”プライベートGPUクラウド”によるLLM＋RAGの導入例ご紹介

カゴヤ・ジャパン株式会社

鶴岡 謙吾

INDEX

- 会社概要
- 「プライベートGPUクラウド」とは
- LLM＋RAG の導入例
- まとめ

COMPANY PROFILE

KAGOYA JAPAN

社 名	カゴヤ・ジャパン株式会社
設 立	1983年9月22日
資本金	1億円
代表者	代表取締役会長CEO 北川 貞大 代表取締役社長COO 岡村 武
事業内容	・データセンター運営 ・クラウドサービス事業

- 20年を超えるレンタルサーバーの事業実績 (31,000社/60,000人以上の導入実績)
- 自社データセンター保有 (2006年04月 A棟竣工、2014年04月 B棟竣工)
- データセンター内稼働サーバー数　8,000台以上
- 各地方の電力系通信事業者への自社サービスのOEM提供を始め、多数のWeb制作事業者やアプリケーションベンダーへのOEM提供の実績あり(7年以上)
- 24H365Dの自社エンジニアによる運用監視＋一部365日無休の電話サポートセンターあり　※一部プランは平日（10時00分～17時00分）のみ電話対応

自社所有のデータセンターを関西学研都市で運用

■ 稼働サーバー台数
8,000台以上

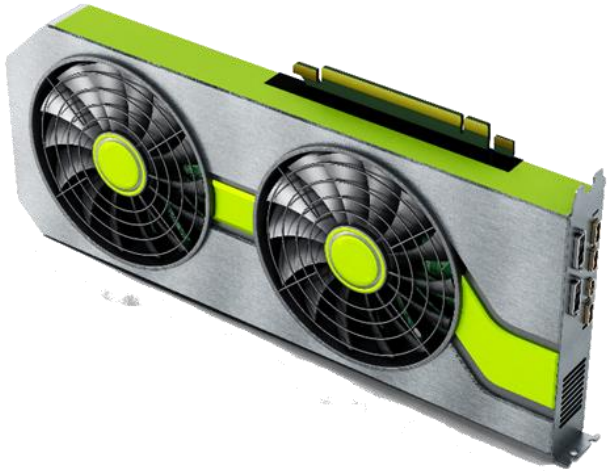
■ 導入実績
31,000社以上

INDEX

- 会社概要
- 「プライベートGPUクラウド」とは
- LLM＋RAG の導入例
- まとめ

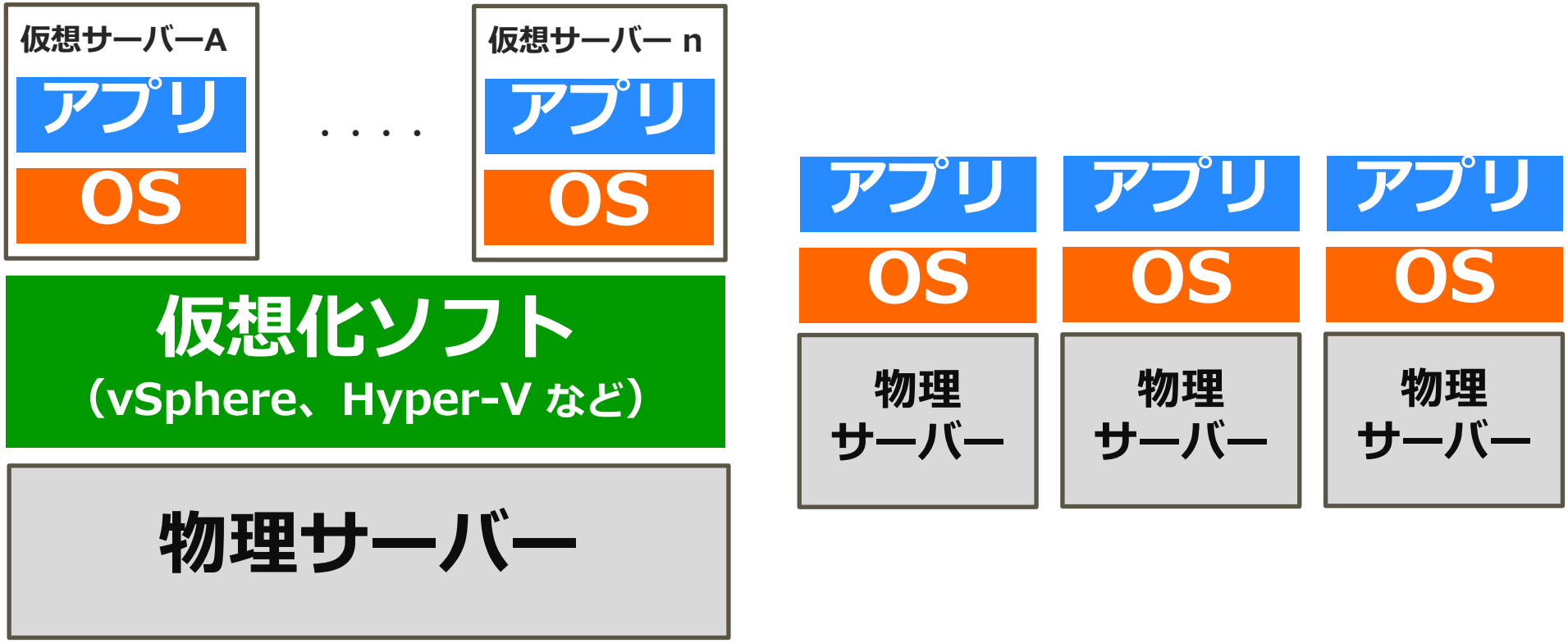
「プライベートGPUクラウド」とは

NVIDIA製GPUを搭載したサーバーを物理環境で専有利用できるクラウドサービスです



https://www.kagoya.jp/cloudplatform/gpu/

仮想環境と物理環境の違い



仮想環境と物理環境の比較

