インパッケージ光インターコネクトで新たな生成AIアーキテクチュアが実現

ウラジミール・ストヤノヴィッチ

情報スーパーハイウェイとも言える「インパッケージ光I/O」は、AIトラフィックを高速かつ効率的に処理できる。

2022年7月に「インパッケージ光I/OでAIの可能性を最大限に牽引」という記事[1]を執筆して以来、多大な変化があった。2022年11月に米オープンAI社（OpenAI）はチャットGPT（ChatGPT）を公開し、それ以来、生成AI（ジェネレーティブAI）はテクノロジーとビジネスの分野で注目されている。組織がAIで課題を解決し、世界を再構築できる可能性が明確になるにつれ、AIは人々の関心を集めている。

最近の調査でもこの傾向を裏付けている。例えば、2023年7月に米ベンチャー・ビート社（VentureBeat）が実施した調査では、組織の55%が自社のニーズに合うかどうかを確認するために、小規模ながら生成AIを試行していることが判明した[2]。また、2023年8月に米マッキンゼー社（McKinsey）が実施したグローバル調査では、回答者の79%が日常生活や職場で生成AIをある程度利用していると回答し、22%が職場で定期的に利用していると回答している（**図1**）[3]。

しかし、次のような落とし穴がある。誰もが生成AIに飛びつく中、これらのモデルをトレーニングし、クエリーをすべて処理するには、GPU、TPU、ASICなどの高性能なハードウエアと、大容量のメモリやストレージが必要になる。さらに、テクノロジーのデータ処理への飽くなき欲求は増大し続けている。AIのデータ渇望を満たすためには、効率的なデータ転送とリソースの割り当てを維持することが鍵となる。

生成AIの場合、大規模なモデルを使用するため、多数のGPUを1つの巨大なGPUとして機能させる必要がある。このため、GPU間の相互接続には、従来のネットワークで処理できる以上に大きな負荷がかかる。この相互接続を、ある場所から別の場所へデータを移動させる高速道路だと想像されたい。従来のネットワークは単純な路面道路のようなもので、今日の大量のAIトラフィック負荷を高速では移動させられない。そこで解決策となるのが、AIトラフィックを高速かつ効率的に移動させ続けられる、いわばスーパーハイウェイとも言える「インパッケージ光I/O」だ。

これらの相互接続は、生成AIモデルのパフォーマンスを決定する上で重要な役割を果たしている。モデルの処理速度、効率性、スケーラビリティ、変化する需要やユーザーニーズに適応する能力に直接影響する。では、生成AIアーキテクチュアの世界に飛び込み、それらがもたらすネットワーキングの課題、従来の相互接続技術の限界、インパッケージ光I/Oのそういった課題への対処方法について、さらに深く考察していく。

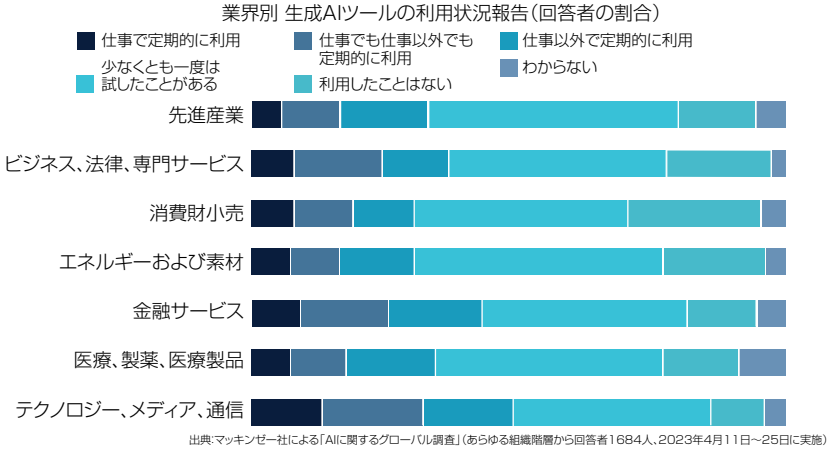

図1 業界別 生成AIツールの利用状況

生成AI拡大の課題への対応

大規模言語モデル（LLM）は、大規模なデータセットを使用してテキストを認識、要約、翻訳、予測、生成できる深層学習アルゴリズムである。LLMは生成AIの中核を成し、急速なペースで成長し続けている。LLMは言語構文を認識することに優れており、検索や視覚技術など他のアプリケーションの基盤となっている（**図2**）。

