SAKURA internet

GPUは「所有」から「活用」へ クラウドで賢く使いこなそう

AI・ディープラーニングに最適な

国産GPUクラウドサービス 高火力

サービス紹介資料

AI 開発において求められる計算量が激増

近年、AI 技術の進歩に伴い、企業や研究機関における大規模機械学習の需要が著しく増加しています。
今後の AI 開発において、GPU の安定供給が、社会と産業の成長を左右する極めて重要な要素となってきています。

1 AI 技術の進歩

- 生成 AI や LLM (大規模言語モデル)、画像生成などの AI 技術は飛躍的に進化
- 企業にとっても AI の活用は競争力強化の鍵となっており、独自の AI モデル開発に取り組む動きが加速

2 計算量の爆発的増加

- AI の活用に伴い、計算量は指数関数的に増加しており、その学習には膨大な GPU が不可欠
- 大規模な AI 開発では、従来のサーバーでは処理が追いつかず、高性能 GPU が前提となる時代に突入している

3 GPU の安定供給が重要

- こうした背景から、GPU の安定供給は AI 開発の成否を左右する最重要課題となっている
- 企業が AI を持続的に活用するためには、中長期的な調達計画などを経営戦略の一環として捉えることが求められている

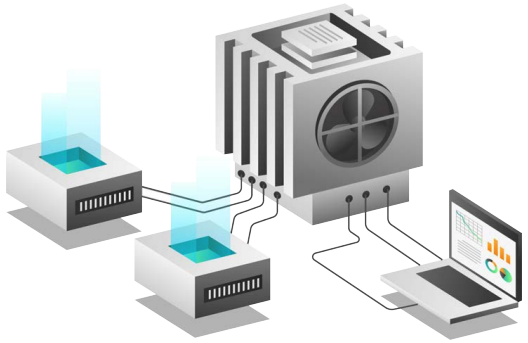
AI 開発で直面する障壁

さまざまな企業・研究機関において、生成 AI や LLM (大規模言語モデル) は実用化されつつあります。
しかし、その一方で AI 開発を阻む壁も顕在化しています。

GPU 調達コスト

- LLM や生成 AI の普及により、多くの計算処理が必要になっている
- モデルサイズや学習データも大きくなり、今ある GPU では処理が追いつかない・学習に時間がかかる

ファシリティ



- オンプレミスの GPU の導入は、設置場所の確保や耐震対策、電力・冷却設備の準備など、時間もコストもかかる

人材 / 運用

- 効率的な利用設計、故障時対応のフロー整備などができる人材が不足している

GPU 調達コスト

GPU 調達コストと供給制約

生成 AI・大規模言語モデル (LLM) の台頭により、AI 開発の複雑度が急速に高まっており、既存の GPU では計算ニーズに対応しきれなくなっています。さらに、世界的な AI 需要の増加によって GPU の品薄状態が続き、価格の高騰や納期の不確定性が企業の AI 開発に深刻な影響を及ぼしています。加えて、GPU のアーキテクチャは進化のスピードが非常に速く、わずか数年で旧世代が陳腐化してしまうことも珍しくありません。

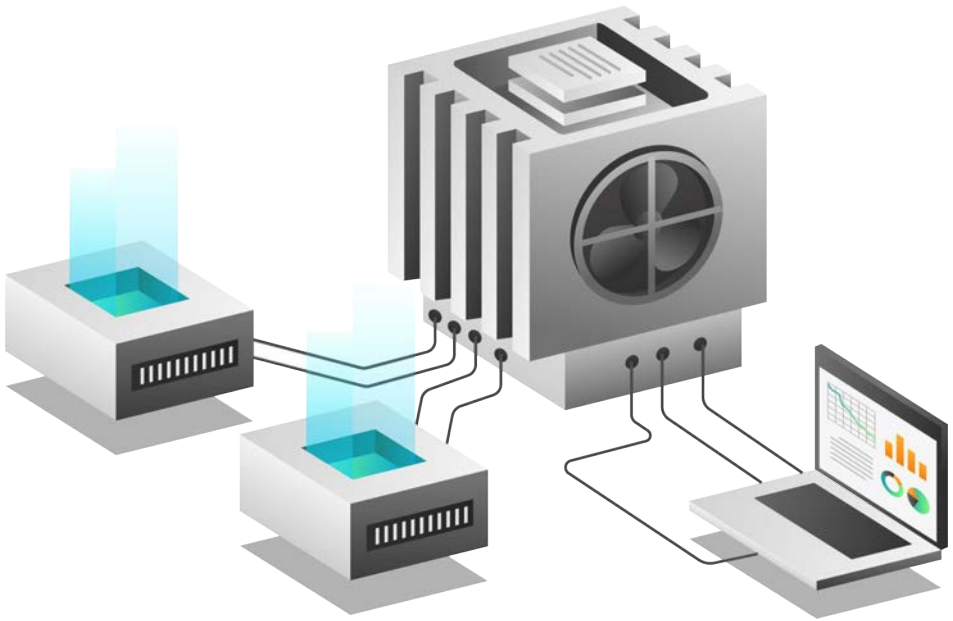
そのため、自社で GPU を購入・保有することは、将来的な技術の進化に対応しづらくなり、資産としての柔軟性を欠くリスクも抱えています。