比較項目	仮想環境	物理環境
リソースの専有性	複数ユーザーで共有	専有利用が可能
パフォーマンス	オーバーヘッドが発生しやすい	フル性能を発揮
セキュリティ	適切な設定強化が不可欠	隔離性が高くセキュリティが強固
障害発生時の影響	他の利用者の影響を受ける	他の利用者の影響を受けない
柔軟性	リソースのスケールが容易	リソース追加時に停止が必要
コスト	コスト効率が良い	高性能だがコストも高め

主な特長

1. **GPUを占有**：国内の自社データセンターに設置されたGPUサーバーを専有し、**プライベートクラウド環境で利用**できます
2. **月額定額料金**：初期費用や月額費用は要望に応じて設定され、データ通信量に応じた追加料金は発生しません
3. **豊富なオプション**：ファイアウォールやルーター、NASストレージなど、多彩なオプションサービスを利用できます
4. **柔軟な構成**：データセンター内のスパコンや他のHPCリソースと高速接続可能です

INDEX

- 会社概要
- 「プライベートGPUクラウド」とは
- **LLM＋RAG の導入例**
- まとめ

LLM＋RAG の導入例

LLM（Large Language Model、大規模言語モデル）は、
膨大なテキストデータを学習し、
高度な自然言語処理を行うAIモデル。

代表例

OpenAI：GPT　Google：Gemini
DeepSeek（中国）

LLMの概念図

Input

プロンプト

あなたは開業医です。

小学生にも理解できる「インフルエンザ予防法」を簡潔に教えて

#設定
#役割り
#依頼内容

チャットアプリ

生成AI

設定 / 役割り / 依頼内容に沿った回答を生成

インフルエンザを予防するためには、次のことを守るといいよ：
1. 手洗い・うがいをしっかりする：外から帰ったら、必ず手を洗ってうがいをしよう。
2. マスクをつける：
・
・
・
これらを守ると、インフルエンザにかかる確率を減らせるよ！

