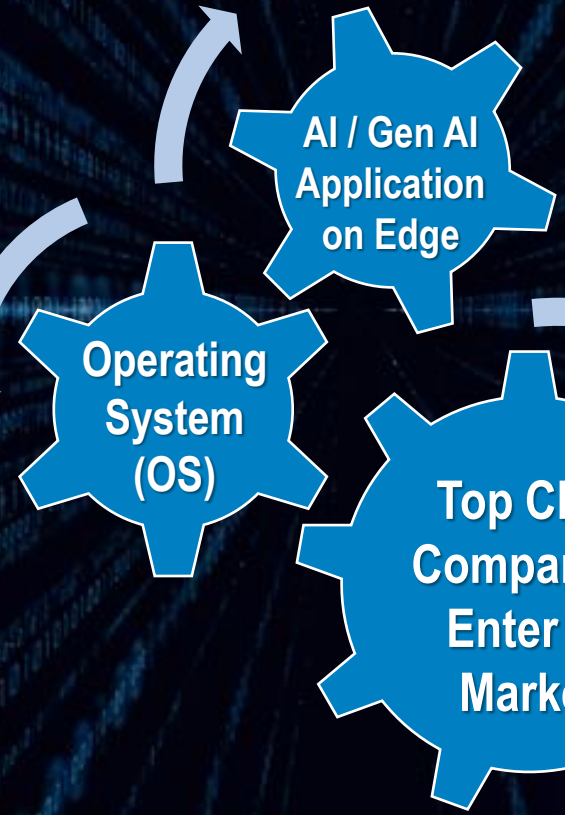
ADVANTECH
Enabling an Intelligent Planet

Build on-premise enterprise AI and continuous fine-tuning with AMD high performance platform

在 AMD 高性能平台上構建可持續優化的產業應用AI系統

Joey Hsu, Director, Edge Intelligent Systems BU, Embedded IoT

Edge AI PC Era is Arrived
邊緣AI PC時代，現在開始



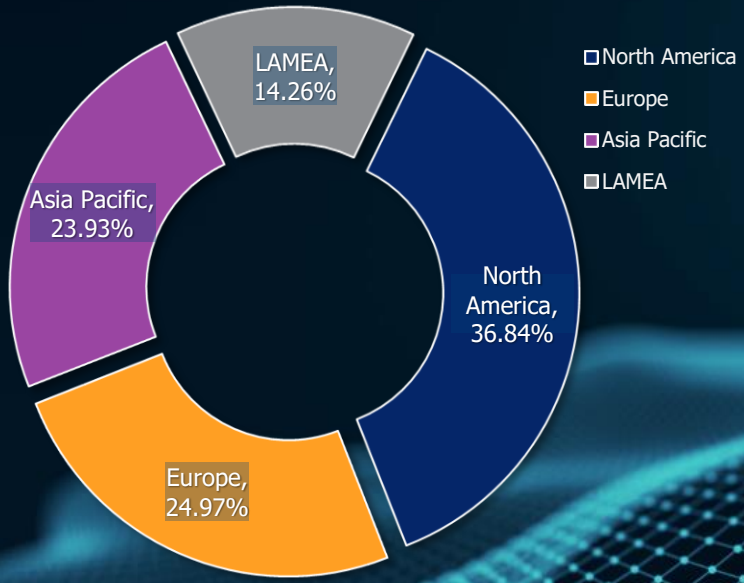
- Edge AI Emerging Market
- Edge AI Inference
- Edge AI Server (LLM)

- WIN 11 / WoA
- WIN Copilot+
- Ubuntu Pro Linux
- Edge AI SDK

- GPU: Nvidia AMD Intel
- NPU: Hailo Axelera
- CPU+NPU : Intel AMD Qualcomm NXP

AI's Exponential Growth: Seizing the Competitive Edge
AI 指數級成長, AI 競爭力即成長力

AI Market Share by Region, 2022 (%)



Global AI Market Revenue, By End User (US$ Billion)

