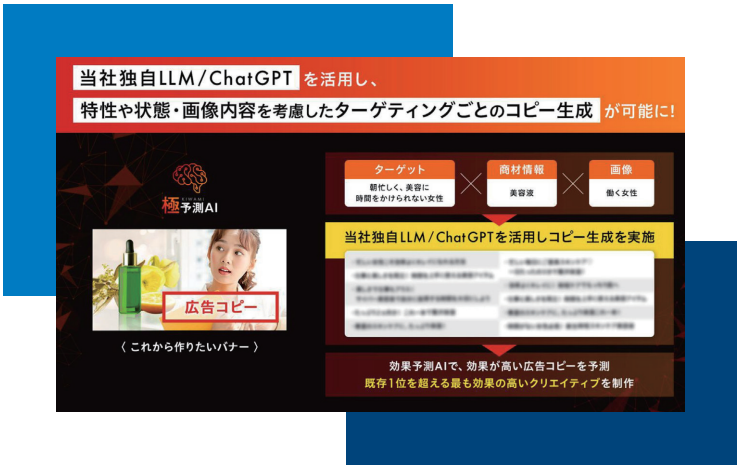
最新GPU搭載のPowerEdge XE9680サーバーで、和製生成AI／日本語LLMの開発や効果の高い広告制作に貢献

お客様 プロフィール

CyberAgent

サービス業 | 日本



生成AI用途にも最適な
NVIDIA® H100 GPUの効果を活かせる
PowerEdge XE9680サーバーが
タイムリーにリリースすることがわかったため、
いち早く導入することにしました。
管理ツールのiDRACが使いやすいことも
PowerEdge XE9680を
採用した理由の1つです

株式会社サイバーエージェント
グループIT推進本部CIU
Solution Architect
高橋 大輔 氏

ビジネス課題

2016年からAIの研究・開発を積極的に行い、広告事業に取り入れている株式会社サイバーエージェントは、新たなGPU基盤として最新のNVIDIA® H100 GPUに注目していた。

導入効果

- NVIDIA® H100 GPUを8基搭載したPowerEdge XE9680をいち早く導入し、以前導入したNVIDIA® A100 GPUを4基搭載のPowerEdge XE8545の約5.14倍の性能向上を実現。
- さらに今後、LLM（大規模言語モデル）含めた特定の計算アルゴリズムを高速化するTransformer Engineへの最適化によって十数倍の性能向上が期待できる
- 最新のデータセットに合わせてAIモデルの高速ファインチューニングが可能に
- 8Uが主流の中で、6Uの筐体で効率的な冷却により省スペース化に貢献

ソリューション

- PowerEdge XE9680サーバー
- ProSupport

約5.14倍～十数倍

NVIDIA® H100 GPUを8基搭載したPowerEdge XE9680は、NVIDIA® A100 GPUを4基搭載したPowerEdge XE8545と比較して、約5.14倍、AIの学習性能が向上し、Transformer Engineへの最適化によって十数倍の性能が期待できる

日本語の大規模言語モデル(LLM)をオープンソースとして公開

国内トップシェアを誇るインターネット広告事業や新しい未来のテレビ「ABEMA」を展開するなど、グループ全体でさまざまな事業を行っている株式会社サイバーエージェント(以下、サイバーエージェント)では、2016年にAI研究組織「AI Lab」を設立し、それ以降、広告クリエイティブの制作をAIで支援する取り組みを行い、積極的にAIの研究開発を行ってきた。「2020年にはAIを活用して広告クリエイティブを制作できる極予測AIを提供し、より広告効果の高いバナー広告のキャッチコピーや画像の組み合わせなどを効率的に制作できるようにしている。

