AI の発展を支える GPU

人工知能（AI）の発展により GPU の需要が拡大している。GPU はもともとグラフィック処理の並列計算用途に使われるプロセッサであったが、近年ではディープラーニング（深層学習)の並列計算にも利用されるようになり、AI の急速な発展を支える技術として不可欠な存在となった。

一例として、製造業では製造ラインにおける工業製品の外観検査で AI の活用が進んでいる。あらかじめ製品の良品・不良品に関する画像データを AI に学習させ、製造ラインに流れる製品をカメラ撮影すると、AI が良品・不良品を自動判定する。大量の画像をディープラーニングで学習させる過程では膨大な量の計算が必要とされ、CPU では処理に数日かかる場合があるが、GPU ならその計算力により数時間で処理が完了する。

その他の AI 活用事例として、自動運転車における物体や人物認識、建設業ではドローン空撮画像を使った地形測量、医療分野では画像を AI が分析する診断支援などがある。また、コールセンターでの電話自動応答に加え、音声データのテキスト化、感情分析、キーワードの自動抽出など、自然言語処理の分野でも GPU が活用されている。学術研究機関では、流体解析や分子動力学などのシミュレーションの用途が挙げられる。

GPU オンプレミスの課題

予算の兼ね合いで AI 開発にゲーミング用の GPU マシンが使われる現場が存在する。ゲーミング用 GPU では計算パワーが足りずに、計算処理が遅延することがある。AI 学習やシミュレーションでは、計算を繰り返すことでモデルの精度を高めていくことが一般的であり、計算処理が遅延した場合、納期に間に合わず成果物の質が落ちることや、最悪の場合プロジェクトを断念する可能性がある。一方で、高性能な GPU マシンをオンプレミスで所有する場合、数百万円の初期コストが必要であり、高額な初期投資が高い壁となる。

柔軟性に欠けることもオンプレミスの課題のひとつだ。繁忙期の GPU リソースの不足し、逆に閑散期に GPU リソースが余剰になる問題がある。リソース不足を防ぐために大規模な GPU 環境を構築するには、新しいサーバールームの確保や電源・空調の増設など、ファシリティ面まで十分な検討をする必要がある。また NVIDIA は約 2 年ごとに GPU のパフォーマンスが大幅に向上していることから、GPU は陳腐化のスピードが早いことが分かる。固定資産の兼ね合いで、型落ちした GPU を 5 年間使い続けることはユーザーにとっては大きなマイナス要因になりうる。

古くなった GPU がソフトウェアのアップデートに対応できない課題も発生する。GPU はハードウェアとソフトウェアのバージョンの組み合わせが非常にデリケートであり、プログラムは動いているが GPU が動かない、パフォーマンスが十分に出ないということがたびたび発生する。一方で、GPU 上で動くソフトウェアは日進月歩でアップデートされている状況で、例えば並列計算ソフトウェアの CUDA やディープラーニング系のライブラリ TensorFlow や PyTorch などがそれである。年月が経つにつれて購入した GPU は古くなり、新しいバージョンのソフトウェアに対応できなくなる限界をどこかで迎えることになる。

GPU クラウドサービスの台頭

オンプレミスが抱える課題を解決する手段として、GPU クラウドサービスが注目を浴びている。AWS や Azure などのメガクラウドが GPU マシンをクラウドサービスとして提供している。初期コストがかからず、使った分だけ料金を支払う従量課金制が採用され、繁忙期はリソースを増やし、閑散期にはリソースを減らすようなコストの最適化を図ることができる。リソースを増やすにあたって、固定資産やファシリティ面の考慮も一切不要である。

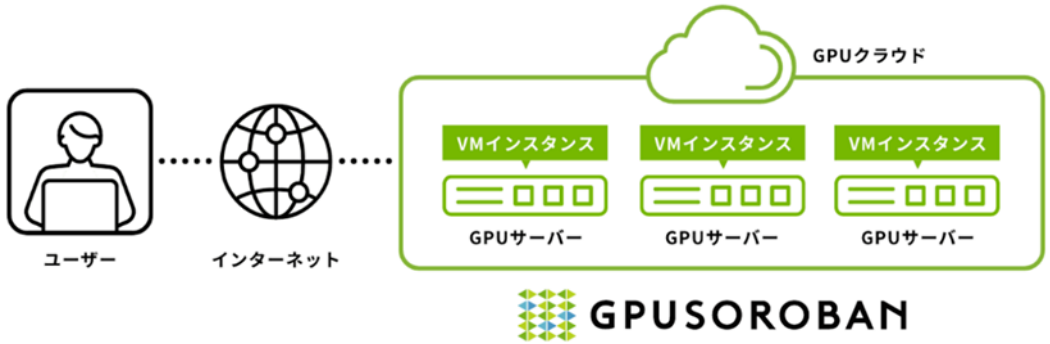
クラウドサービスは GPU が複数ラインアップされ、利用者の要件に応じたスペックを自由に選択できる。新しい GPU もラインアップしているため、前述した GPU の陳腐化やソフトウェアのバージョンアップ問題にも対応する。

