2026年10月27日

DEEP INSIGHT
Monthly Intelligence Brief

次世代AIインフラを制するのは誰か
―GPU保有競争からインフラ最適化競争へ―

米国三井物産
シリコンバレー支店
辻 理絵子

Summary

- 生成AIの需要が急増する中、サービスの基盤となるAIインフラにおいて、トークン処理の低価格化や効率化、安定化、低遅延化へのニーズが急速に高まっている。
- 計算資源が協調動作し、AI処理の効率を高めたAIインフラの提供を担う新興事業者「ネオクラウド」が登場した。AI計算には、演算集約型とメモリ集約型の処理があり、各処理に強みがあるAIチップ企業も注目を浴びる。両者は、おのおのの強みを起点にサービス領域を拡大中で、今後競争が激化する。
- 今後の競争軸は、単一チップの性能やGPU保有量だけでなく、処理ごとに異なるハードウェアを適切に活用するための制御や、AIインフラ全体を横断して最適化する能力にも広がるとみられる。

1. AIインフラの概要と計算処理の流れ

AI活用が実世界で広がるにつれ、サービスの処理を担うインフラへの需要が高まっており、AIに特化したクラウド事業者や、チップを開発するスタートアップが急増している。本稿では、AIインフラの特性や、同領域のスタートアップの特徴等について、筆者がシリコンバレーで収集した動向を踏まえて解説する。

1-1. AIインフラとトークン

AIインフラとは、GPU、メモリ、ストレージ、ネットワークなどの計算資源、電力、関連ソフトウェア等を組み合わせ、AIモデルの学習・推論を支える基盤を指し、AIサービスにトークン処理能力を提供する。本稿では特に、大規模言語モデル（LLM）の推論を支えるAIインフラに焦点を当てる（図表1）。トークンとは、LLMが入出力を処理する際の単位で、文字列を分割したものなどが相当する。AIエージェントでは、AIモデルの呼び出しを繰り返すことなどにより、利用者に見える回答以上のトークンを消費する。Google社の製品やAPI[1]などで処理されるトークン量は、2026年5月時点で前年同月の約7倍、月間3,200兆トークン[2]超となった。このような状況を受け、AIインフラには、トークン処理の価格を下げつつ、応答速度、一定

[1] API（Application Programming Interface）とは、異なるソフトウェア同士が情報や機能をやり取りするための接続口のこと。
[2] Google, “I/O 2026: Welcome to the agentic Gemini era” 2026年5月19日（Google公式ブログ）。

時間に処理できる量、効率、安定性[3]を高めることが求められる。実現には、ソフトウェア、ハードウェア、電力等が相互に影響するが、特にソフトウェアやチップの技術進展による波及効果は大きい。

図表1：AIインフラの全体像



出所：米国三井物産作成

1-2. AI計算処理の流れ

現在のAIモデルで広く採用されている自己回帰型LLM[4]は、入力と自ら生成した内容を基に次のトークンを予測・生成する仕組みで、この処理を繰り返して文章を作る。図表2は、その推論処理の流れを模式化したものである[5]。まずAIモデルの計算の前に、AIモデルをGPUに配置するなどの準備（0：初期化）、入力のトークン化（1：トークナイズ）、処理の受け付け判断と計算資源の割り付け（2：アドミッションコントロールとルーティング）、処理の順番と実行のタイミング決め（3：スケジューリングとバッチング）を行う。その後、AIモデルの計算として、入力内容をまとめて計算（4：プリフィル）と、トークン単位で回答の生成（5：デコード）が行われる。そして、生成結果を順次テキストに変換し利用者へ逐次送信（6：デトークナイズとストリーミング）、回答生成が終了次第、使用していた計算資源を解放（7：終了処理と資源解放）するのが一連の流れである。スケジューリング、デトークナイズとストリーミングは、回答の生成と並行して繰り返し行われる。

[3] AI計算処理で重要な指標は主に3つある。第1は、利用者が回答を待つ時間である応答速度。第2は、システム全体が一定時間に処理できる要求やトークンの量である処理能力。第三は、限られた設備や電力からどれだけ多くの処理量を引き出せるかという効率性である。これらにはトレードオフがあり、例えば、多数の要求をまとめて処理すれば設備の利用効率は上がるが、1件ごとの回答が始まるまでの時間は長くなり得る。もちろん、価格や安定性も重要な要素である。