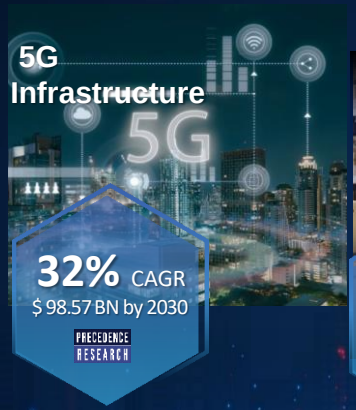
End User	2022	2023	2027	2032
Healthcare 醫療照護	64.33	76.35	152.36	369.22
BFSI 金融保險	72.59	86.13	172.00	416.49
Law 法務相關	15.96	19.02	38.65	95.47
Retail 智慧零售	43.83	52.13	105.03	257.43
Advertising & Media 廣告媒體	63.19	74.97	149.59	362.07
Automotive & Transportation 智動化與交通	45.41	53.84	107.81	260.74
Agriculture 農業應用	29.26	34.78	70.02	171.16
Manufacturing 生產製造	43.44	51.58	103.75	252.81
Others 其他	76.11	89.34	170.89	389.77

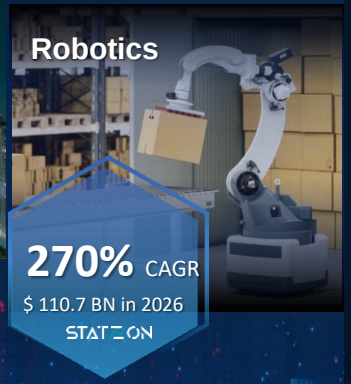
Reference: https://www.precedenceresearch.com/artificial-intelligence-market

The Future of AI and Edge Computing

AI與邊緣運算的完美結合，開創未來新局

Market Trend and Business Opportunities





Focus on Emerging Markets Driven by Edge AI

邊緣AI：新興市場的成長引擎

AI Acceleration Modules

AI-on-Modules

Edge AI

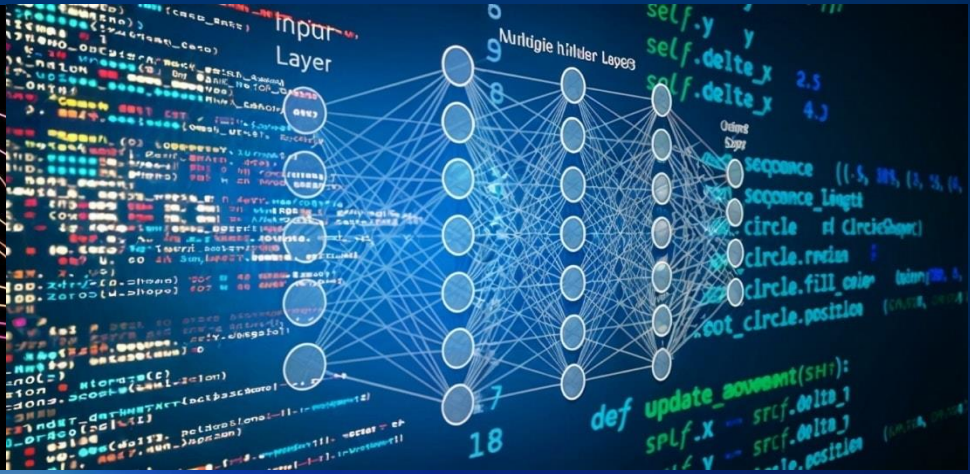
Edge AI SDK

Inference Systems

AI Servers

Edge AI Enterprise SSD

Facing the Challenges of the AI Trend
面對AI趨勢的隱憂



Limited Budget
有限預算

Taring Server = $ 150,000
Inference Server = $30,000

AMD Best GPU Choose
192GB HBM GPU x1= $ 15,000
48GB HBM GPU x1= $ 5,000

Evaluation
學習評估

Training or Inference
Model selection
Efficiency and Productivity
Choice and Flexibility

Implementation
導入實施

Service Coordination
Convenience and Efficiency
Accessibility and Consistency

Maintenance
維護服務

Comprehensive Cost Oversight
Optimized Resource Allocation
Risk Mitigation

Hybrid AI Solution for Section AI Architecture
混合式 AI 解決方案

Pre-Training
Understanding human language

Text
Sentence
Article

Model Creation

GPU Cards Requirements

>1,000 Massive GPUs for High Computing Power

NVIDIA Microsoft Meta Google

Fine-Tuning

Smaller Labelled Datasets
Professional Knowledge

Question Answering
Information Extraction
Instruction Following

Domain Focus Training

4~100 Depends on Memory Size (GPU RAM ≥ 20x Model Size)

AIR-520-L70/33/13B

Inference

Chat
Coding
BizOps
Copywriting
Dev Tools

Practical Application

1 Minimal Requirement

AIR-520

AIR-030

ADVANTECH

Advantech Edge AI Server Solution for LLM & GenAI

研華邊緣AI伺服器，驅動LLM與生成式AI創新

預算自主 支控得宜

1/10 traditional GenAI training Server cost but same Effectiveness

突破限制 無限創新

Innovate Hybrid GenAI solution, fine-turning and inference at the same server. Unlimited innovation.

