技術特集 HOT TECH

知っておきたいHOTキーワード

- ワット・ビット連携
- AIプロセッサ

生成AI時代のエネルギー戦略「ワット・ビット連携」とは

高野 雅晴
株式会社ビットメディア 代表取締役社長/
株式会社MESH-X 代表取締役

2022年のChatGPT登場以来、生成AIはエネルギー・インフラという「古くて新しい問題」を一気に課題の最前線に引き上げた。米NVIDIA創業者ジェンスン・フアン氏は2026年1月の世界経済フォーラム（WEF）でこう断言している。「これは単なるソフトウェアの更新ではない。コンピューティング・スタック全体の再発明であり、人類史上最大のインフラ建設の始まりだ」。

エネルギーとデジタルが交差する時代

PC時代は個人の生産性向上、インターネット時代は情報の検索と共有、モバイル・クラウド時代は常時接続――これらと比べ、生成AIの要求が向かうのは常時接続を前提とした膨大な量の推論の連鎖だ。非構造化データを即時処理し、人間の意図に答えるこの演算は、既存のデータセンター（DC）とは桁違いの電力を必要とする。米国電力研究所（EPRI）の推定では、ChatGPTへの1回の問い合わせの電力消費は約9.2Wh――従来の検索エンジン（約0.3Wh）の約30倍だ。この要求が大量にかつ絶え間なく行われるパラダイムである。

NVIDIAが提示する「生成AIの5層スタック」でいえば、第1層のエネルギーこそが全体を支える基盤である（図1）。多くの注目が集まる第4・5層（モデル・アプリ）の下位で、数兆ドル規模の投資が第1～3層の「AIファクトリー」建設に向けられている。

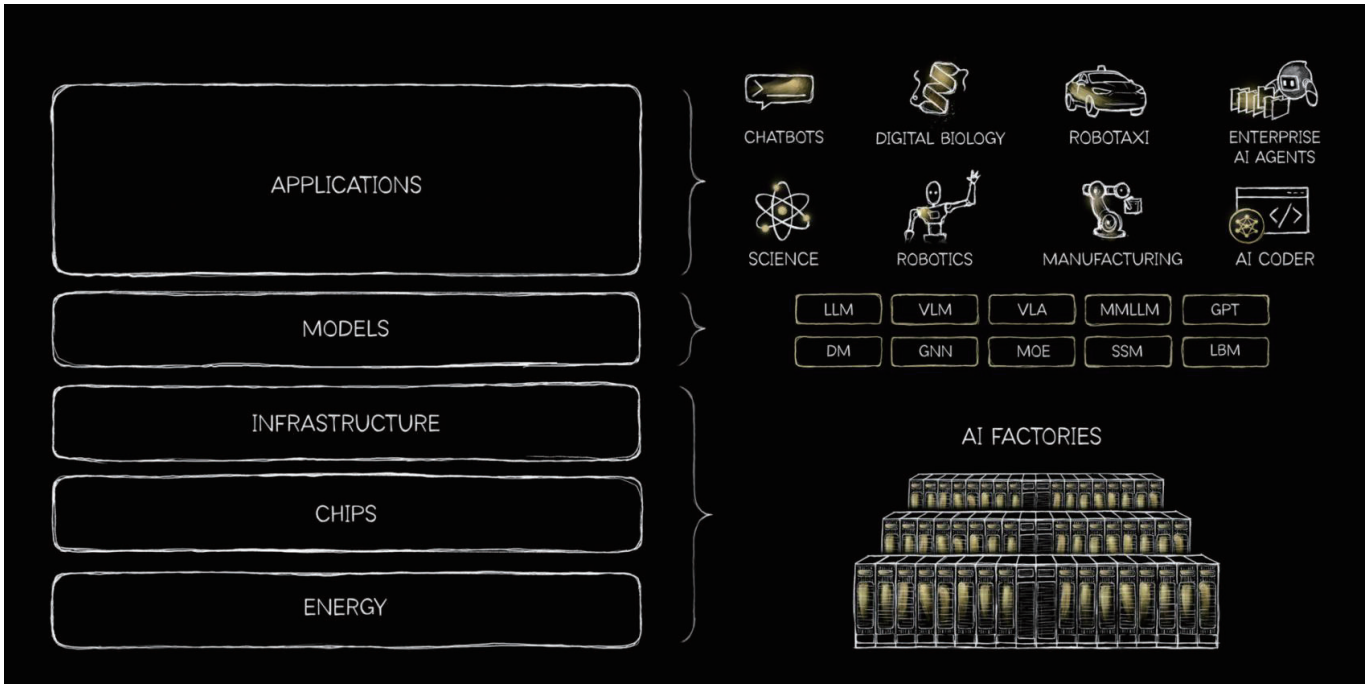
日本の国家戦略ワット・ビット連携

こうした潮流の中で注目すべきコンセプトが「ワット・ビット連携」である。電力（ワット）と通信・情報（ビット）のインフラを有機的に結びつけ、AI時代のデジタル社会基盤を再設計しようという日本の国家戦略だ。2026年2月の国会施政方針演説において高市早苗首相が「ワット・ビット連携の促進」を明言し、政策の最前線に浮上した。

現在、国が推進するワット・ビット連携を一言でいえば、「電力網と情報通信網を同一の戦略インフラとして捉え直し、電力の発電・需給状況に応じてAI処理の実行場所を最適化する」アーキテクチャである。

図1 生成AIを支える「5層スタック」

多くの注目が集まる第4・5層（モデル・アプリ）の下位で、実は第1〜3層が巨大産業基盤として整備されつつある



https://blogs.nvidia.co.jp/blog/ai-5-layer-cake/

LLM (Large Language Models：大規模言語モデル)
VLM (Vision-Language Model：視覚言語モデル)
VLA (Vision-Language-Action model：視覚言語行動モデル)
MMLLM / MLLM (Multimodal Large Language Model：大規模マルチモーダルモデル)
GPT (Generative Pre-trained Transformer)
DM (Diffusion Models：拡散モデル)
GNN (Graph Neural Networks)
MoE (Mixture of Experts：混合エキスパートモデル)
SSM (State Space Models：状態空間モデル)
LBM (Lattice Boltzmann Method：格子ボルツマン法)

3つの技術が柱

既存のデータセンター（DC）は電力会社から安定した電力を調達し、決まった場所でコンピューティングを提供する構造だった。しかし生成AIのGPUクラスターは膨大な電力を消費する上に、再生可能エネルギーの普及により電力の余剰・不足という時間変動が大きくなってきた。つまり「どこで・いつ」演算するかを電力状況に合わせてフレキシブルに変える必要が生じている。ワット・ビット連携を実現するには三つの柱となる技術が存在する。すなわち、

