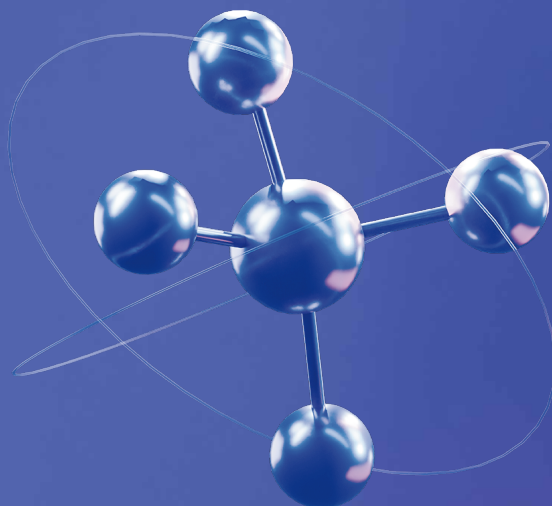


SERVICE LEAFLET

大規模国産 GPUクラウド



FPT GPU Cloud - AI 開発を、最速で本番へ。

国内最速級※の NVIDIA H200を計1,000基超、国内データセンターに設置。
InfiniBand による高速・低遅延通信、DDN Lustre による高速並列ファイルシステムをはじめとする NVIDIA リファレンスアーキテクチャを採用し、大規模なマルチノード環境でも高いパフォーマンスを実現します。

NVIDIA H200 を **1,000基超**

国内H200 商用クラウドで最速級

※当社調べにおいて、
国内H200商用クラウドで最速



スパコン世界ランキング (TOP500)

50位

2026年6月発表データ

1,000基超NVIDIA H200を
国内DCに設置**990円/時間~**1GPU単位で
必要な分だけ利用**96%超**32ノード時の
スケーリング効率※**秒単位~**GPU単位の
オンデマンドから
月単位の専有利用まで

1,000基超の NVIDIA H200 を、国内に

計1,000基超の H200 を国内データセンターに設置。InfiniBand による高速・低遅延通信や、DDN Lustre による高速並列ファイルシステムをはじめとする NVIDIA リファレンスアーキテクチャを採用し、大規模なマルチノード環境においても高いパフォーマンスを実現します。**スパコン世界ランキング (TOP500)** では **50位** (2026年6月発表データ) にランクインしています。

こんなことでお困りの方へ

大規模に、止まらず使いたい方へ

❗ 課題

複数ノードを束ねると性能が頭打ち。
クラスタ構築の手間も大きい。

✓ 解決

NVIDIA リファレンスアーキテクチャを採用し、マルチノード環境でも高い性能を発揮。
クラスタ構築から運用まで代行し、大規模AI学習をスムーズに実行できます。

まず1枚から、手軽に試したい方へ

❗ 課題

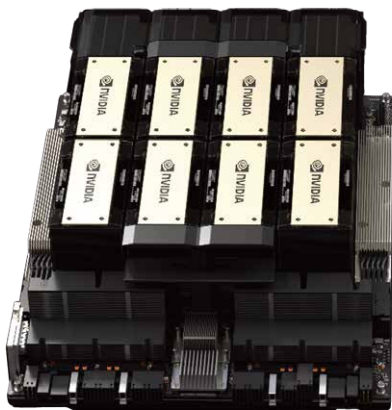
1枚だけ使いたいのにサーバー単位の契約で割高。借りたい時に借りられない。

✓ 解決

1GPU・990円/時間から利用可能。GPU単位での課金により必要な分だけ使え、コストを抑えてスモールスタートできます。

圧倒的な計算リソースと、柔軟な提供形態

LLM・VLMの学習・推論から、シミュレーション、CAE、創薬などのHPC用途まで幅広く対応。



スペック01 GPU

NVIDIA H200 /計1,000基超

スペック02 ネットワーク

GPU間は NVLink/NVSwitch、ノード間は InfiniBand で高速・低遅延

スペック03 スケーリング

8ノードで100%、32ノードでも96%超のスケーリング効率を実現

※当社環境におけるベンチマーク結果です。性能は利用環境やワークロードにより異なります。

スペック04 提供形態

GPU単位・秒単位でのオンデマンドから、ノード単位・月単位での専有利用まで

FPT Token Factory をはじめとする推論基盤を提供

GPUインフラに加え、推論APIを提供するAIプラットフォームまで、ワンストップでご利用いただけます。

FPT Token Factory

NVIDIA AI Enterprise

NVIDIA NGC

NVIDIA NeMo

NVIDIA NIM

用途に合わせて選べる、4つのGPUサービス

物理サーバーを丸ごと専有する「メタルクラウド」から、GPU単位・オンデマンドで使える「GPUコンテナ」まで。開発フェーズと運用要件に合わせて最適な形態をお選びいただけます。