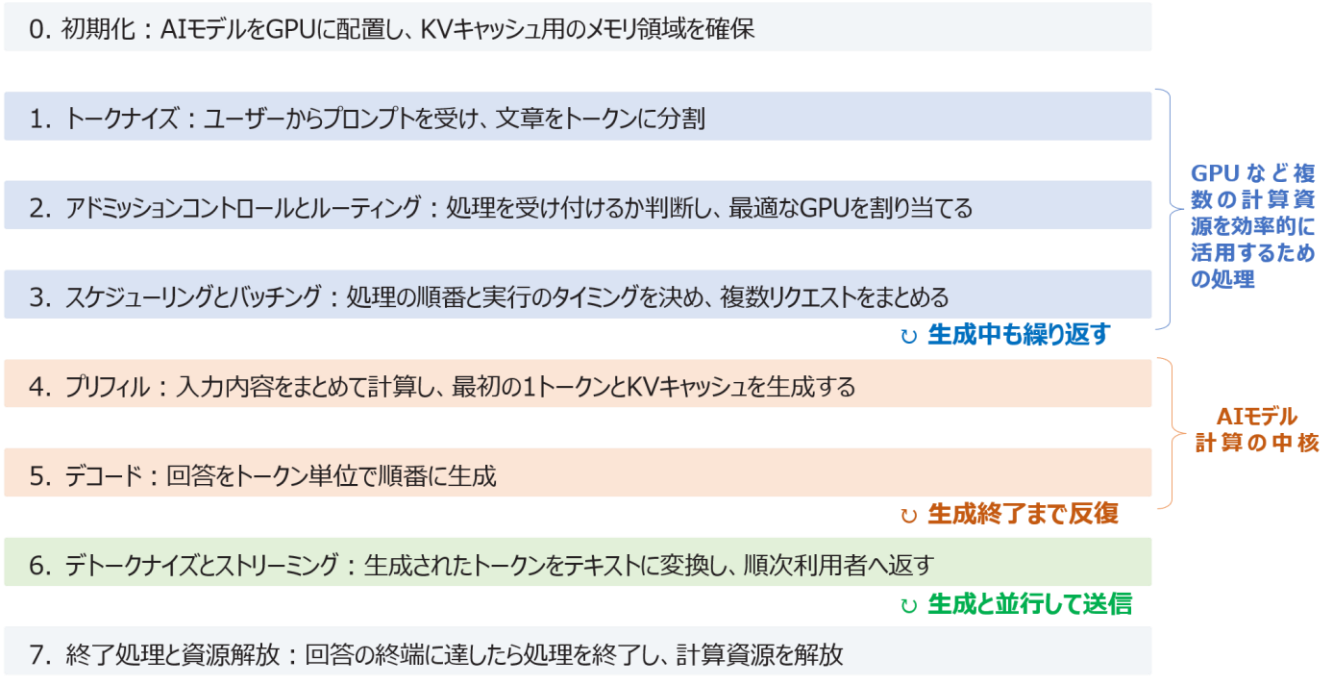
[4] 自己回帰型LLMとは、入力とそれまでに生成した内容を基に、次のトークンを予測・生成する大規模言語モデルのこと。自ら生成したトークンを次の予測にも用い、この処理を繰り返して文章を作る。

[5] 本工程説明において、プロンプトはユーザーの指示テキスト、リクエストはシステムへの要求全体でプロンプトやパラメータ設定などを含むもの、入力内容はAIモデルに直接入力されるトークン化されたデータを指している。

AIモデル計算の中核であるプリフィルとデコードでは、必要な性能が異なる。プリフィルでは、入力トークンをまとめて計算し、回答の最初のトークンを生成し、その過程で後続の計算に使うデータをKVキャッシュ[6]として保存する。入力トークンを並列に計算するため、一般的にはGPUの演算性能への依存が大きい。一方、デコードでは、KVキャッシュを参照・更新しながら、回答の続きのトークンを1つずつ生成する。AIモデルの重みやキャッシュを繰り返し読み込むため、一般にメモリ容量やデータ転送速度が処理性能を左右する。