① APN（All Photonics Network）、
② 分散DC（マイクロDC）、
③ WLS（Workload Location Shifting）

技術である。

APN（All Photonics Network）は光信号のまま情報を伝送する次世代ネットワーク。NTTのIOWNが先行するが、九州電力などもAPN構築に取り組んでいる。複数拠点の分散DCを超低遅延・低ジッターで接続することで、物理的な場所を意識せず一体運用する分散型デジタルインフラを可能にする構想だ。

分散DC（マイクロDC）は、DCを大都市に集中させるのではなく、地方の再エネ産地や余剰電力地域にGPU搭載の小規模DCを分散配置する。たとえば九州電力の取り組みでは、バイオマス・地熱・水力・太陽光などが分布するエリアに小規模DCを点在させ、光ファイバー網で結合。地理的に分散した計算資源を、論理的に一つのAI処理システムとして機能させる。

WLS（Workload Location Shifting）技術は、電力価格・再エネ余剰・系統混雑度など（ワット情報）とGPU負荷率・処理キュー長・NW遅延など（ビット情報）を統合評価し、計算負荷を最適拠点へ動的に配置・移動させる技術である。ビットメディアとMESH-Xが開発を主導する「SmartPowerプラットフォーム」が、この機能の実装例である。東京大学と北海道大学の情報基盤センターを活用して実施した広域ワークロードシフト実証（2025年8〜11月）では、JEPX電力価格を参照し、電力価格の安価な地域にAI推論処理を動的転送することに成功している（図2）。

なお、国家戦略としての政府の構想は「ワット・ビット連携 取りまとめ1.0」で提示されており、対応を三段階で整理している（表1）。単なるインフラ整備にとどまらず、国土強靱化と地方創生を一体的に実現する構想と位置づけられている点が特徴である。

図2 WLS技術の実証例

2025年8月・11月の実証実験概要（北大と共同研究しているMESH-Xが実施）
- 構成：東大情報基盤センター（柏・mdx I） <-> 北大情報基盤センター間をSINETで接続
- 仕組み：JEPX（日本卸電力取引所）の北海道・東京エリア価格をリアルタイム参照
- 対象：AI推論タスクを最エネ余剰で電力価格が安価な地域へ動的転送

https://www.soumu.go.jp/main_content/001050378.pdf

組込み業界にもインパクト

生成AIの進化は現在、三つの段階が重層的に進行している。

● エージェント型AI（Agentic AI）：単なるチャットボットから、推論し計画するシステムへ。ハルシネーションが減少し、多段階タスクを自律的に実行可能になりつつある。
● オープンモデル（Open Models）：DeepSeek等の登場による独自モデルからオープンモデルへのシフト。企業や研究機関が取り組む専門領域（ドメイン）に特化したAIの開発が加速している。
● フィジカルAI（Physical AI）：言語だけでなく物理法則・流体力学・タンパク質構造などを理解するAI。デジタル空間から物理世界（製造現場・材料開発）へ進出する。

NVIDIAのフアン氏が強調するように、「ソフトウェアがソフトウェアを書く」時代が来ることでロボットのプログラミングが劇的に容易になる。AIはもはやテキスト生成だけではなく、物理法則を理解して機械を制御する段階に入った。主戦場は製造業の現場となり、ここに組込みシステム業界が果たすべき役割がある（図3）。

フィジカルAIとロボットに不可欠

組込みエンジニアの得意領域となるリアルタイムOS、省電力アーキテクチャ、エッジAI推論エンジン、センサーフュージョン――これらはすべて、フィジカルAIを現場に実装する際の不可欠な要素である。ワット・ビット連携は、データセンターだけの話ではない。エッジデバイスからクラウドまでを一つの「電力×情報」連続体として設計し直す試みともいえる。組込み分野でいえば「RTOSのリアルタイムスケジューリング」の発想に近いものであり、優先度の高いタスク（緊急処理）を適切な資源（プロセッサ・電力）に割り当てるように、AI推論タスクを電力価格・系統状況・計算資源の状態に応じてダイナミックに配置するのが当たり前になる。

また、IoTセンサーやロボット、EV（電気自動車）のエッジデバイスが生成する膨大なデータは、「ビット」であると同時に物理世界という「ワット」のリアルタイム反映でも

表1 ワット・ビット連携の3段階戦略（ワット・ビット連携取りまとめ1.0より整理）

フェーズ	戦　略	主な施策
①短期	足元のDC需要への対応	既存インフラ最大限活用・省エネ技術開発・DC運用の柔軟化
②中長期	新たなDC集積拠点の実現	GW級巨大拠点を東京・大阪圏から分散配置、電力通信インフラ整備
③継続的	DC地方分散・高度化推進	WLS技術による運用高度化・中小規模DCの全国展開・地域共生

https://www.meti.go.jp/shingikai/economy/watt_bit/pdf/20250612_1.pdf

ある。このデータをクラウドに集約せず、エッジで処理しながら電力系統の調整力として活用するアーキテクチャが、次世代の組込みシステムに求められているといえよう。

九州大学伊都キャンパス周辺で進む糸島サイエンス・ヴィレッジ（SVI）のまちづくりはこうしたトレンドを見据えた実践例だ。太陽光・水素・蓄電池・EVを組み合わせたマイクログリッドに直流給電ネットワークを配備し、そのうえでローカル５Ｇインフラや IoT機器やロボットを活用したさまざまな実証が進行中だ。その大きなねらいは2030年のまちびらきに向けて「災害時にも電力も通信も止まらない強靭なまちづくり」を実現することである。

エッジでの推論を支える

こうした動向を裏付けるように、エッジコンピューティング市場の急拡大が世界的に確認されている。米調査会社IDCの2026年2月の予測によれば、世界のエッジコンピューティング支出は２０２５年の2,650億ドルから年率約15%で拡大し、2029年には4,500億ドルに迫る見通しだ。日本国内に目を向けると、2025年の支出は前年比１２.９%増の１.９兆円に達し、2028年には約2.6兆円規模に拡大すると予測されている。

この成長を牽引するのは、AIの推論処理がクラウドの大規模データセンターからエッジの現場へと移行し始めているという構造転換だ。技術面では、LLMを小型化・特化させたSLM（Small Language Models）の実用化が進み、量子化・知識蒸留などの手法によって数十億パラメータのモデルがエッジデバイス上で動作可能になりつつある。経済面でもDeloitteの分析によれば、2026年には全AIコンピューティングの約3分の2を推論ワークロードが占める見込みで、24時間365日稼働する推論インフラの効率化が企業のIT投資判断を動かし始めている。

業種別に見ると、小売・サービス業では店舗内の映像解析や在庫最適化、製造・資源セクターではAIによる品質検査や予知保全、金融サービスでは不正検知やリスク分析でのリアルタイム処理需要が高まっている。なかでも製造現場においては、不良品の検出がわずかな遅延でもライン停止コストに直結するため、クラウドへの通信往復遅延が許容されない場面が多く、エッジ推論の必然性は組込みシステムの世界とそのまま重なる。

