

VM型GPUクラウドサービス

パート

高火力 VRT

- サービス説明資料 -

目次

- i. 高火力 VRTとは
- ii. 高火力 VRTの 3 つの特徴
- iii. 高火力 VRTが選ばれる理由
- iv. AIワークロードを支える、NVIDIA H100のパフォーマンス
- v. ユースケースのご紹介
- vi. 事例のご紹介
- vii. 料金プラン・仕様
- viii. 構成例
- ix. ご利用の流れ
- x. 料金体系とお支払いについて
- xi. さくらインターネットのGPUクラウドサービスとは
- xii. よくあるご質問・問い合わせ
- xiii. 会社概要

こんなお悩みありませんか？

No.1

GPUを使いたいけど、物理サーバーを持つのは大変・・・



No.2

予算が限られていて、高額な初期投資は難しい・・・



No.3

用途に応じて、環境を自由にカスタマイズしたい・・・



手軽さも、コストも、柔軟性も。

高火力 VRT

なら、すべて解決。

高火力 VRT とは

NVIDIA製の高性能GPUを搭載した **仮想マシン (VM)** を、クラウドならではの手軽さで利用できる **VM型GPUクラウドサービス**です。

生成AIの学習・推論に最適化されたVMを **時間単位** で利用でき、大規模な計算処理も **高速に実行** できます。

1

ハイパフォーマンスなGPUをVM（仮想マシン）で提供

2

初期費用不要で 990円 から利用可能

※一定期間の利用分をまとめてお支払いいただく割安プランもあり

3

柔軟なシステム構築が可能



幅広いAIモデルの学習や、画像生成・自然言語処理など**リアルタイム実行に最適**！

高火力 VRT の3つの特徴

1

ハイパフォーマンスなNVIDIA H100をVM（仮想マシン）で提供

機械学習やディープラーニングといった大量の計算処理が求められるタスクや、リアルタイム性が求められるあらゆる分野のAIワークロードも高速に処理



2

初期費用不要で990円から利用可能

※一定期間の利用分をまとめてお支払いいただく割安プランもあり

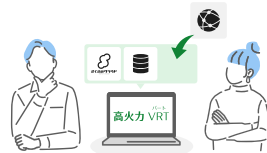
初期費用不要で NVIDIA H100 を1時間990円からご利用いただけます。
処理内容に合わせて最適なプランを選択し、使いたいときに使いたい分だけの利用で無駄なコストを削減



3

柔軟なシステム構築が可能

さくらのクラウドのサーバー、ストレージ、ネットワークと組み合わせることで柔軟な環境構築が可能。
開発期間を短縮



高火力 VRT が選ばれる理由



高性能なGPU

NVIDIA H100 を搭載
計算リソースを大量に必要とする
タスクに最適

リアルタイム処理 が求められる
AIモデルの推論も
高速に実行 できる、**専有環境** を提供

クラウドならではの手軽さ

さくらのクラウドと
シームレスに接続
柔軟なシステム構築を実現

さくらのクラウドの
コントロールパネルから、
簡単に管理・操作が可能

無駄なコストを抑える

使いたいときに使いたい分だけ、
最短1時間だけ利用で
無駄なコストを削減

長期利用にメリットのある、
期間コミットプランも提供

AIワークロードを支える、NVIDIA H100のパフォーマンス

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なら、高速かつ効率的に処理できる。

1 リアルタイムのディープラーニング推論を実現

NVIDIA H100は、様々な AI アプリケーションのパフォーマンスを飛躍的に向上させる革新的な製品です。
AI を活用したビジネスの成功を目指す企業にとって、NVIDIA H100 GPUは最適なソリューションとなっています。

2 エクサスケール ハイパフォーマンス コンピューティング

NVIDIA H100の新しい画期的なAI 機能は、HPC+AI のパワーをさらに増幅し、世界の最重要課題の解決に取り組む科学者や研究者にとって、発見までの時間が加速されます。

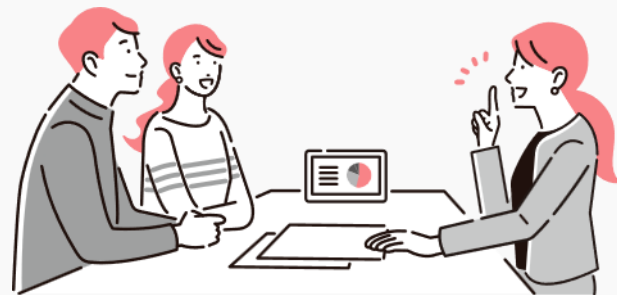
3 データ分析を高速化

多くの場合、データ分析工程はAI アプリケーション開発の大半を占めます。
NVIDIA H100は膨大なデータセットに対するハイパフォーマンスとスケール性でデータ分析を高速化できます。



出典 : <https://www.nvidia.com/ja-jp/data-center/h100/>

ユースケース・事例



高火力 VRT 活用例

こんな使い方ができます

1 小規模学習環境

オープンウェイトモデルのファインチューニングに必要な環境を簡単に切り出すことができます。

2 異常検知システム（動体検知・パターン検知）

設備の各種センサーからクラウドに収集されたデータを活用し、故障や異常の兆候を検出することも可能です。

3 対話型AIアバター・AIチャットボット・コールセンターAI

音声認識・音声合成・画像生成に加え、周囲の状況やユーザーの発言、問い合わせ内容など、多くのデータを並列処理する必要がある場面でも、複数の仮想マシンを組み合わせることで柔軟に対応できます。



事例のご紹介 (株式会社インフォメーション・ディベロップメント / SBI R3 Japan株式会社)

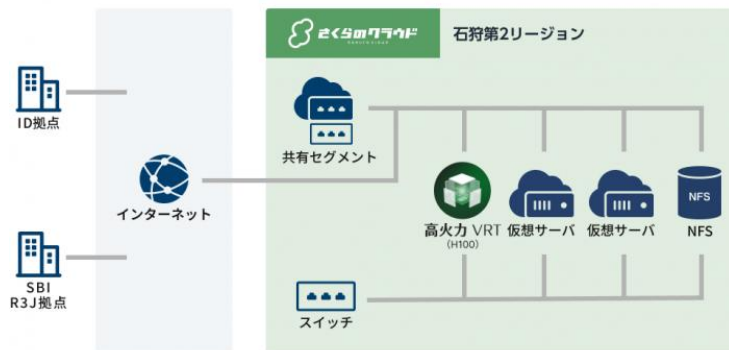


SBI R3.
Japan

お客様：株式会社インフォメーション・ディベロップメント / SBI R3 Japan株式会社

事業内容：監査性・透明性・逐次学習を重視した国産AI基盤

用途：金融・行政分野での利用を想定した次世代AIエージェント「SAPHI」の学習・再学習および性能検証を行うためのGPU基盤として利用されています。ローカル環境では時間がかかっていた大規模なAIモデルの学習やテストを高速に実行しつつ、データを国内に保持したまま厳格なセキュリティとガバナンスを確保する目的で採用されました。