また、サイバーエージェントは、130億パラメータからなる、日本語に特化した独自の日本語大規模言語モデル(LLM: Large Language Model)を開発。2023年5月には、和製生成AI開発基盤用に、最大68億パラメータの日本語LLMを、OpenCALM(オープンカーム)という名称で商用利用可能なオープンソースのLLMとして公開。多くのLLMが英語を中心に学習されている中、日本語に強いLLMとして大きな注目を集めている。CALMは、CyberAgent Language Modelsの略で、たとえばChatGPTはチャットができるようにチューニングされているが、OpenCALMは汎用の日本語モデルとなっており、使う人が用途に合わせてファインチューニングすることが可能だ。サイバーエージェントとしては、クローズドに日本語LLMを開発するよりもオープンにしたほうが多くの人のフィードバックをもらえ、他社との協業や国内のAI技術発展への貢献ができるというメリットがあると考えて、OpenCALMを公開している。

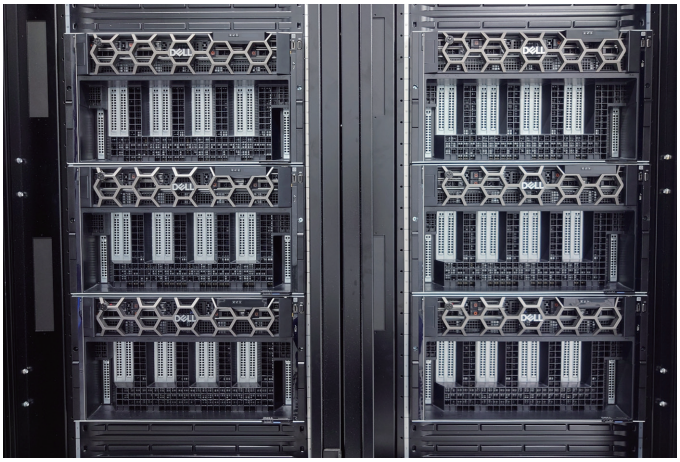
さらに、130億パラメータの日本語LLMはすでに、極予測AI、極予測TD、極予測LPなどのAIを活用した広告クリエイティブ制作領域のサービスにおいて活用を始めているという。130億パラメータの日本語LLMはさまざまな場面で使える汎用のAIモデルとして作られているが、各広告媒体のユーザー層に適したキャッチコピーを作れるようにファインチューニングされ極予測AIで使われ、極予測AI専用のAIモデルである「効果予測AI」を開発して極予測AIの広告効果を予測するために使っている。将来的にサイバーエージェントは、日本語LLMだけでなく、画像なども扱えるマルチモーダルAIの開発を目指している。

NVIDIA® H100 搭載のGPUサーバーを求めてPowerEdge XE9680を採用

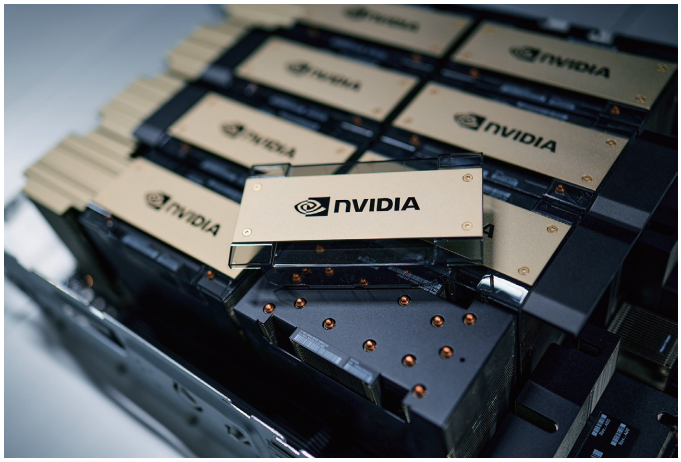
サイバーエージェントがAI研究組織「AI Lab」を設立した2016年の段階では、研究者が1人1台のワークステーションを使って研究・開発を行っていた。しかし、2020年のコロナ禍でオフィスにあるGPUワークステーションのメンテナンスを行いづらい状況となり、当時最新のNVIDIA® A100 GPUが出てきたことで、データセンターにGPUサーバーを置いて機械学習基盤として運用していくことを考え始めた。