ただし、エッジ環境でのAI運用には固有の課題もある。分散した物理デバイスのオーケストレーション、モデルの更新管理、設置場所ごとに異なる電力・冷却・通信環境への対応が必要となり、現時点でエッジAIプロジェクトが本格的な本番稼働に到達したのは全体のおよそ11%にとどまるという調査結果もある。技術的可能性と運用成熟度の間にある隔たりを埋める実装力こ

図3 日本の勝機は製造業とフィジカルAIの融合

そが、組込みエンジニアに求められる今後の差別化要因になるだろう。

「ワット・ビット」が拓く日本の道筋

最後に日本がワット・ビット連携を活かしていくための戦略的ポジションは何かを三つの軸で考えたい。

①製造業×フィジカルAIによる新産業創出

ソフトウェアで主導権を取れなかった日本が、ものづくりの現場に根ざした「物理世界のAI」で独自の競争優位を築く——そのインフラとしてワット・ビット連携は機能する。工場の生産ラインデータ、ロボットアームの制御ログ、センサー群が生み出す物理現象のストリームは、フィジカルAI学習のための貴重な「ビット」資産だ。これを自国の「ワット」インフラで処理することで、AIの国産化と産業競争力を同時に実現できる。

組込みエンジニアの役割はここで決定的に重要になる。フィジカルAIを製造現場・インフラ等に実装するための「最後の一マイル」は、エッジデバイスの設計・実装にかかっているからだ。

②ソブリンAI——国家インフラとしてのAI

「道路・電力・通信と同様に、AIは国家の必須インフラである」——この認識が世界的に広まっている。データの主権（Data Sovereignty）の観点から、自国の言語・文化・産業データは自国のインフラで処理すべきという考え方だ。他国のAIを輸入するのではなく、自国データを自国のインフラで精製する——日本独自のワット・ビット連携プラットフォームを整備することは、AIのデジタル主権確保に直結する。

③エネルギー地産地消×AI処理分散配置

北海道の風力、九州の地熱・太陽光、東北の水力など、日本各地に分散する再生可能エネルギーの余剰電力を活用した分散DCネットワークは、首都圏への過集中リスクを分散しつつ地方創生を実現する有望な戦略となる。「余剰電力をコンピューティングに変換する」という発想は、これまで出力制御で無駄にしていたエネルギーを経済価値に転換する新しいビジネスモデルでもある。

「ワット」と「ビット」——かつて別々の産業として発展してきた電力とデジタルが、生成AIという触媒によって不可分の連続体へと変貌しつつある。

組込みシステムの世界で長年培われてきた「リアルタイム制御」「省電力設計」「分散アーキテクチャ」「物理インタフェース」のノウハウは、この新しい時代の主役となりうる技術資産だ。

ワット・ビット時代は、私たちの業界にとって受け身の変化ではなく、積極的に主導できる変革の舞台である。フィジカルAIが製造現場・社会インフラ・モビリティに実装されていく過程で、組込みシステム技術は「デジタル×エネルギーの産業革命」の最前線に立つことになる。

AI時代のプロセッサ像を探る

木村 優之

AI時代の計算基盤を、GPU、専用アクセラレータ、CPUの優劣を単純に比較するのではなく、実効性能とシステム設計の観点から整理する。まずGPUが主役となった理由を確認し、次に専用アクセラレータが優位を持ちうる条件を示し、最後にCPUを含むシステム全体の役割分担を考える。

情報システムは、AIを「付け足す」段階から、AIを前提に「設計する」段階へ移りつつある。生成AIや推論機能は、もはや一部の研究用途にとどまらず、クラウド、端末、組込み機器まで含めて広く組込まれるようになった。

この変化に伴い、計算基盤に求められる評価軸も変わってきている。重要なのはピーク性能だけではない。電力、コスト、供給、メモリ容量・帯域、相互接続、ソフトウェア互換性まで含めた実効性能が、AIシステム全体の競争力を左右する。

現在、その中心にいるのがGPUである。GPUは大量並列演算に強く、深層学習と相性がよいだけでなく、並列処理プラットフォームCUDA（Compute Unified Device Architecture）を核とするソフトウェア資産によって、研究開発から運用までの標準的な計算基盤となった。

ただし、GPUがあらゆる用途に対して最適とは限らない。推論や電力制約の強い環境では、TPU（Tensor Processing Unit）やNPU（Neural Processing Unit）のような専用アクセラレータが有利になる場合もある。また、それらを活用するためには、CPUがデータ供給、I/O、前後処理、スケジューリングを担うシステム全体の設計が欠かせない。

本稿では、AI時代のシステム設計に必要とされるGPU、専用アクセラレータ、CPUなどのプロセッサ像を探る。

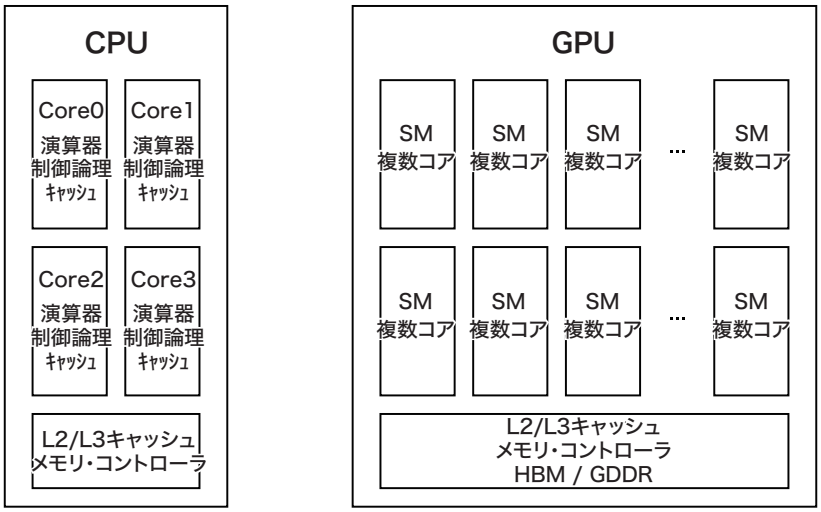
AI時代に必須のハードウェアGPU

AI時代の計算基盤を語るうえで、GPUは避けて通れない存在である。とくに米NVIDIA製GPUは、学習・推論の両面で広く使われており、現在のAIシステムを支える代表的なハードウェアとなっている。GPUがここまで重要になった理由は明快である。深層学習の中核となる処理が、行列積をはじめとする大量の積和演算で構成されており、GPUの大量並列アーキテクチャと強く適合したためである。