ファシリティ

導入にかかる設備投資・納期の遅延

AI 開発環境をオンプレミスで構築する場合、単に設備を揃えるだけでは不十分です。高性能 GPU を活かすには、電力・冷却に優れた外部データセンターの確保が不可欠ですが、実現は容易ではありません。

ファシリティには高度な設計と、故障時の対応力が求められます。
さらに、自社構築では監視や緊急対応、専門人材の確保など、
初期投資以外にも多くの見えないコストが発生します。これらの
負担は、専門サービスを活用することで大きく軽減でき、本来の
AI 開発に集中できます。



人材 / 運用

MLOps 人材の不足

GPU サーバーを構築・管理し、AI モデルの学習を安定して稼働させるには、MLOps (Machine Learning Operations) の体制強化が必要です。しかし、専門性の高い MLOps エンジニアやインフラエンジニアは依然として需要が高く、十分な人材確保が難しい状況です。こうした課題に対しては、高性能 GPU 環境を提供するクラウドサービスへのアウトソースが有効です。これにより、自社でインフラを保有・管理する必要がなくなり、技術的な運用負荷も大幅に軽減できます。

24 時間体制などの運用負荷

24 時間体制の監視や、障害発生時の即時対応は大きな負担となり、本来やるべき開発やデータ活用に十分な工数を割けず、事業の成長スピードを速められないといったケースも見受けられます。

高火力シリーズで課題を一挙解決

高火力シリーズなら「 GPU 調達コスト 」「 ファシリティ 」「 人材 / 運用 」の 3 つの課題に対応可能です。
また、国内運用・国内法令に準拠しているため、海外クラウドと比較して、法的リスクの所在が明確です。

1 最新の GPU を、用途に応じて柔軟に選択可能

- 技術の進化が早い GPU も、GPU クラウドなら最新モデルを利用でき、長期的な設備投資のリスクを避けられます。
- 利用用途に応じて最適な GPU サービスを柔軟に選べるため、開発内容やフェーズに合わせた効率的なリソース活用が可能です。

2 初期投資ゼロ・導入もスピーディー

- オンプレミスのような高額な初期投資は不要。Web 上からの簡単な操作で、GPU 環境をすぐに利用できます。
- クラウドならではの利便性と高性能 GPU のパフォーマンスを両立し、直感的な UI で導入から運用までスムーズに進められます。

