INFINITIX NVIDIA Partner

BC Best Choice AWARD

Enterprise Computing Resources & AI Infra Management Solutions

NVIDIA's Solution Advisor | AMD GPU Ecosystem Development Partner

AI-Stack

AI Infrastructure Resource Management

Computing resources management:

- Supports GPU / NPU / Phison aiDAPTIV+ Heterogeneous compute resource management
- CTAs/Partitioning/Aggregation/Cross-node computing
- Visual monitoring dashboard

Multi-Tenant

Storage management

Model Training and Inference Deploy

Elastic Distributed Training (EDT)

Machine Learning Service (MLS)

Rapid Container Service (RCS)

Task Management and Scheduling

ixCSP

AI Cloud Operations Solution

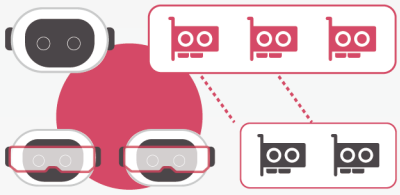
GPU-as-a-Service

Token-as-a-Service

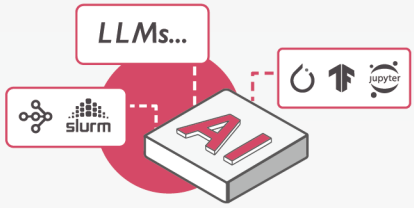
Model-as-a-Service

Billing Management

Model marketplace

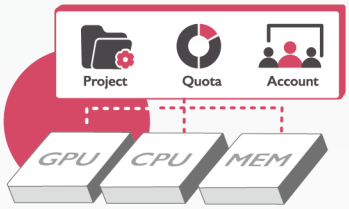


Enable precise, efficient GPU resource management

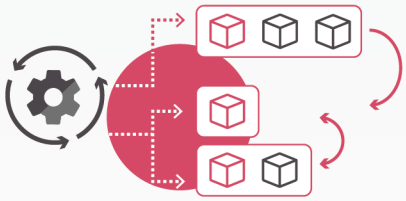


Seamlessly integrate with popular AI development tools

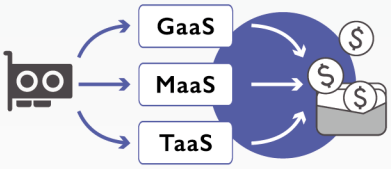
No complex cloud infrastructure required



Ensure resource isolation, access control, and quotas



Automate single and batch job



Monetize existing compute resources GaaS/MaaS/TaaS

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.

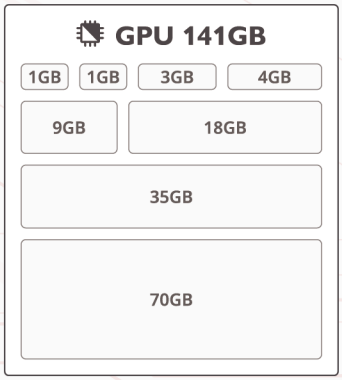
10x ↑

Enhanced ROI

10x ROI Boost - Maximize Your AI Investment.

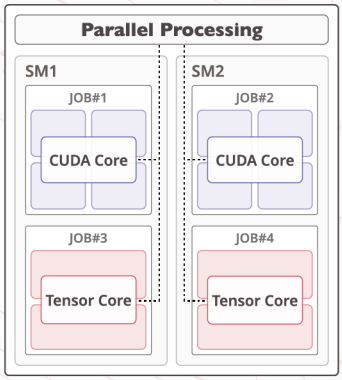
AI-Stack

The 4 Core Technologies for GPU Resource Management



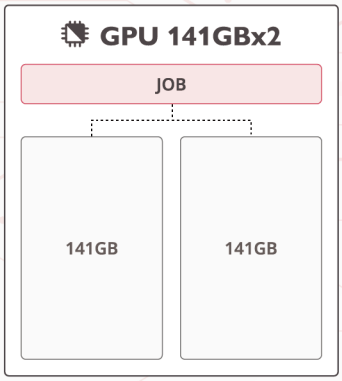
GPU Partitioning

Split a single GPU into multiple segments to support diverse model sizes, boosting utilization to over 90% and minimizing resource waste.



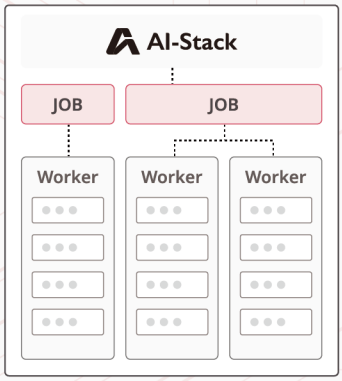
GPU Core Type Aware Scheduler (CTAs)

Dynamically distinguishes CUDA and Tensor cores, co-scheduling complementary workloads on a single GPU to unlock true parallel performance.



