

AIインフラ選びに迷わない！ 国産GPUクラウドと導入事例から学ぶ AI開発成功のヒント

会社紹介

会社紹介

データセンターを中心とした各種クラウドサービスをご提供
社会と顧客のDXを支援

- 1996 ○ さくらインターネット創業**
1996年12月に現社長の田中邦裕が、舞鶴高専在学中に学内ベンチャーとして創業
- 1999 ○ 株式会社を設立 / 最初のデータセンター開設**
1999年8月に株式会社を設立。10月には、第1号となるデータセンターを大阪市中央区に開設
- 2005 ○ 東証マザーズ上場**
2005年10月に東京証券取引所マザーズ市場に上場
- 2011 ○ 石狩データセンター開設**
2011年11月、北海道石狩市に国内最大級の郊外型大規模データセンターを開設
- 2015 ○ 東証一部に市場変更**
2015年11月に東京証券取引所市場第一部に市場変更
- 2021 ○ 創業25周年**
2021年12月、創業25周年
- 2022 ○ 東証プライム市場へ移行**
東京証券取引所 新市場区分のプライム市場へ移行

本社所在地	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
創業年月日	1996年12月23日
設立年月日	1999年8月17日
資本金	112億8,316万円
上場年月日	2005年10月12日（マザーズ） 2015年11月27日（東証一部） 2022年4月（東証プライム）
従業員数	997名（連結）



石狩データセンター

GPUが注目される背景と課題

急増するGPUの需要

1. AI技術の進歩

- 飛躍的に進化するAI技術
- 企業独自の AI モデル開発に取り組む動きが加速

2. 計算量の爆発的增加

- 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、高性能 GPU が前提となる時代に突入している

3. GPU供給の逼迫

- NVIDIAに代表される高機能GPUは供給が追いつかない状況
- 価格の高騰や入手困難な状況が現在でも続いている

AI技術でGPUが重要視される理由

1 優れた並列計算能力

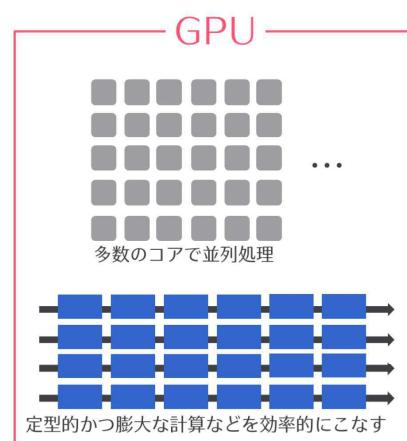
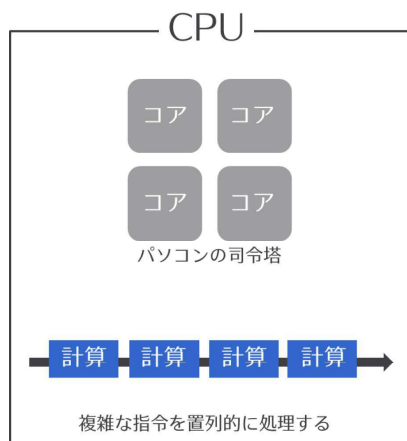
2 ディープラーニング計算との親和性