「GPUを使うだけならパブリッククラウドという選択肢もありますが、パブリッククラウドでは最新のGPUがいつ提供されるようになるかがわかりません。また、GPUを使いたいときに在庫を確保できるという保証もないため、オンプレミスにGPUリソースを確保した上で、パブリッククラウドと使いやすいほうをユーザーに使ってもらえるようにしました。パブリッククラウドとプライベートクラウドをユーザーが行き来しやすいように、できるだけパブリッククラウドの仕様に近づけてユーザーインターフェイスなども工夫しています」とグループIT推進本部CIU, Solution Architectの高橋大輔氏は説明する。機械学習基盤として、NVIDIA® A100 GPUを搭載したデル・テクノロジーズのPowerEdge XE8545サーバーも採用している。

NVIDIAの最新のGPUであるNVIDIA® H100 GPUにもリリース前から注目していたという高橋氏は、「単純に計算性能が向上するだけでなく、特定の計算アルゴリズムを高速化するTransformer Engineなどの仕組みも魅力だと考えていました」と話す。NVIDIAによれば、Transformer Engineによって、前世代のNVIDIA® A100 GPUと比較して、大規模言語モデルのAI学習を最大9倍、AI推論を最大30倍高速化することが可能だという。「最新のGPUを求める社内ユーザーは一定数いるため、NVIDIA® H100を搭載したGPUサーバーについても調べていきました。一方で、コストパ



データセンター内の6台のPowerEdge XE9680サーバー
［写真提供：株式会社サイバーエージェント］



8基のNVIDIA® H100 GPU を搭載可能なPowerEdge XE9680サーバー
［写真提供：株式会社サイバーエージェント］

フォーマンスも含めて、最新のGPUでなくてもよいという需要もあるため、ユーザーの求めるGPUインフラを用意していくことが我々の使命だと考えています」と話す。

最終的にサイバーエージェントでは、8基のNVIDIA® H100 GPUを搭載したPowerEdge XE9680を採用しているが、その理由を高橋氏は次のように話してくれた。「サーバーメーカー各社の話を聞いている中で、生成AI用途にも最適なNVIDIA® H100 GPUの効果を活かせるPowerEdge XE9680サーバーが、タイムリーにリリースすることがわかったため、いち早く導入することにしました。NVIDIA® H100 GPUが発表された後、同GPU搭載のPowerEdge XE9680がどのような構成になるかなどの情報をデル・テクノロジーズと密にやり取りできたことも良かったと思います。なるべく少ない台数で稼働率を高めたいと考えていたため、デル・テクノロジーズが4時間オンサイトなどの高い保守レベルをリーズナブルな価格で実現してくれたのはうれしいポイントでした。過去導入したPowerEdge XE8545などを安定して保守してくれていたことや、管理ツールのiDRACが使いやすいこともPowerEdge XE9680を採用した理由の1つです」。

2023年3月に発注して1か月強先のゴールデンウイーク明けには納品されたことも、高橋氏は高く評価している。「コロナ禍でサプライチェーンが混乱している中で、デル・テクノロジーズは比較的安定したサプライチェーンを持っていることも安心でき、短納期で提供してくれるのはうれしいですね」。また、納品後の構築では、さまざまな工夫を行ったと高橋氏は振り返る。「パラメータ数が多い大規模言語モデルでは、複数のGPUを使う必要があるため、各サーバーに400Gbpsのネットワークカード（NIC）を8枚挿し、RDMA（Remote Direct Memory Access）の技術を使ってサーバー間を高速につなぐインターコネクトを構築し、大規模な深層学習のモデルを学習する際のボトルネックを減らすようにしています。GPUサーバーは発熱量が多いため、効率的に冷却できる設計であることが重要ですが、冷却のために8Uのサーバーが主流となっている中で、PowerEdge XE9680は6Uのサイズでしっかりと冷却できる設計になっているのも評価できます。それに加えて、データセンターの部屋全体を冷やすのではなく、ラック背面に水冷式のリアドア空調機を取り付けることで効果的な冷却を行えるよう、データセンターもリアドア空調機が使える新たな場所に移転しました」。