3 運用・監視の負担ゼロ

- GPU サーバの構築・管理や、監視・障害対応はすべてサービス提供側にお任せ。
- 専門人材がいなくても、AI 開発や研究に集中できます。

特長1：最先端 GPU を提供

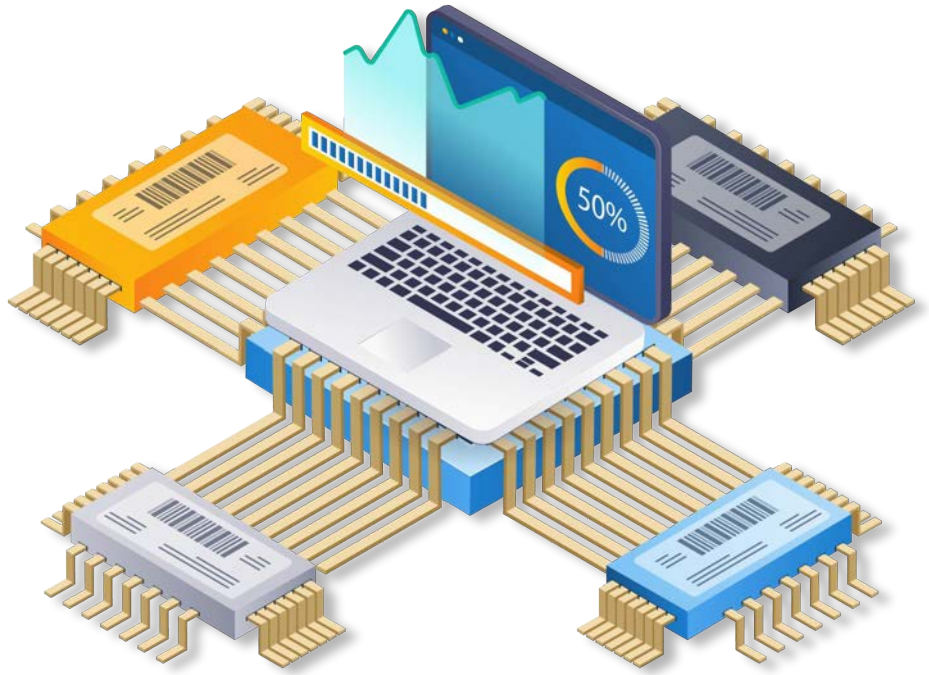
高火力シリーズでは、最新のモデルや学習手法に対応可能な最先端の GPU を提供します。
従来の GPU では難しかった大規模な学習や推論処理にも対応できます。

NVIDIA H100 / H200/ B200※ のパフォーマンス

※ H200、B200はPHYのみ

従来の CPU では対応が難しかった膨大なデータ処理や複雑な計算も、NVIDIA H100 / H200/ B200 なら、高速かつ効率的に処理できるようになります。

データ分析の多くの場合、AIアプリケーションの開発時間の大半を占めます。NVIDIA H100 / H200/ B200 は膨大なデータセットに対処するハイパフォーマンスとスケールでデータ分析を高速化できます。

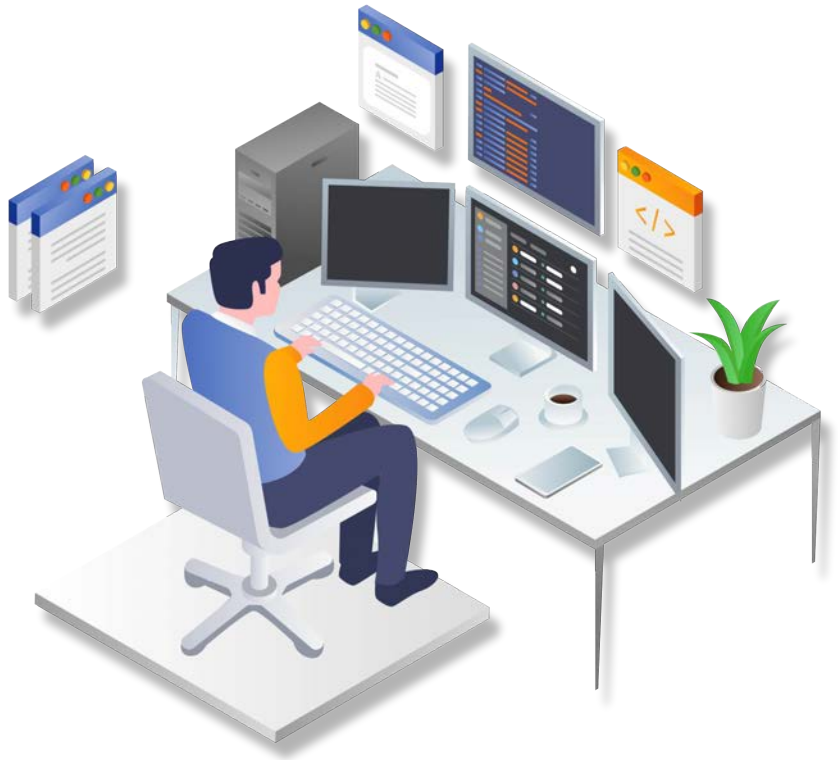


特長2 : 使いたい時にすぐ使える

高火力シリーズなら、必要なときにすぐ使える GPU 環境をご用意しています。
オンプレミスや他社クラウドでの GPU が枯渇していても、タイミングを逃さずに作業を進められます。

