2026 年 3 月 17 日

Press Release

アカマイ・テクノロジーズ合同会社

Akamai、4,400 超のエッジ拠点での分散推論に向けた「AI Grid」インテリジェントオーケストレーションを発表

業界初、Akamai Inference Cloud への NVIDIA AI Grid 実装により、
エッジからコアまで AI ワークロードを動的にルーティングし、
レイテンシー、コスト、パフォーマンスを最適化

※本リリースは 2026 年 3 月 17 日(現地時間) 米国マサチューセッツ州ケンブリッジで発表されたプレスリリースの抄訳版です。

オンラインビジネスの力となり、守る、サイバーセキュリティおよびクラウドコンピューティング企業、Akamai Technologies（NASDAQ：AKAM）は、本日、AI の進化における重要なマイルストーンとなる、NVIDIA® AI Grid リファレンスデザインのグローバル規模の初の実装を発表しました。Akamai のインフラに NVIDIA AI インフラを統合し、ネットワーク全体でインテリジェントなワークロードオーケストレーションを活用することで、孤立した AI ファクトリー中心のモデルから、AI 推論のための統合された分散型グリッドへと業界を進化させることを目指します。

今回の発表は、昨年 10 月に発表された Akamai Inference Cloud の進化における大きな一歩となります。AI Grid を初めて実用化した Akamai は、NVIDIA RTX PRO 6000 Blackwell Server Edition GPU を数千基規模で展開しています。これにより、ローカルコンピューティングのようなレスポンスの良さと、グローバル規模のスケーラビリティを兼ね備えた、エージェント型 AI やフィジカル AI を実行できるプラットフォームを企業に提供します。

Akamai の Chief Operating Officer 兼 General Manager, Cloud Technology Group である Adam Karon は次のように述べています。「AI ファクトリーは、トレーニングやフロンティアモデルのワークロード向けに構築されており、集約型インフラは今後もそれらのユースケースに最適なトークノミクス（トークンあたりのコスト効率）を提供し続けるでしょう」

「しかし、リアルタイム映像やフィジカル AI、多数のユーザーに高度にパーソナライズされた体験を同時に提供するアプリケーションには、中央集約型クラスターへの往復処理ではなく、ユーザー接点に近い場所（エッジ）での推論が必要です。AI Grid のインテリジェントなオーケストレーションは、コンテンツ配信に革命をもたらした分散アーキテクチャを活用し、世界 4,400 拠点にわたって AI ワークロードを最適なコストとタイミングで ルーティングすることで、AI ファクトリーを外側へとスケールさせることが可能になります」

「トークノミクス」の最適化を支えるアーキテクチャ

AI Grid の中核となるのは、AI リクエストをリアルタイムで仲介するインテリジェントオーケストレーターです。アプリケーションパフォーマンス最適化における Akamai の知見を AI に適用したこのワークロード認識型のコントロールプレーンは、トークンあたりのコスト、初回 トークン取得時間、スループットを劇的に改善し、「トークノミクス」を最適化します。

Akamai の大きな差別化要因は、比類のないスケールでグローバルに展開するエッジ基盤を通じて、ファインチューニング済みモデルやスパース化（軽量化）されたモデルにアクセスできる点にあります。これは、ロングテールの AI ワークロードにおいて、コストとパフォーマンスの両面で圧倒的な優位性をもたらします。

具体的なメリットは以下の通りです。

- **大規模環境でのコスト効率：**ワークロードを最適なコンピューティング層に自動的にマッチングさせることで、推論コストを大幅に削減できます。オーケストレーターはセマンティックキャッシングやインテリジェントルーティングなどの技術を活用し、リクエストを適切なサイズのリソースに振り分けます。これにより、高価な GPU サイクルを、本当に必要とするワークロードに優先的に割り当てられます。Akamai Cloud は、オープンソースインフラを基盤とし、大量のデータ処理を必要とする大規模な AI 運用を支えるのに十分なデータ転送枠を備えています。
- **リアルタイムのレスポンス：**ゲームスタジオは、プレイヤーの没入感を損なうことなく、ミリ秒単位のレスポンスで AI による NPC（ノンプレイヤーキャラクター）との対話を提供できます。金融機関は、ログインから最初の画面表示までのわずかな時間の中で、パーソナライズされた不正検知やレコメンドを実行可能です。放送事業者はコンテンツのトランスコーディングや吹き替えをリアルタイムで行い、世界中の視聴者に届けることができます。これらは、4,400 以上の拠点を持つ Akamai の分散型エッジネットワークによって実現されます。このネットワークにはキャッシュ、サーバーレスのエッジコンピューティング、高性能な接続が統合されており、ユーザーに最も近い場所でリクエストを処理することで、オリジン依存のクラウドで発生する往復遅延を回避します。
- **コアを支える本番環境レベルの AI：**大規模言語モデル（LLM）や継続的なポストトレーニング、マルチモーダル推論には、専用インフラのみが提供できる持続的かつ高密度なコンピューティングが必要です。NVIDIA RTX PRO 6000 Blackwell Server Edition GPU を搭載した Akamai の数千基規模の GPU クラスターは、最も負荷の高い AI ワークロードに集約された計算能力を提供し、中央集約型のスケールメリットで分散型エッジを補完します。

