annual Plan per GPU (มีข้อมูลราคาเริ่มต้นใน AWS Marketplace)	มักเป็นแบบ Custom Quote/Enterprise Pricing (ไม่มีราคาเผยแพร่ทั่วไป) และเน้นการขายพร้อม Hardware หรือเป็นส่วนหนึ่งของโซลูชันรวม
จุดเด่นเฉพาะตัว	การเป็นส่วนหนึ่งของ NVIDIA Ecosystem ทำให้มีการสนับสนุนเทคโนโลยี GPU ล่าสุดและได้รับความน่าเชื่อถือในตลาดโลก	One-Click NVIDIA MIG Partitioning/Restoration และการออกแบบที่เน้น ความง่ายในการจัดเตรียมสภาพแวดล้อม (1 นาที)

© INFINITIX 2024 All rights reserved.

The Comparison of AI-Stack with Together.ai

INFINITIX

Main module	Sub module	AI-Stack	Together.ai	說明
Resource management	MIG management	O	X	Supports NVIDIA's Multi-Instance GPU technology for efficient resource utilization
	Various Scheduling Strategies	O	X	Includes Gang Scheduling, BinPack, and TOR-aware scheduling for optimized performance
	GPU Shared and Dedicated Modes	O	X	Offers flexibility in resource allocation
	GPU Auto Reclamation	O	X	Automatically reclaims unused GPU resources to maximize efficiency
Image and Container Management	Integration with CI/CD Workflows	O	X	Streamlines development and deployment processes
	Zero Configuration	O	O	Automatically deployed with minimal configuration
	Container Security Scanning	O	X	Ensures containers are free from vulnerabilities
	Cluster Checkpoint Mechanism	O	X	Supports periodic saving of model states (checkpoints), including model parameters and optimizer states.
	Training Recovery	O	X	Provides a simple interface to resume training from a specified checkpoint
	Public and Private Repositories	O	X	Provides flexibility in storing and accessing images
	Custom Defined Images	O	O	Bring your own custom images to match your unique needs
	Container Backup/Restore	O	X	Save and restore container state, files, and runtime environments
	CUDA Version Switching	O	X	Facilitates compatibility with various CUDA versions
	Rapid Container Services (RCS)	O	O	Enables quick deployment of containerized applications
Monitoring and Toolkits	Real-Time Data & Alerts	O	O	Instant alerts based on live system metrics and anomaly detection
	MUF (Models Untilization FLOPS) support	O	X	Measures model performance for optimization
	LLM toolkits	O	O	Tools to fine-tune, quantize, test, and deploy large language models easily
	Training toolkits	O	O	Assist in data prep, model training, and monitoring
	Inference toolkits	O	O	Improve model serving speed, scalability, and efficiency
	IDE toolkits	O	O	AI dev tools for coding, debugging, and visualization, like SSH, JupyterLabs, etc
GPU type	Multi-GPU type support	O	X	Manages and schedules NVIDIA & AMD GPUs concurrently

© INFINITIX 2024 All rights reserved.