DECODERの最近のレポートによると、オープンAI社の最新のGPT-4言

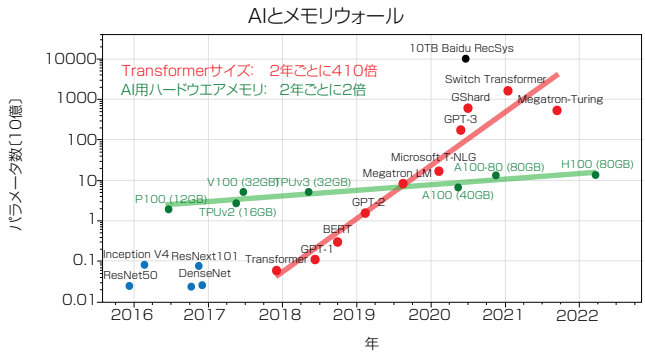

図2　SOTA(最先端)モデルのパラメータ数の経年進化とAIアクセラレーターのメモリ容量(緑色の点)。大規模なトランスフォーマー(Transformer)モデルのパラメータ数は2年ごとに410倍の割合で指数関数的に増加しているが、単一のGPUメモリは2年ごとに2倍の割合でしか増大していない(画像提供:https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8)

語モデルは合計1兆7600億個のパラメータを有しており、同モデルはそれぞれ2200億個のパラメータを有する8つのモデルで構成されている、と報告されている[4]。比較してみると、2020年のGPT-3のパラメータは1750億個であるが、2019年のGPT-2のパラメータはわずか15億4000万個であった。パラメータ数が増加し続けるにつれて、高スループットがますます重要になる。これは、ノードを追加するか、各ノードの速度を上げることで実現可能だ。

この規模の生成AIモデルは、大量のリソースを必要とする。推論には数十台のGPU、ファインチューニング(微調整)には数百台のGPU、トレーニングには数千台のGPUが必要だ。モデルの大規模化と複雑化が加速するにつれて、GPUの必要台数はさらに急増加し、ますます効率性が要求されるだろう。

データセンター事業者は、規模や電力面で限界があるため、戦略的に検討しなければならない。100MWのデータセンターを例に挙げよう。GPUの台数は限られている(GPU1台あたり約1キロワットと仮定すると、1データセンターあたりおよそ10万台)ため、そのGPUが担うべき特定のタスクに応じて、運用上の決定が左右される。推論は最も一般的なタスクであり、これらの制限下では、最大1万台×10または1000台×100のGPU推論システムが利用可能だ。一方、ファインチューニングには、1システムあたり100台から1000台のGPUが必要である。その結果、10万台のGPUの占有面積内に収められるシステム数に限界が生じる。生成AIアーキテクチュアの計算要件としては、さまざまなタスクへさらに柔軟にリソースを割り当てるためにディスアグリゲーション(分割)が必要とされている。つまり、システムが必要に応じてGPUを動的に割り当てることが可能になる、ということだ。

このようなシステム機能するためには、生成AIシステム全体における通信には、GPU間で低テイテンシかつ高帯域幅の相互接続が必要になる。このような接続により、多数のラック間で高速かつ効率的にデータ転送が可能になるのだ。

低レイテンシ自体も重要だが、システム内のGPU台数を増加させるのと同様に、レイテンシを均一にすることも重要だ。レイテンシが均一でない場合、一部のGPUが通信のボトルネックになったり、他のGPUが効率的に動作しているのにアイドル状態になったりすることで、システムのスケーラビリティが制限される。システム全体でレイテンシを均一に維持することは、スケーラビリティと効率性の両面で極めて重要であり、最終的にはソフトウエアやプログラミングモデルの性能を左右することになるのだ。