PHY：リモートメディアマウント
VRT：プライベートアーカイブ／スタートアップスクリプト
DOK：JupyterLab 「ノートブック」機能

これらを活用し、任意のイメージからセットアップを行うことができます。



特長3 : IaaS で IT インフラをトータル提供

高火力シリーズは GPU をコントロールパネルや API といった管理ツールを合わせて提供しています。
これによりインフラエンジニアが不在の状況でも構築や構成変更などを行いやすくなります。

コントロールパネル

専用のグラフィカルな Web 管理画面を提供、クリック操作などでリソースやタスクの制御が可能であり、状態の把握なども簡単にできます。

API

コントロールパネルと同様の制御が可能な RESTful API を提供しています。一般的な Web システムから操作を自動化することもできます。

特長4 : データは国内保管

クラウドサービスが社会経済の重要インフラとなるなか、
さくらインターネットは純国産クラウド事業者として安全性の高いサービスの提供を目指しています。

ISMAP 取得

「 さくらのクラウド 」を対象に、ISMAP※クラウドサービスリストへ登録 (2021 年 12 月)。

➡ 政府機関はもちろん、民間企業や公共機関も安心して利用いただけるサービスへ

※ 政府情報システムのためのセキュリティ評価制度 (Information system Security Management and Assessment Program : 通称、ISMAP (イスマップ)) 政府が求めるセキュリティ要求を満たしているクラウドサービスを予め評価・登録することにより、政府のクラウドサービス調達におけるセキュリティ水準の確保を図り、もってクラウドサービスの円滑な導入に資することを目的とした制度

