FPT AI FACTORY

FPT Smart Cloud, 2025



Table of Content

1 About FPT Smart Cloud

2 FPT AI Factory

3 Service Offerings

4 Use Cases & Case Studies

About
FPT Smart Cloud



The Global ICT Corporation

FPT is a **pioneer in digital transformation** and a leader in consulting, providing and deploying technology and telecommunications services and solutions.

Revenue	**$2.42 billion (FY2024)**
No. of Employees	**50,000+**
Global Presence	**30 countries & territories**

Technology Sector

FPT Smart Cloud

Flagship technology firm that provides:

- **#1 AI Platform in Vietnam**
- **World-class Cloud Infrastructure and Services**
- **#1 world's supercomputers (Top 36 in Japan & Top 38 in Vietnam)**

FPT Software

FPT Digital

FPT Information System

Telecommunications Sector

FPT Telecom

FPT Online

Education Sector & More

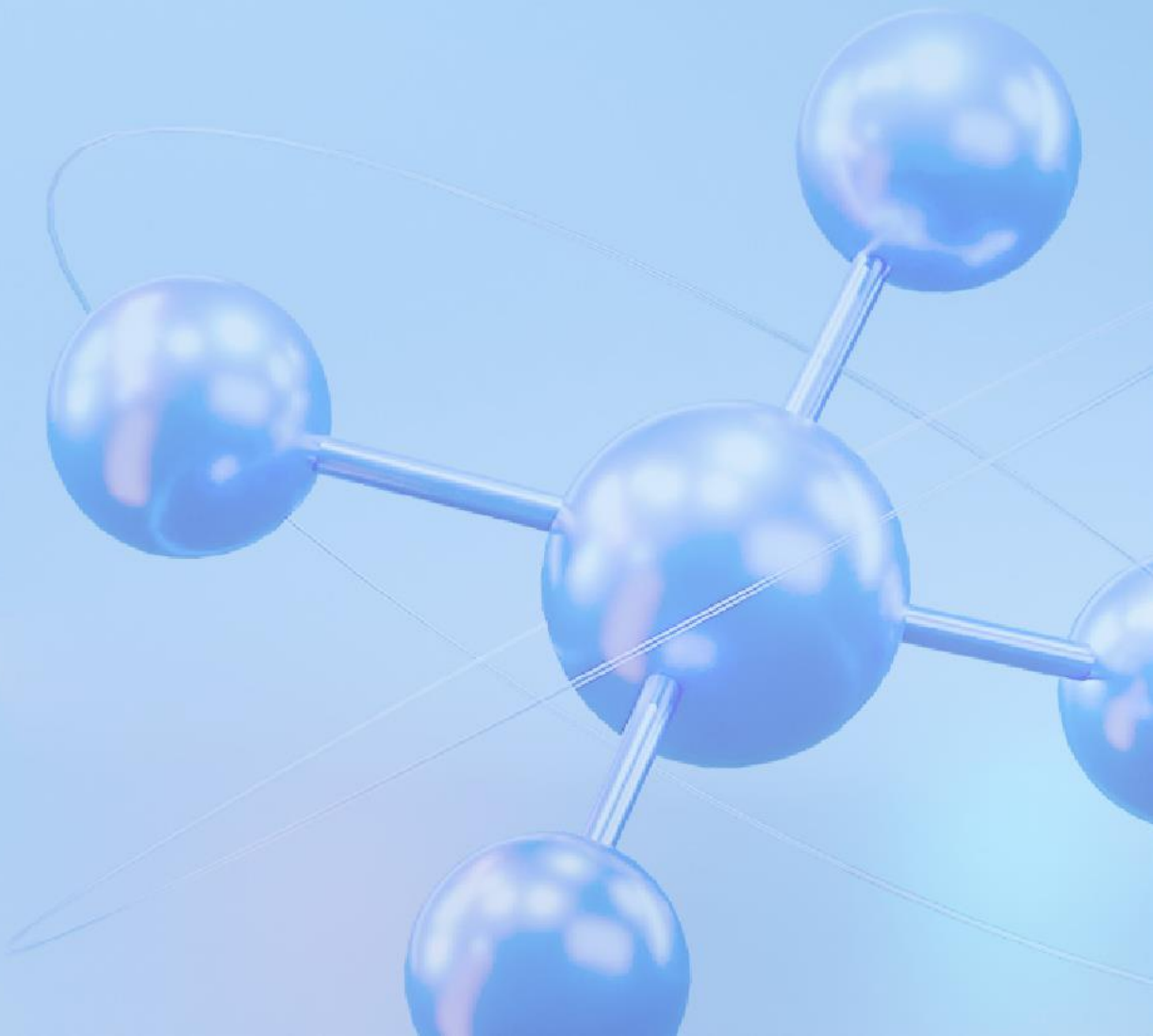
FPT Education

FPT Investment

FPT Smart Cloud

FPT Smart Cloud - a subsidiary of **FPT Corporation**, inherited a solid foundation of 35 years of developing core technologies with more than 72,000+ global talents, leading the industry in AI and Cloud, generating an immense leap in productivity and agility in business operations.

FPT Smart Cloud is **a world-class AI and Cloud provider** with the mission to accelerate digital transformation for business through agility and innovation, aiming to create breakthroughs in operations with first-rate services. We are also a trusted technology firm with more than 3,000 major global customers across various industries, including finance, banking, and retail.



The Leading AI & Cloud Platform in Southeast Asia

FPT provides a comprehensive AI platform, world-class Cloud services, and an advanced Data platform to enable process automation and effective human-machine collaboration, thus generating an immense leap in productivity and agility.