大規模なモデルでは、モデルが単一のGPUのメモリに収まらず、複数のGPUを用いる場合が多く、どの計算資源に要求を送るか、KVキャッシュをどこに置くか、GPU間でどう通信するかといった資源管理と実行の制御は、トークン処理のコスト、効率、応答速度を左右し、ネオクラウド事業者の差別化要因となりつつある。

図表2：AIインフラにおけるAIモデル計算の大まかな手順



出所：米国三井物産作成

2. AIインフラの競争地図

AIインフラの担い手として近年注目を浴びるのが、ネオクラウドと呼ばれる新興事業者と、AI推論向けのチップや関連技術を開発するスタートアップである。AI推論処理の効率向上に向けて、それぞれソフト

[6] KVキャッシュとは、過去のトークンについて計算したキー（K）とバリュー（V）を保存し、生成のたびに同じ計算をやり直すことを防ぐためのもの。効率的な生成を支える一方、長文コンテキストや同時利用者が増えるほどメモリ容量を圧迫する。

ウェア的なアプローチとハードウェア的なアプローチで取り組む。

図表3 ネオクラウド各社の提供サービスと管理範囲

主な強みの起点	企業	GPU VM・クラスタの提供	GPU基盤の運用管理	モデル・推論の運用管理	代表的な提供内容・工夫
① 計算基盤の確保・運用	Lambda	◎ GPU VM・専用クラスタ	○ Kubernetes・Slurm	未確認	• GPU環境を事前整備 • クラスタの監視・復旧・更新も管理
	IREN	◎ 専用GPUクラスタ	○ クラスタ管理等	未確認	• 電力・データセンターからGPU基盤まで展開 • Mirantisでソフトウェアを強化
	CoreWeave	◎ GPUクラウド・専用クラスタ	○ Kubernetes等	○	• GPU・高速ネットワークとクラスタ管理を提供 • 専用推論にも対応
	Crusoe	◎ GPU VM・専用クラスタ	○ Kubernetes等	○	• 電力・データセンターから展開 • 管理推論で実行・KVキャッシュを最適化
	Nebius	◎ GPU VM・GPUクラスタ	○ Kubernetes等	○	• GPU基盤と管理推論を併せて提供
	GMI Cloud	◎ GPUクラウド・専用クラスタ	○ GPU・クラスタ管理	○	• GPU基盤に加え、モデルの推論APIを提供
② モデル展開・推論計算の最適化	Together AI	○ GPUクラスタ（VM構成含む）	○ Kubernetes・Slurm等	◎ 推論API・専用推論	• GPUクラスタと管理推論を展開 • 独自エンジン等で推論を最適化
	Baseten	△ 専用推論環境のGPU	○ 複数クラウド等を統合・管理	◎ Model APIs・専用推論	• 複数クラウドのGPU基盤を統合・管理 • 独自カーネル等で推論を最適化

◎：主力・重点領域 ○：提供 △：一部・条件付き提供 未確認：公開資料では特定できず
注：本表は本稿の整理であり、性能順位を示さない。提供範囲は重なり、管理範囲はサービス・契約により異なる。
出所：各社公式資料等から米国三井物産作成（2026年9月までの公表資料を参照）

2-1. ネオクラウド

ネオクラウドを、その強みの起点によって2つに整理する。第1は、入手が困難であったGPUの確保やGPUクラスタ[7]の運用に強みを持つ企業群である。米Crusoe、米CoreWeave、米Lambda、蘭Nebius[8]、豪IREN、米GMI Cloudなどが該当する。NVIDIAと提携してGPUを大量に調達することで競争力を出してきた経緯がある企業も少なくない。第2は、AIモデルを実行環境に配置する仕組みや推論計算の最適化に強みを持つ企業群で、米Basetenや米Together AIなどが相当する。このグループは、他社のクラウドや提携事業者の計算基盤も活用しながら、独自の推論ソフトウェアでサービスを提供する。Basetenは複数クラウドを横断した推論運用を特徴とし、Together AIは大規模GPUクラスタの構築・運用と顧客への提供も展開する。これらの企業は、アドミッションコントロール（処理の受け付け判断）、ルーティング（処理先のサーバーやモデル実行基盤への振り分け）、スケジューリング（処理の順序や実行タイミングの決定）などを効率化するソフトウェアでも差別化している（図表3）。計算基盤を起点とする前者の企業群も推論サービスを強化し始めており、両群の提供範囲は重なってきている。

