FPT AI Factory

エンドツーエンドのAI開発ライフサイクル
を支える包括的なスイート

高速・安全・スケーラブルな形で先進的なAIソリューションを革新できるよう設計された、高性能インフラとインテリジェントなプラットフォームを活用

TOP500 世界のスーパーコンピュータランキング「TOP500」で第36位にランクイン

日本およびベトナムにおける商用クラウドサービス第1位

FPT AI Factory

FPT AI Infrastructure

スケーラブルで高信頼なモデル開発と性能実行のための、GPUアクセラレーションによるクラウドコンピューティングを提供

- GPU コンテナ
- GPU クラスター
- メタルクラウド
- GPU仮想マシン

FPT AI Studio

AIモデルの構築・事前学習・ファインチューニングを包括的に支援するマネージドサービスを提供

- データハブ
- AI ノートブック
- モデルテスト
- モデルのファインチューニング
- モデルハーブ

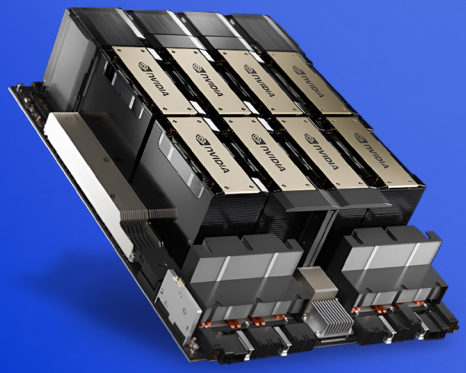
FPT AI Inference

AIモデルの規模および使用回数に応じた展開とスケーリングを実現

- 専用推論
- サーバーレス推論

- 高性能AIスーパーコンピュータ
- NVIDIA認定アーキテクチャ
- 100以上の手頃な価格の製品・サービス
- ハイパースケーラーに対抗できる競争力のある価格設定
- セキュリティとコンプライアンスを標準装備

大規模なAI/機械学習ワークロードに対応するため、NVIDIA H100/H200 Tensor Core GPUを搭載したAIスーパーコンピュータの力を活用



› 2つの地域で提供中

- 日本 (H200 GPU)
- ベトナム (H100 GPU)

FPT AI Infrastructure

最高性能の生成AIおよびHPCプラットフォームを加速する
GPUインフラストラクチャ

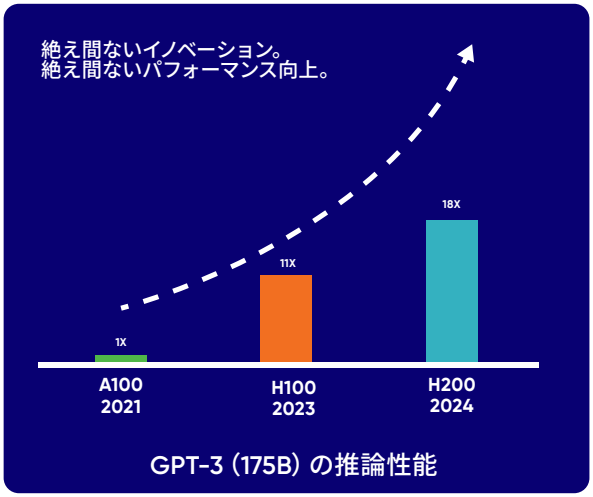
単一ノードから大規模GPUクラスターまで拡張可能なスケーラブルなインフラストラクチャ。モデルの学習から推論までの多様なAIワークロードに最適化されており、迅速な導入を可能にするFPTのAI/MLイメージが事前構成されています。

FPT AI Infrastructureが支援すること

最新のNVIDIA GPUを活用し、モデルの学習・ファインチューニング・推論を効率的に実行。

クラウドの複雑さなしに、3,000台以上のH100およびH200 GPUでシームレスにスケール可能。

導入から運用まで、FPTの専門家が24時間365日対応。



GPT-3 (175B) の推論性能

FPT AI Infrastructureが支援すること

数百のGPUノード間で高性能ストレージを共有し、多様な高品質GPUサービスを提供。

FPT Cloudポータルを活用し、物理サーバーやGPUインスタンスを数分で展開・管理。インフラ全体をシームレスに可視化・制御可能。

専用GPUによって最高のパフォーマンスとセキュリティを実現し、InfiniBand対応の高速ネットワークがGPUクラスターの効率を向上。

FPT AI Studio

アイデアから実現まで、AI開発を力強く支援

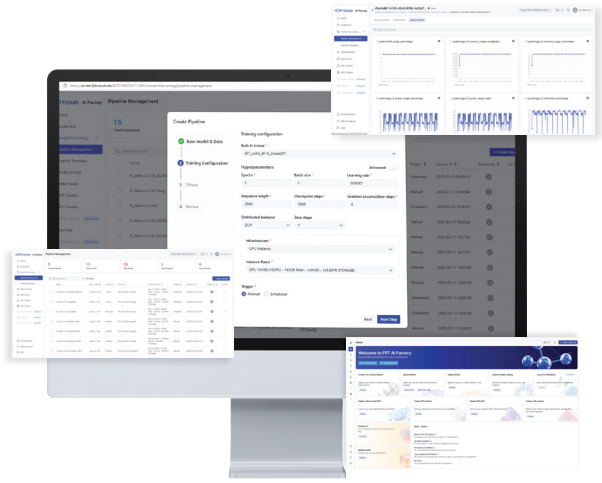
FPT AI Studioは、NVIDIAベースのFPT AI Factory上に構築された、データ準備からモデルのデプロイメントまで、AI開発ライフサイクル全体を支援する統合プラットフォームです。インフラ構築は不要で、柔軟なGPUオプションとセキュリティ機能を標準搭載しており、アイデアからAIの実運用までを、より迅速・安全かつコスト効率よく実現します。