3 学習時間の短縮

AI技術でGPUが重要視される理由

- 1 優れた並列計算能力
- 2 ディープラーニング計算との親和性
- 3 学習時間の短縮

理由①優れた並列計算能力



AI技術でGPUが重要視される理由

- 1 優れた並列計算能力
- 2 ディープラーニング計算との親和性
- 3 学習時間の短縮

理由②ディープラーニング計算との親和性

ディープラーニングは計算量が非常に大きい



何度も同じ処理を繰り返す特徴がある

学習/推論どちらも積和演算が主体。

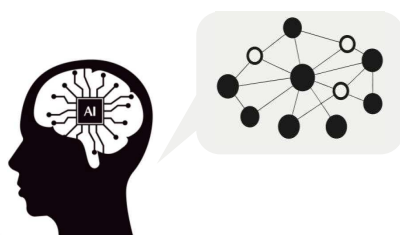


並列処理が得意なGPUを利用して大幅に高速化ができる

NVIDIAの「Tensorコア」



AI用途に特化した演算ユニット。
処理効率が大幅アップ



AI技術でGPUが重要視される理由

- 1 優れた並列計算能力
- 2 ディープラーニング計算との親和性
- 3 学習時間の短縮

理由③学習時間の短縮



圧倒的な計算性能で学習時間を大幅短縮
→開発全体のスピードを加速



仮説をすばやく検証、改良のサイクルを高速で回せる
→数倍、数十倍の試行錯誤が可能に

GPUは開発効率を押し上げる重要な戦略的リソース

GPU選定の基準

機械学習やディープラーニングの学習・開発を効率的に進めるために、
適したGPUを選ぶポイントは、以下のとおりです。

- コスト・価格
- コア数
- メモリ容量・帯域幅
- 冷却性能・冷却方式
- 消費電力
- グラフィックボードの動作環境

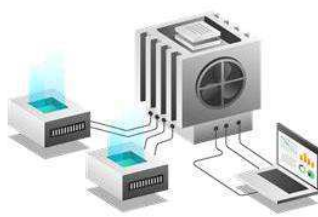
AI開発で直面する障壁

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



ファシリティ



人材 / 運用



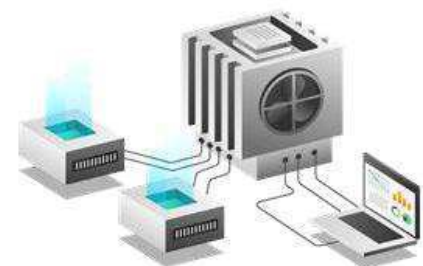
障壁① GPU調達コスト

- 生成AI・大規模LLMの普及により、**計算負荷が増大**
- 世界的なGPU**需要の高まりと供給不足**
- GPU技術の進化スピードが非常に早い



障壁② ファシリティ

- 高性能GPUの活用には、**電力・冷却に優れたデータセンター環境が必要**
- 自前構築では、**監視体制や緊急対応、専門人材の確保が必須**
- インフラ構築に時間と負担がかかる



障壁③ 人材 / 運用

- MLOpsやインフラエンジニアなどの**専門人材**は依然として不足
- **24時間体制での監視・障害対応**には高い負荷と工数がかかる
- インフラ運用をクラウドに委ねることで技術負担を軽減