LLM＋RAG の導入例

RAG（Retrieval-Augmented Generation）は、LLMに外部データを動的に取り込んで応答の精度を向上させる手法です。

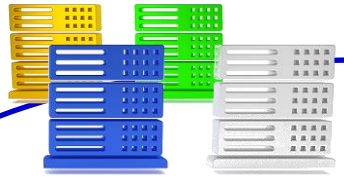
検索で情報を抽出し、それをもとに生成AIが回答を作る技術です。これにより、**AIは未学習の社内情報などからも回答を生成できます。**

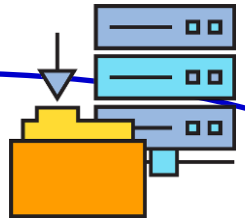
GPUクラウドを活用することで、RAGの検索・生成プロセスを高速化可能です

LLM＋RAGの概念図

社内システム

ファイルサーバー





社内文章

利用者

①質問

②検索

③検索結果

フロント
エンド

クラウド

box
Googleドライブ
Microsoft365
など

A社の企業概要と
過去３年間の
売上実績を教えて

生成AI

④プロンプトに
文章を挿入

⑤回答

閲覧権限等を加味

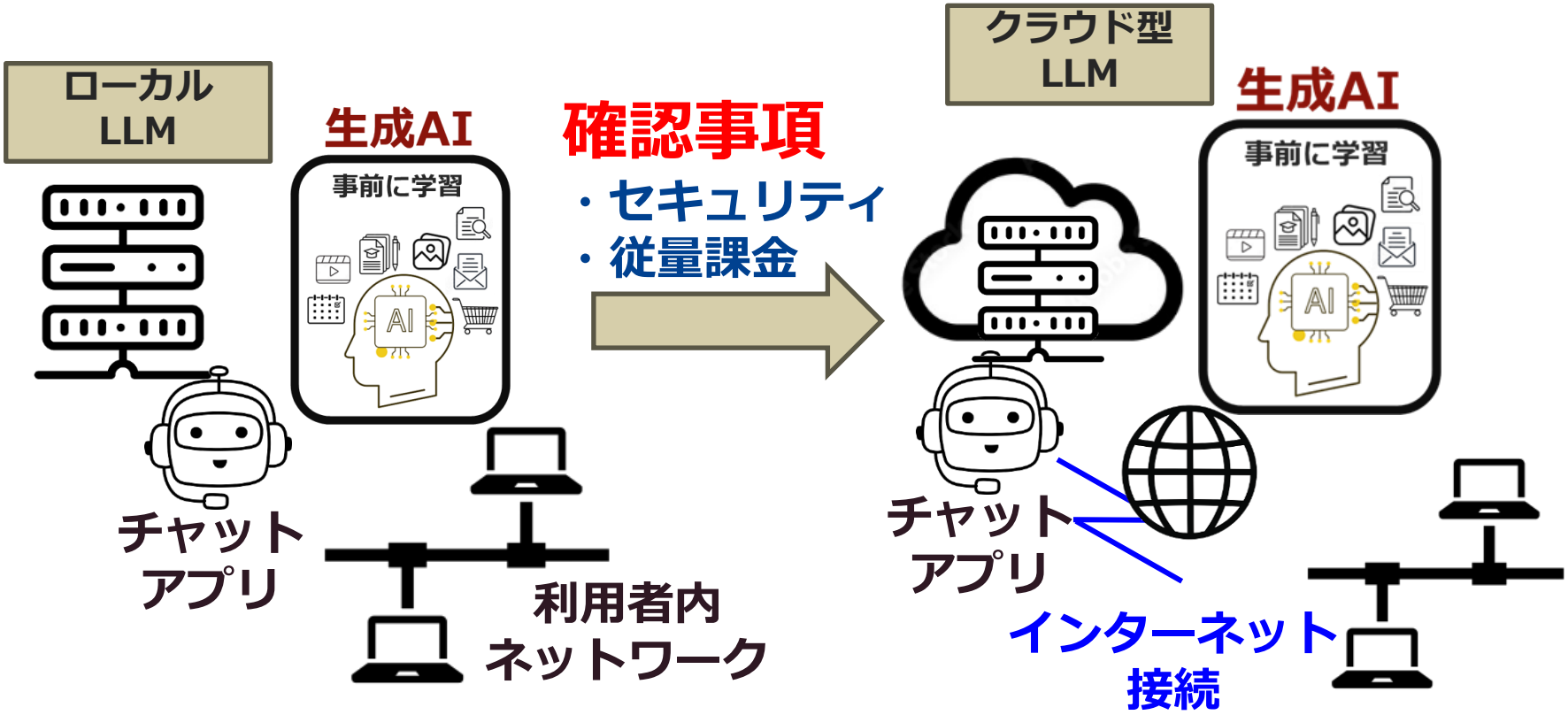
RAGの仕組み

1. 情報検索（Retrieval）：外部データベースや文書から関連情報を検索
2. 文章生成（Generation）：検索した情報をもとに、LLMが適切な回答を生成

RAG利用メリット

- **最新情報を活用**：LLM単体では学習時点までの知識しか持たない
- **企業データ**や専門知識**に基づいた回答**が可能
- 学習コストを抑えつつ、正確性を向上

ローカルLLMをクラウドに移植



自社内でLLMを運用する際の注意事項

1.データのプライバシー管理とセキュリティ

- 機密情報や個人情報が含まれる場合、**データ漏洩防止対策が必須**。暗号化やアクセス管理を徹底すること。
- 機密情報の流出防止：LLMに入力するデータの管理を徹底（**社内データのフィルタリング**が重要）
- アクセス制御：特定ユーザーや部署ごとの**利用制限を設定**