<https://case.sakura.ad.jp/case/1895/>

事例のご紹介（株式会社Tellus 様）



お客様： Tellus 様

事業内容： 日本初の衛星データプラットフォーム『Tellus』を開発・運用

用途： 衛星データ × AIモデル × ITインフラを一体化することで
利用者に衛星データ活用の開発環境を迅速に提供



Tellus AI Playground 3つの特徴

01

DATA

日本最大規模の衛星データ量。
オープンデータはもちろん、
解像度が高い
商用データも提供可。



02

MODEL

複数用意した
衛星データ用 AI モデルから
セレクト



03

CLOUD GPU

AI 学習に適した
高性能 H100 を
VM 形式で提供



<https://case.sakura.ad.jp/case/1622/>

事例のご紹介 (HarvestX株式会社様)



お客様： HarvestX株式会社

事業内容： AIロボティクスにより、自動授粉技術でイチゴを安定生産できる

用途： 栽培システムを提供、人工光型植物工場の完全自動化に挑戦
AIモデルの学習基盤として利用、安心のデータ保管と高速な学習サイクルで学習時間を短縮、授粉に必要な花の検出精度を大幅に向上



<https://case.sakura.ad.jp/case/1599/>

料金プラン・仕様



高火力 VRT ご利用料金 - 通常料金：上限付きオンデマンドプラン -

当社おすすめ

NVIDIA V100プラン

*3 2027年3月31日をもって、提供終了いたします。

時間額	481円
日額	11,550円
月額	231,000円
時間額適用 *1	24時間まで
日額適用 *1	20日まで
最低利用期間	1時間
申込方法	コントロールパネル／API *2

NVIDIA H100プラン

990円
23,100円
385,000円
24時間まで
17日まで
1時間
コントロールパネル／API *2

* 金額はすべて税込みです、VMをシャットダウンしている時間は課金対象外となります

* OSインストールのために別途ディスクの作成が必要です <https://cloud.sakura.ad.jp/products/disk/>

*1 指定時間を超過した場合、自動的により安価な課金額が適用されます <https://cloud.sakura.ad.jp/payment/about/#billing-system>

*2 サービスリリース後、在庫があればお申込みいただけます

*3 サービスラインアップの見直しに伴い、2027年3月31日をもって「NVIDIA V100」プランの提供を終了いたします。

また、2026年12月31日をもって新規作成の受付を終了いたします。詳しくは [こちら](#) をご参照ください。

高火力 VRT ご利用料金 – 特別料金：期間コミットプラン –

期間コミットプランとは、事前に該当期間分の料金をまとめてお支払いいただくことで、
通常よりも割安にGPUリソースを確保し、**長期的かつ安定的にご利用いただけるプラン**です。継続的な高負荷利用が見込まれる、
有期プロジェクトや、ユーザーからのGPU処理要求が継続的に発生するアプリケーションへの組み込み利用に最適です。

期間コミットプラン	
NVIDIA H100プラン	
6か月コミット	365,750円/月額
12か月コミット	346,500円/月額
(参考) 通常料金	385,000円/月額

申込方法

[専用の問い合わせフォーム](#)からお申し込みください。

- * 金額はすべて税込みです
- * 上記以外の期間についても、期間コミットプランを適用可能です。割引価格は原則として6か月または12か月の料金体系が適用されます。
例：8か月コミットの場合6か月コミットに準拠した365,750円/月額 となり、8か月分の2,926,000円を一括請求いたします。
- * OSインストールのために別途ディスクの作成が必要です <https://cloud.sakura.ad.jp/products/disk/>

サーバ提供仕様

	NVIDIA V100プラン ※6 2027年3月31日をもって、提供終了いたします。	NVIDIA H100プラン
GPU	NVIDIA V100 PCIe 32GB x1	NVIDIA H100 SXM 80GB x1
CPUモデル	非公開	非公開
仮想コア (vCPU)	4 コア	24 コア
メモリ容量	56 GB	240 GB
一時領域 ^{*1}	なし	6.9 TiB (NVMe)
グローバルNW帯域 ^{*2 *3}	100 Mbps	100 Mbps
ローカルNW帯域 ^{*2}	2.0 Gbps	10.0 Gbps
ローカルNWインターコネクト	なし	なし
提供ゾーン	石狩第1ゾーン	石狩第1ゾーン
OS ^{*4}	Ubuntu Server 24.04/22.04/20.04 AlmaLinux 9系 Rocky Linux 9系	Ubuntu Server 24.04/22.04/20.04 AlmaLinux 10系/9系/8系 Rocky Linux 10系/9系/8系

*1 一時領域はメンテナンスなどによるサーバーの停止ごとに消去されます、データは必要に応じて別途提供されるディスクやオブジェクトストレージなどに保存いただく必要があります。

*2 帯域はベストエフォートになります。

*3 グローバル回線は別途有償オプションとなるルータ+スイッチを追加することで帯域を増強できます。詳しくは、[マニュアルページ](#)をご確認ください。

*4 エラーなく起動することまでの確認となります、各ミドルウェア・アプリケーションの動作はお客様にてご確認いただく必要があります

*5 既に提供終了しているOSを含みます、提供しているOSの一覧は [提供中のアーカイブ・ISO イメージ一覧](#) をご確認ください

*6 サービスラインアップの見直しに伴い、2027年3月31日をもって「NVIDIA V100」プランの提供を終了いたします。

また、2026年12月31日をもって新規作成の受付を終了いたします。詳しくは [こちら](#) をご参照ください。

高火力 VRTではサーバーとディスクは別々のリソースとして提供しています。

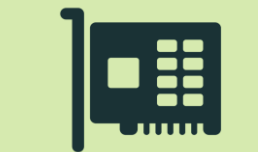
そのため、**利用目的に応じた自由な組み合わせ**が可能です。



サーバー



ディスク



一時領域 (NVMe)

*H100のみ



SSDプラン



標準プラン
(HDD)