さくらのGPUクラウド 「高火力シリーズ」のご紹介

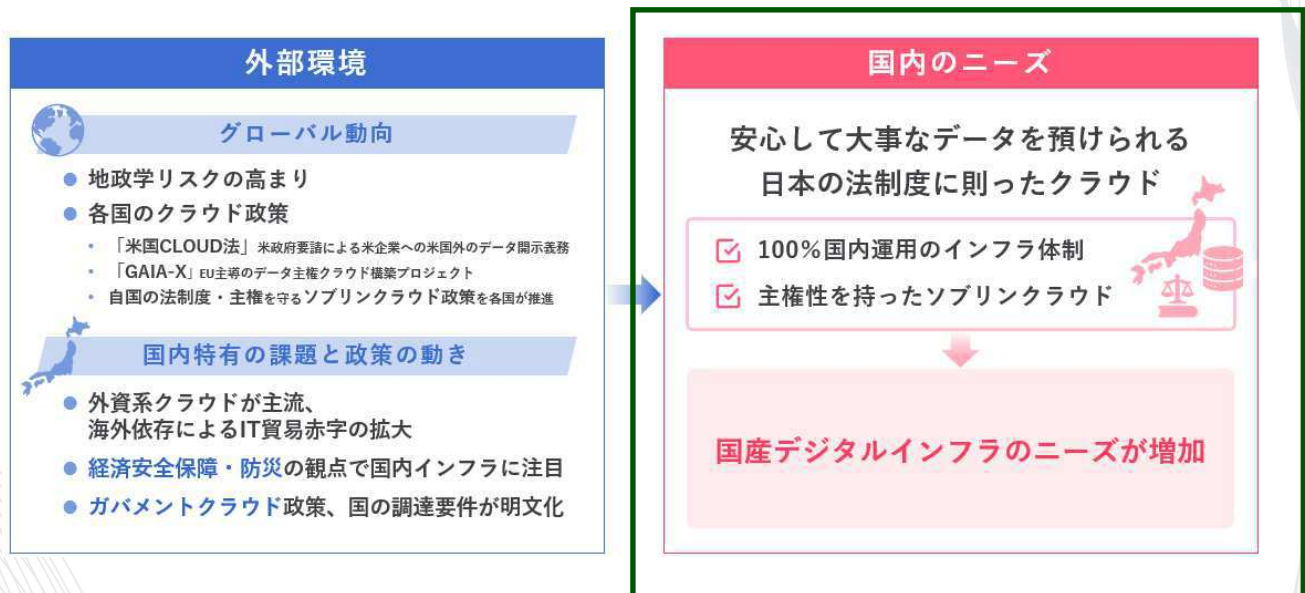
高火力シリーズとは

さくらインターネットが提供しているGPUクラウド
「高火力シリーズ」なら課題を一挙解決

- 1 最新のGPUを、用途に応じて柔軟に選択可能
- 2 初期投資ゼロ・導入もスピーディー
- 3 運用・監視の負担ゼロ

GPUクラウド事業と背景

主権性をもった「ソブリンクラウド」

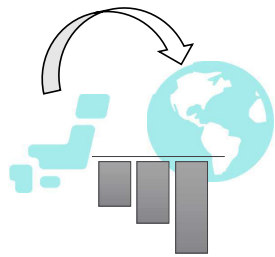


GPUクラウド事業と背景

◆ 経産省の安全保障政策認定に伴い、クラウドプログラムの安定供給の確保

- ✓ 生成AIが急速に発展し世界を席捲。国外サービス利用による国富の流出、貿易赤字の拡大を懸念
- ✓ AIプログラムの安定供給のため、経産省がAI基盤となるクラウドインフラ事業者へ助成

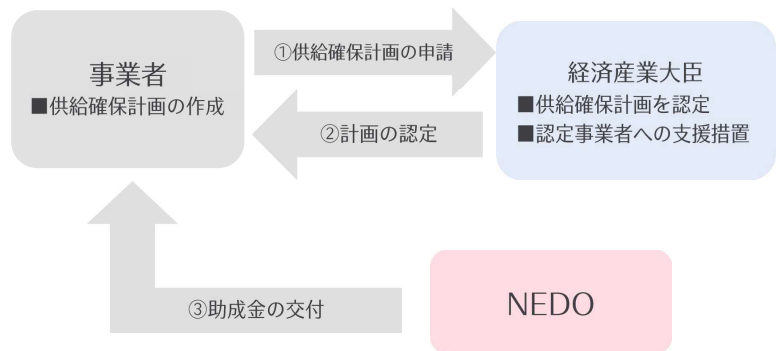
国外サービス利用による
デジタル貿易赤字の拡大を懸念



2022年 **4.7兆円**のデジタル赤字 (※)