GPU Aggregation

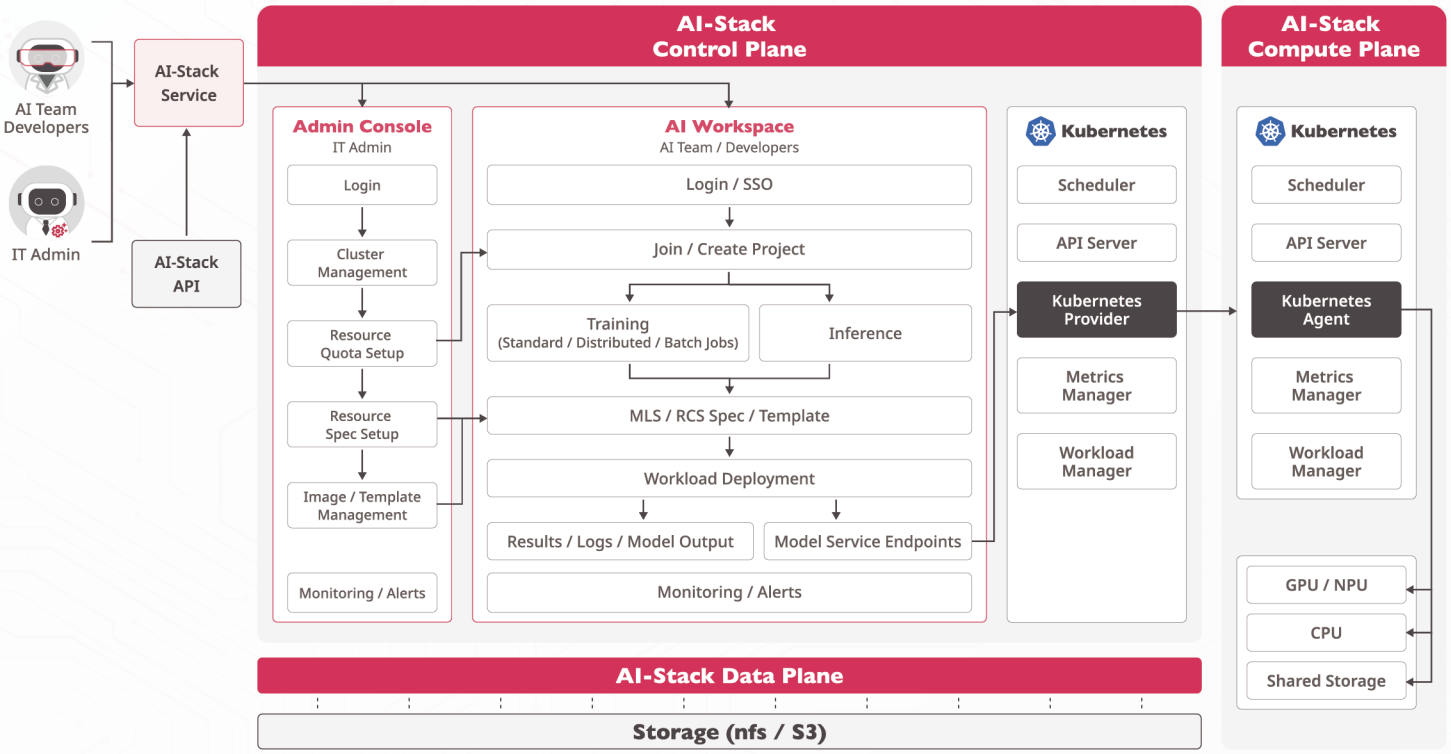
Combine multiple GPUs to power large AI/ML training, accelerating performance and improving development efficiency.



Cross-Node Computing

Distribute training across multiple nodes for parallel processing of large datasets, cutting model training time.

AI-Stack System Architecture



Admins | IT Managers

- Enterprise-grade security and efficient resource use
- Project access control and quota management
- Unified management of heterogeneous resources
- Professional dashboards for real-time GPU monitoring

Users | Data Scientists & AI Researchers

- Deploy containers in 1 minute
- Focus on model development and training
- Automate task scheduling
- Quickly deploy inference models with simplified workflows

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 · Qwen3:235b · 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 / OCP 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
- CTAs

Workload Orchestration

kubernetes

Infrastructure Cluster Layer

A unified platform to manage heterogeneous compute resources (NVIDIA, AMD GPU/NPU/PHISON aiDAPTIV+), enabling high-performance AI computing with NFS and MinIO storage for efficient data flow.