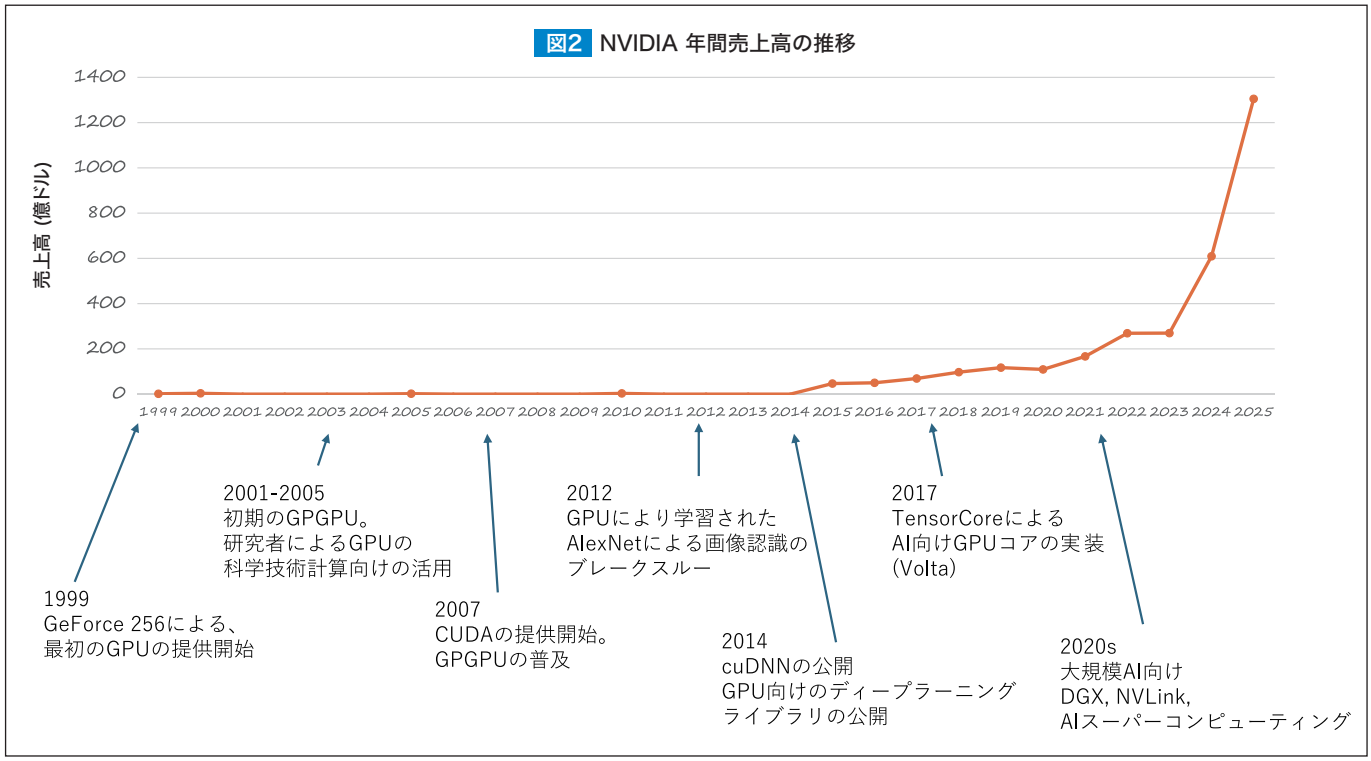
GPUの大きな特徴はその演算性能だ。AIの性能はそれを実行するハードウェアに大きく依存し、より高性能なAIサービスを提供するためにはより高性能なGPUハードウェアが必要となる。そのため、米国は敵対する国に対して高性能なGPUの輸出を禁止したりするなど、もはやGPUは国家戦略的な意味合いを持つ重要なハードウェアとなっている。

GPUがここまでAIにとって大きな影響を与える理由は、演算性能と、AIに必要な処

図1 CPUとGPUのアーキテクチャ模式図



SM:
Streaming Multiprocessor
GPUの基本的な計算ユニット



理がGPUのハードウェアとマッチしていたから、という点だ。AI処理の基礎をなすニューラルネットワークは、数学的には行列演算として落とし込まれる。行列演算は行列積、つまり行列の各要素に対する複数の同じような加算と乗算を実行することにつながる。つまり、AIの性能を上げるためには、同じような積和演算をひたすらこなしていくことが重要となる。

これらの処理は、もちろんCPUを使っても実現可能だ。しかし、異なる行列要素に対して同じような大量の演算を適用するという処理に対してCPUとGPUは異なるアーキテクチャで実現している。CPUは、分岐や依存関係の多い処理、レイテンシに敏感な処理を効率よく実行できるよう、制御機構やキャッシュ階層を厚く持つ。一方GPUは、多数の演算器を並べ、同種の演算を大量のデータに並列適用することで、高いスループットを実現する設計になっている。

図1は、CPUとGPUの違いを概念的に示している。CPUはダイ全体に対して演算器の比率が少ない。その代わり、各種キャッシュや予測器などのハードウェアに多くの面積が費やされている。一方、GPUはダイの大半が演算ユニットに費やされている。各ユニットは各々がシンプルな構成でありながら、大量のデータに対して同じ処理が並列に適用できるように、大量の演算器が並んでいる。

上記で述べたように、AI処理にとって重要なのは「大量の処理を同時にこなす能力」である。AI処理にとって、GPUの方がCPUよりも適しているというのはこのような理由による。

NVIDIA独り勝ちに3つの要因

NVIDIAの優位は、単一の要因で説明できるものではない。図2に見るように過去の蓄積の上に、現在の興隆が成り立っている。技術的には、大きくハードウェアの進化、ソフトウェア基盤の整備、システムとしての提供の3つの要因に整理できる（表1）。

AIと相性の良いハードウェア

まず、GPUそのものが深層学習と極めて相性のよいハードウェアだった。深層学習で中核となるのは、行列積や畳み込みに代表される大量の積和演算である。こうした処理は、多数のデータに対して同種の演算を並列に適用するGPUのアーキテクチャとよく適合する。CPUが分岐や依存関係の多い処理、レイテンシに敏感な処理に強みを持つのに対し、GPUは大量の演算器を並べることで高いスループットを実現できるため、深層学習の学習・推論の両方で優位に立ちやすかった。

しかし、GPUがAI用途で広く使われるようになった理由は、ハードウェア特性だけでは説明しきれない。決定的だったのは、NVIDIAが早い段階からGPUをグラフィックス以外の計算にも活用できるようにし、そのためのソフトウェア基盤を整備したことである。CUDAの提供によって、開発者はGPUを汎用計算資源として扱えるようになり、GPUは単なる描画用プロセッサから、科学技術計算や機械学習にも使える高性能計算基盤へと役割を拡張していった。