METAL CLOUD メタルクラウド 単一テナント専用の物理サーバー（ベアメタル）。ハードウェアを完全制御し、GPUの最大パフォーマンスを引き出します。 こんな用途に 大規模なAI学習/複雑な推論/低レイテンシー・厳格なセキュリティ要件	GPU VIRTUAL MACHINE GPU仮想マシン 専用GPUカードを搭載したVM。迅速な起動・停止と柔軟なリソーススケーリングが可能です。 こんな用途に モデル開発/小～中規模の学習・推論/柔軟・分離・高速起動の環境	GPU CLUSTER GPUクラスター KubernetesでベアメタルGPUサーバーやGPU VMを統合管理。コントロールプレーンはFPTが運用します。 こんな用途に 分散学習/スケーラブルな推論/マルチGPU・マルチノードのオーケストレーション	GPU CONTAINER GPUコンテナ 専用GPU付きコンテナをテンプレートから即時起動。インフラ管理は一切不要です。 こんな用途に ノートブック・開発ツール/推論・軽量の学習/スモールスタート
--	---	--	--

サービス比較表

	メタルクラウド	GPU仮想マシン	GPUクラスター	GPUコンテナ
GPU提供形態	ベアメタルサーバー専用GPU	仮想マシン専用GPU	ベアメタル/VM向け専用GPU	コンテナ専用GPU。GPU仮想化（MIG・タイムスライシング）対応
SLA	99%	99.5%（NVMe VM） 99.95%（ブロックストレージVM）	コントロールプレーン稼働率 99.95%	99.9%
主な特徴	ハードウェア完全制御／GUIセルフサービスでサーバー管理・監視	迅速な起動・停止／GUIでデプロイ・設定・ネットワーク管理	GPU Operator・vGPUスケジューラーによる高速プロビジョニング／複数K8sバージョン対応	すぐ使えるコンテナカタログ（テンプレート付き）／複数ポート・SSH対応
最小構成	1サーバー（GPU 8基）～	GPU 1基／VM～	ベアメタル：GPU 8基～／VM：GPU 1基～	GPU 1基／コンテナ～
課金モデル	専有利用（ノード単位・月単位～）	リザーブド／ 従量課金（時間単位）	リザーブド／ 従量課金 （ワーカーノードはGPU VMに準拠）	従量課金

※ 提供形態は、GPU単位・秒単位でのオンデマンドから、ノード単位・月単位での専有利用まで柔軟に選択いただけます。

マネージド Kubernetes で運用負荷を最小化

GPUクラスターではコントロールプレーンの管理やKubernetesのアップデートをFPTが代行。クラスタ構築から運用までお任せいただくことで、お客様はAI/MLアプリケーションの開発・デプロイ・運用に集中できます。

主要スペックと推奨構成

メタルクラウド仕様（1 ノードあたり）

	Metal Cloud GPU H200（日本）	Metal Cloud GPU H100（ベトナム）
GPU	NVIDIA H200 SXM5 × 8（HBM3eメモリ計 1.1TB）	NVIDIA H100 SXM5 × 8（GPUメモリ計 640GB）
CPU	Intel Xeon Platinum 8558 × 2	Intel Xeon Platinum 8462Y+ × 2
メモリ	2TB（DDR5 4800MHz DIMM）	
ストレージ	30TB（3.84TB NVMe SSD × 8）	
ネットワーク	InfiniBand 400Gbps × 8ポート + BlueField-3 DPU 200Gbps × 2ポート	

※ノードあたりのGPU・CPU・RAM・ストレージは拡張不可。単一ノード/複数ノード構成でご提供。GPU間はNVLink/NVSwitch（900GB/s）で接続。

GPU 仮想マシン & GPU コンテナ

H100

80GBGPUメモリ/
GPU

H200

141GBGPUメモリ/
GPU

B300

288GBGPUメモリ/
GPU

Coming Soon!

いずれも1・2・4・8 GPUのパッケージ構成（vCPU最大256コア/RAM最大1,536GB）。NVMeローカルストレージ（GPU 1基あたり3TB）と、オンデマンド拡張可能なブロックストレージを搭載。**1GPU・990円/時間**から、必要な分だけご利用いただけます。

高性能ファイルストレージ

DDN Lustre による並列ファイルシステム。数百のGPUノード間で学習データを高速共有できます。

容量	オンデマンド・50GB から
性能	ノードあたり最大 書込 2GB/s/ 読取 4GB/s
ネットワーク	200Gbps × 2 ポート (EDR/HDR InfiniBand)
機能	並列ファイルシステム、 NFS & S3 ゲートウェイ
対応サービス	メタルクラウド / GPU クラスター / GPU VM

AI 基盤を完成させる推奨オプション

オブジェクトストレージ（学習データ配布）

ロードバランサー（高可用性）

ファイアウォール（ワークロード保護）

汎用VM（管理・ログインノード）

高速インターネット/専用線接続

提供リージョン：H200 は日本（国内データセンター）、H100・B300 はベトナム（ハノイ/ホーチミン）にてご提供します。

まずは**1GPU**から。お気軽にご相談ください。

PoC・お見積り・技術相談・検証環境のご提供など、担当営業までお問い合わせください。

< Contact Us > <https://factory.fpt.ai/ja/contact-us-ja>