The Comparison of AI-Stack with Together.ai

INFINITIX

Main module	Sub module	AI-Stack	Together.ai	説明
Rental Services: Portal Access	GaaS (GPU as a services) Portal	○	✕	GPU rental via pay-as-you-go portal
	TaaS (Token as a services) Portal	○	○	Token-based compute access porta by API or webpage
	MaaS (Model/Microservice as a services) Portal	○	✕	Subscribe to models or Microservices via portal
	LLMaaS (LLM as a services) Portal	○	○	Delivers language models based on available GPU specs for optimal deployment
	Storage Services Portal	○	✕	Provides integrated storage solutions for data management
	Subscription Accounting Management	○	○	Billing and subscription management system
Rental Functions	GPU Shared mode	○	✕	Allows multiple users to share GPU resources, optimizing utilization
	GPU Dedidate mode	○	○	Provides exclusive GPU access for high-performance needs
	GPU Cluster modes	○	○	Supports clustering of GPUs for large-scale computations
	Elastic Training mode (Deepspeed, Horovod, Megatron)	○	✕	Integrates frameworks like Deepspeed, Horovod, and Megatron for scalable training
	Slurm mode	○	○	Utilizes Slurm for efficient resource management and job scheduling
	Bare Mental mode	○	✕	Offers direct access to hardware for maximum performance
	Token mode	○	○	Implements token-based access for enhanced security
	Models (Microservices) mode	○	○	Facilitates deployment of models as microservices
	Storage mode	○	○	Provides versatile storage options for various data needs
Multi-Tenant Support	SSO, AD, LDAP Integration	○	✕	Enterprise-level identity signon integration
	RBAC (Role Base Access Control)	○	○	Provides Role-based access control policies for users and groups
	User Group Management	○	○	Organize and manage user groups
	Tailored Service	○	○	Offers customized solutions to meet specific tenant requirements
	Tenant isolation	○	○	Ensures data and resource isolation between different tenants

© INFINITIX 2024 All rights reserved.

Preinstall Information

Customer Reply (Please overwrite the example wording or add new rows)									Note
1. Confirmation of GPU host hardware data **(if there are not enough rows, please add your own)**									
	Mainframe Model-1	GPU Model	GPU Memory	GPU Quantity	Memory	CPU Model	CPU Core	CPU Quantity	Please make sure that the number of GPUs matches the information provided in Part 2.
Example	MSI G4201	Nvidia A100	40GB	2	256GB(32GB*8)	Intel Xeon Gold 6338	32	1	
Example	Transport HX TS75-B8252	AMD100	32GB	3	2TB(64GB*32)	AMD EPYC	64	2	
2. Please ensure that you complete each section of the form. Note: Once the environment has been set up, it is important not to make any changes, as this may affect system operation.									
Please confirm whether the installation resource is for one server only	**Yes**	Suggested resource specifications for each node			GPU Model	Quantity	[Host Name +Password]	LAN IP	
AI-Stack Node	AI-Stack	CPU 6 vCore、Memory 8/16GB、Disk 100/200GB			-	-	root / 8@8@8@8@	192.168.10.102	When running the AI-Stack, Master, and Harbor nodes on the same server, they will share the same IP address. Please use the same IP information for each.
Master Node	Master	CPU 4 vCore、Memory 8/16GB、Disk 200/500GB					root / 8@8@8@8@	192.168.10.103	
Harbor Node	Harbor	CPU 4 vCore、Memory 8/16GB、SystemDisk: 100 GB、Data Disk: 500 GB/ 1TB					root / 8@8@8@8@	192.168.10.104	
Worker Node1	GPU Worker	CPU 6 vCore、Memory 8/16GB、Disk 100/200GB			Nvidia A100	2	root / 8@8@8@8@	192.168.10.100	Please fill in the GPU specifications for each worker node based on your actual architecture, as the number of nodes and GPU configurations may differ per node.Same Worker Node required for same GPU model.
Worker Node2	GPU Worker	CPU 6 vCore、Memory 8/16GB、Disk 100/200GB			AMD M100	2	root / 8@8@8@8@	192.168.10.101	
Worker Node3	GPU Worker	CPU 6 vCore、Memory 8/16GB、Disk 100/200GB			AMD M100	1	root / 8@8@8@8@	192.168.10.102	
Worker Node4	GPU Worker	CPU 6 vCore、Memory 8/16GB、Disk 100/200GB							
AI-Stack is installed on a bare-metal server or a virtual machine?					Virtual Machine				
Default Gateway					192.168.160.254				
DNS					8.8.8.8				Internet connection possible
CUDA versions					12.4				
Default NTP Server					time.google.com				Internet connection possible
NFS mount point					192.168.160.39:/nfs				
3. Check Each Item Before Installation (If your response is 'Yes', please tick the checkbox. Otherwise, leave it empty.) Note: Fields in red are required									
Internet	1. Make sure that the LAN IP (Master / GPU Worker) of each of the above nodes is on the same network segment, and you can also establish TCP/IP Telnet.				☐				
Internet	2. Can the environment be connected to the Internet externally? (Some packages and images will download the latest version from Linux and NVIDIA website.				☐				
Internet	3. Is the environment where AI-Stack is built in a firewall architecture? Can the firewall be adjusted? (NO-Firewall cannot be adjusted to disable, please refer to the port number in the table below and make sure that bidirectional TCP can be established)				☐				
	Node	Port	Describe						
Internet	(All)	22,23,80,443	allow access to install ai-stack access and TCP		☐				
	AI-Stack	8080	allow access to AI-Stack web services						
	AI-Stack	61616	allow access to activemq from skyport container						
	Harbor	18080	allow access to harbor from skyport container						
	Master	6443	allow access to k8s from skyport container						
	Master / Worker	2375	allow access to docker api from aistack container						
	Worker	30000:32767	allow access to open training portal						
Infra	3. Confirm that the GPU server is in a clean state, such as by performing a reset or reinstalling its operating system.				☐				
Infra	4. OS: Ubuntu 22.04 Server LTS (kernal 5.15.0). Please do not download the Desktop version; after the installation is complete, and list the non-specified version if it is not the specified one.)				☐				
Infra	5. GPU Server disable Secure Boot in BIOS				☐				
4. Remote Installation/Connection Information (If you require remote support for installing AI-Stack, please fill in the connection information below. If not, you may leave this section empty.)									
Remote Connection Information	Please select and provide one of the following methods to connect to environment				Please enter the relevant information				
	VPN Server Address / Port / Login ID / Password.				IP / Account / Password or Key				
	Anydesk IP / Password				Account / Password				
	Video conferencing and authorization (Webex / Teams)				Teams				
	Team Viewer (enterprise verion is recommended.)				ID/ Password				
5、Contact									
Contact information	If encounter any environmental or technical problems during installation, please provide your technical support contact information.				Email: infinitix@infinitix.co Name: Shing Ko Mobile: 010-1111-2222				

