FPT AI FACTORY

FPT Smart Cloud, 2025

About
FPT Corporation

The Global ICT Corporation

FPT is a **pioneer in digital transformation** and a leader in consulting, providing, and deploying technology services and solutions with **200+ Made-by-FPT products** to drive sustainable growth for organizations and excellent experiences for customers.

FPT

Technology Sector

- **#**1 Vietnamese technology enterprise to reach the **billion-dollar mark in revenue** from global IT services.
- #1 Vietnam provider of IT services, software solutions, and system integration.
- Top 1 Global AI Platform 2023 Award by Software Reviews.

Telecommunications Sector

- #1 Vietnam's Data center provider
- #1 Vietnam's e-newspaper – VnExpress.net with 46M+ frequent readers

Education & Investment Sector

#1 Vietnam's Private University

- 40+ countries of exchange partnership
- 170,000+ students at all levels

Commerce Sector

- #1 Vietnam's pharmacy chain
- #1 Vietnam's IT distributor
- #2 Vietnam's retail chain of technology products

International Recognitions

FORTUNE

- Recognized in Fortune Southeast Asia 500 as the Largest IT Services Company in the Region

Gartner

- Top 50 IT Services Companies in Asia by Market Share 2023
- Gartner Market Guide for Public Cloud Managed and Professional Services, Asia/Pacific 2024

IDC

- Major Player in the Asia/Pacific Manufacturing Execution Systems 2023

FORRESTER®

- The Application Modernization And Multicloud Managed Services Landscape, Q4 2024
- Modern Application Development Services, 2024

Everest Group®

- Everest PEAK Matrix® Assessment: Major Contender in ACES Automotive Engineering Services 2023

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

80+
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

Software Reviews Champion
Asian Technology Excellence Awards
IDC
SHINRA
ASOCIO Reimagining the Digital Era
Vietnam Digital Awards

20+ International Certificates

PCI DSS Compliant
AICPA SOC
TÜV NORD ISO/IEC 27018
TÜV NORD ISO/IEC 27017
TÜV NORD ISO/IEC 27001
CMMI DEV | ML5 Appraised, Appraisal # 66736 | Exp. Oct 20, 2026
TMMi Certified Level 5
Certified ISO 9001
iBeta Level 1 ISO 30107-3 Compliant

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 SCSKサービスウェア, 3M, TOYOTA, HONDA, mazda, AIRBUS, BOEING, Panasonic, CHUBB, Grab

TOKYO GAS, TOSHIBA, BOSCH, SONY, HONDA, Kawasaki, SEINO, SCHAEFFLER, YAMAHA

Ontinental, Qualcomm, AT&T, PETRONAS, DAIKIN, e.on, AstraZeneca, BAYER, OMRON, MedAdvisor

Domestic Clients

Vietcombank, BIDV, MB, VietinBank

VPBank, Sacombank, VIB, SSI

AGRIBANK, TPBank, SHBFinance Tài chính tiêu dùng

TECHCOMSECURITIES, HDBank, FE CREDIT

BAOVIET, inSmart, Vietnam Airlines, Nhà thuốc Long Châu

FPT AI Factory Service Offerings

FPT AI Factory

07 Zones in 02 Regions in Viet Nam (Hanoi & Ho Chi Minh)
& 01 Region in Japan (Chiba)

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 Agent

MARKETPLACE

- AI Marketplace

AI Platform

FPT AI AGENTS

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

FPT AI STUDIO

- AI Notebook
- Model Hub
- Model Fine-Tuning
- Data Processing for AI
- Model Pre-Training

FPT AI INFERENCE

- Model-as-a-Service
- Model Serving

Cloud Infras. & Platform

DATABASE PLATFORM

- Vector Database

FPT AI INFRASTRUCTURE

- GPU Container
- Managed GPU Cluster
- GPU Virtual Machine
- Metal Cloud
- 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