※日本経済新聞23年2月「[「デジタル赤字」4.7兆円](#)」

経産省 (NEDO) AI基盤となる
クラウドインフラ事業者へ助成



GPUクラウド事業と背景

弊社が経産省から認定を受けたポイント

GPUクラウドサービスの提供ナレッジ

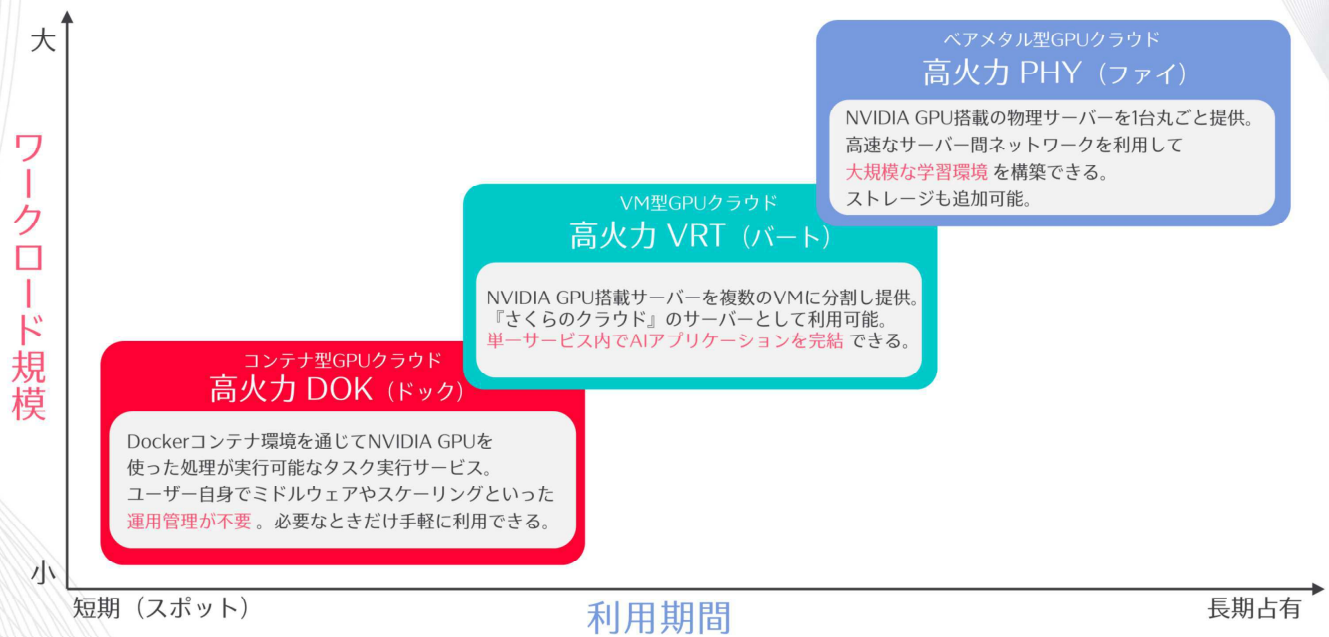
- ✓ 2016年から機械学習を行うための、NVIDIA GPU を使ったGPUクラウドサービスを提供、技術的および事業的な知見あり

GPU運用に最適な環境

- ✓ 石狩データセンターの拡張性、電力キャパシティを評価
- ✓ 再生可能エネルギー100%・CO2排出ゼロ・外気を利用した空調機構



高火力シリーズとは



高火力PHYの概要

AI開発に欠かせない圧倒的な性能をもつ
NVIDIAのGPU を8基搭載したベアメタルサーバー

- 1 NVIDIAの最新GPU「H100/H200/B200」が使える
- 2 大規模・長期間の専有利用で開発リソースを確保
- 3 データセンタースタッフが、24時間体制の監視・管理
- 4 初期費用なし！月額費用のみで利用可能



高火力PHYのユースケース

複雑なデータセットの解析や大規模モデルのトレーニングなど、
高性能なGPUが求められる多様なユースケースに活用可能

大規模言語モデル(LLM)