このようなパフォーマンスの課題の根底には、GPT-3などのAIモデルをトレーニングする際に、1.287GWhのエネルギーを消費し、552トンのCO_2を排出するという事実があり(**図3**)、その費用は約19万3000ドルにも上る。従って、このようなLLMの高速コンピューティング・アーキテクチュアのエネルギー要件をどのように管理するかが問題となる。

米エヌビディア社(NVIDIA)の創業

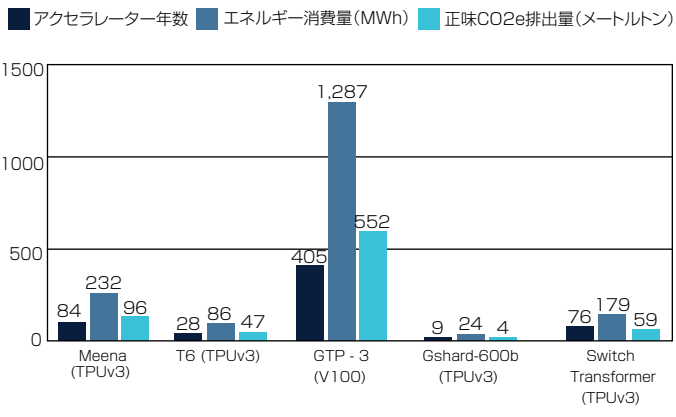

図3　5種類の大規模自然言語処理(NLP)ディープ・ニューラル・ネットワーク(DNN)のアクセラレーターの稼働年数、エネルギー消費量、CO2e排出量(画像提供:https://arxiv.org/ftp/arxiv/papers/2104/2104.10350.pdf)

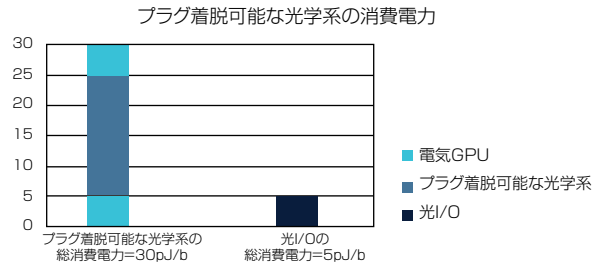

図4　プラグ着脱可能な光学系の消費電力

者兼CEOであるジェンセン・フアン氏(Jensen Huang)は、COMPUTEX 2023の基調講演にてこの課題について言及した[6]。フアン氏の指摘によると、例えば1000万ドルあれば、960台のCPUサーバを購入できるが、1台のLLMをトレーニングするには11GWhものエネルギーが必要となる。一方、同じ1000万ドルで48台のGPUサーバを購入すると、わずか3.2GWhのエネルギーで44台ものLLMをトレーニングできるのだ。言い換えると、40万ドルで2台のGPUサーバを購入すると、1台のLLMをトレーニングするのに必要なエネルギーはわずか0.13GWhだ。つまり、GPUは、同じトレーニング性能に対してコスト効率が25倍、エネルギー効率が85倍も優れている。

興味深いことに、フアン氏は、GPUサーバはもはやコンピューターではなく、データセンターであると指摘した。彼は効率性について言及していたが、相互接続の必要性も強調している。

相互接続の帯域幅密度とスケーラビリティのハードルを乗り越えて

組織は現在、従来の相互接続によるシステム上で、プラグ着脱可能な光学系を組み込んだ電気的I/Oを使用して生成AIタスクを実行している。しかし、この接続技術ではレイテンシと帯域幅のボトルネックが発生し、実用的かつ効率的な次世代の生成AIタスクを実現できなくなる。そのため、このような相互接続ソリューションは時代遅れになりつつある。