Preinstall Information

```
root@k8s-master:/home/ubuntu# kubectl get svc -A
NAMESPACE              NAME                                        TYPE           CLUSTER-IP       EXTERNAL-IP     PORT(S)                                                  AGE
default                kubernetes                                  ClusterIP      10.96.0.1        <none>          443/TCP                                                  32d
ingress-nginx          ingress-nginx-controller                    ClusterIP      10.106.172.183   192.168.10.235  80/TCP,443/TCP                                           32d
ingress-nginx          ingress-nginx-controller-admission          ClusterIP      10.108.139.97    <none>          443/TCP                                                  32d
istio-system           istio-ingressgateway                        LoadBalancer   10.96.29.118     <pending>       15021:31966/TCP,80:30294/TCP,443:31970/TCP               32d
istio-system           istiod                                      ClusterIP      10.109.227.50    <none>          15010/TCP,15012/TCP,443/TCP,15014/TCP                    32d
kepler                 kepler                                      ClusterIP      10.108.149.214   <none>          9102/TCP                                                 26d
kube-system            dcgm-exporter                               ClusterIP      10.97.161.221    <none>          9400/TCP                                                 32d
kube-system            gpu-isolation-ixgpu-device-plugin-monitor   NodePort       10.103.205.214   <none>          31992:31992/TCP                                          11d
kube-system            gpu-isolation-ixgpu-device-plugin-monitor-amd NodePort     10.107.37.34     <none>          31994:31994/TCP                                          11d
kube-system            gpu-isolation-ixgpu-scheduler               NodePort       10.101.59.216    <none>          443:31313/TCP,31993:31993/TCP                            11d
kube-system            ix-nvidia-gpu-exporter-monitor              ClusterIP      10.102.178.43    <none>          9835/TCP                                                 32d
kube-system            kube-dns                                    ClusterIP      10.96.0.10       <none>          53/UDP,53/TCP,9153/TCP                                   32d
kube-system            kube-state-metrics                          ClusterIP      None             <none>          8080/TCP,8081/TCP                                        32d
monitoring             node-exporter                               ClusterIP      10.104.199.197   <none>          9100/TCP                                                 32d
monitoring             prometheus-service                          NodePort       10.97.70.33      <none>          9090:30909/TCP                                           32d
node-feature-discovery nfd-master                                  ClusterIP      10.101.145.18    <none>          8080/TCP                                                 32d
project1736150177365   service-anychat0203                         NodePort       10.96.103.101    <none>          22:32559/TCP,6006:31973/TCP,8889:31753/TCP,8888:30686/TCP  20h
seldon-system          seldon-webhook-service                      ClusterIP      10.96.215.122    <none>          443/TCP                                                  32d
```