データ分析やリスク予測などの
LLMのトレーニング

生成AI



画像、文章、音声など、
多様なコンテンツの生成

先進運転支援システムや 自動運転AIの開発



大量の走行データの学習、
大規模シミュレーション

科学シミュレーション



複雑な科学的計算や
シミュレーション

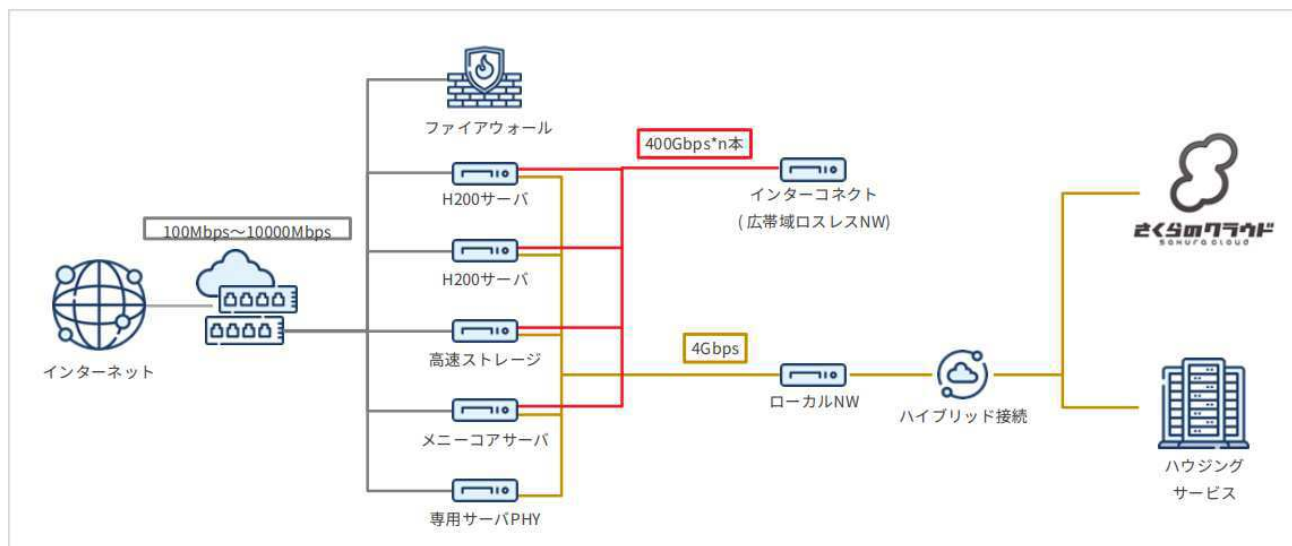
画像解析



リアルタイム画像処理などの
高度な画像解析

高火力PHY(構成例)

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



高火力VRTの概要

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**

1 ハイパフォーマンスなGPUをVM（仮想マシン）で提供

2 柔軟なシステム構築が可能

3 初期費用不要で 990円 から利用可能



高火力VRT(ユースケース)

チャットボットサービス(自然言語処理)



ユーザーが入力したテキストを即座に解析し、適切な応答を返す必要がある場面で活躍します。

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。

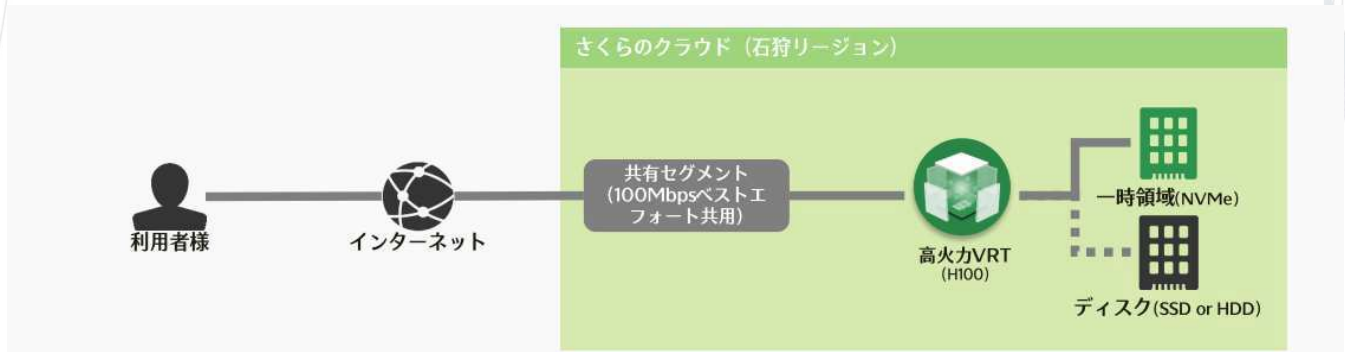
対話型 AI アバター(音声認識・音声合成・画像生成)



周囲の状況やユーザーの発言など、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使うことで対応できます。

高火力VRT(構成例)

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。

高火力DOKの概要

NVIDIAのGPUを使用して、お客さまが事前に用意した
「Dockerイメージ」を、毎回同じ環境で実行ができる
コンテナ型GPUクラウドサービス

- 1 NVIDIA H100をはじめとしたGPUを低価格で提供
- 2 初期費用なし！従量課金でコスト最適化を実現
- 3 面倒なインフラ管理不要で運用負荷軽減