電気的相互接続において、距離はいわばアキレス腱のようなものである。距離が長くなるほど信号劣化が発生し、

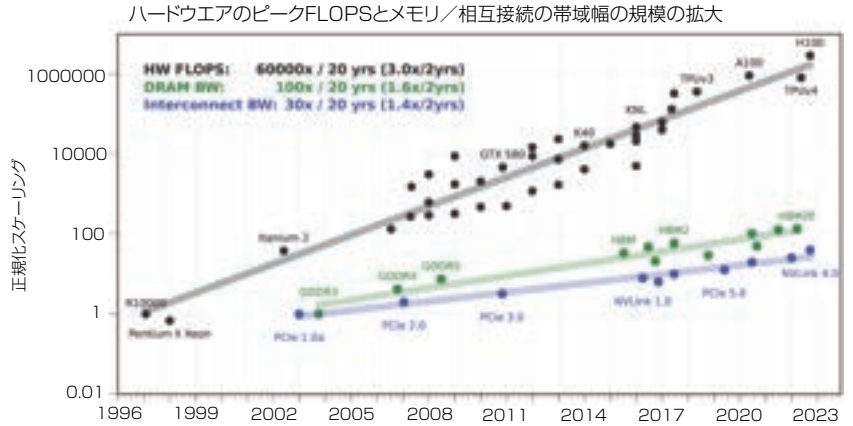

図5　さまざまな世代製品の相互接続およびメモリの帯域幅と、スケーリングと、ピークFLOPSの規模の拡大。本図に見られるように、帯域幅は非常にゆっくりと増加している（画像提供：https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8）

現在の100Gリンク、さらには今後見込まれる200Gリンクにおいてさえ、単一のシャシー内でしか機能しない。接続性の要求がシャシーを超えてクロスラック規模に、さらにはマルチラック規模に拡大するにつれてプラグ着脱可能な光接続が不可欠となってくる。

つまり、すべてのGPUから他のすべてのデバイスでシステム全体における通信を試行する場合、従来の相互接続は、ラック内の電気的相互接続とクロスラックのプラグ着脱可能な光接続で構成されていた。

しかし、ここで本質的な問題に目を向ける必要がある。システムが大規模化するにつれ、システムファブリックを介し、あるGPUから別のGPUへ、さらに多大な高帯域幅の接続を構築しなければならない。一方、従来の相互接続では、プラグ着脱可能な光学系が消費するエネルギーと物理的な占有面積が膨大であるため、そのような接続は実現的ではない。生成AIの高速性とニーズを十分に満たすまでには、効率的に拡大させるのは困難なのだ。

また、プラグ着脱可能なGPU間のリンクは1ビットあたり30ピコジュール（pJ/b）を消費するため、プラグ着脱可能なGPUにとって電力消費が弊害となる。一方、パッケージに直接接続するインパッケージ光I/Oソリューション（これについては別途詳述する）の場合、消費電力はわずか5pJ/b未満だ。しかしそれには費用面の課題もある。現在のプラグ着脱可能なGPUのコストは、1Gbit毎秒につき1～2ドル程度もかかり、生成Aの費用対効果を高めるには、およそ10分の1にコストを削減する必要がある（**図4**）。

もう1つの欠点は、プラグ着脱可能な光学系が大型であることだ。そのエッジ帯域幅密度はインパッケージ光I/O

Gen（世代）＝アヤール･ラボ社製 TeraPHY光I/Oチップレットの世代 *は現在の製品世代を示す
I/F ＝インタフェース
Mod ＝モジュール

Gen	電気インタフェース（先端パッケージ）				光インタフェース（CW-WDM）			光チップレット帯域幅（送受信）	オフパッケージ光I/O帯域幅（1パッケージあたり4～8チップレット）	オフパッケージ基数（ポート数）
	I/F	Mod	送受信I/O数	データ速度［1光I/Oあたりの Gbit毎秒］	ポート数	1ポートあたりの波長［λ］数	データ速度［1波長あたりの Gbit毎秒］			
1	AIB	24	20/20	2	8	8	16	2 Tbps	8-16 Tbps	32-64
2*	AIB	16	80/80	2	8	8	32	4 Tbps	16-32 Tbps	32-64
3	UCIe	16	32/32	8	8	16	32	8 Tbps	32-65 Tbps	32-64
4	UCIe	16	64/64	8	16	16	32	16 Tbps	65-131 Tbps	64-128
5	UCIe	16	64/64	16	16	16	64	32 Tbps	131-262 Tbps	64-128