サーバー仕様

- サーバープランに付属、追加料金なしで利用可能
- 電源断（ホストサーバーの障害含む）時に データは消去
- VMごとに専有、ハードウェアを活かした高速なI/Oを提供
- 容量は 固定 (6.9TiB)、追加不可
- 課金は 起動中のみ 発生、シャットダウンで課金停止

ディスク仕様

- サーバープランとは別に選択、別途料金が発生
- 冗長構成により、リソース削除まで データを保持
- さくらのクラウド全体で 共有されるため、容量・プランにより 性能上限あり
- 容量はプランごとに 複数用意
- ブートディスクとして最低1本の構成が必須、最大3本まで追加可能
- 課金は リソースが存在する限り 発生、リソース削除で課金は停止

高火力 VRT では、安定したサービス提供を目的として、初期設定では GPU の利用上限を 1 基に制限しています。

ただし、サポート窓口へお申し込みいただくことで、**リソース上限は緩和可能**です。

リソース上限変更のお申し込み方法

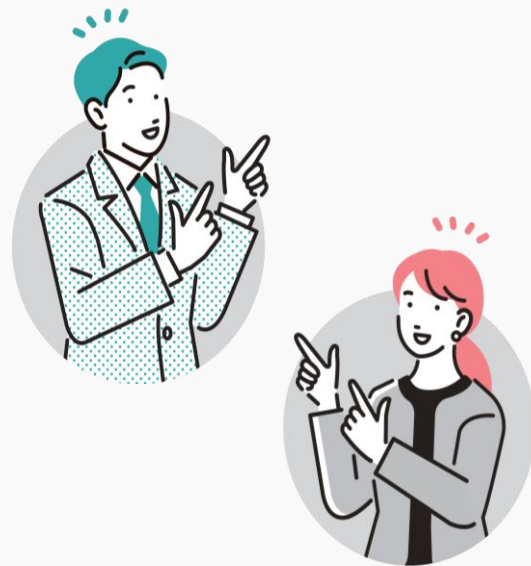
以下の内容をご記入のうえ、サポート窓口へご連絡いただきますようお願い申し上げます。

- 宛先：support@sakura.ad.jp
- 件名：さくらのクラウド リソース上限変更依頼
- 本文：ご希望の内容を本文にご記入のうえ、ご連絡いただきますようお願い申し上げます。

1. 会員 ID（例：XXX000000）
2. 会員名
3. プロジェクト名（例：Project-name）
4. 利用ゾーン
5. 変更対象の項目（GPU 数、メモリ総容量、サーバー数、ディスク総容量 など）
6. 変更後の希望上限値
7. 利用用途（背景・目的・想定する利用内容）
8. 消費ペース（直近 6 か月間におけるリソース利用増加の状況）