2.計算リソースの確保と運用環境整備

- LLMは計算資源を大量に消費するため、十分な**GPUリソース確保**や**インフラ整備**の必要がある（大容量電源、発熱対策）

「プライベートGPUクラウド」の利用メリット

餅は餅屋に任せて本来のタスクに集中！

オンプレミス

- LLM + RAG
- SDK・ライブラリ
- OS
- サーバー × GPU
- ストレージ
- ネットワーク

自社で運用管理

カゴヤが運用管理

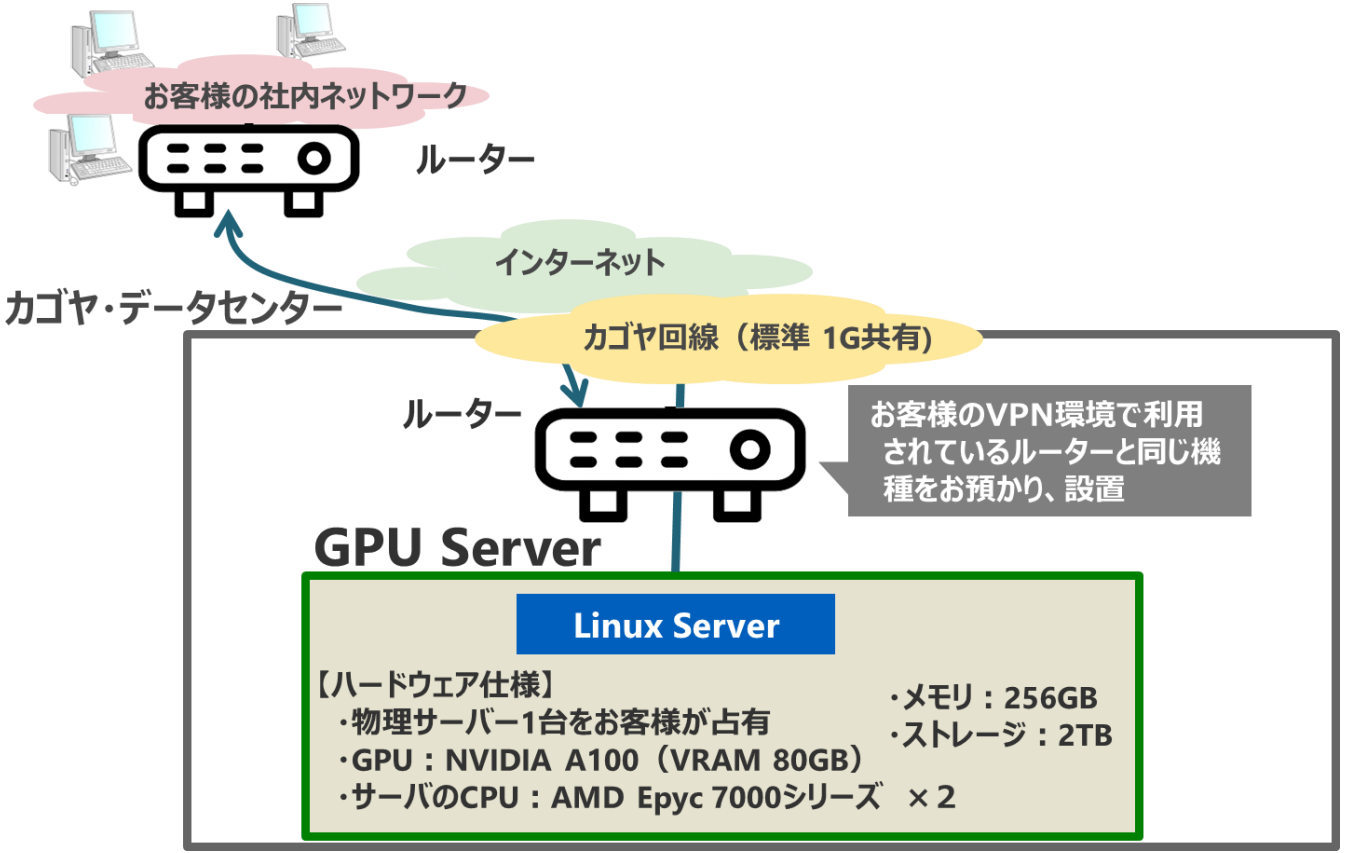
Cloud

HPCサービス
「プライベートGPUクラウド」

- LLM + RAG
- SDK・ライブラリ
- OS
- サーバー × GPU
- ストレージ
- ネットワーク

クラウドサービスを利用

クラウド接続イメージ（例：VPNとの接続）



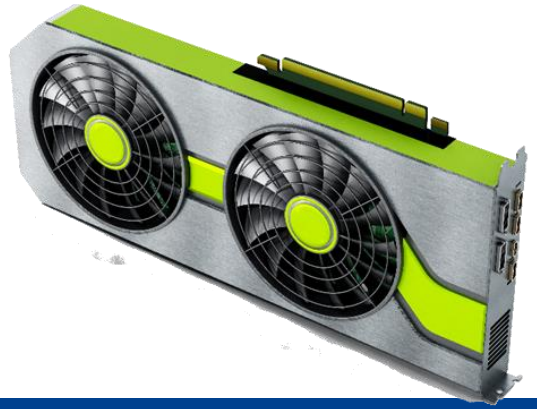
月額定額料金で利用できるGPUクラウドの一例