04
mega Tier III Data Centers

100+
AI products & Cloud services

1000+
AI Professionals

20M+
AI users with 200M+ interactions/month

80+
International partners

Strategic Partners

NVIDIA Preferred Partner
Mila
LANDING AI
AITOMATIC

Certification & Awards for AI, Cloud Services

Prestigious Awards

20+ International Certificates

International Clients

J.P.Morgan, HSBC, standard chartered, Deutsche Bank, HOME CREDIT, DAI-ICHI LIFE, MIRAE ASSET Finance Company, Yuanta Securities, Daiwa Securities Group Inc., Liberty Mutual, PRUDENTIAL, Sun Life, AIA, FWD, BANK OF AMERICA, MIZUHO, FUJITSU, Fubon Life, SCSK, 3M, TOYOTA, HONDA, mazda, AIRBUS, BOEING, Panasonic, CHUBB, Grab, TOKYO GAS, TOSHIBA, BOSCH, SONY, HONDA, Kawasaki, SEIKO, SCHAEFFLER, YAMAHA, Continental, Qualcomm, AT&T, PETRONAS, DAIKIN, E.ON, AstraZeneca, BAYER, OMRON, MedAdvisor

Domestic Clients

Vietcombank, BIDV, MB, VietinBank, VPBank, Sacombank, VIB, SSI, AGRIBANK, TPBank, SHBFinance, TECHCOMSECURITIES, HDBank, FE CREDIT, BAOVIET, inmart, VietnamAirlines, NHÀ THUỐC LONG CHÂU

About
FPT AI Factory



FPT AI Factory System Architecture Certified by NVIDIA

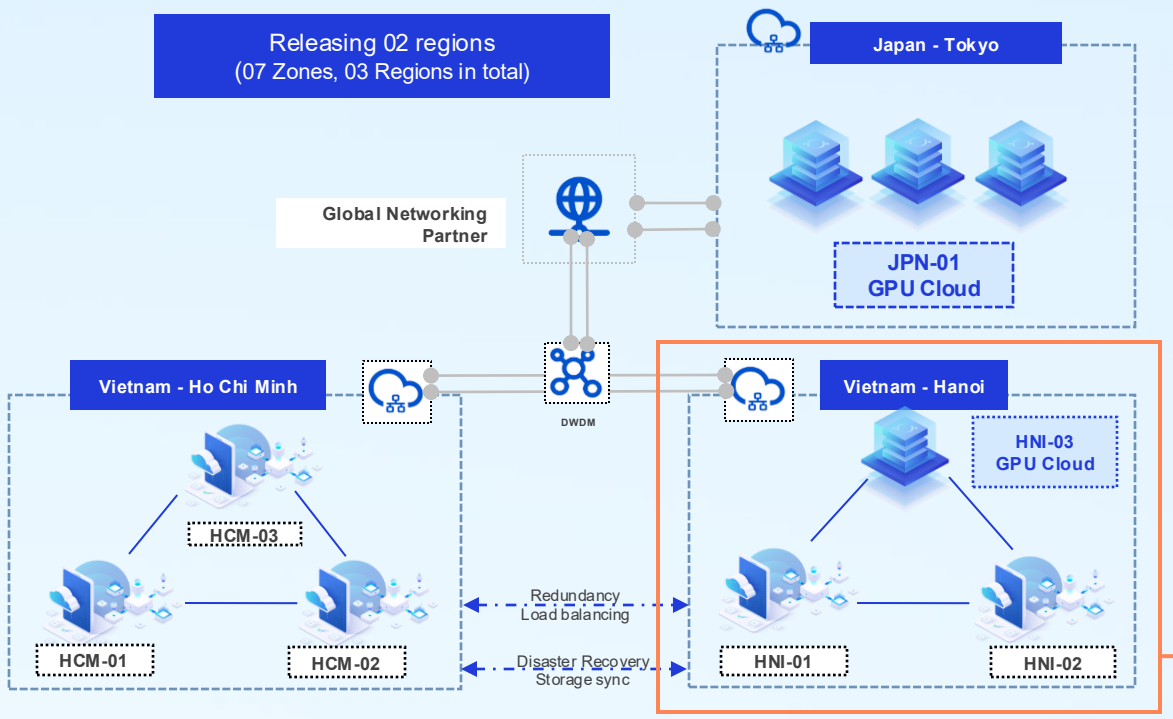
FPT AI Factory services are deployed in multi-regions – Vietnam and Japan. Particularly, the architecture of FPT AI Factory complies with NVIDIA's global standards, and the implementation and operations are certified by NVIDIA.

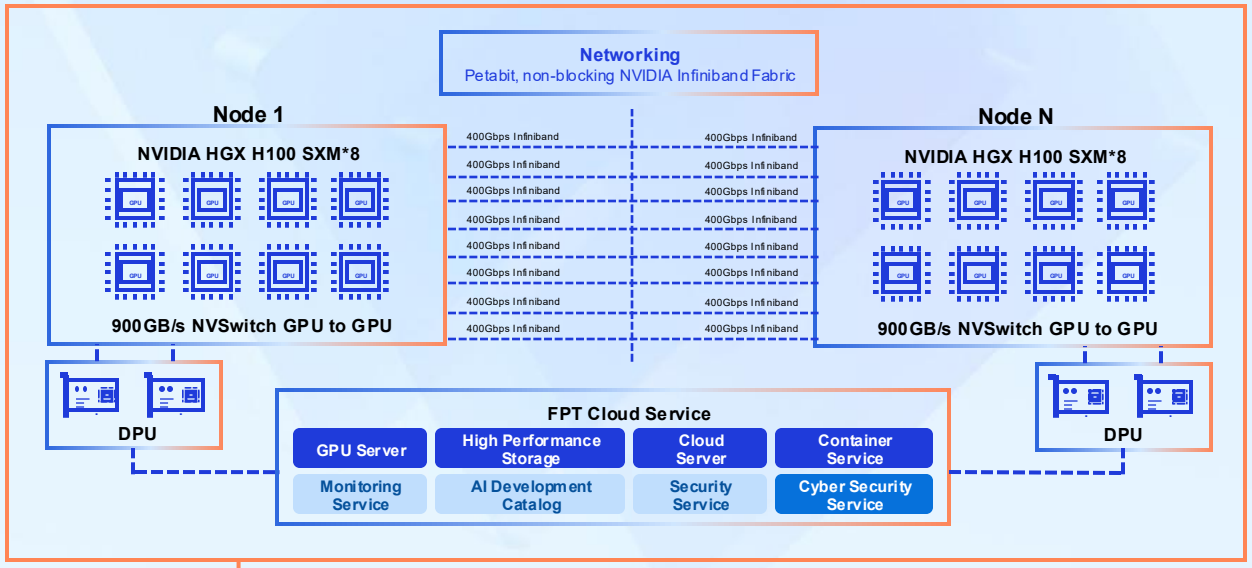
- A supercomputing cluster for large-scale requirements, with **thousands of NVIDIA Hopper GPUs (H100, H200…)**.
- Ultrafast networking includes non-blocking **InfiniBand fabric and Data processing unit (DPU)** for built-in hardware acceleration.

Integration with **NVIDIA AI Enterprise** software platform offering 100+ frameworks, pretrained models, and libraries.