この土台の上で、深層学習ブームが到来したことが大きかった。2012年の画像処理コンテストで圧勝したモデル

表1 NVIDIA優位を支えた要因

要因		内容
ソフトウェア基盤の整備	CUDA	GPUを汎用計算に使うための基盤。開発者が簡単にGPUを使えるようになった
	ライブラリ	cuBLAS、cuDNN、TensorRTなど。高性能実装をすぐ利用できるようになった
	フレームワーク対応	PyTorch、TensorFlowなどとの統合。AI開発の標準環境になった
ハードウェアの進化	AI向けハードウェア進化	高速演算ユニットTensor Core、高性能メモリHBM、高速相互接続。学習・推論性能が伸びた
システムとしての提供	システム化	プラットフォーム（DGX、HGX）、GPU間接続（NVLink、NVSwitch）。大規模導入しやすい
	エコシステム	ツール、知見、研究コードの蓄積。他社へ乗り換えにくい強い基盤になった

「AlexNet」以降、ニューラルネットワークの学習を現実的な時間で実行するためにGPUが広く用いられるようになり、研究実装や論文コード、主要な機械学習フレームワークがNVIDIA GPUを前提に整備されていった。つまり、NVIDIAは単に「速いチップ」を作ったのではなく、研究開発から実運用までを通じて使われる環境全体を先に押さえることに成功したのである。

整備されたソフトウェアとシステム

さらに、CUDAだけでなく、cuBLASやcuDNNのような高性能ライブラリの整備も大きな意味を持った。深層学習で頻出する行列演算や畳み込み演算が最適化済みライブラリとして提供されたことで、多くの利用者は低レベル実装を自前で書かなくてもGPU性能を引き出せるようになった。PyTorchやTensorFlowといった主要フレームワークも、こうしたライブラリ群を前提に発展しており、その結果としてNVIDIA GPU上での開発・実行が事実上の標準となっていった。

ハードウェア面でも、NVIDIAは継続的にAI向けの改良を進めてきた。Tensor Coreの導入によって、深層学習に多用される行列演算を低精度で高速に処理できるようになり、学習・推論の性能を大きく高めた。加えて、高性能メモリHBM（High Bandwidth Memory）の採用やメモリ帯域の拡大、NVLinkやNVSwitchによるGPU間接続の強化によって、単体GPUの性能だけでなく、大規模システムとしての拡張性も向上した。AIシステムでは、演算性能だけでなく、メモリ帯域やGPU間通信が実効性能を左右するため、この点で先行した意義は大きい。

加えて、NVIDIAはGPUを単体部品として売るだけではなく、DGXやHGXのような形でシステムプラットフォームとして提供してきた。これにより、顧客はGPUそのものだけでなく、ソフトウェア、通信、冷却、運用を含めた形で導入しやすくなった。AIインフラの構築では、個々のチップ性能だけでなく、すぐに使える形で提供されることが重要であり、このシステム化の戦略もNVIDIAの優位を支えた要因である。

要するに、NVIDIAがGPU独り勝ちの状況を作れた理由は、単一の技術要素ではなく、GPUアーキテクチャの深層学習との適合性、CUDAを中核とするソフトウェア資産、AI向けハードウェアの継続的進化、そしてシステム全体としての提供能力を一体で揃えた点にある。AI時代におけるNVIDIAの強さは、GPU単体の性能ではなく、ハードウェア・ソフトウェア・システムの3層を同時に押さえたことによって成立している。

CUDA優位は今後も続く可能性大

当面の見通しとして、GPGPU（General-Purpose computing on Graphics Processing Units）とCUDAの優位はなお続く可能性が高い。重要なのは、単にCUDAという言語仕様そのものではなく、その上に積み上がった膨大なソフトウェア資産である。主要フレームワーク、最適化ライブラリ、研究コード、運用ノウハウが蓄積しているため、既存資産を重視する限り、GPU優位は容易には崩れない。

ただし、この優位が絶対的であるわけではない。たとえば、フレームワークやランタイムの抽象化がさらに進み、開発者がハードウェア差を意識せずに済むようになれば、バックエンドの置き換えは容易になる可能性がある。また、行列積中心ではない新しいアルゴリズムが主流になった場合には、現在のGPUアーキテクチャと異なる構造を持つハードウェアが優位に立つ余地もある。

GPU代替を狙うAI専用ハードウェア

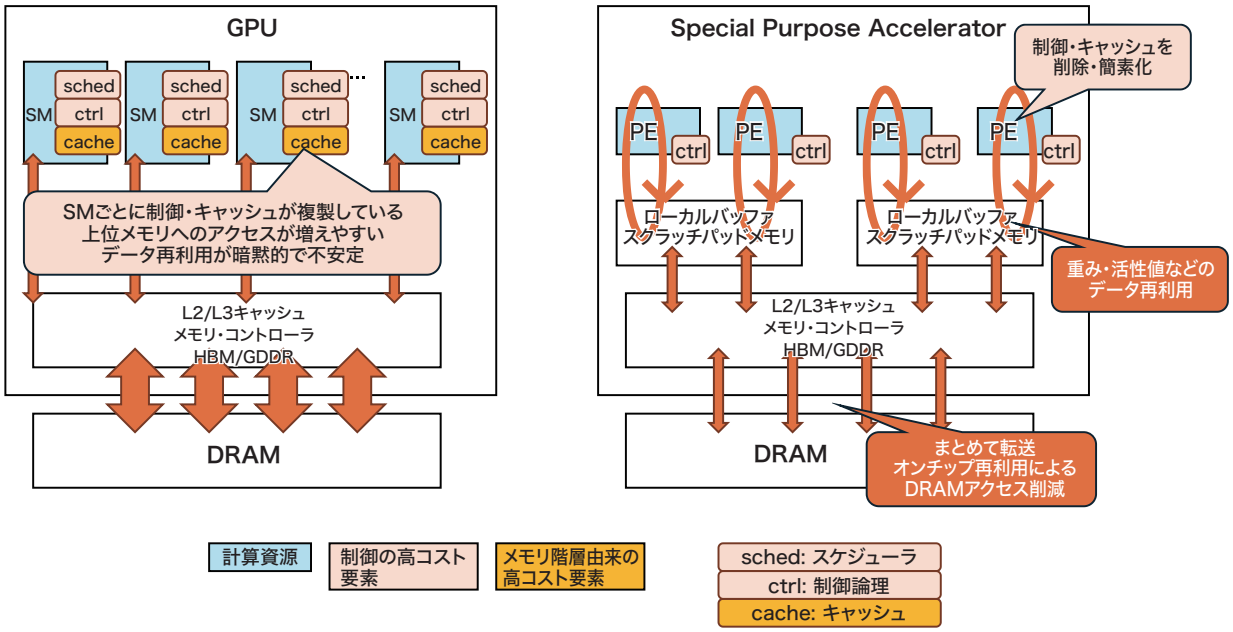
このように、NVIDIAのGPUはAIハードウェアの一強時代を築いてきた。しかし、それを置き換える、あるいは少なくとも一部の用途で代替することを狙うハードウェアも登場している。主な論点は、コストと電力効率である。