コンピューティングの連続性：コアからファーエッジまで

NVIDIA AI Enterprise を基盤とし、ハードウェアによるネットワークとセキュリティの高速化を実現する NVIDIA Blackwell アーキテクチャおよび NVIDIA BlueField DPU（データ処理ユニット）を活用することで、Akamai はエッジとコアの全拠点にわたり複雑な SLA を管理できます。

- **エッジ（4,400 拠点以上）：**フィジカル AI や自律型エージェントに迅速なレスポンスを提供します。セマンティックキャッシングや、WebAssembly ベースの Akamai Functions、EdgeWorkers といったサーバーレス機能を活用し、ユーザーとの接点でモデルの親和性と安定したパフォーマンスを実現します。
- **Akamai Cloud IaaS と専用 GPU クラスター：**コアとなるパブリッククラウドインフラは、大規模ワークロードにポータビリティとコスト削減をもたらします。NVIDIA RTX PRO 6000 Blackwell GPU を搭載したポッドは、高負荷のポストトレーニングやマルチモーダル推論といった AI ワークロードを支えます。

NVIDIA の Global VP - Business Development - Telco である Chris Penrose 氏は「新しい AI ネイティブアプリケーションは、地球規模のスケールで予測可能なレイテンシー（遅延）と優れたコスト効率を求めています。NVIDIA AI Grid を実用化することで、Akamai は生成 AI、エージェント型 AI、フィジカル AI をつなぐ接続基盤を構築し、インテリジェンスをデータのある場所に直接近づけることで、リアルタイムアプリケーションの次代の波を解き放とうとしています」と述べています。

リアルタイム AI の次代の波を牽引

Akamai Inference Cloud は、コンピューティング負荷が高くレイテンシーに敏感な業界において、すでに強力な初期導入が進んでいます。

- **ゲーム：**AI による NPC やリアルタイムのプレイヤーインタラクションのために、50 ミリ秒未満の推論を導入。
- **金融サービス：**顧客がログインする重要な瞬間に、高度にパーソナライズされたマーケティングや迅速なレコメンドを提供。
- **メディアと動画：**AI によるトランスコーディングやリアルタイム吹き替えに分散ネットワークを活用。
- **小売およびコマース：**店舗内 AI アプリケーションや POS システムでの生産性向上ツールに採用。

また、エンタープライズからの需要に後押しされ、主要なテクノロジープロバイダーとの間で、メトロエッジの AI 専用データセンターにおける数千基の GPU クラスターに関する 4 年間で 2 億ドルのサービス契約も締結されています。

AI ファクトリーの拡張：集約型から分散型へ

AI インフラの第一波は、トレーニングに最適化された、一握りの中央集約型拠点に設置された大規模 GPU クラスターによって定義されました。しかし、推論がワークロードの主流となり、あらゆる業界で AI エージェントの構築が焦点となるにつれ、その集約型モデルは、かつてメディア配信やオンラインゲーム、金融取引、そして複雑なマイクロサービスアプリケーションなどでインターネットインフラが直面したのと同様のスケーリングの制約に直面しています。

Akamai は、分散ネットワーク、インテリジェントオーケストレーション、そしてコンテンツとコンテキストをユーザーに最も近い場所に届ける専用システムという、これまでと同じ基本的なアプローチでこれらの課題を解決しています。その結果、このモデルを採用した企業では、ユーザー体験の向上と ROI の改善が実現されています。

Akamai Inference Cloud は、この実績あるアーキテクチャを AI ファクトリーにも適用し、コアからエッジまで高密度のコンピューティングを分散させることで、次なる成長とスケールを可能にします。

これは企業にとっては、コンテキストを理解しながら柔軟に応答する AI エージェントをデプロイできるようになることを意味します。業界全体にとっては、AI ファクトリーが孤立した設備からグローバルに分散されたユーティリティへと進化する道筋を示す青写真となります。

提供開始時期

Akamai Inference Cloud は、本日より法人のお客様を対象に提供を開始します。詳細およびアクセスリクエストについては、https://www.akamai.com/ja/products/akamai-inference-cloud-platform をご覧ください。

また、2026 年 3 月 16 日～19 日にサンノゼ・コンベンションセンターで開催される NVIDIA GTC 2026 の Akamai ブース（Booth 621）にて、デモンストレーションを実施いたします。

Akamai について

Akamai は、オンラインビジネスの力となり、守るサイバーセキュリティおよびクラウドコンピューティング企業です。当社の市場をリードするセキュリティソリューション、優れた脅威インテリジェンス、グローバル運用チームによって、あらゆる場所でエンタープライズデータとアプリケーションを保護する多層防御を利用いただけます。Akamai のフルスタック・クラウド・コンピューティング・ソリューションは、世界で最も分散化されたプラットフォームで高いパフォーマンスとコストを実現しています。多くのグローバル企業が、ビジネスの成長に必要な業界最高レベルの信頼性、拡張性、専門知識を提供できる Akamai に信頼を寄せています。詳細については、akamai.com および akamai.com/blog をご覧いただくか、X や LinkedIn で Akamai Technologies をフォローしてください。