NVIDIA AI Enterprise

- NVIDIA NGC
- NVIDIA Inference Microservices
- NVIDIA NeMo
- NVIDIA Base Command Manager





FPT AI Factory

In Vietnam & Japan

The latest NVIDIA GPUs (GPU H200, H100)

Provision of highly flexible services

AI Product

FPT AI APPLICATION

- AI Mentor
- Digital Customer Onboarding
- Intelligent Document Processing
- AI Contact Center Enhancement
- Conversational AI & AI Agents

MARKETPLACE

- AI Marketplace

AI Platform

FPT AI AGENTS

- Agent Studio
- Planning
- Memory
- Tools
- Guardrails
- Fine-Tuned LLM models

FPT AI STUDIO

- AI Notebook
- Data Hub
- Model Fine-Tuning
- Model Testing
- Model Hub

FPT AI INFERENCE

- Serverless Inference
- Dedicated Inference

Cloud Infras. & Platform

FPT AI INFRASTRUCTURE

- GPU Container
- GPU Cluster
- GPU Virtual Machine
- Metal Cloud (Bare Metal)
- File Storage

Physical Infrastructure

- NVIDIA H100/H200
- High-End Networks
- High-End CPU
- High-Performance Storage

NVIDIA AI Enterprise

- NVIDIA NGC
- AI Solution Workflows, Frameworks & Pretrained models — NVIDIA NeMo
- AI & Data Science Development Tools — NVIDIA Inference Microservices
- Cloud Native Management & Orchestration — NVIDIA Base Command Manager
- Infrastructure Optimization
- Infrastructure Management

Offerings for Customers

Managed by FPT

Accelerating a Successful AI Journey with FPT AI Factory

The Finest Approach for AI Advancement & Sovereignty

Top-notch AI Infrastructure



Access powerful supercomputers equipped with the first-rate NVIDIA GPUs, ready for any use case with instant provision and scale.

Diverse, Affordable Products & Services



Explore a variety of ready-to-use AI services, applications, and solutions for instant results. Suitable for any size of business with reasonable pricing tiers.

Streamlined Resource Management



Archive up to 45% better TCO through predictable and optimized resource management.

Built-in Security & Compliance



Run AI/ML securely with enterprise-grade protection, ensuring data sovereignty and safety.

FPT AI Factory

FPT AI Factory in Japan
Top 36/500 June 2025


TOP 500 CERTIFICATE
The List.

FPT AI Factory Japan - HGX H200, Xeon Platinum 8558 48C 2.1GHz, NVIDIA H200 SXM5 141 GB, Infiniband NDR400, Ubuntu 22.04.5 LTS

FPT AI Factory Japan, Japan

is ranked

No. 36

among the World's TOP500 Supercomputers

with 49.85 PFlop/s Linpack Performance

in the 65th TOP500 List published at the ISC2025

Conference on June 14, 2025.

Congratulations from the TOP500 Editors

Erich Strohmaier, NERSC/Berkeley Lab
Jack Dongarra, University of Tennessee
Horst Simon, NERSC/Berkeley Lab
Martin Meuer, Prometeus

FPT AI Factory in Vietnam
Top 38/500 June 2025


TOP 500 CERTIFICATE
The List.

FPT AI Factory Vietnam - ASUS HGX H100, Xeon Platinum 8462Y+ 32C 2.8GHz, NVIDIA H100 SXM5 80GB, Infiniband NDR400, Ubuntu 22.04.5 LTS

FPT AI Factory Vietnam, Vietnam

is ranked

No. 38

among the World's TOP500 Supercomputers

with 46.65 PFlop/s Linpack Performance

in the 65th TOP500 List published at the ISC2025

Conference on June 14, 2025.

Congratulations from the TOP500 Editors

Erich Strohmaier, NERSC/Berkeley Lab
Jack Dongarra, University of Tennessee
Horst Simon, NERSC/Berkeley Lab
Martin Meuer, Prometeus

About
Service Offerings



FPT AI Infrastructure

	Metal Cloud	GPU Virtual Machine	GPU Cluster	GPU Container
Definition	• **Dedicated physical servers (Bare Metal) for a single tenant** • Direct access to physical hardware	• **Virtual machine (VM) with dedicated GPU Cards** • Flexible resource scale and deployment	**Manage a multiple** of Bare Metal GPU servers or Virtual machines **by Kubernetes**	• **Containers with dedicated GPU cards** • No need to handle the infrastructure
Purposes	• **Deep training at scale** • **Complex inference needs** • **Custom orchestration** • Low latency and strict security requirements	• **Model development** • **Small- to medium-scale training or inference** tasks • Flexible, isolated, fast start	• **Distributed training** • **Scalable inference** • Multi-GPU and multi-node orchestration	• **Notebook or development tools** • **Inferencing and lightweight AI training** • Lightweight, portable environments
SLA	99%	99% for VM use NVMe storage 99.5% for VM use Block storage	99.95% for Control Plane	99.9%
GPU Models	**Dedicated GPUs** for a Bare Metal server	**Dedicated GPUs** for a Virtual machine	**Dedicated GPUs** for a Bare Metal server or Virtual machine	• **Dedicated GPUs** for a container • **Support GPU virtualization** (MIG, Timeslicing)
Key Features	• Fully control over hardware for customization • Maximum performance of GPUs • GUI self-service to manage servers and monitor	• Quick spin-up/down • GUI self-service to deploy & configure VMs, manage networking	• Faster provisioning with GPU softwares (GPU Operator, vGPU scheduler) and GPU drivers • Ready-to-use muptiple Kubernetes versions	• Scalable, on-demand GPU • Container catalogs (Ready-to-use templates) • Multi-ports and SSH support
Storage options	• **Included NVMe Local storage** • High-performance Storage (Optional)	• **Offered 01 in 02 types**: NVMe (Ephemeral) or SSD (Block) • High-performance Storage (Optional)	• Depends on Worker node types • High-performance Storage (Optional)	• **Included NVMe Temporary storage** • Persistent Storage (Optional)
Package	From 01 server (8x GPUs)	From 1x GPU/VM	8x GPUs with Bare metal or 1x GPU with Virtual machine	From 1x GPU/Container
Billing plans	Reserved	• Reserved • Pay-as-you-go (hourly)	• Reserved • Pay-as-you-go (hourly, Worker node is GPU VM)	Pay-as-you-go (Per-minute usage charge)

Metal Cloud Offerings

Our packages
(Single node or Multi nodes)