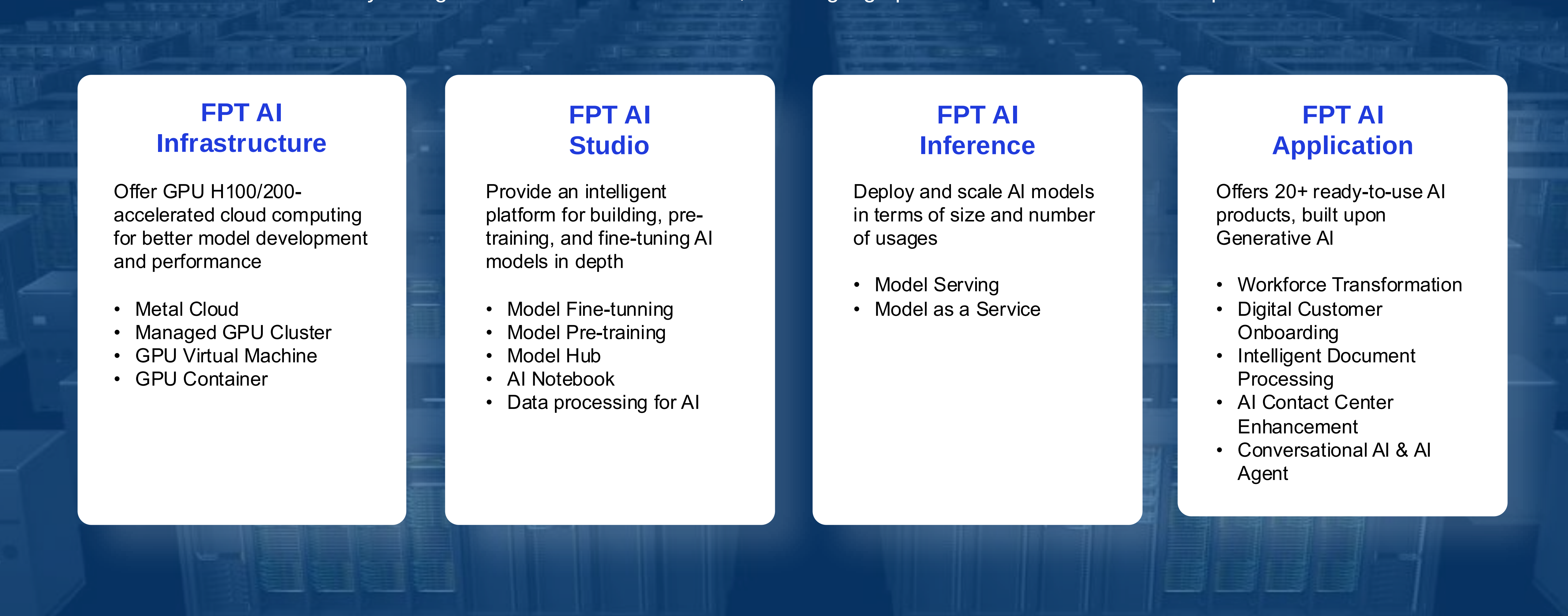
▨ ▨ ▨ Offerings for Customers   □ Managed by FPT

FPT AI Factory

NVIDIA Certified

A Comprehensive Suite for Accelerating AI Innovation and Sovereignty

FPT AI Factory is an all-inclusive stack of High-Performance Infrastructure, Platforms, and AI applications for end-to-end AI product lifecycle, boosting with the latest NVIDIA technologies to drive innovation and efficiency in large-scale AI and ML workloads, ensuring high performance and effortless adoption.



FPT AI Infrastructure

Offer GPU H100/200-accelerated cloud computing for better model development and performance

- Metal Cloud
- Managed GPU Cluster
- GPU Virtual Machine
- GPU Container

FPT AI Studio

Provide an intelligent platform for building, pre-training, and fine-tuning AI models in depth

- Model Fine-tunning
- Model Pre-training
- Model Hub
- AI Notebook
- Data processing for AI

FPT AI Inference

Deploy and scale AI models in terms of size and number of usages

- Model Serving
- Model as a Service

FPT AI Application

Offers 20+ ready-to-use AI products, built upon Generative AI

- Workforce Transformation
- Digital Customer Onboarding
- Intelligent Document Processing
- AI Contact Center Enhancement
- Conversational AI & AI Agent

FPT AI Infrastructure

Powerful infrastructure with the ability to scale from a single node to clusters of hundreds of GPUs.
Well-suited for any ever-evolving AI workloads, from training of large generative AI models to hosting various models for quick inferencing, in a scalable and cost-effective way.