Transformer Engineへの最適化でキャッチコピーの大幅な精度向上に期待

PowerEdge XE9680を導入することによって、同社はさまざまなメリットを実感しつつある。「パフォーマンスが大幅に向上したことで、日本語LLMの更新をより早く頻繁に行えるようになると期待しています。日本語LLMの進化速度も向上していくでしょう。また、NVIDIA® A100 GPU搭載のPowerEdge XE8545に比べて約5.14倍の性能向上が実現できていて、今後Transformer Engineへの最適化を行うことで十数倍の性能向上ができることを期待しています。最新のデータセットに合わせた機械学習モデルのファインチューニングなどもかなり高速に行えるようになっているので、サービスを進化させるという要望に応えやすくなっていますし、キャッチコピーの候補の精度を上げることができるようになると思います」と高橋氏は語る。

また、PowerEdge XE9680をベースとした機械学習基盤は、ユーザーにも高評価を得ていると高橋氏は続ける。「パブリッククラウドでGPUを確保できなかったり、長期の利用で課金額が大きくなってしまうこともありましたが、社内のリサーチャーからは、より大量のリソースを確保できて課金額を気にせずに安心して利用できるという話が出ています。ユーザーがビジネスインパクトを出せるように、インターコネクトも含めてハイスペックなインフラを提供することができたこともメリットであると言えます」。

以前から使ってきたデル・テクノロジーズの管理ツールiDRACによって、管理の負荷が低減されていることも高橋氏は評価している。「データセンターに我々は常駐していないので、リモートでさまざまなことができるiDRACは便利ですね。OSに入らなくてもGPUの温度や状態を確認することができたり、ファームウェアの更新などもできます」。

今後もGPUの最新情報に注目してインフラを提供していく

今後、サイバーエージェントでは、2023年5月に公開したOpenCALMに対して集まってきたさまざまなフィードバックを改善に活かし、社内向けの大規模言語モデルにも反映させていく方針だ。また、OpenCALMを通じて、広告以外の業種の企業や団体との協業も模索しており、小売りなどリテール業界や金融業などと連携し、それぞれの固有データを学習し、その業界で使えるような「業界特化型のLLM」を構築するような議論が始まっているという。

その上で今後もGPUやその市場には注目し、新たな技術とそれがどのように製品化されていくかを見ていきたいと高橋氏は説明する。「現在のところ、GPUの選択肢はNVIDIAが非常に有力となっていますが、積極的にGPU市場に参入してきている他のGPUベンダーの動きにも注目しています。NVIDIAが実現しているようなソフトウェアのエコシステムをどれだけ他のベンダーが作っていけるかも楽しみですね。また、GPUの性能を出すためにPCIeバスがボトルネックとなる場合もあるので、CPUとGPUをつなぐNVLink-C2Cや新たな規格のCXL（Compute eXpress Link）などが製品に実装される動きにも興味があります。CPU、GPU、メモリを筐体内でどのように接続するかで性能が変わっていくので、デル・テクノロジーズには、新たな技術をスピード感を持って取り入れ、高い性能が出る設計を今後も行ってくれることを期待しています」。

最新のGPUからコストパフォーマンスの高いGPUまで、ユーザーの求める機械学習基盤を常に提供することでサイバーエージェントのAIの研究・開発は進化し続けていく。また、日本語LLMのさらなる開発などで、サイバーエージェントは自社の広告事業のみならず、日本のAI市場で大きく注目され続けていくのは間違いない。

株式会社サイバーエージェント
グループIT推進本部CIU
Solution Architect
高橋 大輔 氏