2-2. AIチップ企業

現在のAIインフラでは、NVIDIA製GPUが広く使われ、多くのAI推論ソフトウェアも同社GPU向けの環境で

[7] クラスタとは、多数のGPUやCPUなどの計算機を接続し、ひとまとまりの大きな計算システムとして使う仕組みのこと。
[8] Nebiusは、旧Yandex N.V.がロシア事業を売却後、2024年に社名変更して発足したオランダ企業。ロシア事業との関係は売却完了時に解消したと発表している。

あるCUDA[9]を前提にシステムが成立している。先に挙げたネオクラウド事業者の多くも、CUDAを活用したAI処理基盤を提供する。一方、AI処理に特化したチップにより処理効率を向上しようとするスタートアップもある。各社の方向性の違いにより、主に3つに分けて整理する。

第1は、チップ内の高速メモリ（SRAM[10]）の大容量化やSRAMへの演算機能の組み込みにより、演算器の近傍にデータを保持する方法である。チップ外のメモリへのアクセスやデータの移動距離を減らすことで、デコード処理の高速化と、大電力を消費するデータ移動における電力効率改善を図る。米Groq、米Cerebras、米d-Matrixが代表例である。

第2は、主にDRAM[11]の高帯域化や大容量化により、データの読み出し効率を改善する方法である。日Preferred Networks、米Kepler、米Positron AIが取り組む。単位時間当たりに演算器へ供給できるデータ量を増やすことで、データ待ちによる速度の低下を抑える。また、メモリ容量を拡張し、モデルの重みやKVキャッシュをより多く保持できるようにすることで、大規模モデルや長文の処理への対応を図る。

第3は、計算とデータの流れを制御する方法で、米SambaNovaや米Tenstorrentが代表例である。演算を進めながら次に必要なデータを準備するなど、演算とデータ転送を並行させ、演算器の待ち時間を減らす。これにより、チップの演算能力を有効に活用し、推論の高速化を図ると同時に、データ転送時の消費電力の削減も見込む（図表4）。このように各社は、それぞれのアプローチで遅延の低減や消費電力の削減といったAI処理の効率向上に取り組んでいる。ただし、各社は複数の工夫を組み合わせており、チップの特性には重なりがある。

[9] CUDAとは、NVIDIAが開発した、GPUを画像描画以外の汎用的な計算に活用するためのプラットフォーム。GPUの得意とする大量の並列処理をプログラムから直接実行できるようにし、AIの学習や科学技術計算などを高速化する。

[10] SRAM（Static Random Access Memory）とは、トランジスタで構成する極めて高速な静的メモリのこと。データ保持のための定期的な再書き込み（リフレッシュ）が不要で、主にプロセッサのキャッシュメモリとして処理の高速化を支えている。オンチップSRAMとは、CPUやGPU、SoCなどの半導体チップの内部に直接組み込まれたSRAMのことを指す。

[11] DRAM（Dynamic Random Access Memory）とは、小さなコンデンサに蓄えた電荷によってデータを保持する動的メモリのこと。電荷が時間とともに漏れるため、データ保持にはリフレッシュが必要となる。SRAMに比べて大容量を低いコストで実現しやすく、主にパソコンやサーバーの主記憶、GPUの大容量メモリとして利用されている。HBMやLPDDRもDRAMの一種である。