Name	Metal Cloud GPU H100	Metal Cloud GPU H200
GPU	8* NVIDIA H100 SXM5 640GB Memory (8* 80GB)	8* NVIDIA H200 SXM 1.1TB HBM3e Memory (8* 141GB)
CPU	Dual Intel Xeon Platinum Processor 8462Y+	Dual Intel Xeon Platinum Processor 8558
Memory	2TB via 4800MHz DDR5 DIMMs	2TB via 4800MHz DDR5 DIMMs
Storage	30TB (8 x 3.84TB NVMe SSD)	30TB (8 x 3.84TB NVMe SSD)
Network	• 400Gbps* 8 ports InfiniBand • 200Gbps* 2 ports BF3 DPU	• 400Gbps* 8 ports InfiniBand • 200Gbps* 2 ports BF3 DPU

- *GPU, CPU, RAM, and NVMe storage per node are not resizable.*
- *H100 is available in Vietnam, H200 is in Japan*

Recommendation (In addition)

- **File Storage - High-performance**
  for sharing training data between multiple nodes
- **Managed GPU Cluster** with **Kubernetes** Orchestration
- **Virtual Machine (General purpose)**
  For the management node and login node of a GPU cluster
- **Object Storage**
  To distribute your training data files
- **Load Balancer**
  For a highly available infrastructure
- **Firewall**
  To secure your AI workload
- **Networking**
  High-speed Internet or Direct Connect

GPU Virtual Machine Offerings

GPU packages (per Virtual machine)

Name	GPU	vCPUs	Memory (RAM)
1xH100	1x GPU 80GB Memory	16 Cores	192GB
2xH100	2x GPUs 160GB Memory	32 Cores	384GB
4xH100	4x GPUs 320GB Memory	64 Cores	768GB
8xH100	8x GPUs 640GB Memory	128 Cores	1536GB
1xH200	1x GPU 141GB Memory	12 Cores	192GB
2xH200	2x GPUs 282GB Memory	24 Cores	384GB
4xH200	4x GPUs 564GB Memory	48 Cores	768GB
8xH200	8x GPUs 1.1TB Memory	96 Cores	1536GB

Storage (Choose 01 in two below types per Virtual machine)

Types	Specification
Ephemeral storage NVMe	**3 TB per GPU** (Total capacity depends on GPU count on GPU package)
Block storage SSD	**Free 250GB per GPU, Scale up from 1GB on demand** (Total capacity depends on GPU count on GPU package)

Recommendation (In addition)

- **Managed GPU Cluster** with **Kubernetes** Orchestration
- **File Storage - High-performance**
  for loading the training data across hundred of GPU nodes
- **Virtual Machine (General purpose)**
  For the management node and login node of a GPU cluster
- **Object Storage**
  To distribute your training data files
- **Load Balancer**
  For a highly available infrastructure
- **Firewall**
  To secure your AI workload
- **Networking**
  High-speed Internet or Direct Connect

- *GPU, CPU, RAM, and NVMe storage per VM are not resizable.*
- *H100 is available in Vietnam, H200 is in Japan*

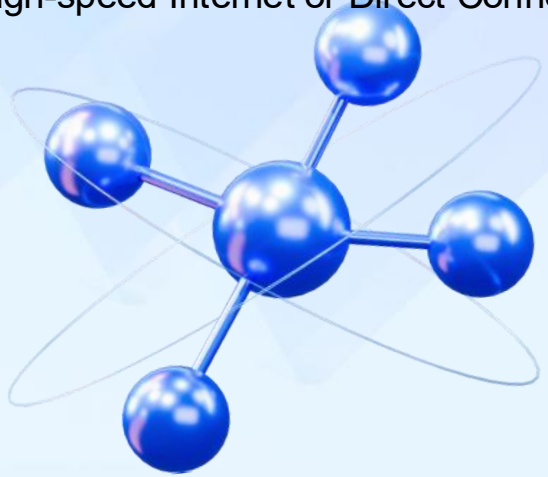
GPU Cluster & File Storage Offerings

Our packages (per a GPU cluster)	
Control plane	Managed by FPT, For cluster & resource management
Load Balancer	Managed by FPT, Built-in for Kubernetes service management
Worker Nodes	Bare Metal or Virtual Machines
Node pricing	The service is billed by packages of virtual machines or bare metal servers (Metal Cloud).
Nodes per Worker group	From 01 bare metal GPU server

File Storage High Performance (for loading the training data across hundered of GPU nodes)	
Capacity	As demand, from 50GB
Performance	Max Throughput: 550 GBps
Network	200Gbps * 2 ports (EDR/HDR InfiniBand)
Feature	Parallel File system, NFS and S3 Gateways
GPU services	Supports Metal Cloud, GPU Cluster, GPU VM

Recommendation (In addition)

- **Add High Availability (HA) Control plane**
- **Load Balancer**
  for Applications or Workloads
- **Floating IP address**
  for Load Balancer
- **Object Storage**
  To distribute your training data files
- **Networking**
  High-speed Internet or Direct Connect



GPU Container Offerings

GPU, CPU, RAM per container cannot be expanded.
H100 is available in Vietnam, H200 is in Japan

Container Packages (GPU instance per container)

Name	GPU	CPU	RAM	NVME Storage
1xH100	1x GPU 80GB Memory	15	250GB	1TB
2xH100	2x GPUs 160GB Memory	30	500GB	2TB
3xH100	3x GPUs 240GB Memory	45	750GB	3TB
4xH100	4x GPUs 320GB Memory	60	1000GB	4TB
5xH100	5x GPUs 400GB Memory	75	1250GB	5TB
6xH100	6x GPUs 480GB Memory	90	1500GB	6TB
7xH100	7x GPUs 560GB Memory	105	1750GB	7TB
8xH100	8x GPUs 640GB Memory	120	2000GB	8TB

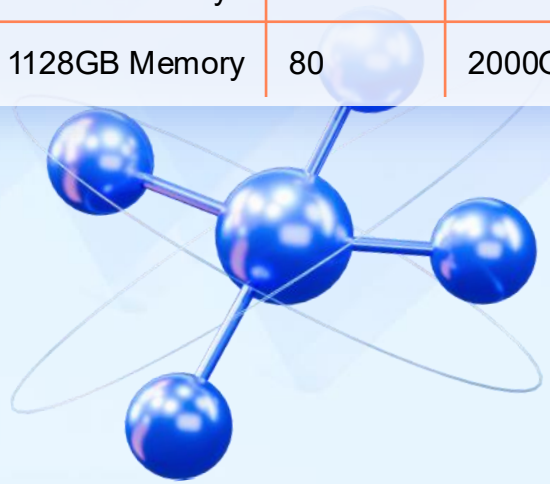
Container Packages (GPU instance per container)