	Metal Cloud	Managed GPU Cluster	GPU Virtual Machine	GPU Container
Definition	• **Dedicated physical servers (Bare Metal) for a single tenant** • Direct access to physical hardware	**Manage a cluster of Bare Metal GPU servers or GPU Virtual Machine by Kubernetes**	• **Virtual machine (VM) with dedicated GPU Card** • Flexible resource scale and deployment	• **Container with dedicated GPU card** • No need to handle the infrastructure
Purposes	• **Model training at scale, Complex inference needs, Custom orchestration** • Strict latency and security requirements	• **Intensive AI workloads need multi-GPU nodes** • Inherit advantages of Metal Cloud	• **Inference, Lightweight AI training, Moderate data processing**	• **Inferencing and lightweight AI training**
SLA	99%	99%	99.9%	99.9%
GPU Operational Model	**Dedicated GPUs** for a Bare Metal server	**Dedicated GPUs** for a Bare Metal server or a Virtual machine	**Dedicated GPUs** for a Virtual machine	• **Dedicated GPUs** for a container • **Support GPU virtualization** (MIG, Timeslicing)
Key Features	• Fully control over hardware for customization • GUI self-service manage resource	• Faster provisioning and more productive management. • Auto update Kubernetes version • Slurm support	• Scalable, on-demand GPU • Auto scaling • Seamless integration with FPT Cloud ecosystem.	• Scalable, on-demand GPU • Auto scaling, scale-to-zero • Intuitive container execution on GUI
Package	From 01 server (8x GPUs)	From 1 nodes (8x GPUs)	From 1x GPU/VM	From 1x GPU/Container
Billing Model	• Reserved • Pay-as-you-go (hourly)	• Reserved • Pay-as-you-go (hourly)	• Reserved • Pay-as-you-go (hourly)	• Reserved • Pay-as-you-go (hourly)

FPT AI Infrastructure - Metal Cloud Specification

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, Storage per Node cannot be expanded.*
*** H100 is available in Vietnam, H200 is in Japan*

Essential for Multi nodes - High Performance Storage

Capacity	As demand, A minimum block 50TB for an increment
Network	Connection to Server: 200Gbps * 2 ports (EDR/HDR InfiniBand) Up to 550 GBps throughput
Feature	Parallel File system, NFS and S3 Gateways

For a cluster of distributed training

Available **NVIDIA InfiniBand** to connect the network between multiple nodes

Recommendation (In addition)

- **Workload Management Software**
  NVIDIA
  Base Command Manager
- **Virtual Machine (General purpose)**
  To setup head nodes manage multiple nodes
- **Managed GPU Cluster with Kubernetes**
- **Object Storage**
  To distribute your training data files
- **Load Balancer**
  For a highly available infrastructure
- **Firewall**
  To secure your AI workload
- **Networking**
  Internet Bandwidth or Direct Connect

FPT AI Studio

Offer a comprehensive set of intelligent tools and services for AI/ML development, evaluation, and deployment, empowering businesses to transform AI models into real-life solutions, even without deep expertise. This approach accelerates AI innovation, simplifies operations, and enhances overall performance.

