生成AIと半導体開発競争

生成AIを巡って大手テック企業が競争を繰り広げているのは、大規模言語モデルや対話型AIといったソフトウェアだけではない。言語モデルの学習や推論処理などに使用される「AIチップ」の開発を巡っても競争が始まっている。

巨大テック企業の次なる競争の舞台は生成AI

2022年11月末にOpenAIがChatGPTを公開して以来、大規模言語モデルとそれを活用した対話型AIが大きな注目を集めている。

主要プレイヤーとしてはOpenAIのほか、OpenAIに多額の出資をして提携しているマイクロソフト、ChatGPTに対抗して対話型AI「Bard」を発表したグーグルの動きが目を引く。しかし、独自の大規模言語モデル「Amazon Titan」と、API（アプリケーション・プログラミング・インターフェース）経由で他社の大規模言語モデルを利用可能にするクラウドサービス「Amazon Bedrock」を発表したアマゾンや、オープンソースとして商用利用可能な大規模言語モデル「LLaMA 2」を公開したメタ、さらには詳細は不明であるが、「Apple GPT」と呼ばれる対話型AIの開発を進めていると報じられたアップルなど、巨大テック企業はいずれも生成AI領域に注力し始めており、次なる競争の舞台は生成AIだといってよいだろう。

同時進行するAIチップの開発競争

こうしたソフトウェア領域での競争の水面下では、「AIチップ」と呼ばれる半導体の開発を巡っても競争が繰り広げられている。AIチップとは大規模言語モデルの学習や推論処理に使用される半導体チップのことを指す。

大規模言語モデルでは数千億のパラメータを使用し、次世代モデルでは数兆に及ぶ。モデルの学習には膨大な計算能力を必要とするため、高性能なAIチップの使用が不可欠である。たとえば、ChatGPTの商用化には非常に高度な並列処理能力を持つ1個数万ドルもするNVIDIAのハイエンドGPU「A100」が3万個以上も必要になったと推測されている。旺盛な需要に生産が追いついておらず、A100の後継製品である「H100」はインターネットオークションサイトのeBayでは4万ドル超で販売されるほど、価格が高騰している。

現在、AIチップの供給に関してはNVIDIAが市場シェア9割近くを占める圧倒的なリーダーとなっており、AIチップを必要とする企業は同社に依存している状態である。生成AIブームが強烈な追い風となり、2023年5月30日には、同社の時価総額が初めて1兆ドルを突破したことでも話題となった。

しかし、今後しばらくは生成AIが競争の舞台になる以上、いつまでもNVIDIAに依存するわけにはいかず、巨大テック企業はAIチップの内製化を始めている。

たとえば、マイクロソフトは2019年からコードネーム「Athena（アテナ）」と呼ばれるAIチップの開発を進めており、2024年早々にもマイクロソフト社内、及びパートナーシップを結んでいるOpenAIが使用する可能性があるということだ。

ソフトウェア業界の巨人であるマイクロソフトが、自社でAIチップを開発する目的はコスト削減である。同社はオープンAIの大規模言語モデル「GPT-4」を活用した対話型AI「BingAI」以外にもオフィススイート「Microsoft 365」やOS「Windows 11」、さらにはCRM/ERPソフトの「Dynamics 365」などにも大規模言語モデルを組み込み始めている。そのため、AIチップをNVIDIAから購入せずに自給自足できれば、相当な

コスト削減が期待できるというわけだ。

ハードとソフトの統合が競争優位に直結

もっとも、マイクロソフト以外の大手テック企業は以前からAIチップの独自開発を進めている。グーグルは2013年から機械学習に特化した自社開発のAIチップ「TPU（Tensor Processing Unit）」の開発を開始し、2015年に社内運用を開始している。

TPUは「TensorFlow」というグーグルが開発したオープンソースの機械学習フレームワークに最適化されており、従来のCPUやGPUと比較し、高速処理の実現を目的として開発された。当初は機械学習の推論フェーズに特化していたが、2017年に導入を開始した第二世代の「TPU v2」は、推論に加えて学習でも利用できるようになった。そして、2018年に発表された第3世代の「TPU v3」は、v2の倍の演算性能を持つようになり、学習でも高速処理が可能となった。さらに2021年に発表された第4世代の「TPU v4」は、v3に比べて学習時の浮動小数点演算能力は約1.7倍、推論時の整数演算能力は約27倍と大幅に向上している。TPUは、2017年から一般提供も開始され、Google Cloud Platform上で利用できるようになっている。

一方、アマゾン（AWS）は2018年に機械学習の推論処理に最適化した独自開発のAIチップ「Inferentia」を発表、2019年にクラウドサービス「Amazon EC2 Inf1インスタンス」として一般提供を開始している。さらに2020年には第2世代のAIチップとして、学習処理に最適化した「Trainium」を発表、2022年10月からは「Amazon EC2 Trn1インスタンス」として一般提供を開始している。同社はNVIDIAのGPUを搭載したインスタンスを提供する一方、より高い性能、より高いコストパフォーマンスを持つ幅広い選択肢をユーザーに提供するために独自設計のカスタムチップを開発したと説明している。

メタは2023年5月、2025年を目標に学習と推論の両方で使用可能な「MTIA（Meta Training and Inference Accelerator）」と呼ばれる自社製AIチップの開発に着手していることを発表した。実は、メタは2022年半ばまでAIの学習にCPUを、推論には「NNPI（Neural-Network Processor for Inference）」と呼ばれる独自開発のチップを使用してきたが、GPUの方が効率が良いと分かったため、それ以降は学習と推論の両方でNVIDIAのGPUを使用している。

それなのになぜ？という疑問が湧くが、これにはAIのモデルサイズが指数関数的に大きくなっており、GPUでさえも推論の処理効率が上がらなくなってきたことが背景にある。

このように、巨大テック企業は次の主戦場となる生成AIを巡り、半導体の開発でもしのぎを削っている。AIチップの設計と製造を自社で行い、コストを下げつつ、ハードウェアとソフトウェアを統合することで、一般的なGPUよりも高速かつ効率的にAIワークロードを処理できる。これはここまでしなければ、熾烈な競争を勝ち抜くことはできないことを意味しているといえよう。

Writer's Profile

城田 真琴 Makoto Shirota
デジタル社会研究室
プリンシパル・アナリスト
専門は先端技術、先端ITビジネスの調査研究
focus@nri.co.jp