図表4 AIチップと関連技術の特徴

着目点	企業・製品/技術	仕組みと備考	主に適する処理、目指す方向性
	NVIDIA等 GPU	多数の演算器で並列に計算し、メモリ・複数チップを組み合わせる	• プリフィル・デコード • 幅広い用途
	Google TPU	行列計算向けの専用回路で、多くの計算をまとめて実行する	• 大規模な行列計算（プリフィル等）
① チップ内でのデータ移動を減らす	Groq LPU	オンチップSRAMと静的スケジューリング	• 低遅延・高速な応答生成（デコード）
	d-Matrix DIMC・Corsair	Digital In-Memory Compute（デジタル・メモリ内演算）でメモリと演算を一体化し、データ移動を減らす	• 応答生成（デコード）の高速化を目指す
	Cerebras WSE-3	演算器と高速メモリをウェハー上の巨大チップに統合し、通信やメモリ帯域の制約を緩和する	• プリフィル・デコード
② メモリの接続・容量・読み出しを改善	Preferred Networks MN-Core Lシリーズ	DRAMを演算器の上に積層し、大容量化と高速な読み出しを目指す	• デコード
	Kepler AI向けメモリ	3D積層と新材料で、AI向けメモリの帯域・容量・電力効率の向上を目指す	• デコード
	Positron AI Atlas/Asimov	Atlas：メモリ読み出しの効率を重視 Asimov：大容量メモリを計画（2027年提供計画）	• 主に応答生成（デコード）
③ 計算とデータの流れを制御	SambaNova RDU	RDUでデータの流れを組み替え、階層型メモリで速度・容量が異なるメモリを使い分ける	• プリフィル・デコード • 大規模モデル・肥大したKVキャッシュに対応
	Tenstorrent Tensix・RISC-V	公開命令セットRISC-V・独自コア・公開ソフトで開放性と計算制御の柔軟性を重視	• プリフィル・デコード

注：本表は本稿の整理であり、各社の技術は複数の要素を含む。性能はモデル・入力長・同時処理数等で異なる。
出所：各社公式資料等から米国三井物産作成（2026年9月までの情報）

また、AIチップ企業には、チップの提供に加え、その性能を引き出すコンパイラやランタイム、ドライバ、演算ライブラリ[12] などのソフトウェア基盤の整備が求められる。CUDAはNVIDIAのGPU向けの基盤であり、使い慣れた環境からの乗り換えの難しさもNVIDIAの強みとなっている。他社のAIチップ企業には、CUDA環境で構築された既存のAIモデルや開発環境を、自社チップでも利用しやすくすることが重要である。

3. 各社の提供範囲とAIインフラの方向性

本章では、LLM推論のソフトウェア技術を主な役割に応じた4つの機能に整理し、ネオクラウド事業者とAIチップ企業の提供範囲を俯瞰したうえで、今後の方向性を展望する。

3-1. AIインフラのソフトウェア技術と各社のカバー範囲

4つの機能とは、①計算手順を工夫するアルゴリズム、②処理の割り当て・順序を決めるスケジューリング、③複数チップでの計算、KVキャッシュ、通信などを管理するランタイム、④チップ上で行列演算などを動かすカーネルである（図表5）。これは説明のための機能別整理であり、実装では機能が重なり、一つの製品が複数の領域を担う。これらのソフトウェアの指示を受け、計算、保存、転送を行うGPU、メモリ、

[12] コンパイラは、プログラムやAIモデルの計算手順を対象チップで実行できる命令に変換するソフトウェア。ランタイムは、実行時の演算開始、メモリ確保、データ転送などを管理するソフトウェア。ドライバは、OSなどからチップを制御するソフトウェア。演算ライブラリは、行列演算などを効率よく実行するための機能をまとめたソフトウェア。

ネットワークなどがハードウェアに相当する。本項で触れてきたネオクラウド事業者やAIチップ企業ごとに提供範囲や運用形態は異なる。計算資源の提供、推論ソフトウェア、チップ開発という異なる強みを起点に、各社の提供範囲は重なってきており、ハードウェアとソフトウェアを組み合わせ、インフラ全体を効率よく動かす能力が競争力を左右しつつある。