	Data Processing for AI	AI Notebook	Model Pre-training	Model Fine-tuning	Model Hub
Definition	Transform raw data into a form suitable for building and training AI models.	Offer a fully managed JupyterLab, utilizing the integrated development environment (IDE) for notebooks, code, and data; leveraging FPT GPU Instances.	Managed service for pre-training and continually pre-training AI models with customer data. Train a model from scratch or adapt a pre-trained model to a specific domain with ease.	Managed service for fine-tuning AI models using customer's data and based on wide range of base models. No code required. No long-term GPU reserved capacity.	Centralized place for storing custom models and their respective crucial information; managing model life-cycle; and collaborating through out AI projects.
Key Features	• Data Collection • Data Cleansing • Data Transformation • Data Monitoring and Evaluation	• Serverless • Seamless integrated • No limit running time • Preinstalled AI Frameworks	• Multi-node with multi-GPU • Integrated training pipeline • Support built-in and bring-your-own algorithm • Data quality enhancement	• Multi-GPUs container • Integrated fine-tuning pipeline • Support built-in and bring-your-own algorithm	• Upload your own model or choose from catalog • Fast transfer to model serving • Sharing within the org • Keep track of all versions, evaluation results, and serving history
Benefits	• Enhanced data security • Scalability and flexibility • Cost efficiency • High performance and reliability	• Leveraging cutting-edge GPU models • On-demand, self-service	• Eliminate complex and repetitive engineering tasks for training a model • Shorten GPU occupation time • Better model quality	• Eliminate complex and repetitive engineering tasks for fine-tuning a model • Shorten GPU occupation time • Embrace continuous fine-tuning	• Effortless management and collaboration based on model life-cycle

FPT AI Inference

Provide hundreds of foundation and FPT-developed models, offered as a service to effectively deploy and scale models in terms of size and number of usages, and design unique solutions hosted on NVIDIA-certified AI infrastructure

	Model Serving	Model-as-a-Service
Definition	Provide a unified interface (APIs or other interfaces) that models can be accessed by applications or users to infer outcomes based on input data.	Provide AI models as APIs, allowing companies to integrate pre-trained AI models into applications, without needing to develop, train, or manage AI models.
Key Features	• All-in-one page of model deployment • Model Security: encryption and access controls • Monitoring and Logging: Tracks model performance, latency, errors, and input/output data • Version Control: Allow multiple versions of a model to coexist	• Pre-trained models: LLM models, VLM models • Easy Integration via API • Auto-scale based on demand • Continuous updates to improve performance and adapt to changing needs • Fine-tuning in FPT AI Studio
Benefits	• Efficient Model & Resource Utilization • Ease of Deployment: Flexibility and deployment in minutes • Seamless integration to existing workflows or systems	• Faster Time to Market: ease of integration & deployment • Cost efficiency: Token-based pricing • Scalability with minimal risk

Why FPT AI Factory

Offering a One-stop Shop for AI Development

As an NVIDIA Cloud Partner and Service Delivery Partner (SDP), FPT offers **a comprehensive suite for end-to-end AI development** with powerful GPU infrastructure, platforms, applications, and professional services, with the support of a seasoned AI expert group, holding 9,363 NVIDIA DLI Certificates and 103 NVIDIA Certified Associates.

Leveraging the Latest NVIDIA Technologies with Speed & Compliance

FPT **utilizes the most advanced technologies and frameworks certified by NVIDIA**, including next-gen GPUs and the NVIDIA AI Enterprise software platform.

NVIDIA Preferred Partner

Enabling Significant Cost Savings with First-rate Performance

Optimizing the total cost of ownership (TCO) for excellent results compared to hyperscalers through effective resource and process management.