Name	GPU	CPU	RAM	NVME Storage
1xH200	1x GPU 141GB Memory	10	250GB	1TB
2xH200	2x GPUs 282GB Memory	20	500GB	2TB
3xH200	3x GPUs 423GB Memory	30	750GB	3TB
4xH200	4x GPUs 564GB Memory	40	1000GB	4TB
5xH200	5x GPUs 705GB Memory	50	1250GB	5TB
6xH200	6x GPUs 846GB Memory	60	1500GB	6TB
7xH200	7x GPUs 987GB Memory	70	1750GB	7TB
8xH200	8x GPUs 1128GB Memory	80	2000GB	8TB

Storages

Persistent Storage

- **Recommend,** Suitable for storing data, models, or files that need to be retained even the instance is halted or rebooted.
- **Capacity: As demands, from 01GB**

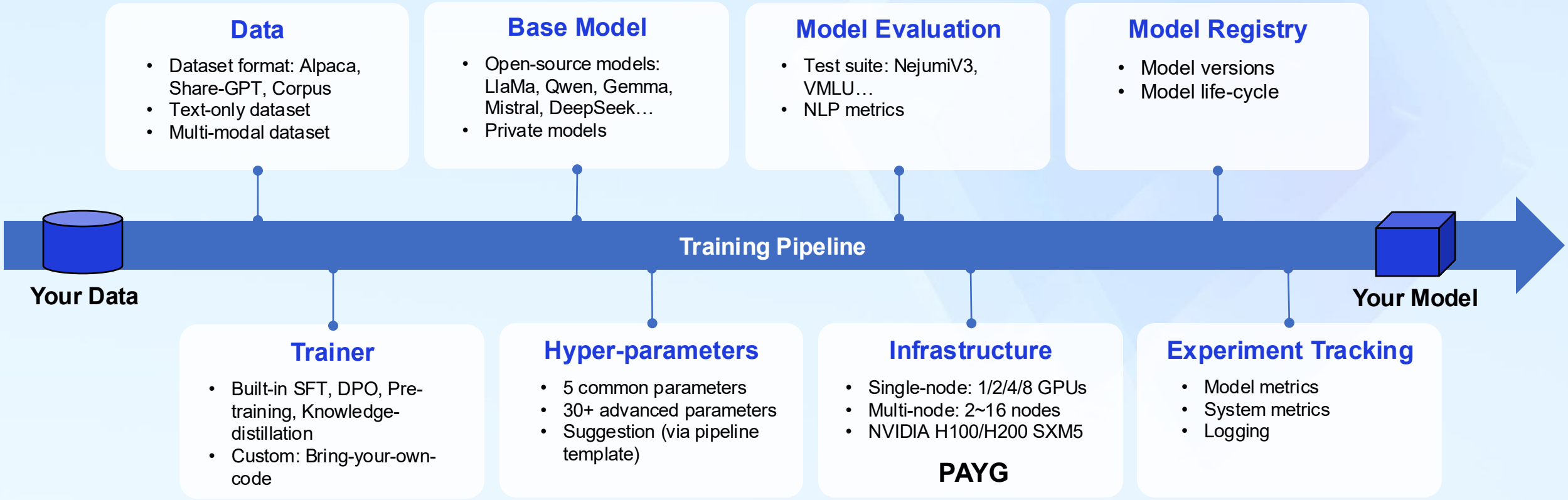


FPT AI Studio – The Core Concept

FPT AI Studio empowers corporate customers to master AI by providing an integrated tool suite for Generative AI development.

With FPT AI Studio, AI engineers can:

- Start training LLM/VLM models very quickly – without worrying about infrastructure, setup, or cost
- Tweak every aspect of model training & evaluation – perform as many deep experiments as needed
- Take full control of the model training process – bring your own models & source code



FPT AI Studio

	Model Fine-tuning	Model Hub	AI Notebook	Data Hub
Definition	Managed service for fine-tuning of AI models, using customer's data and a wide range of base models.	Managed service for centrally storing custom models and their respective crucial information.	Managed service for integrated development environment (IDE), backed by GPU container.	Managed service for processing raw dataset, making high quality dataset.
Purposes	Tune or adapt a pre-trained model for specific tasks or domains. • Model fine-tuning. • Continual pre-training.	• Managing model life-cycle; • Collaborating through out AI projects. • Integrate with model fine-tuning and model serving as source and sink.	• Data exploration. • Model training experiment. • Other tasks that require coding.	• Pulling training data from various data sources. • Make data ready for model training. • Make data reusable across training jobs.
SLA	99.9%	99.9%	99.9%	99.9%
GPU Operation Model	• **Dedicated GPUs** for each model training job. • Each training job is performed in its own training containers.	No GPU required	**Dedicated GPUs** for each container.	GPU is optional.
Key Features	• **No-code**: user only need to provide train/eval/test data. • **Low-code**: user can bring their own training container. • User-friendly GUI.	• Upload your own model or choose from model catalog. • Secured storage and fine-grained permission control. • Sharing within the organization • Keep track of all model versions, evaluation results, and serving history.	• Serverless. • Jupyter Notebook. • Control of long running container. • AI frameworks pre-installed.	• On-demand CPU and GPU • Data connector • Data transformation, deduplication • Data encryption
Package	From 1xGPU / training job.	By storage quota.	From 1xGPU / container.	By Infrastructure resource
Billing Model	• Pay-as-you-go (hourly) • Reservation	• Reservation	• Pay-as-you-go (hourly)	• Pay-as-you-go (hourly) • Reservation

FPT AI Inference

	Dedicated Inference (Model Serving)	Serverless Inference
Definition	Allowing companies to deploy models as dedicated endpoints with custom hardware and scaling configurations, without managing the underlying infrastructure	• Provide AI models as APIs, allowing companies to integrate pre-trained AI models into applications, without needing to develop, train, or manage AI models.
Purpose	• Consistent, predictable performance, unaffected by other users' load • No rate limits, with a high maximum load capacity • More cost-effective under high utilization • Access to a broader selection of models	• Faster Time to Market: ease of integration & deployment • Cost efficiency: Pay as you go • Scalability basd on demand
SLA	Availability: 99.95%	Availability: 99.95%
GPU Operation Model	Serverless - Customers can select custom hardware and scaling options	Serverless
Key Features	• Supporting OS, FPT's and User's own models • Orchestral inference: allow using the same endpoint & API Keys for all models • Easy-to-use Deployment & Scaling configuration • Real-time Usage monitoring • Isolated endpoint: enhancing security & allowing personalized configuration	• OS & FPT's models: LLM, VLM, Multimodal, Embeddings, Text to speech, Speech to text • Easy Integration via API • Auto-scale based on demand • Continuous updates to improve performance and provide SoTA models • Fine-tuning in FPT AI Studio
Package	On demand	On demand
Billing Model	Pay as you go: Inference minute/ hour	Pay-as-you-go: Token/Character/Image/Length-based pricing (depending on the model type)