データの保管場所

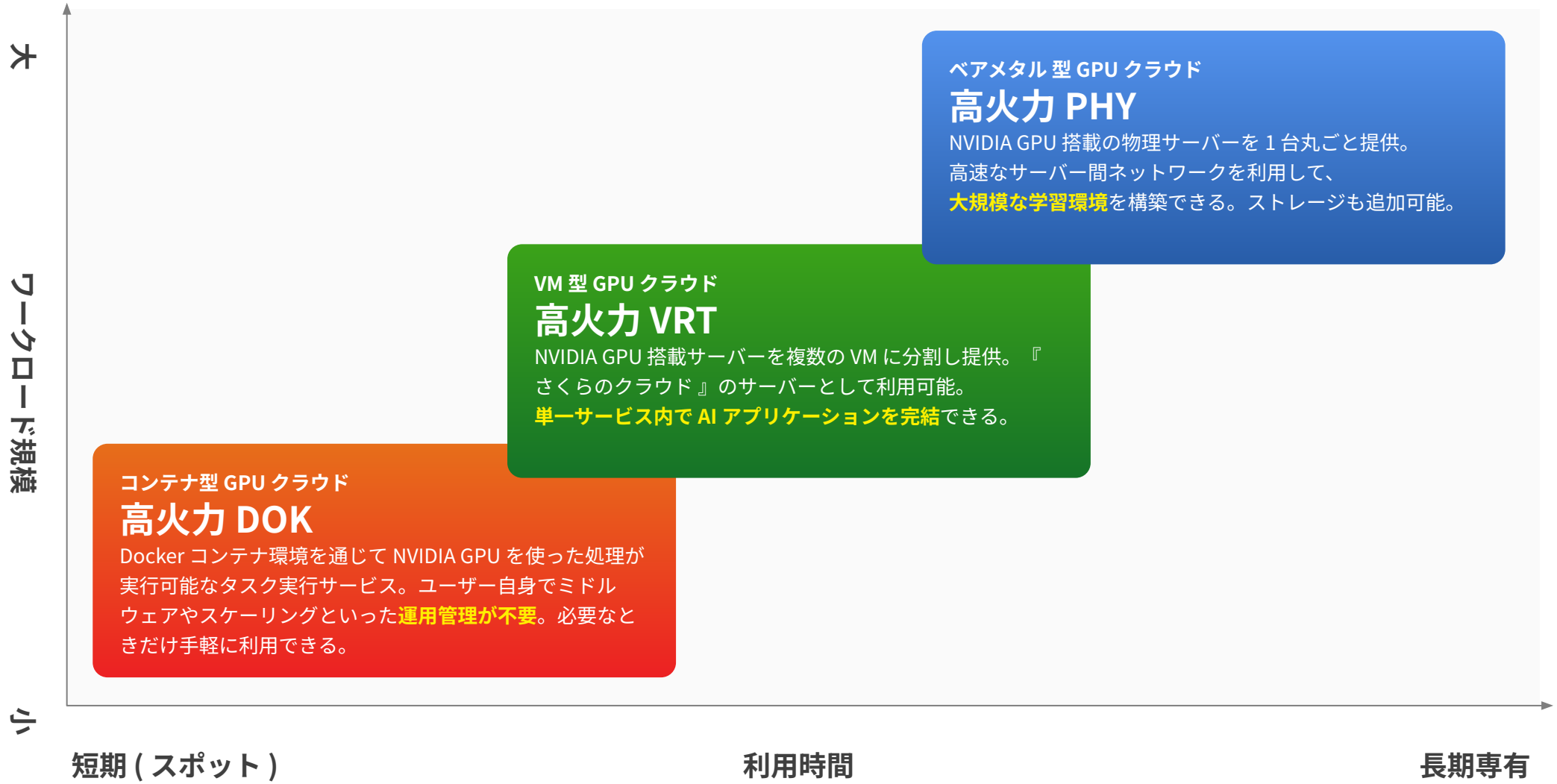
データセンターはすべて国内にあり、堅牢なセキュリティと設備環境を備えています。

サービス棲み分け比較表

実行する学習の大きさに合わせてお選びいただけます。

	高火力 PHY	高火力 VRT	高火力 DOK
概要	**本格運用に使えるベアメタル型**	**柔軟に使えるVM型**	**手軽に使えるコンテナ型**
搭載 GPUモデル	• NVIDIA B200 (8GPU) • NVIDIA H200 (8GPU) • NVIDIA H100 (8GPU)	• NVIDIA H100 (1GPU)	• NVIDIA H100 (8GPU) • NVIDIA H100 (1GPU) • NVIDIA V100 (1GPU)
料金 (税込) (H100モデル)	• お問い合わせください	• 990 円 / 時間 / GPU • 23,100 円 / 日 (963 円 / 時間 / GPU) • 385,000 円 / 月 (533 円 / 時間 / GPU)	• 0.28 円 / 秒 (1,008 円 / 時間 / GPU) ※H100(8GPU)の場合、0.83円 / 秒
課金対象	• 固定 ※ 利用有無不問	• VM 起動時間 ※ 起動中の利用有無不問	• コンテナ実行時間 ※ 実行前後は課金対象外
マネージド or アンマネージド	• アンマネージド ※ HW 構成検討やOS 管理、スケール対応が必要	• アンマネージド ※ OS 管理やスケール対応が必要	• マネージド ※ コンテナイメージの準備のみ
利用者像	• 大企業 (メーカーなど) • 研究機関 ※ 大規模なワークロード • AI メガベンチャー	• 大企業 (メーカーなど) • 研究機関 ※ 小規模なワークロード • AI アプリ / サービス提供者 ※ リアルタイム性の高い用途	• ABCI ユーザー • AI アプリ / サービス提供者 ※ リアルタイム性のないコスパ重視用途

GPUクラウド「高火力」の比較表



高火力 PHY の特徴

高火力 PHY (ファイ) は、AI 開発に欠かせない圧倒的な GPU 性能をもつ「 NVIDIA H100 /H200/ B200 Tensor コア GPU 」を 8 基搭載 。
ベアメタルサーバー (物理的なサーバー) 1 台を丸ごと提供するサービスです。