完整配置 立即可用

Llama3-70B and more than 10 models supported. Free dataset, fine-turning, monitor, validation, inference GUX, easy to use.

安全可靠 保密防諜

Local fine-turning **w/o internet.** 100% protect your valuable know-how and keep outstanding

落地 自主 安全 可用

T	R	U	S	T
Timely 即時性	Robustness 穩健性	Useful 實用性	Security 安全性	Technology 技術整合

IEC 62443 TAIWAN EXCELLENCE ADVANTECH

How many GPUs are needed for model training?
模型訓練需要多少GPU 資源?

LLaMA 70B Training

1.4TB memory space required

Data Data Data Data Data Data Data
Data Data Data Data Data Data Data

Data cut into slices

Current
AI Computing Architecture

Limited by GPU HBM

18 H100 GPU to run LLaM2 (70B)

(Requires **DGX x3 Chassis**)

DGX H100 **10.2kw**

CPU

Data Data
Data Data
Data Data
Data Data

DGX H100 **10.2kw**

CPU

Data Data
Data Data
Data Data
Data Data

DGX H100 **10.2kw**

CPU

Data Data
Idle Idle
Idle Idle
Idle Idle

Advantech AIDE
AI Computing Architecture

Unlimited Model Size

Scale # GPU to match budget

(Requires **AMD W7900 GPUs + AIDE Solution**)