AI Inference – Serverless Inference

Model	Model category	Quantization	Context length	Max output token	Price USD/ M tokens
Llama-Guard-3-8B	Guardrail	FP8	8K	NA	0.08
LLama 3.1 8B instruct	Lange Language Model	FP8	128K	128K	0.08
LLama 3.3 70B instruct	Lange Language Model	FP8	128K	128K	0.36
Llama-3.1-Swallow-70B-Instruct-v0.1	Lange Language Model	FP8	128K	128K	0.36
Deepseek R1	Lange Language Model	FP8	164K	164K	3
DeepSeek-R1-Distill-Llama-70B	Lange Language Model	FP8	128K	128K	0.36
DeepSeek-R1-Distill-Qwen-32B	Lange Language Model	FP8	128K	128K	0.3
DeepSeek-R1-Distill-Llama-8B	Lange Language Model	FP8	128K	128K	0.08
DeepSeek-R1-Distill-Qwen-1.5B	Lange Language Model	FP16	128K	128K	0.06
QwQ (alternative to Deepseek R1)	Lange Language Model	FP8	128K	128K	0.3
Qwen2.5-Coder-32B-Instruct	Lange Language Model	FP8	128K	8K	0.3
Qwen 2.5 7B	Lange Language Model	FP8	128K	8K	0.08
SaoLa LLama 3.1 Planner (8B)	Lange Language Model	FP8	128K	128K	0.08
SaoLa3.1-medium (32B)	Lange Language Model	FP8	128K	128K	0.3
Qwen2.5-VL-7B-Instruct	Multimodal – Text & Image	FP16	33K	8K	0.7 *
FPT.AI-KIE-v1.7	Vision Language Model	FP16	33K	8K	0.7 *
FPT.AI-Table-Parsing-v1.1	Vision Language Model	FP16	33K	8K	0.7 *
FPT.AI-e5-large	Text Embeddings	FP16	8K	1024-dimensional	0.02
FPT.AI-gte-base	Text Embeddings	FP16	8K	768-dimensional	0.016
Vietnamese_Embedding	Text Embeddings	FP16	8K	1024-dimensional	0.008
Llama-4-Scout-17B-16E	Multimodal – Text & Image	FP16	128K	128K	Coming soon

* *With monthly minimum 1B token package*

AI Inference provides AI models as API endpoints, allowing companies to integrate pre-trained AI models into applications without the need to develop, train, or manage models.

- **AI models**: LLM, VLM, Embedding, Multimodal, ASR, Text to speech, Speech to text and more
- **Fast inference**: <3s Time to first token
- **Stability & High availability:** Auto-scale resources based on demand

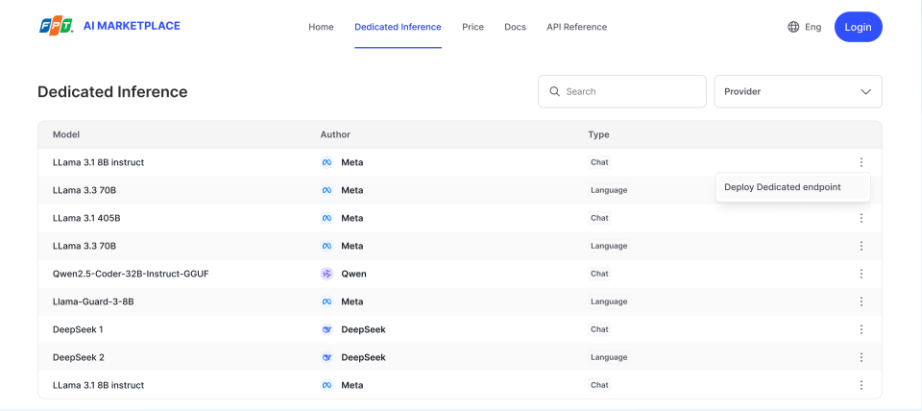
Other services:

- **Model customization**: Tailoring Foundation models for your specific use cases securely with the support of our AI expert team

Dedicated Inference

Dedicated endpoint allows companies to deploy models as dedicated endpoints with custom hardware and scaling configurations, without managing the underlying infrastructure.

Supported Models	Difference from Serverless Inference	Features
• Open-source models: LLM, VLM, Embedding, Multimodal, ASR, Text to speech, Speech to text and more • Fine-tuned models (including LORA) • Other custom models	• Consistent, predictable performance, unaffected by other users' load • No rate limits, with a high maximum load capacity • More cost-effective under high utilization • Access to a broader selection of models	• Orchestral inference: allow using the same endpoint & API Keys for all models *(both Serverless & Dedicated inference)* • Easy-to-use Deployment & Scaling configuration • Real-time Usage monitoring • Isolated endpoint: enhancing security & allowing personalized configuration
How to use	**Pricing model**	**SLA**
• GPU Operation Model: Serverless, no need to reserve any resource in advance or manage infrastructure.	• PAYG • Pricing Unit: Minute/ Hour/ Day/ Month • Price range variation: depending on Model, Hardware & Scaling configuration	• Monthly Uptime >= 99.95%



AI Factory Billing & Payment models

	Billing plans		Payment terms		Payment methods	
Options	**Reserved**	**Pay-as-you-go (PAYG)**	**Prepaid**	**Postpaid**	**Online via Cards**	**Offline via sales/AM**
Definitions	• Resource is dedicated for 01 customer. • Fixed size and duration of use.	Billed based on actual usage	**Partially or fully paid in advance** before using services	• **Pay after using services** • Long-term commitment	Buy credits by cards before using services	Bank transfer with guidance from the Sales team.
Metal Cloud	✅ >=1 node Contract Contract >= 1month	❌	✅	❌	❌	✅
GPU Virtual machine	✅ From 1 card Contract >= 1month	✅ From 1 card	✅	✅	❌	✅
GPU Cluster (K8S)	✅	✅	✅	✅	❌	✅
GPU Container	❌	✅	✅	✅	✅	✅
File storage – High performance	✅	❌	✅	❌	❌	✅
AI Studio	❌	✅	✅	✅	✅	✅
Serverless Inference	❌	✅	✅	✅	✅	✅