GPGPUは汎用計算も視野に入れた設計であるため、AI以外の処理に対応するための機構も多く含んでいる。そのため、AI用途だけを考えた場合には、必ずしも完全に最適化された構造とはいえない。さらに、近年の高性能GPUは非常に高価であり、供給不足もあって導入コストが大きい。こうした背景から、GPGPUよりも低コストかつ高い電力効率を目指すAI専用ハードウェアが注目されるようになった。

専用ハードで高い電力効率を実現

専用アクセラレータがGPGPUより高い電力効率を達成しやすい理由は、大きく分けて2つある。第1に、データ移動を減らせることである。一般に、半導体では演算そのものだけでなく、データをメモリから運ぶことにも大きな電力を要する。ニューラルネッ

図3 消費電力の主因の違い：GPUと専用アクセラレータ



トワークでは、入力、重み、活性値を繰り返し参照するため、これらを毎回深いメモリ階層から取り出す構成では、電力効率が悪化しやすい。これに対して専用アクセラレータは、ローカルバッファやシストリック・アレイのような構造を用いて、データをチップ内で繰り返し再利用するよう設計できる。その結果、外部メモリアクセスを減らし、データ移動に伴う電力消費を抑えやすい。

第2に、制御を簡素化できることである。GPGPUは汎用性を持つがゆえに、命令フェッチ、デコード、動的スケジューリング、分岐処理などの制御機構を必要とする。一方、専用アクセラレータでは命令の種類や実行パターンを狭く限定できるため、制御回路を簡素化し、その分を演算効率や電力効率の向上に振り向けやすい。つまり、専用化によって削減できるのは、単なる「無駄な演算」ではなく、データ移動と制御のオーバーヘッドなのである。図3が示すように、差の本質は演算器の有無ではなく、データがどこをどれだけ往復するか、そしてそれを制御する回路がどれだけ重いかにある。

新アーキテクチャでGPU代替を狙う5社

GPUは現在のAI計算基盤において圧倒的な存在感を持っているが、先に述べたようにその優位が絶対的なものとは限らない。AI処理では、演算性能そのものだけでなく、データ移動、メモリ帯域、チップ内外の通信、レイテンシのばらつき、ソフトウェア最適化コストが実効性能を大きく左右する。そのため、GPUと同じ土俵で単純に演算器の数を競うのではなく、こうしたボトルネックそのものを別の設計思想で解消しようとする専用ハードウェアが登場している。

こうした新しいアーキテクチャの狙いは、大きく分けて四つある。①メモリ帯域の制約を緩和し、データ移動を減らすこと、②チップ内・チップ間通信を効率化し、大規模計算時のオーバーヘッドを抑えること、③実行レイテンシの揺らぎを小さくし、推論時の応答性を安定させること、④複雑な低レベル最適化をコンパイラやランタイムで吸収し、ソフトウェア開発の負担を下げること、である。

この観点から見ると、各社の専用ハードウェアは、それぞれ異なる方法でGPUの弱点を突こうとしている。

たとえば、英Graphcoreはタイル型アーキテクチャとローカルメモリを重視する。多数の計算タイルがそれぞれ局所メモリを持つことで、データをできるだけチップ内に留め、外部メモリアクセスを減らそうとする設計である。これは、計算とデータ配置を近づけることで、データ移動に伴う遅延や電力消費を抑える発想だといえる。

米SambaNova Systemsは、計算グラフ全体をデータフローとして捉え、それをコンパイラによってハードウェアに写像する方向を取っている。狙いは、開発者が個々の演算カーネルやメモリアクセスを細かく最適化しなくても、システム側で効率的な実行を引き出せるようにすることである。これは、ハードウェア単体の性能だけでなく、プログラミング容易性や運用性を含めた実効性能を高めようとする試みだといえる。

米Groqは、推論用途における低レイテンシと決定論的実行を重視している。一般的なGPUでは、スケジューリングやメモリアクセスの揺らぎによって実行時間が変動しやすいが、Groqは演算、通信、メモリアクセスのタイミングをあらかじめ厳密に計画することで、その揺らぎを抑えようとしている。これは、単なる平均性能ではなく、応答時間の予測可能性が重要な用途に向いた発想である。

米Cerebrasはさらに極端な方向を取り、

ウェハースケールの巨大チップによって、通常ならチップ間通信として発生するオーバーヘッドを極力減らそうとしている。巨大な単一チップ上に多数の計算資源と通信網を載せることで、データ移動を可能な限りチップ内部で完結させようとする設計である。これは、メモリ帯域や通信の問題を、アーキテクチャのスケールそのものを変えることで解決しようとするアプローチだといえる。

米Tenstorrentは、タイル型計算、ローカルSRAM、RISC-V系の制御要素、そして比較的オープンなソフトウェアスタックを組み合わせることで、データ移動の抑制と制御の柔軟性を両立させようとしている。これは、専用ハードウェアでありながら、ある程度の拡張性やソフトウェアの自由度も確保しようとする方向性として位置づけられる。

これらに共通しているのは、GPUと同じSIMT（Single Instruction, Multiple Threads）型アーキテクチャの延長線上で正面から競争するのではなく、AI計算で支配的になりやすいボトルネックそのものを別の方法で減らそうとしている点である。つまり、競争の軸を「どれだけ多くの演算器を並べられるか」から、「どれだけ無駄なデータ移動や制御を減らせるか」へと移そうとしているのである。

実用化が進むTPUとNPU

こうした専用アクセラレータの中でも、実用化の面で特に重要なのが、TPUとNPUである。前者はデータセンタやクラウド基盤におけるAI専用ハードウェアの代表例であり、後者はエッジ機器やSoC内部に広く組込まれるAIアクセラレータの代表的なカテゴリである。

なお、TPUが特定ベンダによる代表的実装であるのに対し、NPUはより広いカテゴリ名であり、主としてエッジやSoCに組込まれる専用推論器を指す。

米GoogleのTPUは、行列積を高効率に実行することを主目的として設計されたAI専用ハードウェアである。特徴的なのは、行列演算を規則的に流すシストリック・アレイ型の構造を採用し、データを局所的に再利用しながら高い演算効率を実現している点である。これは、演算器を大量に並べるだけでなく、データ移動を抑えながら行列計算を高速に処理するという、専用アクセラレータの典型的な設計思想を体現している。

また、TPUは単体のチップとしてだけでなく、Googleのソフトウェア基盤やクラウドサービスと一体で提供されている点にも特徴がある。TensorFlowやJAXなどのフレームワークから利用しやすい形で整備されており、ハードウェア、コンパイラ、ランタイム、クラウド基盤を含めた総合的な設計によって実用化が進められてきた。TPUは、専用ハードウェアが単なる高性能チップではなく、システム全体として成立して初めて広く使われることを示す代表例だといえる。