FPT AI Studio支援すること

データ準備、モデル学習、テスト、リポジトリ管理、デプロイメントまでを網羅し、MLOps管理をシンプルにします。

多様な基盤モデル、大規模データセットへの柔軟なGPUオプション、およびマルチモーダル機能に対応した組み込み型サポートを提供します。

さまざまなビジネスニーズに合わせた柔軟な請求・支払いオプションを備えた、従量課金制の料金モデルを提供します。



主な特長

インフラのセットアップは不要で、必要なのは学習・評価・テスト用のデータのみ。

多様なニーズに対応する柔軟なGPUインスタンスオプション。

安全性とコスト効率に優れたインフラ。

FPT AI Inference

次世代AI推論プラtットフォームを公開

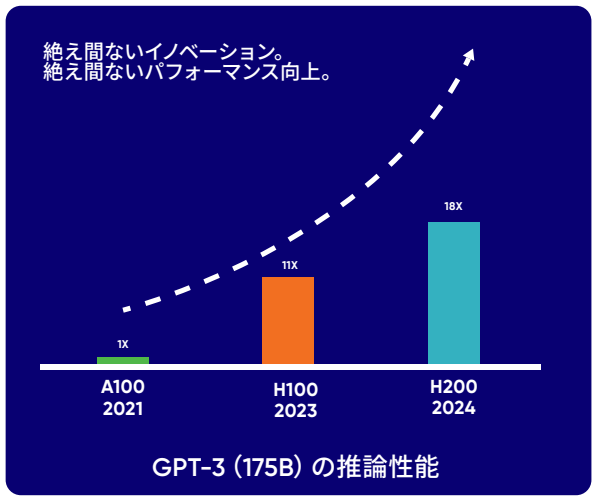
APIエンドポイントとサーバーレスGPUを活用したサーバーレス推論および専用推論により、リアルタイムで高性能なアプリケーション向けに、安全性・スケーラビリティ・効率性に優れたAI推論を、柔軟な従量課金モデルで提供します。

FPT AI Inferenceが支援すること

AIエンドポイントとサーバーレスGPUを活用しAIモデルを手間なくホスティング可能。

AI推論の最適なパフォーマンスを実現するため、自動スケーリングアルゴリズムを実装。

強力なAIインフラストラクチャに、データローカライゼーションおよびコンプライアンス管理機能を組み込み。



主な特長

シームレスな統合とデプロイにより、タイムトゥマーケットを短縮。

柔軟な従量課金モデルによる高いコスト効率。

あらゆる需要に対応するダイナミックなスケーラビリティ。

事例紹介

スピードと効率性を兼ね備えたVisual AIプラットフォームの拡張

ユースケース

- Generative AI / LLMs
- Visual AI / Agentic AI

商品

- H100 Metal Cloud

アプリケーション

- 大規模なファインチューニングや強化学習のために、NVIDIA H100 GPUを搭載したMetal Cloudを活用。
- スムーズな統合、安定した運用、高可用性を実現するための継続的なサポートを提供

利点

視覚タスクの汎化性能を強化
多様な視覚タスクにおいて、より高い精度と高速な応答性を実現

顧客向け機能の展開を3倍高速化
新機能のリリースを加速し、エンドユーザーへの価値提供までの時間を短縮

コスト効率を備えたかつてないスピード
最適化されたコストで、タイムトゥマーケットの短縮とエンタープライズレベルのパフォーマンスを実現

LandingAI

> Thanks to the services provided by FPT AI Factory and the dedicated support from the FPT team, we were able to rapidly develop and integrate large-scale AI models into our solutions, significantly improving operational efficiency while optimizing both time and cost.

LandingAI CEO ダン・マロニー氏

事例紹介

カスタマーサポートと人材育成の強化

ユースケース

- Generative AI / LLMs
- Conversational AI / NLP

商品

- NVIDIA H100 Tensor Core GPUs
- NVIDIA A100 Tensor Core GPUs

結果

- データ処理を3倍、音声合成処理を4倍高速化し、モデル学習のスピードを大幅に向上。
- 18言語に対応し、音声合成モデルの出力精度と表現力を100倍に向上。
- 前世代と比べて、大規模言語モデルの学習速度を30倍に高速化。

顧客成果

01. カスタマーサービスの強化

カスタマーサービス、テレセールス、セルフサービスなど、さまざまな業務を自動化するオムニチャネル型バーチャルアシスタントにより、優れたカスタマーエクスペリエンスを実現。

1,200万件 の通話／月	98% の顧客問い合わせを解決	顧客満足度 4/5

02. 人材の活用強化

従業員のスキルとパフォーマンスを向上させるバーチャルアシスタント。

1万人 の月間アクティブユーザー	従業員の能力を 15%向上	従業員研修の工数を 80%削減

お問い合わせ

Hanoi: FPT タワー, 10 Pham Van Bach 通り, Cau Giay 区

Ho Chi Minh: PJICO タワー, 186 Dien Bien Phu 通り, Xuan Hoa 区

東京：東京都港区三田3-5-19 住友不動産東京三田ガーデンタワー33階

aifactory.fptcloud.com 1900 638 399 ai.factory.contact@fpt.com