SQ aiDAPTIV+ AI **1kw**

CPU

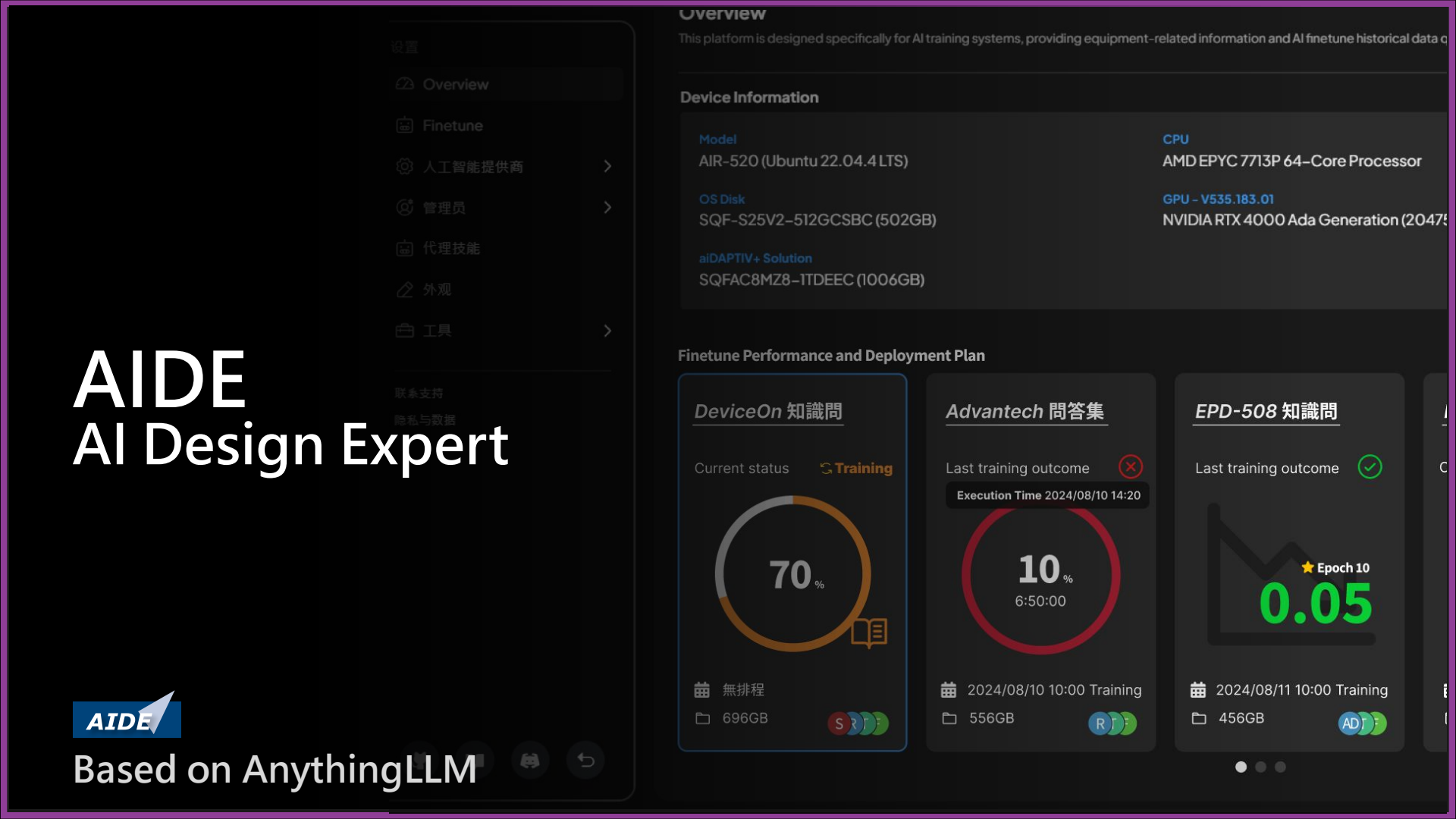
Data Data Data Data

Quantization

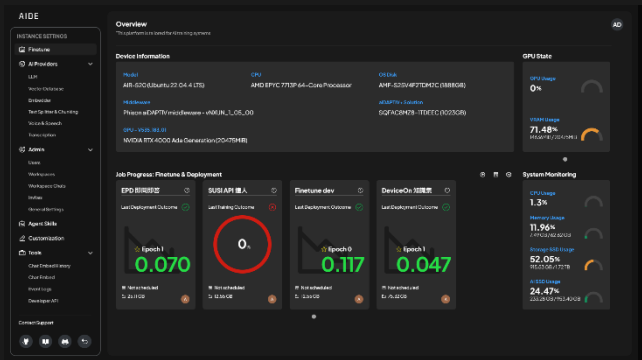
ADVANTECH AIDE

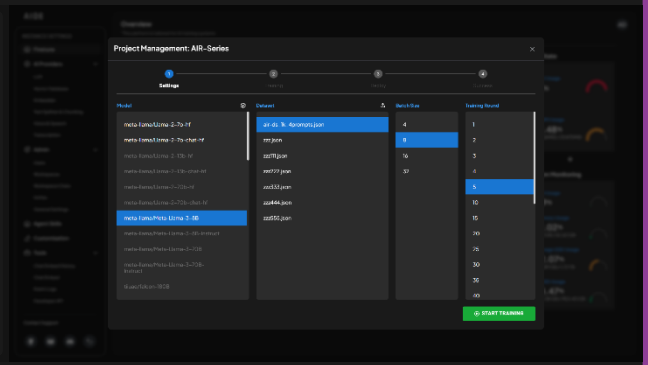
Data Data Data Data Data
Data Data Data Data Data

PENDING SLICES

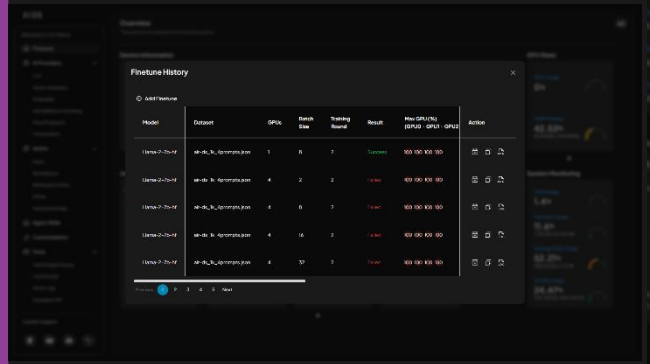

Overview
This platform is designed specifically for AI training systems, providing equipment-related information and AI finetune historical data c
Device Information
Model
AIR-520 (Ubuntu 22.04.4 LTS)
CPU
AMD EPYC 7713P 64-Core Processor
OS Disk
SQF-S25V2-512GCSBC (502GB)
GPU - V535.183.01
NVIDIA RTX 4000 Ada Generation (2047
aiDAPTIV+ Solution
SQFAC8MZ8-1TDEEC (1006GB)
Finetune Performance and Deployment Plan
DeviceOn 知識問
Current status
Training
70 %
無排程
696GB
Advantech 問答集
Last training outcome
Execution Time 2024/08/10 14:20
10 %
6:50:00
2024/08/10 10:00 Training
556GB
EPD-508 知識問
Last training outcome
Epoch 10
0.05
2024/08/11 10:00 Training
456GB
AIDE
AI Design Expert
AIDE
Based on AnythingLLM