1 専有物理サーバーで安定した高性能 GPU を確保

- 物理サーバー (ベアメタルサーバー) を専有利用できるため、他のユーザーとリソースを共有せず、
  安定した高性能 GPU をフル活用できます。

2 安定運用とスケーラビリティ

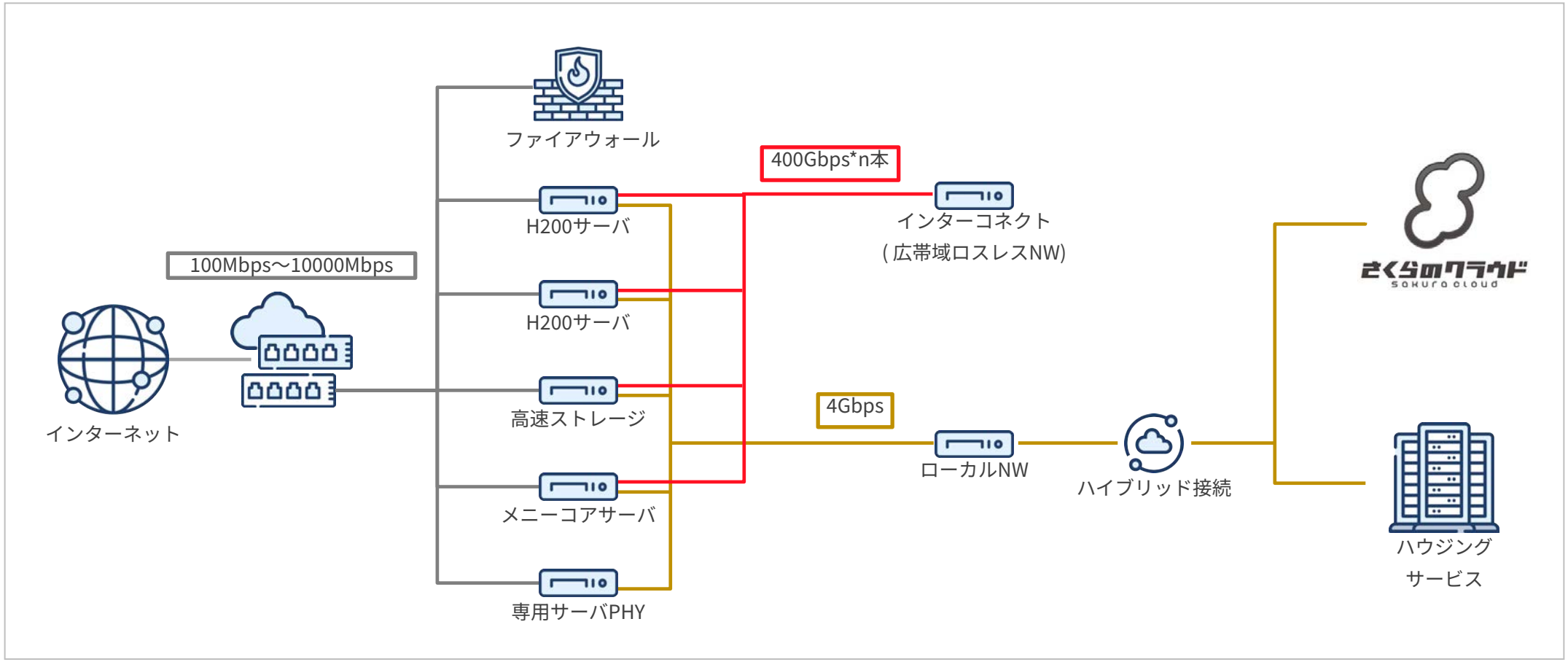
- 高速なインターコネクトが標準装備されており、大規模並列処理や複数ノードを活用した計算がスムーズに実行できます。

3 初期投資を抑えて月額料金で利用可能

- 転送量に応じた追加課金のない月額固定料金制で、月々のコストが把握しやすい。
- 予期しない費用の発生を避けられるため、安心して運用できます。

高火力 PHY の利用イメージ

高火力 PHY は、高性能な物理サーバーの特性を活かしつつ、
柔軟なネットワーク構成が取れる特長を活かしながら、クラウドサービスと組み合わせることも可能です。



高火力 PHY のユースケース

高火力 PHY は、高度な計算やシミュレーションを強力にサポートする技術として活用いただけます。

高解像度映像のレンダリング（VFX・アニメーション）

映画やアニメ、ゲーム制作などの現場では、4K や 8K といった高精細な映像のレンダリングに膨大な GPU パワーが必要とされます。

高火力 PHY なら、数十台規模の GPU クラスタをすぐに立ち上げ、筐体をまたいだ高速なデータ転送にも対応。ピーク時の負荷にも柔軟に追従し、長期の制作プロジェクトを安定して支えます。