一方、NPUはニューラルネットワーク向けに最適化された演算ハードウェアの総称であり、とくにスマートフォン、PC、組込み機器、自動車向けSoCなどに組込まれる小規模アクセラレータを指すことが多い。こうした環境では、データセンタ向けGPUのような絶対性能よりも、低消費電力、低発熱、小面積、リアルタイム性が重視される。そのためNPUは、低精度演算、専用バッファ、用途特化回路などを活用し、限られた電力枠の中で効率よく推論を実行する役割を担う。

このように見ると、TPUとNPUは規模や利用環境こそ異なるものの、いずれも「AI処理に必要な演算とデータ移動に設計を集中させる」という専用アクセラレータの考え方を共有している。TPUはクラウド側での大規模AI処理の代表例であり、NPUは端末・エッジ側でのAI実行の代表例として、専用ハードウェアがGPU以外の形で広がっていることを示している。

エコシステムやツールチェインが鍵

もっとも、専用ハードウェアが高い潜在性能を持っていたとしても、それだけで広く普及するとは限らない。実際の普及を左右するのは、ハードウェア性能そのものに加えて、使いやすさ、最適化のしやすさ、導入と運用のしやすさを支えるエコシステムとツールチェインである。

まず重要なのは、PyTorchやJAXなど主要フレームワークに自然に統合できることである。多くの開発者はアクセラレータ固有の命令を直接扱うわけではなく、既存のソフトウェア資産をなるべく保ったまま利用したい。そのため、既存フレームワークとの接続性は普及の前提条件になる。

次に、コンパイラ最適化の充実が必要である。GEMM、Attention、LayerNorm、Softmax、KVキャッシュ更新など、実運用で頻出する処理をどれだけ効率よく実行できるかが、実効性能を大きく左右する。理論性能が高くても、コンパイラやランタイムが未成熟であれば、実アプリケーションでは十分な性能を引き出せない。

さらに、デバッグとプロファイリングのしやすさも重要である。単に速いだけではなく、「どこが遅いのか」「何がボトルネックなのか」を可視化できなければ、実運用の場では扱いにくい。開発者が性能問題を切り分け、継続的に改善できる環境が整っていることは、導入のしやすさに直結する。

最後に、供給性、納期、障害対応、ソフトウェア更新の安定性まで含めた運用面も見逃せない。どれだけ性能/Wに優れていても、必要な台数を調達できない、サポートが不十分、更新の互換性が低いといった問題があれば、基盤として採用しにくい。AIインフラは研究用途だけでなく、事業継続性が求められる運用基盤でもあるため、エコシステムの成熟度は極めて重要である。

要するに、GPU代替を狙う専用ハードウェアが本当に広く使われるためには、新しいアーキテクチャや高い性能だけでは足

りない。ハードウェア、コンパイラ、ランタイム、フレームワーク統合、運用支援まで含めた総合力が求められるのである。

性能はメモリと帯域が握る

GPUとCPUを搭載したシステムの性能は、個々の演算器のピーク性能だけでは決まらない。とくにAIやHPCでは、性能を制約するのは演算器そのものよりも、しばしばデータをどれだけ速く供給・転送できるかである。CPU、GPU、メモリ、インターコネクトを含むシステム全体で、必要なデータを十分な速度で流せなければ、高い演算性能を備えていてもその能力を十分に引き出すことはできない(図4)。

このとき重要になるのが、メモリ階層である。GPUは大量の演算器を並列動作させるため、データを継続的に供給する高帯域メモリを必要とする。HBMのような高帯域メモリは、GPUの演算性能を支える重要な要素であり、キャッシュやオンチップバッファも含めたメモリ階層全体の設計が実効性能を左右する。逆に言えば、メモリ容量や帯域が不足すると、演算器が待たされ、ピーク性能は実アプリケーションでは発揮されにくい。

さらに、システム全体では通信帯域も重要である。単一GPUの内部だけで処理が完結するとは限らず、実際のAI処理ではCPU-GPU間のデータ受け渡しや、複数GPU間の同期・通信がしばしば発生する。たとえば学習では、パラメータ更新や勾配同期のためにGPU-GPU間通信が性能を左右しやすい。一方、推論では、モデル重みやキャッシュの参照、CPU側での前後処理との連携がボトルネックになることがある。そのため、PCIe、NVLink、NVSwitchのような相互接続技術も、システム性能を決める重要な要素になる。

また、最適化の方向は用途によって異なる。単発の応答時間が重要な処理ではレイテンシの低減が重視され、大量のジョブを継続処理する用途ではスループットの最大化が重視される。したがって、AIシステムの性能向上を考える際には、GPU単体の演算性能だけを見るのではなく、高速メモリ、帯域幅、通信方式、そしてレイテンシとスループットのどちらを優先するかを、システム全体の観点から整理することが不可欠である。

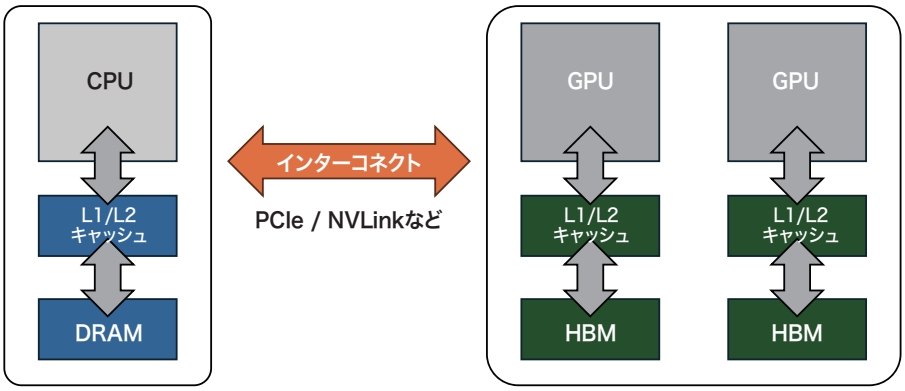


図4 CPU-GPUシステムにおけるメモリ階層と通信帯域の概念図

AI時代のCPU

ここまで、GPUやNPUなどの専用ハードウェアを中心に、AI時代の計算基盤を見てきた。では、そのような時代にCPUは役割を失うのだろうか。結論から言えば、そうではない。むしろCPUは、専用ハードウェアが主役となる時代において、自らもAI処理能力を高めつつ、システム全体を支える中核として、これまで以上に重要な存在になる。

AI時代におけるCPUの進化には、大きく二つの方向性がある(図5)。一つは、CPU自身がAI向け命令や行列演算機能を取り込み、単独でも一定のAI処理を担えるようになる方向である。もう一つは、GPUやNPUなどの専用アクセラレータを支えるホストプロセッサとして、システム全体を制御・統合する方向である。前者はCPUそのもののAI化であり、後者はAIシステムの司令塔としてのCPUの強化である。