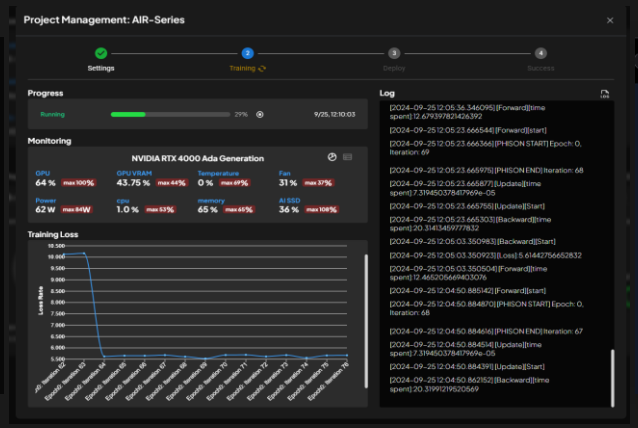
AIDE (AI Design Expert)

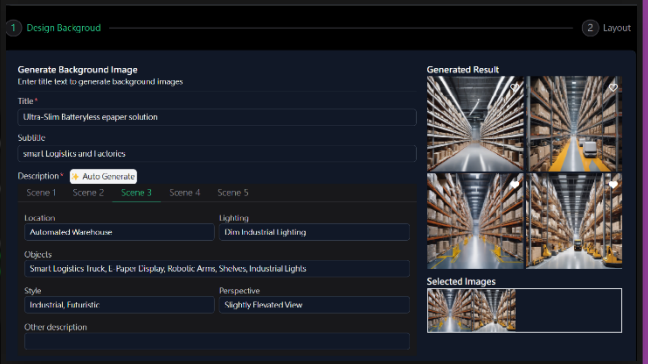




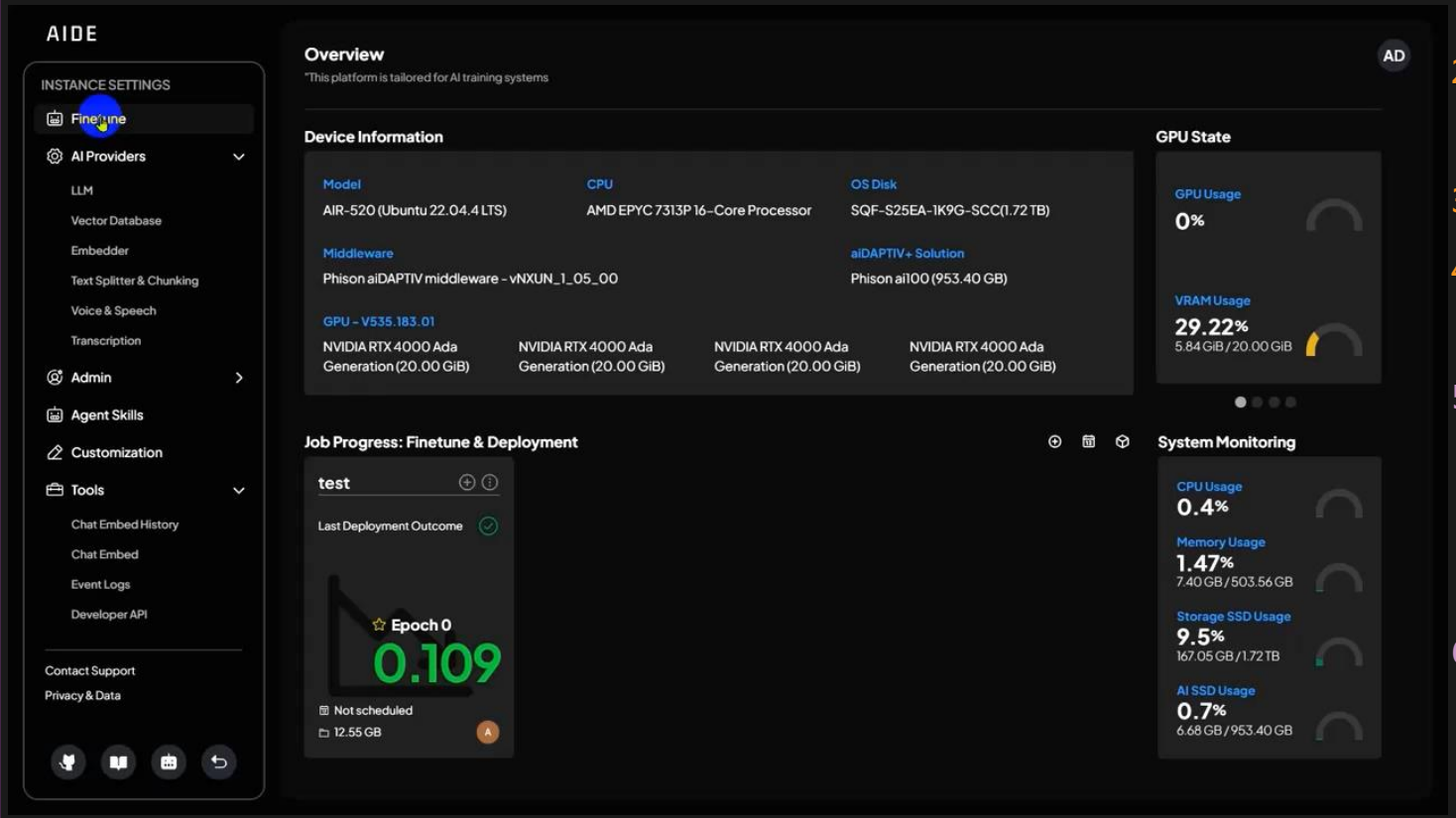
Support foundation model downloading, fine-tuning, quantization, inference chatbot, RAG .....
(Under developing, will release by the end of 2024)