図表5 LLM推論を効率化する4つのソフトウェア機能とハードウェア

ソフトウェア 4つの機能 SOFTWARE	アルゴリズム （計算方法の工夫）	計算手順を工夫し、推論を効率化する 複数トークンを先行生成し、まとめて検証するなど
	スケジューリング （処理の割り当て・制御）	どの処理を、いつ、どの計算資源で実行するか決める 受け付け、処理の振り分け、まとめ方、負荷分散など
	ランタイム （計算の実行・資源管理）	計算を実行し、そのための資源を管理する 複数チップでの実行、KVキャッシュの管理、通信など
	カーネル （チップ上の個別演算）	行列計算などをチップ上で動かすプログラム チップの特性に合わせて、個々の計算を効率化する
ソフトウェアとハードウェアの境界		
ハードウェア HARDWARE	計算・保存・転送を行う機器	ソフトウェアの指示を受けて動く物理的な機器 GPU・AIチップ、メモリ、ネットワークなど

注：本表は本稿の整理であり、多様な切り口がある。
出所：米国三井物産作成

3-2. 次世代AIインフラはどこで差がつくのか

これらを踏まえ、AIインフラ事業者が取り得る2つの技術アプローチと今後の技術開発の方向性を整理する。

① プリフィルとデコードの分離制御

一般的に、プリフィルは演算性能、デコードはメモリ帯域・容量が重要となる。この違いを踏まえ、プリフィルとデコードを別々のチップやサーバーに割り当て、並列度やバッチサイズなどを各処理に合わせて調整する分離推論が進んでいる。同種GPUでも実施でき、GPUと専用チップなど異種のハードウェアを使い分ける構成も試みられている。将来は、異なる種類のハードウェアがクラスタ内で混在し、処理に応じて使い分ける構成が広がる可能性がある。ただし、プリフィルで生成したKVキャッシュをデコード側へ渡す際、ネットワーク転送が新たな遅延とコストの増加要因となり、短い入力や低負荷時には、同一サーバーで処理する方が速い場合もある。入力の長さ、負荷、既存のKVキャッシュを再利用できる割合を見ながら、ルーティングと実行基盤が連携し、分離の要否や処理の割り当てを調整する制御が重要になっていく見込みだ。

② AI処理アルゴリズムの進化

Speculative Decoding（投機的デコード）は、BasetenやTogether AIなどで採用が進む高速化手法である。小型モデルや専用の予測機構などの軽量な処理で複数トークンを先行生成し、メインモデルの予測に基づいて各トークンの採否をまとめて判定する。適切な採択・補正手順を用いれば、メインモデル単独の生成と同じ出力の質を保ちつつ、1回の実行で複数トークンを確定できる。モデルの実行や重みの読み出しを複数トークン分で共有することで、デコード時間を短縮できる場合がある。効果は候補の採択率、予測処理の負荷、同時処理数などに依存するため、今後も予測精度の改善や処理方式の研究が進むと考えられる。

3-3. 今後の展望

AIインフラ競争では、単一GPUやAIチップの性能・保有量に加え、GPU、AIチップ、メモリ、ネットワークを組み合わせ、ソフトウェアで統合制御し効率を競う段階へ移っている。こうした最適化やモデルの小型化、アルゴリズムの進歩により、AI回答を生成するためのトークン当たりのコストは、急速に低下している。一方、長文処理やAIエージェントの普及で処理量が増えれば、インフラ全体の需要や支出は拡大し得る。今後は、計算資源や電力を確保する力と、一定の品質・応答速度を保ちながらトークン当たりのコストを下げる力を併せ持つ事業者が優位に立つと考えられる。

アンケートご協力のお願い

本レポートをご覧いただきありがとうございました。
今後の改善のため、ぜひご意見をお聞かせください。

👉 **アンケートはこちら**

当レポートに掲載されているあらゆる内容は無断転載・複製を禁じます。当レポートは信頼できると思われる情報ソースから入手した情報・データに基づき作成していますが、当社はその正確性、完全性、信頼性等を保証するものではありません。当レポートは執筆者の見解に基づき作成されたものであり、当社および三井物産グループの統一的な見解を示すものではありません。また、当レポートのご利用により、直接的あるいは間接的な不利益・損害が発生したとしても、当社および三井物産グループは一切責任を負いません。レポートに掲載された内容は予告なしに変更することがあります。