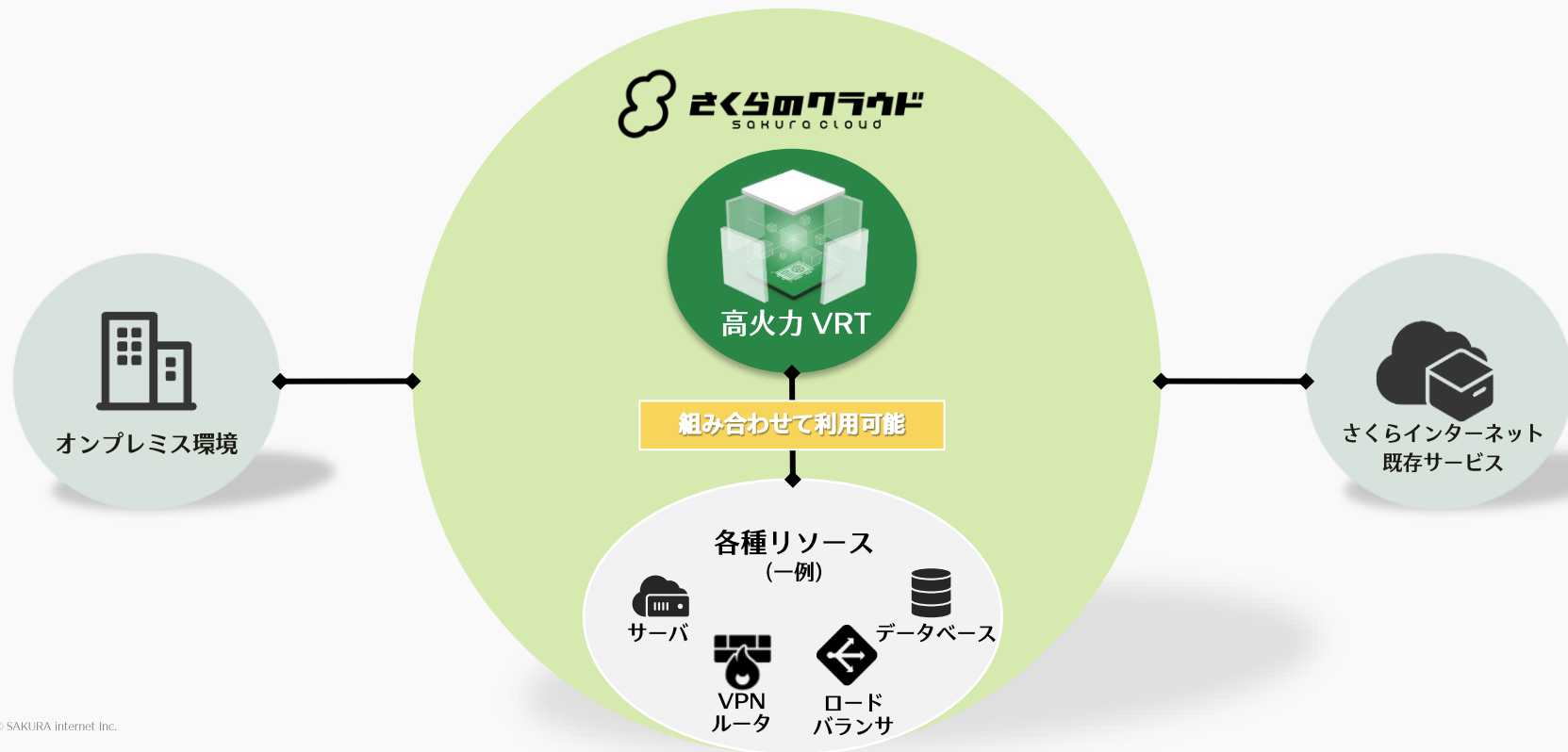


構成例

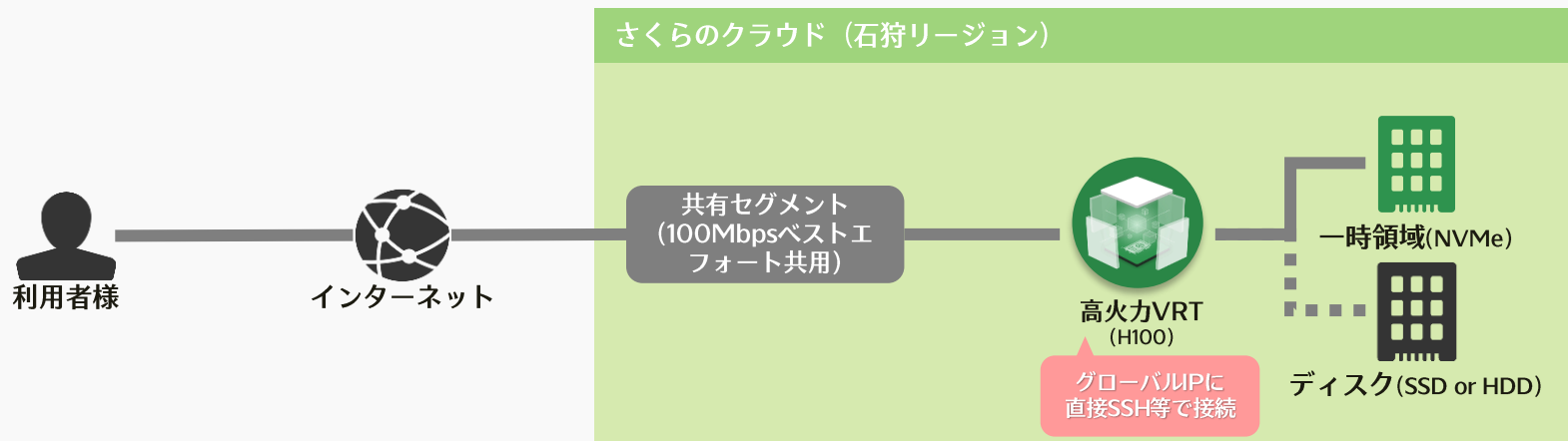


高火力 VRTとの連携

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。
GPU非搭載のサーバーだけでなく、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。

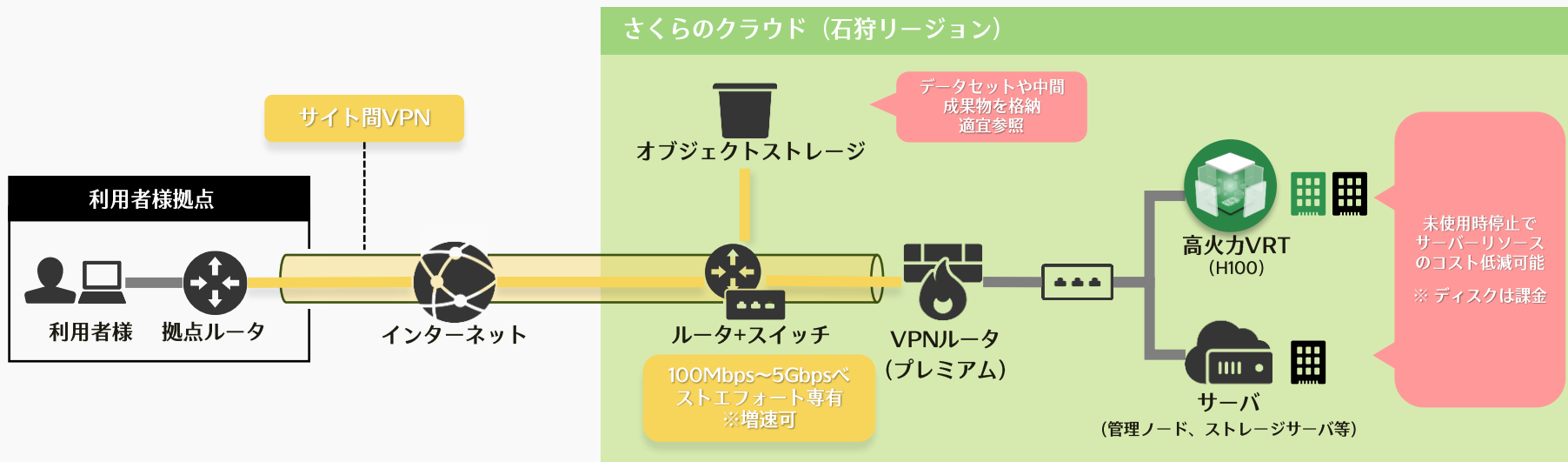


構成例①：インターネットに高火力VRTを接続するシンプルな構成



- * サーバへの接続はインターネット経由でSSH等を利用して接続する想定です。
- * 一時領域はメンテナンスなどによるサーバーの停止ごとに消去されます、データは必要に応じて別途提供される**ディスク**や**オブジェクトストレージ**などに保存いただく必要があります。
- * サーバ料金以外に別途ディスクの料金が発生します。また、ディスクはサーバを停止しても課金対象となりますのでご注意ください。
- * 共有セグメントは100Mbpsベストエフォートの他ユーザとの共有回線となります。単一ユーザによる帯域の利用や、100Mbps以上の帯域をご希望の場合には、**ルータ+スイッチ**をご利用ください。
- * パケットフィルタ機能によりサーバの仮想NICに着信するパケットを指定した条件でフィルタすることが可能です。
- * 高火力VRTをインターネットに直接接続する場合には上記パケットフィルタ機能やOS上でファイアウォールを利用するなど、セキュリティ対策にはご注意ください。
- * その他詳細は**マニュアル**をご参照ください。

構成例② サイト間VPNによるセキュリティを意識した構成



- * サイト間VPNにより拠点側のローカルNWとさくらのクラウドのローカルNWを接続し、高火力VRTを利用する想定です。
- * 一時領域はメンテナンスなどによるサーバーの停止ごとに消去されます。データは必要に応じて別途提供されるディスクやオブジェクトストレージなどに保存いただく必要があります。
- * サーバ料金以外に別途ディスクの料金が発生します。また、ディスクはサーバを停止しても課金対象となりますのでご注意ください。
- * さくらのクラウド上からオブジェクトストレージへのデータ転送量は無料となります。
- * 構成図上のサーバは管理ノードやストレージサーバとして利用する想定です。
- * 高火力VRTのローカルNW帯域はV100の場合は2Gbps、H100の場合は10Gbpsで固定となりますが、通常のサーバの場合は搭載メモリによってローカルNW帯域が変動します。
- * ルータ+スイッチの帯域は100Mbps~5Gbpsの間で選択が可能です。また、ルータ+スイッチは停止を伴わずに帯域の拡張が可能です。
- * VPNルータを利用することで、サイト間VPN機能や、ステートフルなアクセス制御機能が利用できます。詳細はマニュアルをご確認ください。
- * VPNルータ⇄拠点ルータ間でサイト間VPNを構成する場合、拠点ルータに接続しているグローバル回線に固定のグローバルIPが必要となります。
- * その他詳細はマニュアルをご参照ください。