AIDE Application Architecture



1. Foundation Model 下載
   - Hugging Face

2. Model Finetune
   - Phison
   - LORA

3. RAG 資料管理 & 即時更新

4. Inference
   - API Provider
   - Deployment

5. 內建 EIoT Assistance
   -產品主視覺生成器
   - Demo Room AI 自助導覽
   - DeviceOn, EPD 知識庫
   - SUSI 開發專家
   - EIoT 硬體推薦大師
   - aiDAPTIVinbox

6. 系統管理
   - Multi-User, Language
   - Monitoring
   - Scheduling
   - Chat Embed

7. Core Containers

Rapid AI SDK Development Toolkit

A single toolkit streamlines cross-platform AI inference

Edge AI SDK

TensorRT OpenVINO ROCm HailoRT

eIQ Neural Processing SDK NeuroPilot

Inference Benchmark Tool

Inference Runtime SDK

Inference Deployment

AMD intel Qualcomm MEDIATEK NXP Texas Instruments HAILO kinara AXELERA NVIDIA PHISON

Product Selection Guide

AMD Edge AI SDK
Ubuntu DeviceOn

Solution SKUs	AIR-520-L13A1	AIR-520-L33A1	AIR-520-L70A1
LLM Size	13B	33B	70B
CPU	AMD EPYC 7313P 16C	AMD EPYC 7313P 16C	AMD EPYC 7543P 32C
Memory (RDIMM)	128 GB	256 GB	512 GB
SQ AI SSD	1 TB	2 TB	4 TB
GPU Card VRAM	16 GB	48 GB	96 GB
Training Time (hrs)*	~ 4 hrs	~ 7.5 hrs	~ 10 hrs
Power Consumption	**~ 276 W**	**~ 427 W**	**~ 573 W**
Model Purpose	Chatbot service Marketing article	Domain-specific applications Research and development (R&D)	High-complexity tasks Enterprise-grade data analytics

Model Supported

- Breeze-7B
- CodeLlama-7B, 34B, 70B
- Llama-2-7B,13B,70B
- Llama-3-8B, 70B
- Mistral-7B-Instruct-v0.2
- Mistral-8x7B, 8x22B
- TAIDE-7B
- Vicuna-33B
- Whisper-V2, V3

***Testing Record Note**

- Training time and token/sec are based on the test using **5M** dataset with Llama-2 13B/70B and Vicuna-33B models
- 4 x RTX 4000 Ada have better training efficiency than 2 x RTX 6000 Ada