MFG.	TYPE
NVIDIA	H100 94GB ⇦
NVIDIA	A100 80GB ⇦
NVIDIA	RTX A5000 24GB
NVIDIA	RTX A6000 48GB

大容量VRAM：

より大きなモデルを処理可能にし、学習効率や推論性能を向上

これにより、LLMのパフォーマンスやスケールが大幅に向上



月額定額料金で利用できるGPUクラウドの一例

税込価格

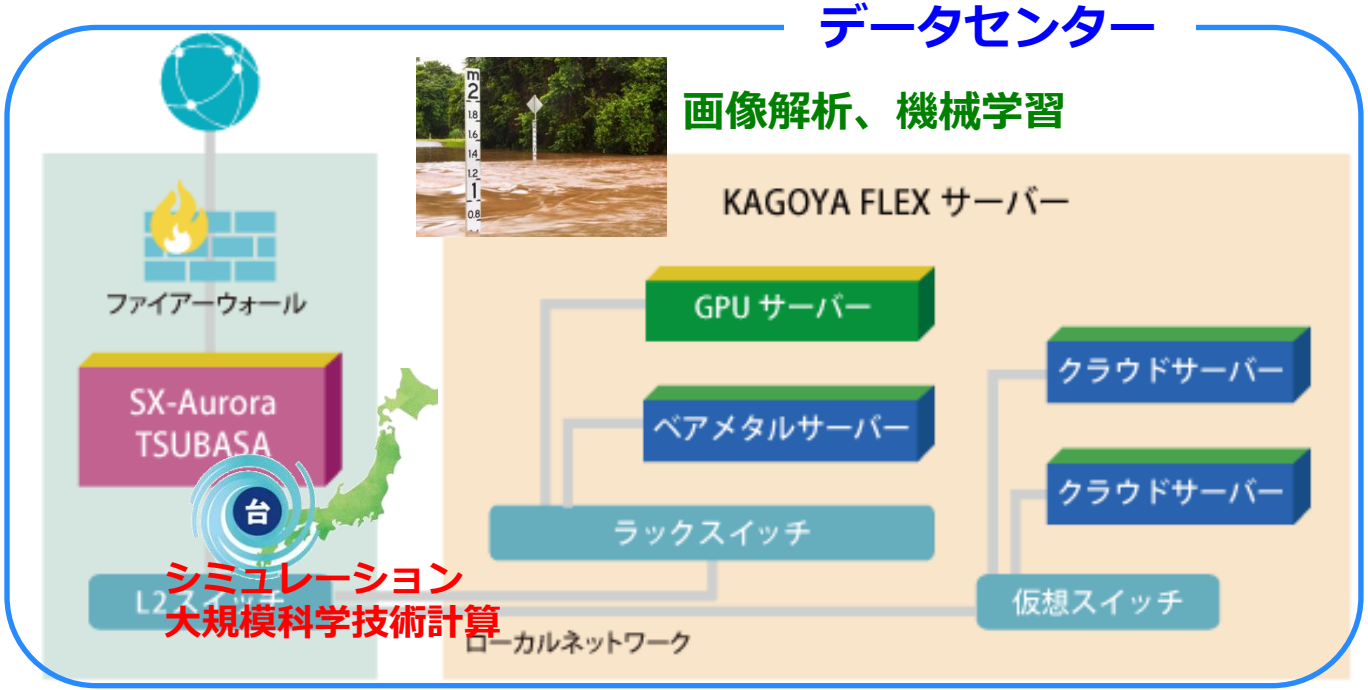
NVIDIA H100 94GB	NVIDIA A100 80GB	NVIDIA RTX A5000 24GB	NVIDIA RTX A6000 48GB
AI・機械学習に最適化された圧倒的な性能	大規模データ処理とAIワークロードに対応	CADや3Dレンダリングを強力にサポート	VR、AR、CGI制作リアルタイムシミュレーション
初期費用 902,000円	初期費用 902,000円	初期費用 297,000円	初期費用 297,000円
月額料金 858,000円	月額料金 297,000円	月額料金 97,900円	月額料金 159,500円

※ 最低利用期間：2カ月

※ ご利用期間やオプションの組合せにより価格は変動します。

科学技術計算と機械学習をハイブリッドで利用

参考：スパコンとGPUサーバーを接続した災害対策インフラ構築



INDEX

- 会社概要
- 「プライベートGPUクラウド」とは
- LLM＋RAG の導入例
- まとめ

まとめ

1. **GPUを占有**：GPUサーバーを専有し、プライベートクラウド環境で利用できます
2. **月額定額料金**：データ通信量に応じた追加料金は発生しません
3. **柔軟な構成**：データセンター内のスパコンや他のHPCリソースと高速接続可能です

秘匿性が高いデータを扱う企業様や研究開発業務に最適

HPCサービス
についてのお問い合わせ窓口

contact@kagoya.jp

カゴヤ・ジャパン　フィールドセールスチーム

お気軽にご連絡ください