ご利用の流れ



ご利用の流れ

STEP

1

新規会員登録

会員IDをお持ちでない方は、まずはじめに会員IDを作成します。詳細は [さくらインターネット会員IDの取得](#) をご参照ください。

※ ご登録には電話番号が必要です。(電話番号認証を行います)

※ すでにさくらインターネットの会員IDをお持ちの方はSTEP2にお進みください。

STEP

2

プロジェクトの作成

コントロールパネルにログイン後、さくらのクラウドのプロジェクトを作成します。詳細は [さくらのクラウドプロジェクト登録](#) を参照ください。

※ 「さくらのクラウド」をご利用いただく際にはクレジットカード登録が必要です。さくらインターネットの他のサービスでクレジットカードを登録されている場合、追加登録は不要です。

※ すでに「さくらのクラウド」のプロジェクトを作成済の方はSTEP3にお進みください。

STEP

3

サーバー作成画面からご利用開始

サーバー作成画面から高火力VRTのプランを選んでご利用ください。

※ STEP1、STEP2で取得した情報を入力し、ログインしてください。

※ 期間コミットプランは[専用フォーム](#)よりお問い合わせください。

※ 詳しいご利用方法は次のスライドでご確認ください。

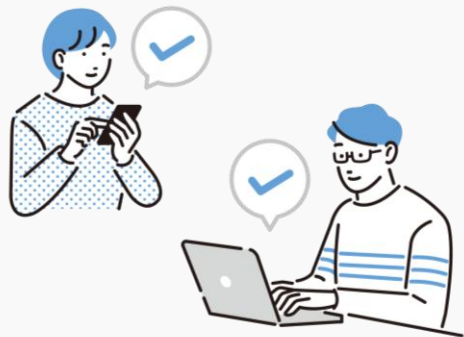
事前申請なしで即利用可能

高火力VRTの通常料金プランは事前申請不要。

オンラインサインアップ後、すぐに利用を開始できます！

NVIDIA H100などのハイエンドGPUも、コントロールパネルからの

簡単な操作で迅速に起動でき、スムーズに開発を始められます。



サーバ作成からGPUの利用開始までの流れ

1 コントロールパネルにログインし、「サーバ追加」画面に移動

さくらのクラウドのコントロールパネルにログインし、操作対象とするプロジェクトを選択し、「さくらのクラウド(IaaS)」をクリック。左メニュー「サーバ」から「追加」を選択してください。

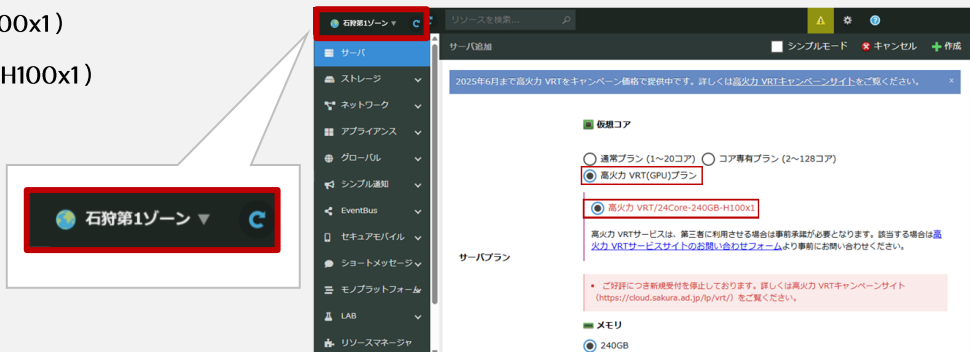
▼ 注意点

画面上部の「シンプルモード」チェックボックスを外し、**通常モード**に切り替えてください。GPUプランは通常モードでのみ作成できます。

2 高火力 VRT (GPU) プランを選択

石狩第1ゾーンを指定し、サーバプランから「高火力 VRT(GPU)プラン」を選択し、ご希望のGPUを選択してください。

- NVIDIA V100プラン (高火力 VRT/4Core -56GB -V100x1)
- NVIDIA H100プラン (高火力 VRT/24Core -240GB -H100x1)

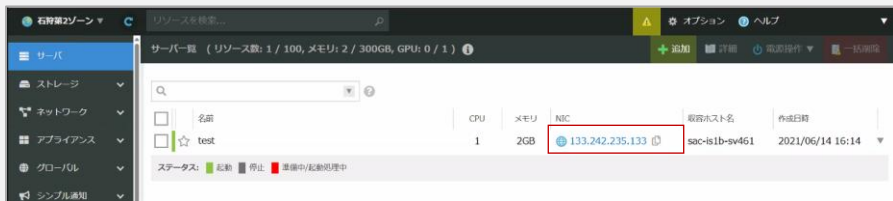


3 各種設定（ディスク、ネットワーク、サーバー名など）を入力し、「作成」をクリック

ディスクの作成、NICの接続、サーバー名の入力など、必要な項目を順に設定してください。
画面下部には料金の概算も表示されますので、構成内容と費用を確認のうえ「作成」ボタンを押してください。



数分後には、OSイメージが展開されサーバーが起動状態になります。サーバーにはグローバルIPアドレスが付与されますので、SSHなどで接続し、ログインすることができます。

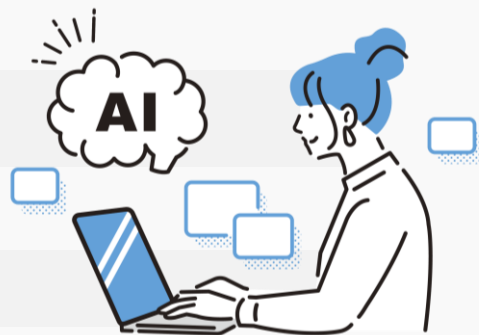


4 GPUドライバーをインストール

サーバー作成後は、お客様側でOSへのGPUドライバーのインストールが必要となります。
マニュアルにドライバーのインストール手順を記載しておりますので、ご確認ください。