AI-Stack Cluster Engine

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

Server Cluster

GPU Server Cluster

NVIDIA · AMD · PHISON

Storage Server Cluster

NFS, MinIO

AI-Stack Application Industries

AI Data Center · Semiconductor · Manufacturing · Finance · Healthcare · Academic · Energy Sector · Government · Transport · Telecom

ixCSP AI Cloud Operations Solution

INFINITIX's ixCSP enables enterprises to turn GPU server resources into revenue—without complex development—and instantly offer GPU-as-a-Service (GaaS), Model-as-a-Service (MaaS), and Token-as-a-Service (TaaS) to global users.

Operation Center

Operation Managers

Catalog Management
Centralize model and service management for easy organization and updates.

Resources Configuration
Tailor compute, storage, and templates to any workflow.

User Management
Easily onboard, organize, and manage user access.

Operations Dashboard
Real-time monitoring with dashboards and alerts.

Billing & Payment
Full transparency with automated tracking, billing, and flexible payments.

Audit Logs
Enterprise-grade security with detailed audit logs.

Dev Console

AI Developers

On-Demand Resources
Launch and manage AI compute instantly, pay-as-you-go.

GenAI Studio
Access and integrate GenAI models via OpenAI-compatible APIs.

One-Click Deployment
Deploy and update apps in seconds, from dev to production.

Collaboration & Accounts
Seamless collaboration with secure account controls.

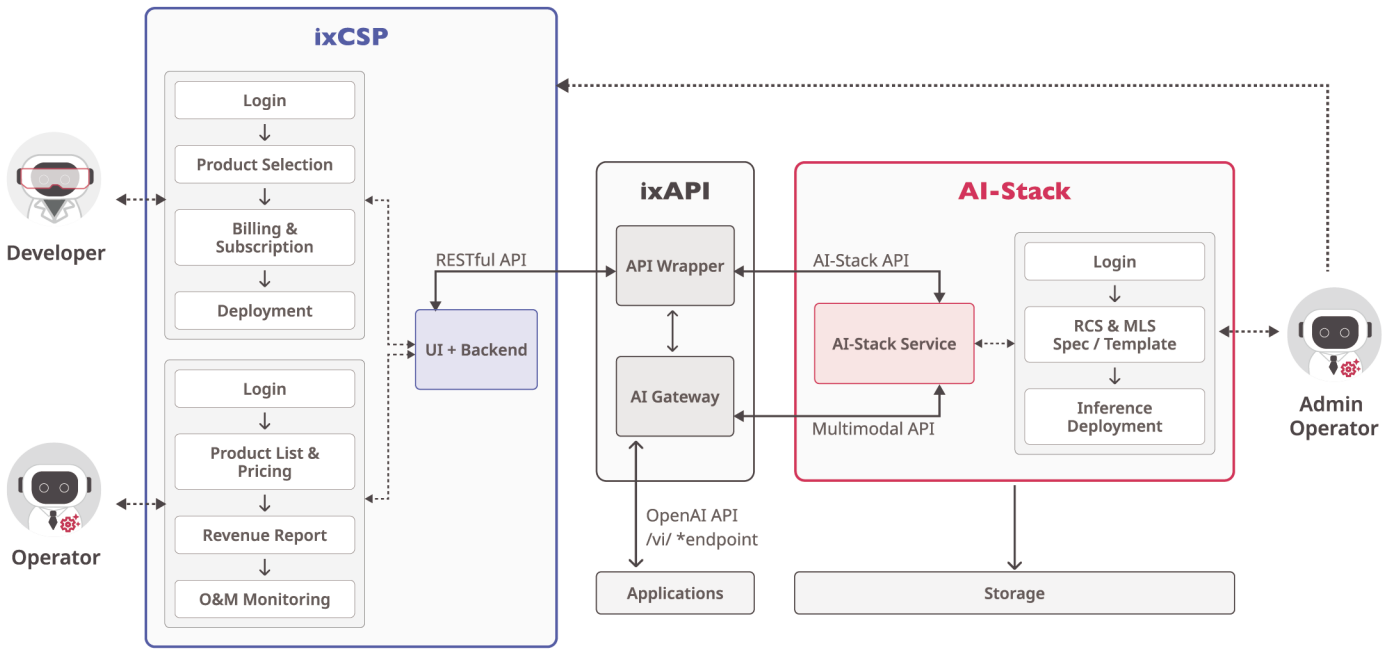
Three Business Models of ixCSP

GaaS (GPU-as-a-Service)
On-demand, pay-as-you-go GPU compute resources.

MaaS (Model-as-a-Service)
Upload the model to the One-Click Deployment store.

TaaS (Token-as-a-Service)
Real-time token tracking with pay-as-you-go AI inference.

ixCSP System Architecture



INFINITIX

NVIDIA Partner

Since 2017, INFINITIX has specialized in AI infrastructure management, delivering advanced GPU resource solutions that accelerate enterprise transformation and help build a thriving global AI ecosystem.

infinitix.ai