分子動力学シミュレーション

分子構造解析における分子シミュレーションは、精密な計算を膨大に繰り返すため、GPU による並列処理と高速なノード間通信が不可欠です。

高火力 PHY は、研究に必要な GPU リソースを柔軟に組み合わせて利用でき、研究の進行や検証段階に応じた環境を迅速に用意できます。

自動運転 AI の大規模シミュレーション・学習

自動運転の開発では、大量のセンサーデータを用いた AI モデルの学習と、仮想空間での走行シミュレーションが同時に行われます。

レンダリングと推論処理が並行するこの高負荷な環境において、高火力 PHY は GPU クラスタを即時に構築でき、継続的な学習と検証のサイクルを効率的に支えます。

高火力 VRT の特徴

高火力 VRT (バート) は、高性能 GPU を搭載した VM 型 GPU クラウドサービスです。
生成 AI の学習・推論に最適な環境を、時間単位で手軽に利用できます。

1 高性能 GPU を時間単位で柔軟に利用可能

- NVIDIA 製のハイパフォーマンスな GPU を、時間単位で利用できます。
- クラウドの使い勝手そのままに利用することが可能です。

2 リアルタイム推論で遅延を最小限に

- 専有の GPU を割り当てることで、ユーザーの任意のタイミングで処理を行います。
- これによりリアルタイム性が求められる AI ワークロードも高速に処理できます。

3 コスト効率の良い AI 開発環境が構築できる

- 必要なリソースを必要なときに利用可能。初期費用不要で、1 時間 990 円からご利用いただけます。
- スケールアップ・スケールダウンも容易で、無駄なコストを削減できます。

高火力 VRT の利用イメージ

ガバメントクラウド* に認定される IaaS「 さくらのクラウド 」のリソースとして提供されます。
ストレージやネットワークなどの機能を組み合わせて単一のサービスの中で AI アプリケーションを構築することができます。

*2025年度末までに技術要件をすべて満たすことを前提とした条件付きの認定

高火力 VRT のユースケース

リアルタイム処理に適した専有環境の仮想マシン (VM) をご利用いただけます。
お客様の処理負荷に合わせて自由に台数を増減させることもできます。

チャットボットサービス

ユーザーの入力に対して瞬時に応答するチャットボットのようなシステムでは、リアルタイム性の高い処理が求められます。

VRT は、GPU による高速な推論処理を通じて、入力テキストをすばやく解析し、自然な応答を返すための環境を提供します。

異常検知システム

設備等の様々なセンサーからクラウド上に収集されたデータを用いた、故障や異常の兆候の検出にもVRTは有効です。
リアルタイムでの異常検知を実現するには、推論速度に優れた軽量なディープラーニングモデルの活用が重要とされており、VRT はこうしたモデルを安定して継続的に処理できる GPU 環境を提供します。

対話型 AI アバター

周囲の状況やユーザーの発言など、複数のデータを組み合わせて処理する対話型 AI では、並列処理が求められる場面も少なくありません。

VRT なら、用途に応じて複数のインスタンスを柔軟に展開できるため、音声認識・感情推定・応答生成といったマルチモーダルな処理にも対応可能です。

高火力 DOK の特徴

高火力 DOK (ドック) は、お客様が事前に準備した Docker イメージをクラウド上で手軽に実行できる、
マネージドなコンテナー型 GPU クラウドサービスです。

1 小規模・短時間処理に適した柔軟な GPU 環境

- Docker コンテナーを活用したジョブ実行型の GPU 環境で、処理の完了と同時に自動で停止するため、無駄なリソース利用を抑え、コストの最適化が可能です。
- 小規模な学習や、繰り返し実行される推論処理に適しています。