AI処理能力を高める

まず前者について見ると、CPUは専用ハードウェアに比べて、大量のデータを単純な規則で一括処理する能力では不利である。しかしその一方で、CPUはベクトル拡張や行列演算拡張を取り込むことで、AI処理に必要なデータ並列性をある程度まで高めることができる。実際、x86系ではAVX-512に加えてIntel AMXのような行列拡張が導入され、Arm系ではSVE2やSMEがAI・機械学習向けの機能として整備されている。RISC-VでもRVV(RISC-V Vector Extension)を中心に、ベクトル処理や行列演算を強化する方向で標準化が進められている。

この方向性において、CPUはGPUの代用品になるというより、小規模モデルの推論、前処理・後処理、制御の多い処理、メモリ容量が効く処理などにおいて独自の役割を担う。たとえば、モデル本体の演算量がそれほど大きくない場合や、入力データの整形、条件分岐を多く含む推論処理では、CPUの方が扱いやすく効率的な場合がある。近年のサーバ向けCPUは、AI向け命令拡張に加えて、ライブラリやランタイムの最適化も進んでおり、PyTorchやONNX Runtimeのような主要ソフトウェアとの統合を通じて、CPU単体でのAI処理能力を底上げしている。

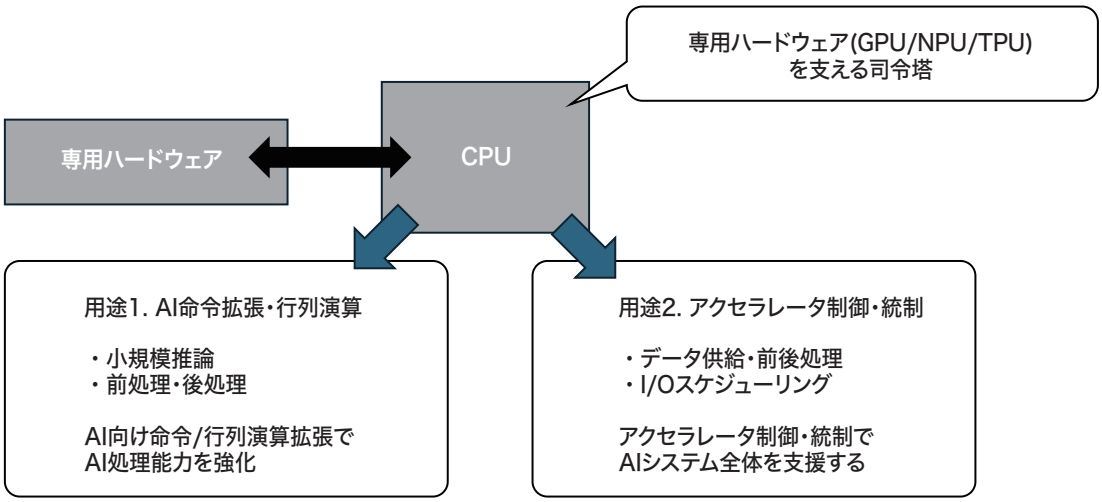


図5 AIシステムにおけるCPUの役割

キメ細かい処理でシステム性能を高める

一方で、AIシステム全体の観点から見たとき、CPUのもう一つの役割はさらに重要である。GPUやNPUのような専用アクセラレータは、大量のデータを規則的に処理することには優れるが、細かな分岐、複雑な制御、OSやランタイムとの連携、I/O処理、ジョブ管理、通信制御のような処理には必ずしも向いていない。こうした役割は、依然としてCPUの得意領域である。

実際のAI処理では、モデル本体の計算だけが存在するわけではない。その前後には、トークナイズ、データ整形、特徴量変換、バッチング、キャッシュ管理、ネットワーク処理、スケジューリングといった多くの処理が存在する。こうした処理は、演算密度の高いコア計算とは性質が異なり、むしろCPUの方が適している場合が多い。そのため、AIシステム全体では、専用アクセラレータが計算の主役であっても、CPUが司令塔としてデータ供給、前処理、後処理、I/O、通信、実行管理を担う構成が基本になる。

さらに、今後のトレンドとしては、SoCやサーバプロセッサの中にNPU系機能を内蔵する流れが強まる可能性が高い。すでに各ベンダは、ベクトル拡張や行列演算拡張をISA(命令セットアーキテクチャ)に取り込みつつあり、CPUとアクセラレータの境界は徐々に曖昧になりつつある。ただし、巨大モデルの学習や大規模推論においては、今後もしばらく専用アクセラレータが主役であり続けるだろう。その意味でCPUは、専用ハードウェアに置き換えられる存在ではなく、専用ハードウェアが最大限に性能を発揮するための基盤として重要性を増していく。

ISAの観点から見れば、x86は依然としてデータセンタ分野で強いソフトウェア資産を持ち、IntelやAMDは既存エコシステムの上にAI拡張を積み増している。Armはクラウドからエッジまで広い展開力を背景に、SVE2やSMEを通じてAI性能の向上を図っている。RISC-Vは現時点ではソフトウェア資産の面でx86やArmに及ばないが、RVVや行列演算拡張、オープンな実装の広がりによって、中長期的にはAI向け計算基盤として存在感を増す可能性がある。

要するに、AI時代のCPUは、専用ハードウェアに主役の座を譲る一方で、役割を失うわけではない。むしろ、自らもAI処理能力を高めながら、同時に専用アクセラレータを束ねるAIシステムの中核として振る舞う、より複合的な存在になっていくと考えられる。

本稿では、AI時代の計算基盤を、GPU、専用アクセラレータ、CPUの役割分担という観点から整理した。現在の主役はGPUであるが、その優位は単なる演算性能の高さによるものではない。深層学習と親和性の高いアーキテクチャに加え、CUDAを中核とするソフトウェア資産、さらにシステムとしての提供体制を含めた総合力が、その競争力を支えている。

一方で、AI処理の実効性能は演算器単体ではなく、メモリ帯域、データ移動、チップ間通信、電力効率、ソフトウェア最適化を含むシステム全体で決まる。そのため、TPUやNPUをはじめとする専用アクセラレータは、GPUとは異なる設計思想によって、特定用途における高効率化を追求している。ただし、それらの普及には、ハードウェア性能だけでなく、コンパイラ、ランタイム、フレームワーク対応、運用性を含むエコシステムの成熟が不可欠である。

また、AI時代においてCPUの役割が失われるわけではない。CPUは、自らもAI向け機能を取り込みながら、同時にGPUやNPUを束ねるホストプロセッサとして、データ供給、前後処理、I/O、スケジューリング、通信制御を担う中核であり続ける。したがって、今後の計算基盤は、単一の万能プロセッサへ収束するのではなく、GPU、専用アクセラレータ、CPUが相互補完しながら、システム全体として最適化される方向へ進むと考えられる。