光I/Oチップレットの世代ロードマップ

のわずか10分の1であり、その面積密度は100分の1未満にすぎない。このため、GPUからシステムの残りの部分への利用可能な帯域幅が制限され、将来の高パラメータの生成AIタスクを実行不可能にするボトルネックとなってしまう。

つまり、物理的なスペースと消費電力の制約に、スケーラビリティの低さが相まると、特に生成AIモデルのニーズに応える場合、プラグ着脱可能な光学系の実装は困難を極める。

生成AIのデータ集約的な側面、つまり大規模なデータセットとさらに大規模なパラメータセットに対処するには、そのタスクに見合う相互接続が要求される。光I/Oの場合、GPU間で必要とされる高帯域幅容量を実現できる。米アヤール・ラボ社（Ayar Labs）の光I/Oソリューションは、高度な光源技術と組み合わせたインパッケージ光I/Oチップレットで構成され、毎秒4Tbitの双方向帯域幅を実現することで、この需要に応えられる。

しかし、光I/Oの利点は帯域幅の高性能化だけにはとどまらない。光I/Oはまた、パッケージレベルの指標に影響を与え、コンピュートソケット帯域幅の拡大を可能にする性能を有している。UCIe、CW-WDM MSA、マイクロリングベースの光I/O技術に基づくアーキテクチュアを活用することにより、毎秒100Tbit超のオフパッケージ光I/O帯域幅を利用可能にし、1パッケージあたり最大128ポートの接続に対応できるのだ（下記の表を参照）。

この高基数とポートあたりの高帯域幅に、光I/Oリンクの低レイテンシ[10nm未満＋ToF（飛行時間）]の効果が加わることで、システム全体を光ファブリックで接続して設計する場合、前例のないレベルの柔軟性が実現する。この柔軟性は、生成AIの将来を支える大規模分散コンピューティングシステムのファブリックネットワークには不可欠である。このような機能により、低く均一なレイテンシと高スループットを特徴とする大規模システムのファブリックネットワークが設計可能になり、コンピュートノードを最大限に活用した状態を維持できる。

光I/O相互接続の高帯域幅が分散システムにもたらす影響は革新的であり、注目に値する。過去20年間にピークFLOPS（演算性能）が6万倍に急増したのに対し、オフパッケージ相互接続の帯域幅はその間にわずか30倍しか増加していない。生成AIの性能を引き出すために不可欠な、真にスケーラブルな分散システムを実現するには、相互接続の帯域幅におけるボトルネックの課題に取り組む必要がある（**図5**）。

生成AIは近年、最も破格的な技術開発の1つとして台頭しているが、これには十分な理由がある。数々の点において、その使いやすさと多用途性はインターネットの導入に匹敵する。そして今日、多くの組織の将来は、最先端を維持するために、生成AIを大胆かつ創造的に、新たな方法で組み込むことにかかっている。

残念ながら、従来の相互接続は、生成AIアーキテクチュアのニーズに応えるには不十分だ。しかし、インパッケージ光I/Oは、高帯域幅、低レイテンシ、高スループット、エネルギー効率という差し迫ったニーズに応える有望なテクノロジーである。最も重要なのは、光I/Oを導入して、生成AIのスケーラビリティを実現し、かつてない規模で産業を革新する新たな可能性への扉を開くことだ。

参考文献

(1) See www.laserfocusworld.com/14278346.
(2) See http://tinyurl.com/yc3vt76h.
(3) See http://tinyurl.com/2w6wbxez.
(4) See http://tinyurl.com/4a8wjs2z.
(5) D. Patterson et al., arXiv:2104.10350v3 [cs.LG] (Apr. 23, 2021); https://doi.org/10.48550/arXiv.2104.10350.
(6) See https://youtu.be/i-wpzS9ZsCs.

著者紹介

ウラジミール・ストヤノヴィッチ（Vladimir Stojanovic）は、米アヤール・ラボ社（Ayar Labs）の最高技術責任者（CTO）。e-mail: info@ayarlabs.com https://ayarlabs.com

LFWJ