2 使った分だけの料金

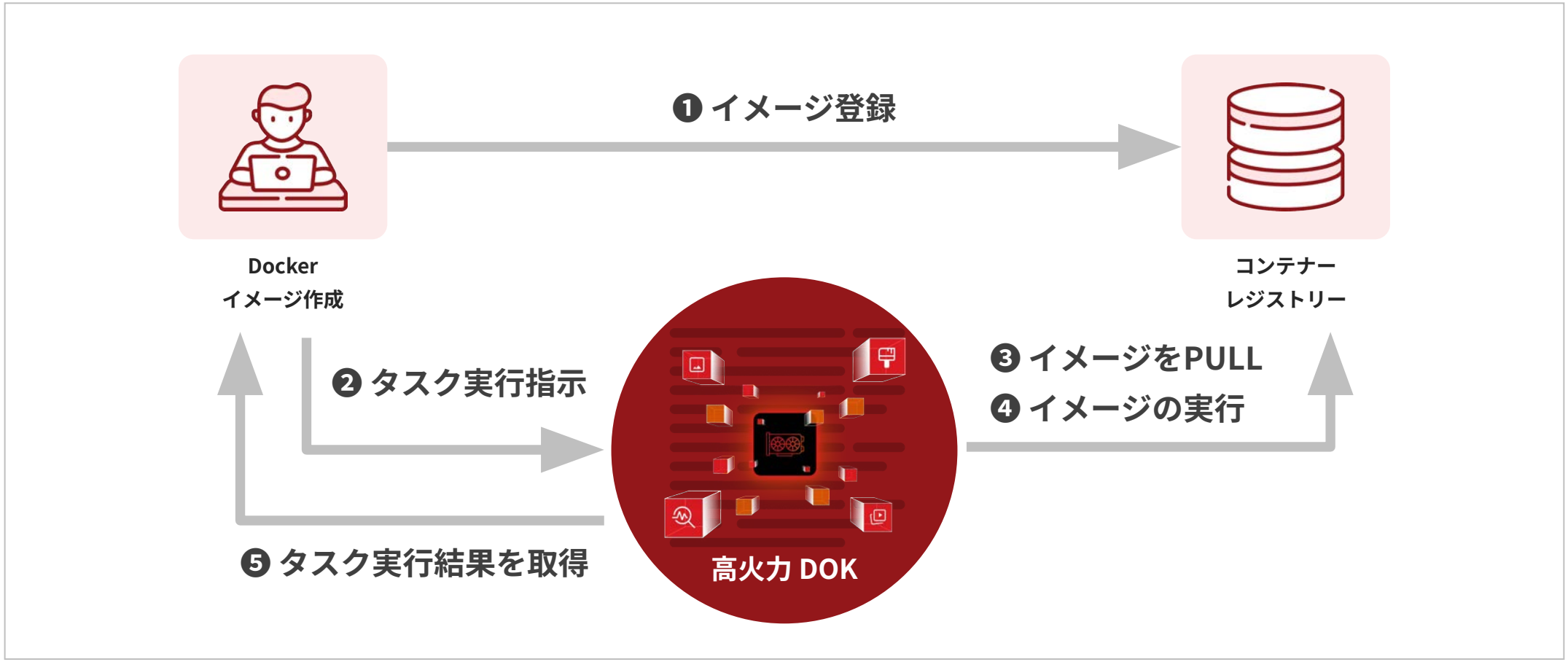
- 秒単位課金で、NVIDIA の GPU が 0.016 円 / 秒からご利用いただけます。
- 初期費用もなく、スモールスタート利用に最適です。

3 JupyterLab をワンクリックで起動可能

- コントロールパネル上から JupyterLab をワンクリックで起動できます。
- 必要なソフトウェアがあらかじめインストールされており、Docker の知識がなくても、すぐに GPU を活用した開発を始められます。

高火力 DOK の利用イメージ

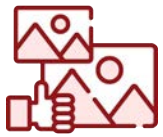
事前に作成いただいた Docker イメージをコンテナーレジストリーに登録し、高火力 DOK にタスクの実行を指示します。
タスクの実行完了後、所定の領域に保存された実行結果を取り出して活用します。



高火力 DOK のユースケース

さまざまな場面で開発の効率化や運用負荷軽減につなげることができます。

高精細画像の多数生成



膨大な画像データから学習した AI の力を借りれば、オリジナリティ溢れる高解像度な画像を高速に生成できます。

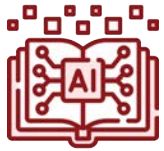
コンテンツ制作や研究開発など、クリエイティブの新たな可能性を切り拓きます。

ファインチューニング

AI モデルのファインチューニングに最適な NVIDIA 製 GPU で、お手持ちのモデルを短時間でチューニングできます。

従量課金制だから必要な分だけ利用可能。モデルのパフォーマンス向上にお役立てください。

AI の学習・推論



リアルタイムの画像分類や物体検出、音声解析など、AI を利用したサービスに活用できます。

GPU を必要なときだけ利用できる柔軟性で、サービスの成長に合わせたスケールが可能です。

資料ダウンロード・お申し込み

もっとお知りになりたい方向けに、多数のダウンロードコンテンツをご用意しております。
下記リンクより、ご確認ください。

ダウンロードはこちら

https://www.sakura.ad.jp/document/?service=koukaryoku#documents

高火力シリーズをくわしく見る

https://ai.sakura.ad.jp/gpu/