© INFINITIX 2024 All rights reserved.

Tier GPU Card

AI-Stack GPU Tier Support List

Tier	Product
1	**NVIDIA:** **Telsa:** B200, B100, H200, H100 **AMD:** **Instinct:** MI325X, MI300X, MI300A
2	**NVIDIA:** **Tesla:** A100, V100 **Quodra:** 6000 Ada, 5880 Ada, 5000 Ada, L40S, L40 **AMD:** **Instinct:** MI250X, MI250, MI210
3	**NVIDIA:** **Tesla:** L4, A2, A10, A16, A30, A40, A400 **Quadro:** 2000 Ada, 4000 Ada, 4500 Ada, A1000, A2000, A4000, A4500, A5000, A5500, A6000, RTX 4000, RTX 5000, RTX 6000, RTX 8000, T4, T400, T600, T1000, P100, P400, P600, P620, P1000, P2000, P2200, P4000, P5000, P6000, GP100, GV100 **GeForce:** RTX 5090, RTX 5090 D, RTX 5080, RTX 5070, RTX 5070 Ti, RTX 4090, RTX 4090 D, RTX 4080, RTX 4080 Super, RTX 4070, RTX 4070 Super, RTX 4070 Ti, RTX 4070 Ti Super, RTX 4060, RTX 4060 Ti, RTX 3090, RTX 3090 Ti, RTX 3080, RTX 3080 Ti, RTX 3070, RTX 3070 Ti, RTX 3060, RTX 3060 TI, RTX 3050, RTX 2080, RTX 2080 Super, RTX 2080 TI, RTX 2070, RTX 2070 Super, RTX 2060, RTX 2060 Super, TITAN RTX **AMD:** **Instinct:** MI100, MI60 **Radeon:** RX 7900 XTX, RX 7900 XT, RX 7900 GRE, RX 7800 XT, RX 7700 XT, RX 7600 XT, RX 7600, RX 6950 XT, RX 6900 XT, RX 6800 XT, RX 6800, RX 6750 XT, RX 6750 GRE, RX 6700 XT, RX 6700, RX 6650 XT, RX 6600 XT, RX 6600, RX 6500 XT, RX 6400, RX 6300, RX 5700 XT, RX 5700, RX 5600 TX, RX 5600, RX 5500 XT, RX 5500, RX 5300 XT, RX 5300

1.All GPUs in the list are CUDA/ROCm-supported.
2.GPUs in the list are supported by Infinitix AI-Stack.
3.AI-Stack may work with GPUs not listed herein, but at users' oen risks.
4.Infinitix reserves the rights to modify this Tier Table with a 30-day notice. All changes will not be retrospectively applied.

© INFINITIX 2024 All rights reserved.

Connect with us

FB

LinkedIn

- Somrath Tojumroen (ART)
- Mobile : 0888999692

© INFINITIX 2024 All rights reserved.