Optimizing AI Innovation with Highly Flexible Services

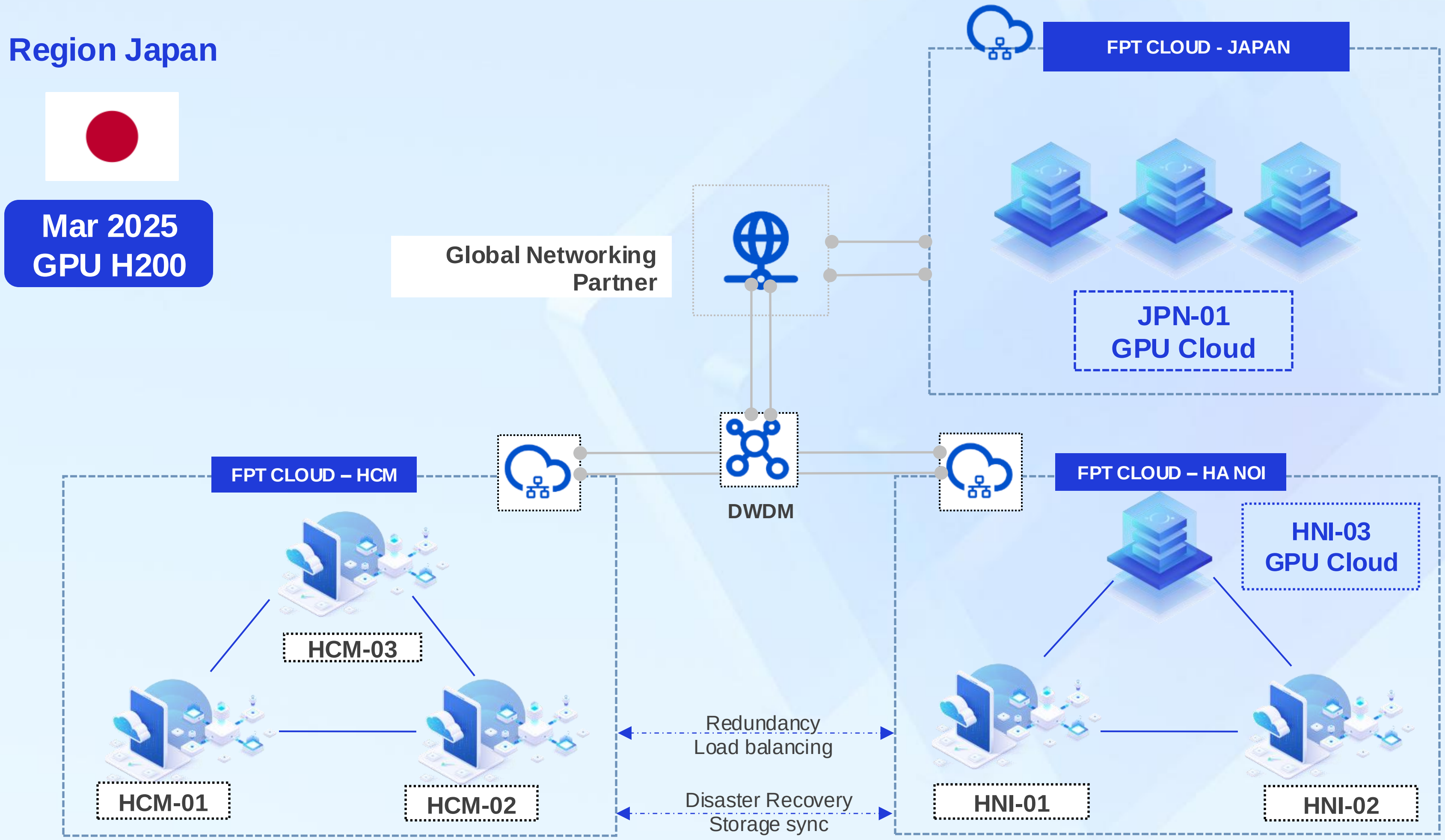
Delivering a diversified range of services with **optimal service models** tailored to customer needs, including the purpose of use, scope, cost, etc.

The Current Status of FPT AI Factory Service Deployment

Offering FPT AI Factory services in two regions – Vietnam and Japan

Region Hanoi	Region Japan
Jan 2025 GPU H100	Mar 2025 GPU H200

The architecture of FPT AI Factory complies with NVIDIA's global standards, and the implementation and operations are certified by NVIDIA.



Be the First to Experience FPT AI Factory

Exclusive Offerings Available for Pre-orders

Enjoy priority availability of FPT AI Infrastructure at special rates

Get a first experience of FPT AI Factory's premium features, boosted by NVIDIA technologies

Earn Cloud Credits to access the full portfolio of AI products and Cloud services

Have tailored-made consultations with seasoned AI & Cloud experts and engineers

Join the Forefront of AI Innovation with Exclusive Benefits

https://aifactory.fptcloud.com

Thank you!