Security & Compliance



Information Security and Data Protection Principles

- **FPT Smart Cloud maintains the infrastructure on the best security platforms and architectures to provide comprehensive solutions, ensuring the highest security for customers.**

Comply with international standards and certifications

- **ISO/IEC 27001:2013**: Information Security Management Systems (ISMS)
- **ISO/IEC 27017:2015**: Code of practice for information security controls based on ISO/IEC 27002 for cloud services
- **ISO/IEC 27018:2019**: Code of practice for protection of personally identifiable information (PII) in public clouds acting as PII processors
- **PCI DSS**: Payment Card Industry Data Security Standard
- **SOC 2**: Systems and Organization Controls 2 by American Institute of Certified Public Accountants (AICPA)

Standards of Information security at the Cloud Service level

Standards of Information security at the Data Center level

- ✓ A team of security experts with advanced AI technology continuously monitors the system and proactively warns customers 24x7x365
- ✓ Integrate and utilize advanced technologies from the world's leading security partner
- ✓ Dedicated and Secure Infrastructure: A private cloud environment exclusively for the Customer, with isolated network, servers, and systems accessible only by authorized Customer personnel and designated FPT employees via VDI.
- ✓ End-to-End Encryption: All data in transit (Metal Cloud, GPU Managed Cluster, GPU Container) is secured with TLS 1.3, while stored data leverages AES-256 encryption. Data at Rest is seamlessly encrypted and decrypted with 256-bit AES-XTS encryption.
- ✓ Secure Data Erasure: Data is permanently deleted from the GPU infrastructure and storage following NIST SP 800-88 standards when no longer needed.

Case Studies



Case Study: Enhance Customer Support and Workforce Development

Use Cases

Generative AI / LLMs
Conversational AI / NLP

Products

NVIDIA H100 Tensor Core GPUs
NVIDIA A100 Tensor Core GPUs

Results

- Reduced data processing by 3X and speech synthesis processing by 4X, resulting in faster model training
- Improved speech synthesis model output in dimension and aptitude by 100X for 18 languages
- Sped up large language model training by 30X compared to the previous generation

Customer Outcomes

Customer Service Enhancement

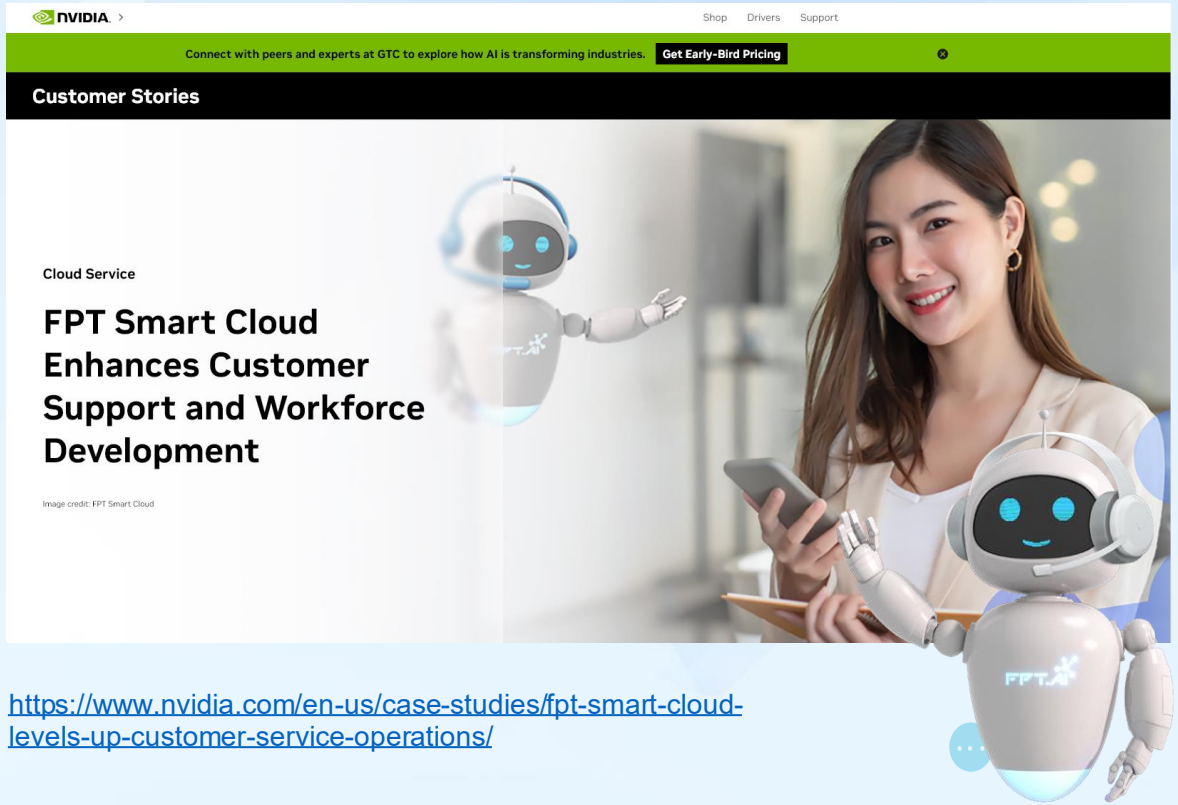
Omni-channel virtual assistants to foster excellent customer experience
Tasks: Customer service, telesales, self-service…
Results:

- **12,000,000** calls/month
- Resolve **98%** of customer inquiries
- Customer satisfaction: 4/5

Human Resource Empowerment

Virtual assistant to improve employee competency and performance

- Participation: **10,000+ monthly active users**
- Employee competency increases by **15%**
- Save **80%** effort for employee training



https://www.nvidia.com/en-us/case-studies/fpt-smart-cloud-levels-up-customer-service-operations/

Case Study: Accelerate Intelligent Inspection Solutions for Smart Cities

Use Cases

Generative AI / LLMs
Conversational AI / NLP

Products

NVIDIA Metropolis
NVIDIA AI Blueprint
NVIDIA NeMo Guardrails

Applications

- Suitable for various smart city infrastructures, including **traffic management, urban safety,** and **automated insurance processing**
- Automate the detection of asset degradation, vandalism, and safety hazards in real-time, enabling authorities to quickly gather precise data without needing on-site personnel

Benefits

Intelligent Inspection (I2)

- **Real-Time Accident Detection & Damage Assessment:** Faster and accurate claim processing.
- **Cost Efficiency:** Reduced need for on-site inspections, optimizing workforce utilization.
- **Better quality with NVIDIA AI Blueprint Integration:** Enhanced video processing capabilities, seamless deployment, and parallel processing for faster insights.
- **Accelerates time to solution:** NVIDIA AI Blueprints eliminate months/years of development effort by removing the heavy lifting of using various models and NIM microservices together.