5 ドライバーの動作確認

インストールが完了したら、nvidia-smi を実行。GPU名・ドライバー情報などが表示されれば準備完了です。



料金体系とお支払いについて



通常料金プランの料金体系について (1)

1 初期費用

新規導入時の初期費用はありません。

2 課金対象

さくらのクラウドの料金は作成からコントロールパネル上から削除するまでの、サービスが存在した時間の総計で課金されます。計算対象の時間は1秒単位で集計されます。またご請求金額を算出する際は1時間を課金単位とし、1時間未満の端数は1時間として切り上げされます。

サーバーやデータベースの料金は、起動から停止している時間の総計で計算されます。サービスの種類によって、課金対象の期間が異なるものがあるため、詳細は、さくらのクラウドマニュアルの「[ご利用料金計算方法](#)」をご参照ください。

※ データベースは、冗長化オプションのないプランの場合のみ、起動状態の時間の総計が課金対象となります。

※ 2024年9月26日(木)より前のデータベースは、作成から削除までの存在した時間によって請求されます。

課金対象

さくらのクラウドの料金は「起動」から「停止」までの期間を課金対象とする場合と、「作成」から「削除」までを課金対象とする場合があります。



通常料金プランの課金の仕組みについて (2)

3 ご利用料金の月額・日割り・時間割り

高火力 VRTは、24時間以降の利用から日額料金が適用されます。NVIDIA V100は20日以降、NVIDIA H100は17日以降から月額料金が適用されます。

4 請求サイクル

請求はすべて後払いとなり、お申し込み時や利用開始前に請求は発生しません。毎月末締めで翌月に請求を行わせていただきます。また、右の図のように月をまたいだご利用に関しても月末の締め時点で10月・11月の2ヶ月をそれぞれ別の請求に分割いたします。



お支払について

1 クレジットカード

お支払いには、クレジットカードをご利用いただけます。お使いいただけるクレジットカードは以下の種類です。

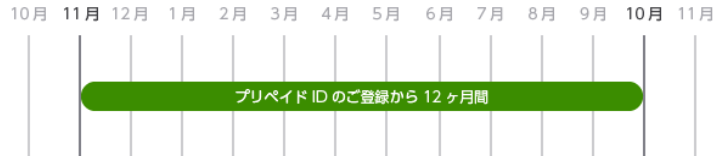


※ クレジットカードのお支払いに問題が発生した場合、または1回の請求額が1,000万円以上となった場合は銀行振込でのご請求をさせていただきます

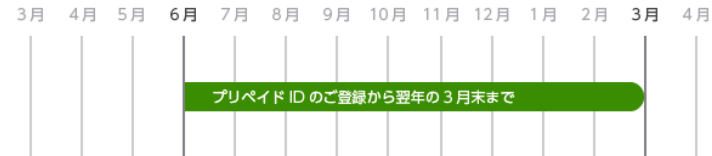
2 プリペイド

プリペイドは、法人・文教のお客様向けの前払いでご利用いただけるお支払い方法です。プリペイドIDのご登録から12ヶ月間のものと、ご登録から3月末までの期間ご利用いただける単年度プリペイドがあります。詳細は[さくらのクラウドマニュアル](#)をご参照ください。

プリペイド利用イメージ



単年度プリペイド利用イメージ



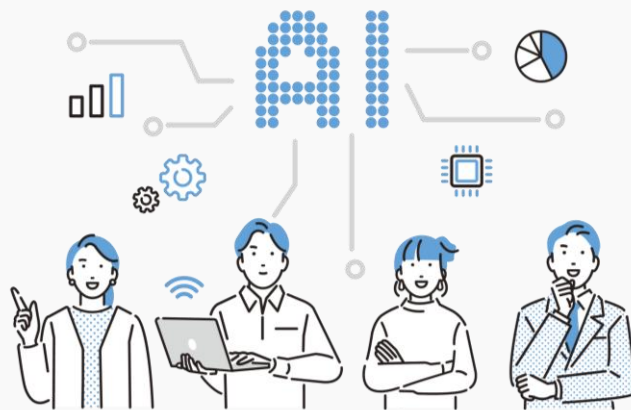
単年度プリペイドの詳細につきましては、お気軽にお問い合わせください。

※ 請求に関するよくある質問は[さくらのクラウドマニュアル](#)をご参照ください。

※ お支払いが確認できない場合、サービスの一部機能を制限いたしますことをご了承ください。

お役立ち情報

さくらインターネットの GPUクラウドサービスとは？



「さくらインターネット」が「GPUクラウド」を提供する理由

IT業界では珍しい**垂直統合型・自前主義**のビジネスモデルにより
顧客の「やりたいこと」を柔軟かつスピーディーに実現できるDXソリューションを提案

強み

事業環境への柔軟性・顧客ニーズへの即応性・バリューチェーンへの最適化



サポート

販売

サービス運用

サービス開発・技術研究

データセンター



クラウドサービス



ブロックチェーン
AI・
ディープラーニング

IoT

エッジ
コンピューティング

さくらインターネットのサービス

クラウドサービス



さくらのレンタルサーバ

手軽な個人向けから安定性重視の法人向けまで、運用・保守はさくらにおまかせ



さくらのVPS

レンタルサーバーの手軽さとクラウドの拡張性を併せ持つ仮想サーバーサービス



さくらのクラウド

インターネット上で多彩なインフラ構成を実現できるパブリッククラウドサービス

物理基盤サービス



さくらの専用サーバ PHY

クラウドの使い勝手と物理筐体の性能を併せ持つ物理専有型サービス



ハウジング リモートハウジング

耐震性を備えた建屋に安定した電力／冷却／ネットワークを備えたコロケーションサービス

GPUインフラストラクチャーサービス



生成AI・機械学習に最適な計算リソースサービス

高火力 DOK

Dockerコンテナを通じてGPU処理が実行できるコンテナ型GPUクラウドサービス

高火力 VRT

さくらのクラウドとして、GPUを1基から利用できるVM型GPUクラウドサービス

高火力 PHY

複数GPU搭載の物理サーバーを1台丸ごと提供するベアメタル型GPUクラウドサービス

マネージドHPCクラスターサービス さくら ONE

高火力PHYと高速＆大容量なストレージで構成された予約利用型のHPCクラスター大規模な研究に求められるGPU数百基の構成を最短30日から利用可能

周辺サービス



オリジナルの画像変換・最適化とキャッシュ・配信までを一括でおこなう画像変換・配信クラウドと、高品質なライブ配信環境を手軽に構築できるマネージドサービス「ImageFlux Live Streaming」を展開