ADVANTECH

機會來的時候
它是不會問你準備好沒

也只有在你遇見機會的時候
你才知道自己懂的會的夠不夠

喬伊

EIoT Edge AI Computing Solutions
EIOT 邊緣AI運算解決方案

High Performance Inference & AI Model Optimization

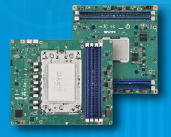
> 500 TOPS

AI Enabled Application Solution

5~100+ TOPS

AI Plug in Modules

25~200 TOPS



Edge AI Server Boards

Edge AI Server Systems

AI on Modules & Boards

Edge AI Application Systems

AI Acceleration Modules

AI Model Training & Inferencing

Data Labeling
Model Training
Model Fine-Tune
Model Compression
Inference Development

AI Runtime Environment

Performance Evaluation
AI SDK Integration

OS & Cloud Services

Enterprise-Grade AI Software Platform

- PAPIDS
- NeMo
- TensorRT

Computer Vision AI Platform

DeviceOn
AI Model Deployment & Management

Edge AI SDK
AI Inference Framework & Tool Kits

TensorRT OpenVINO
Neural Processing SDK
ROCm eIQ HailoRT
NeuroPilot

MicroCloud
ubuntu

Azure OpenAI
Windows

Advantech Product Roadmap – Embedded Boards

High-Performance

AMD EPYC Embedded

SOM-E780
- EPYC™ Embedded 7003 Series
- COM-HPC SERVER E

AIMB-592
- EPYC™ Embedded 7003 Series
- MicroATX

SOM-E781 Q1'25
- EPYC™ Embedded 8000 Series
- COM-HPC SERVER E

AIMB-593 Q3'25
- EPYC™ Embedded 8000 Series
- MicroATX

Mainstream

AMD RYZEN Embedded

AIMB-522
- RYZEN™ Embedded 5000 Series
- MicroATX

AIMB-523
- RYZEN™ Embedded 7000 Series
- MicroATX

Low Power

AMD RYZEN EMBEDDED V / AMD RYZEN EMBEDDED R

SOM-6872 A1/A2
- RYZEN™ Embedded V2000 Series
- COM Express Compact

AIMB-229
- RYZEN™ Embedded V2000 Series
- Mini-ITX

MIO-5376
- RYZEN™ Embedded R2000 Series
- 3.5" SBC

AIMB-228R2
- RYZEN™ Embedded V1000/R1000/R2000 Series
- Mini-ITX

SOM-6873 Q2,'25
- RYZEN™ Embedded 8000 Series
- COM Express Compact

AIMB-2210 Q2,'25
- RYZEN™ Embedded 8000 Series
- Mini-ITX

Available | Developing | Planning

2023 — 2024 — 2025 AMD

* All product specifications are subject to change without notice

Advantech Product Roadmap – System Solutions

Embedded Systems

EPC-T3229

- RYZEN™ Embedded V2000 Series
- 1U THIN Embedded PC

EBC-B3522

- RYZEN™ Embedded 5000 Series
- 3U system, 500W PSU

AIR-520

- EPYC™ Embedded 7003 Series
- 4U System, 1200W PSU

AIR-770 Q2,25

- EPYC™ Embedded 8004 Series
- 4 x dual-slot GPU

Gaming

DPX-S450

- RYZEN™ Embedded V1000/R1000 Series
- 4 DP out

DPX-J100

- RYZEN™ Embedded V1000/R1000 Series
- 6 COM, intr. M.2

DPX-S451

- RYZEN™ Embedded R2000 Series
- DC / ATX power

DPX-M266

- RYZEN™ Embedded R2000 Series
- 4 DP out

Digital Signage

DS-082

- RYZEN™ Embedded V1000/R1000 Series
- 4/3 x HDMI 2.0, 1 x LAN, 1 x COM & 4 x USB

EBC-V001(AEVA) Q4'24

- RYZEN™ Embedded R2000 Series
- Dual EDSFF E1.S accelerator

DS-084 Q2,'25

- RYZEN™ Embedded R2000 Series
- 4/3 x HDMI 2.1 TMDS
- EDID/OOB Option

Available | Developing | Planning

2023 — 2024 — 2025

* All product specifications are subject to change without notice

ADVANTECH × AMD

Download Now