https://fptsoftware.com/newsroom/news-and-press-releases/news/fpt-partners-with-nvidia-to-revolutionize-smart-city-inspections-with-advanced-ai

Case Study: Expand Visual AI Platform with Speed and Efficiency

Challenges

As LandingAI expands the Visual AI platform, an AI agentic framework designed to help users solve complex visual tasks through natural language prompts and media inputs such as images or videos, it faces the main challenges in **managing the need for high computational power** to fine-tune the agents and run reinforcement learning loops for continual performance improvement, all while maintaining **rapid iteration speed** to keep pace with innovation

Solution

- Metal Cloud from FPT AI Factory
- NVIDIA H100 GPUs

Benefits

- **Significant improvements in visual task generalization** with high service stability and availability
- **Faster deployment** of customer-facing features

> “As LandingAI expands our Visual AI technology offerings, FPT AI Factory has provided us with a solid and flexible infrastructure for our large-scale AI development and deployment. Their system's reliability and flexibility have streamlined our Visual AI workflows, significantly reducing iteration time. We've seen improved operational stability in production and cost savings. Their responsive support has made integration seamless.”
>
> *Dan Maloney, CEO of LandingAI*



Case Study: Internal Implementation for FPT.AI

H100 GPU significantly reduces overall training expenses due to shorter training times and accelerates training by a factor of **3.5x** compared to A100.

Metric	16 x H100 (HBM3-80GB)	16 x A100 (PCIe-80GB)	Higher is Better?
Epoch Time (Qwen2.5VL-7B, total_batch_size=16, num_sample=500k)	~12 hours	~ 42 hours	No
GPU Utilization	100%	100%	Yes

Metric	2 x H100 (HBM3-80GB)	2 x A100 (PCIe-80GB)	Higher is Better?
Inference Throughput - Qwen2.5VL-3B (token/sec, PyTorch)	~410	~150	Yes
GPU Utilization	100%	100%	Yes

Resource Optimization

- **Maximize GPU Utilization**: Proper tuning of batch size, sequence length, and caching mechanisms.
- **Parallel Processing Strategies**: Implementing FSDP, ZeRO, and NCCL to improve training speed.

Distributed Training Challenges

- **Data Synchronization:** Efficient GPU communication to avoid bottlenecks.
- **Infrastructure Readiness:** Ensuring power and cooling support for high-energy-consuming H100 clusters.

System Integration & Stability

- **Software Stack Compatibility:** Ensuring seamless operation with PyTorch/XLA, Triton, and TensorRT.
- **Continuous Performance Monitoring:** Regular fine-tuning to maintain optimal efficiency.

Case Study: An FPT AI Factory Client in a Tech Firm

Fine-tuning LORA LLM: Using H100 is faster than 2.5x times of A100

Models	Data size	Time to train 8x A100 (hyperscaler)	Time to train 8xH100 (FPT AI Factory)
Llama3.3-70B-Instruct	550K (~1GB)	32 hrs	13 hrs
Qwen2.5-72B-Instruct	550K (~1GB)	38 hrs	15 hrs

Customer Feedback on the Advantages of the H100 GPU on FPT Cloud

- ✓ The H100 provides a good balance between performance and cost, optimizing AI training workloads.
- ✓ Running H100 on the FPT Cloud platform reduces domestic latency and avoids reliance on international infrastructure like AWS.
- ✓ Well-suited for fine-tuning LoRA models, especially in natural language processing and large-scale data synthesis.

Case Study: Speed Up LLM Development for FPT.AI

Adopting advancements of AI quickly and efficiently

Enabling Technologies

- NVIDIA H100 Tensor Core GPUs
- NVIDIA NGC, PyTorch

Outcomes

- Training parameters grew by **10X** in combination with more sophisticated learning objectives for the model and training data
- The H100 can speed up LLM training by **30X** over the previous generation
- The application of NVIDIA-powered retrieval-augmented generation (RAG) workflows reduces how often pre-trained models need to be updated and enables a wide and flexible range of tasks that can be done with the same pre-trained foundation model

Benchmarks on LLM performance

Model	Number of params	VMLU		lambada_vi		exams_vi		wikipediaqa_vi		comprehension_vi	
		0-shot		0-shot		5-shot		5-shot		0-shot	
		acc	error	acc	ppl	avg. acc	avg. acc_norm	acc	acc_norm	acc	acc_norm
SaoLa-Llama-3.1-planner-8B	8B	48.23	0	37.29	18.753	35.97	39.93	47.60	49.75	67.11	68.56
SaoLa-3-Medium	32B	**74.72**	0	-	-	-	-	-	-	-	-

Benchmarks on LLM (8B params, FP16) inference performance

Model name	Time to first token (ms)	Infer token latency (ms)	Request latency (ms)	Ouput sequence length	Input sequence length	Output token throughput (per sec)	Request throughput (per sec)	GPU
SaoLa3.1-planner	376.96	12.13	14233	1145	550.1	804.81	0.7	**A100**
SaoLa3.1-planner	45.28	7.19	4508.93	601.55	550.05	1221.96	2.06	**H100**

Case Study: Enable Fast & Efficient Speech Synthesis Model for FPT.AI

Developing custom AI voices to foster human-like experience

Enabling Technologies

- NVIDIA H100 Tensor Core GPUs
- NVIDIA A100 Tensor Core GPUs
- NVIDIA A30 Tensor Core GPUs
- NVIDIA Triton Inference Server

Outcomes

- Reduced training time by **2X**, resulting in faster model training and fine-tuning
- Improved speech synthesis model inference capacity by **14X**
- Improved speech synthesis model output in dimension and aptitude by **100X** for 18 languages

H100 GPU significantly reduces overall training expenses due to shorter training times and accelerates training by a factor of **2x** compared to A100.

Metric	H100 (HBM3-80GB)	A100 (PCIe-80GB)	Higher is Better?
Training time (100k steps) (VITS-ISTFT, ~200h audio data)	~45 hours	~ 86 hours	No
GPU Utilization	100%	100%	Yes

Combined with Triton Server's batching abilities, the speedup offered by H100 allows us to serve **14x** more simultaneous calls than with A30.

Metric	A30	H100	H100 w/ Triton	Higher is Better?
Supported simultaneous calls (with latency < 1s)	~50	~150	~700	Yes
GPU Utilization	100%	100%	100%	Yes

Thank you!