ただしメガクラウドの GPU サービスにも課題がある。まずはランニングコストが高い。そのうえ従量課金でかつ料金体系が複雑だ。GPU 以外にもネットワークやストレージのコストなどの見えにくい料金が別途発生するため、事前にコストを試算することが難しい。AI 学習やシミュレーションにおいては、試行錯誤をして計算を繰り返す性質があるため、所要時間が読みづらく、使っているうちにコストが高騰して予算を超過する恐れがある。特に GPU はプロジェクトを頓挫させるほどコストが高騰するリスクがあり、可能な限りコストを抑える策としてハイブリッド構成やマルチクラウドを選択する企業も増えてきている。

メガクラウドはリージョンによって GPU のラインアップやコストの差が大きいことも注意すべき点だ。日本リージョンでは使えない GPU が多数あり、米国と比べて割高である。日本にいながら米国リージョンの GPU を使うこともできるが、ネットワークレイテンシが問題になるケースやセキュリティを懸念する声がある。

国内で使いやすい GPU クラウドサービス「GPUSOROBAN」

メガクラウドがもつ課題を解決するのが、ハイレゾが展開している「GPUSOROBAN」だ。GPUSOROBAN は「高性能 GPU マシンを低価格で提供し、より多くの人に利用してもらう」ことをコンセプトにした GPU クラウドサービスである。



GPUSOROBAN は、メガクラウドが提供する GPU マシンに比べて利用料が安価で、ランニングコストを抑えることができる。例えば NVIDIA A100 などのハイエンド GPU はメガクラウドの 40%～50%の価格に設定されている。

　加えて、GPUSOROBAN は従量課金の他に月額固定料金も選択可能だ。メガクラウドの従量課金がもたらす「コストが高騰するリスク」に対応している。ネットワークとストレージ料金も含めた月額固定料金で追加コストが発生しないため、ユーザーは予算計画を立てやすく安心して利用することができる。

GPUSOROBAN が低価格や月額固定料金を実現できる理由は、データセンターの運営コストを抑えているためだ。データセンターは年間を通じて冷涼な石川県志賀町に建設され、外気を最大限利用するエア

フロー設計を施している。エアコンを一切使用しないため、電力消費量（電気代）を最小限に抑えられる。
　国内のデータセンターからクラウドサービスを提供しているため、海外リージョンと比べてレイテンシやセキュリティ面で安心して利用できる。また国内の GPU エンジニアによる技術サポートを無償で受けられる点もメリットとして大きい。GPU に関する環境構築は膨大な時間を要することがあるが、GPUSOROBAN なら専門エンジニアによるサポートにより環境構築の時間を短縮し、ユーザーは本業の開発にとりかかることができる。

　これまでハイレゾは NVIDIA と密接なパートナーシップを組み、NVIDIA の GPU を中心にクラウドサービスを国内展開してきた。その実績として、2022 年に NVIDIA パートナープログラムの最上位レベル「Elite Partner」に日本で初めて認定された。また、同年 6 月には NVIDIA が展開する「NPN Partner Award」において、最も変革的なインパクトを与えたパートナーに贈られる「Best CSP Partner of the Year」を受賞した。

　GPUSOROBAN は国内で年間 100 件を超える利用実績があり、学術研究機関や大手企業からベンチャー企業まで幅広く利用されている。業界は製造業、建設業、通信業など多種多様である。用途は様々で画像認識・音声認識などの AI 学習や、3 次元画像処理、3DCG レンダリング、分子動力学シミュレーションなどが挙げられる。昨今の AI・データ分析の需要拡大に伴い、GPU を用いた計算ニーズも各業界で高まっている傾向が見られる。

年間100件を越える利用実績

京都大学 様	早稲田大学 様	筑波大学 様
神戸大学 様	立命館大学 様	富山大学 様
エネルギア・コミュニケーションズ 様	株式会社スカイマティクス 様	株式会社ブレインズ 様

ハイレゾは、計算処理が IT 活用のインフラの一つとなる未来を見据え、海外ベンダーの GPU サービスではなく国産で安定したインフラを供給し、日本の AI ディープラーニングの社会実装を加速させることを目指している。今後も国内の企業や研究機関に寄り添ったサービスを展開していく。

株式会社ハイレゾ
〒107-0051　東京都港区元赤坂 1-1-8　赤坂コミュニティビル 5 階
TEL：03-6864-6222　　E-mail：gpu_marke@highreso.co.jp
URL：https://highreso.jp

GPUSOROBAN