さくらのSSL



ウェブアクセラレータ



さくらのIoT

さくらのセキュアモバイルコネクト

クラウド事業者が提供するIoT用途向けモバイルネットワークサービス



antenna-eye

Antenna-eye (エレコム社協業)

さくらのクラウドで稼働、ハイレベルなセキュリティ対策で大切なデータを守るクラウド録画サービス



さくらのAI

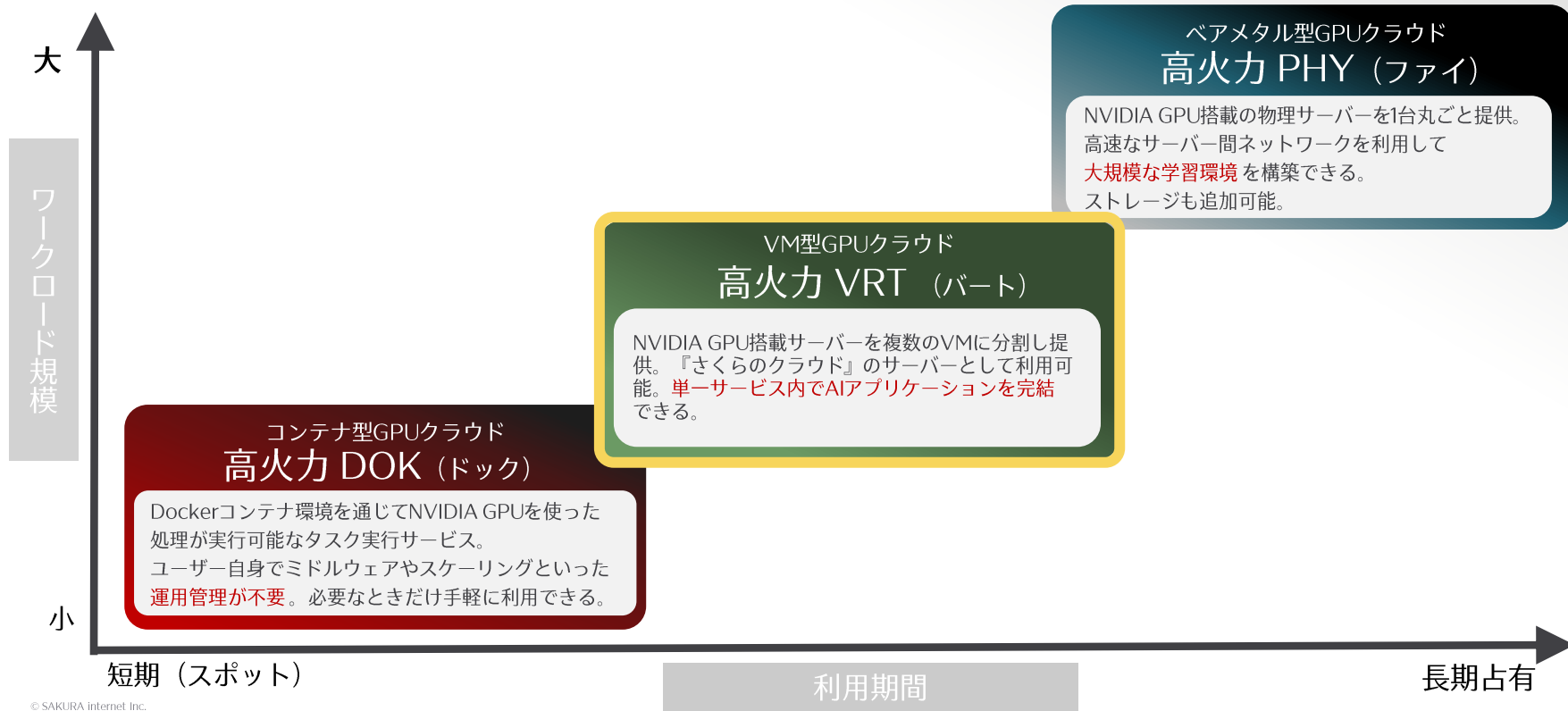
AI Engine / AI ソリューション

業務にすぐ使えるアプリケーションを多数ラインアップする国内完結型基盤

GPUクラウド「高火力」のラインアップ

「高火力シリーズは、ニーズに合わせて選べる **3つのGPUクラウドサービス** を提供します。

最適なプランで、すぐに高性能な計算環境を利用可能です。



高火力シリーズ 比較 (1)



高火力シリーズ 比較(2)

	高火力 PHY	高火力 VRT	高火力 DOK
提供 GPU プラン	・ NVIDIA B200 (8GPU) ・ NVIDIA H200 (8GPU) ・ NVIDIA H100 (8GPU)	・ NVIDIA H100 (1GPU) ・ NVIDIA V100 (1GPU)	・ NVIDIA H100 (8GPU) β版 ・ NVIDIA H100 (1GPU) ・ NVIDIA V100 (1GPU)
料金 (税込) ※ H100 プラン ※ オンデマンド	詳しくは問い合わせください ※コミットプランあり	NVIDIA H100 (1GPU) 990円/時間額 23,100円/日額 385,000円/月額 ※コミットプランあり	0.28円/秒 (1,008円/時間/GPU)
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管理や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	・ AIメガベンチャー ・ 研究機関 ※その中でも大規模なワークロード ・ 大企業 (メーカーなど)	・ 研究機関 ※その中でも小規模なワークロード ・ 大企業 (メーカーなど) ・ AIアプリ/サービス提供者 ※その中でもリアルタイム性の高い用途	・ ABCI (AI橋渡しクラウド) ユーザー ・ AIアプリ/サービス提供者 ※その中でもリアルタイム性のない コスト重視用途
提供開始日	2024年1月31日 (水)	2025年4月16日 (木)	2024年6月27日 (木)