高火力DOK(ユースケース)

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。

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。

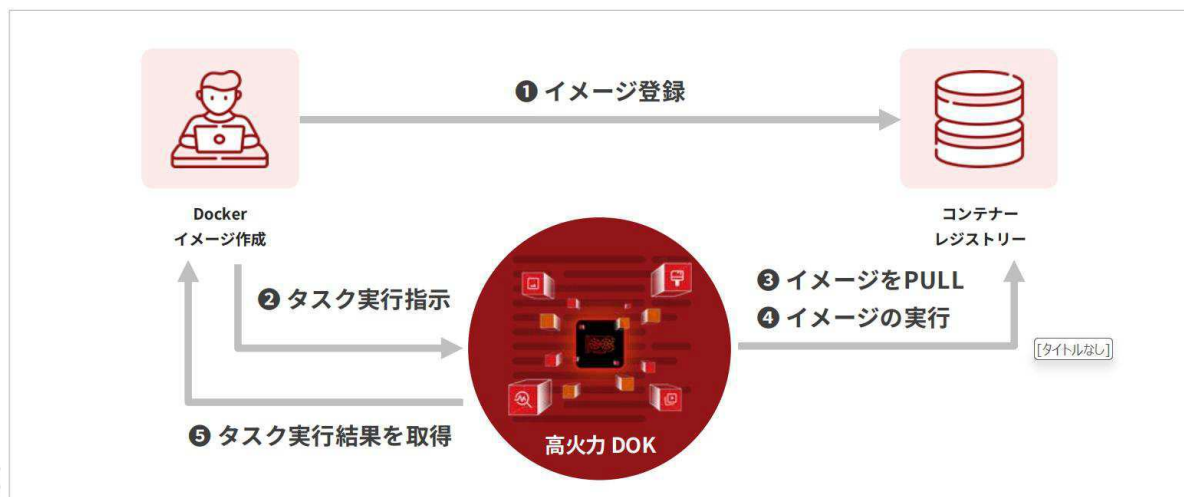
AIの学習・推論



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。

高火力DOK(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。



導入事例

(C) SAKURA internet Inc.

34

事例①



お 客 様 : 株式会社Preferred Networks 様

事 業 内 容 : WEBサービス&アプリ提供者

用 途 : 大規模言語モデル



詳しく見る

導入前

- ✓早急に高性能GPU搭載の機械学習クラスタが必要
- ✓クラウドベンダー任せにした場合の不透明感
- ✓高価な高性能GPUを安価に利用したい



導入後

- ✓「高火力PHY」を国内最速クラスでリリース
- ✓透明性が高く、深い部分の技術要件まで協議が可能
- ✓アグレッシブな価格で利用できる



(C) SAKURA internet Inc.

35

事例②



お客様：ダイキン工業株式会社様
事業内容：空調・冷凍機、化学、油機、特機、電子システム
用途：遠隔支援システム、AI解析モデル



詳しく見る

導入前

- ✓ AI開発において十分なGPUを確保できなかった
- ✓ 高性能なGPUを安価で利用したい
- ✓ 物理サーバーの導入に対し、稼働面で少し不安があった



導入後

- ✓ 高火力PHYの導入で高性能なGPUを十分に活用できるようになった
- ✓ 圧倒的なコストメリットを実感
- ✓ 高火力PHYにより安定稼働を実現



事例③

TIER IV

お客様：株式会社ティアフォー様
事業内容：自動運転関連ソフトウェアの開発・サービス提供
用途：自動運転技術の開発



詳しく見る

導入前

- ✓ 社内には高性能GPUを稼働させる十分な設備が整っていない
- ✓ 実験1回当たりの学習時間をできるだけ短縮しなくてはならない



導入後

- ✓ 高性能GPUを高い経済合理性を兼ね備え、即時に利用可能に
- ✓ 従来2～3週間を要していたすべての実験が、1週間以内で終わるように



そのほか導入企業様一覧（抜粋）