さくらインターネットが提供する、高付加価値クラウドサービス一覧

	さくらONE	高火力シリーズ			さくらのAI	
		高火力 PHY	高火力 VRT	高火力 DOK	さくらの AI Engine	さくらの AIソリューション
サービス種類	マネージドHPCクラスタ	ベアメタル型 GPUクラウドサービス	VM型 GPUクラウドサービス	コンテナ型 GPUクラウドサービス	生成AI向け 推論API基盤	生成AI 業務支援サービス
特徴	高火力PHYと高速ストレージで構成された予約利用型のHPCクラスタ。GPU8基からなるサーバー1台から、GPU数百基構成まで利用可能。	1筐体に8基搭載されたNVIDIA TensorコアGPUを占有できるサービス。物理サーバーを直接利用することで高い性能を発揮。	GPUを1基から時間単位で利用できるサービス。IaaS「さくらのクラウド」のネットワークやストレージと組み合わせる柔軟なシステム構成や拡張も容易。	GPU用に作成されたDockerコンテナを実行できる、マネージドサービス。Jupyter Notebookなどがワンクリックで起動し、すぐにGPUを使える。	国内完結型の安全な基盤で、音声認識から文書検索まで幅広く使えるAIパッケージを提供。	AI導入から運用まで、専門チームがPoC設計から本番導入まで伴走支援。既存環境や他クラウドとの閉域接続にも柔軟に対応。
提供プラン	・ H200ノード ・ B200ノード	・ NVIDIA B200 (8GPU) ・ NVIDIA H200 (8GPU) ・ NVIDIA H100 (8GPU)	・ NVIDIA H100 (1GPU) ・ NVIDIA V100 (1GPU)	・ NVIDIA H100 (8GPU) β版 ・ NVIDIA H100 (1GPU) ・ NVIDIA V100 (1GPU)	・ 基盤モデル無償プラン ・ 従量課金プラン	・ 月額固定料金プラン (定額)
料金 (税込)	・ 別途お問い合わせ	・ 別途お問い合わせ ※コミットプランあり	・ NVIDIA H100 990円/時間額 23,100円/日額 385,000円/月額 ※コミットプランあり	・ NVIDIA H100 0.28円/秒 1,008円/時間 ・ NVIDIA H100 8GPU専有 0.83円/秒 2,988円/時間	・ 月3,000リクエストまで (チャット) などの無償利用枠あり。	・ 別途お問い合わせ
利用者像	大規模な科学計算・シミュレーション・AI研究を行う研究者やエンジニア。	大手企業や研究機関で、大規模な計算やAI学習を行う研究者・技術者。	研究機関や大企業（メーカーなど）で、小規模なワークロードを扱うユーザー、またはAIアプリやサービスを開発・提供する事業者。	ABCユーザーなど研究系クラウド利用者や、リアルタイム性を求めないコスト重視のAIアプリ・サービス提供者。	業務やサービスで生成AIを活用したい企業・自治体・開発者。	AI導入から運用まで伴走支援が必要な企業や、構築リソースを持たない法人に最適。
提供開始日	2025年9月30日 (火)	2024年1月31日 (水)	2025年4月16日 (木)	2024年6月27日 (木)	2025年9月24日 (水)	2025年10月30日 (木)

学習 (開発) 向け

多様なニーズに答える高付加価値サービスラインアップ

推論 (利用) 向け

よくあるご質問・問い合わせ



よくあるご質問

Q1. 「NVIDIA H100」に搭載されているGPUは何ですか？

NVIDIA H100 SXM 80GB ×1 を搭載しております。

Q2. 高火力 VRTを利用するには何が必要ですか？

さくらインターネットの会員ID、「さくらのクラウド」プロジェクト、「さくらのクラウド」ユーザーの作成が必要です。詳しくは以下ページをご参照ください。

<https://cloud.sakura.ad.jp/flow/>

また、高火力 VRTとは別に、さくらのクラウドのディスクが起動領域として必要になります。

<https://cloud.sakura.ad.jp/products/disk/>

Q3. ストレージの容量を増やすことは可能でしょうか？

はい、さくらのクラウドのディスクプランを変更、追加することで容量を増やすことができます。

ただし、特定のプランに搭載される内蔵の一時領域についてはプランごとに固定となり、増やすことはできません。

Q4. 前払いで利用することは出来ますか？

通常料金プランでは原則前払いでの利用はできません。

ただし、法人のお客様については、通常のクレジットカード決済での後払い以外にプリペイド購入によるご利用料金の前払いを選択することができます。ご利用の際に事前の見積りによる予算確保が必要な場合などに便利にご利用いただけます。また、プリペイドは省庁・自治体・学校法人様向けに単年度提供もご致しますのでお問い合わせください。詳細はマニュアルをご確認ください。

また、期間コミットプランでは、該当期間分の料金をまとめてお支払いいただく形で、通常より割安にご利用いただけます。

Q5. 海外から利用は可能でしょうか？

利用可能ですが一部制限がございますので、まずはお問い合わせください。

※ その他詳しい情報は、[マニュアルページ](#)からご確認ください。



まずはお問い合わせください

✓ サービス導入をご検討中のお客様

サービスの新規導入に関するご質問や契約前のご相談を受け付けております。

URL : <https://www.sakura.ad.jp/inquiry/koukaryoku-vrt/>

お問い合わせはこちら



会社概要

商号

・さくらインターネット株式会社

所在地

- ・大阪本社：大阪府大阪市北区大深町6-38
グラングリーン大阪 北館 JAM BASE 3F
- ・東京支社：新宿区西新宿 7-20-1 住友不動産西新宿ビル 3 2F
- ・福岡オフィス：福岡県福岡市中央区大名2丁目6-11 Fukuoka Growth Next内
- ・SAKURA innobase Okinawa：沖縄県那覇市松山1丁目2番13号

設立年月日

・1999年8月17日（サービス開始1996年12月23日）

上場年月日

- ・2005年10月12日（東証マザーズ）
- ・2015年11月27日（東証一部）

資本金

・112億8,316万円

従業員数

・連結 997名（2025年3月末時点）

主要事業

・データセンター、クラウド、IoT関連事業



「やりたいこと」を「できる」に変える

私たちは「熱量」を持って挑戦する人々の夢を全力で応援します。

新しい社会のインフラを支えながら、
最先端のサービスを構築していく。

さくらインターネット株式会社 代表取締役社長

田中 邦裕





さくらインターネットは新たなアイデアの創出に強い熱意と情熱を持って挑戦するお客様をはじめ、
私たちとつながりのあるすべての人たちのために、未来のあるべき姿を思い描きながら

— 「やりたいこと」を「できる」に変える —

あらゆるアプローチを“インターネット”を通じて提供します。

- サービスサイト : <https://cloud.sakura.ad.jp/lp/vrt/>
- お問い合わせフォーム : <https://www.sakura.ad.jp/inquiry/koukaryoku-vrt/>
- オンライン相談予約 : https://www.jicoo.com/t/sakura_internet/e/document_mtg?utm_source=document&utm_medium=whitepaper&utm_campaign=online_meeting