2025 年 4 月 8 日

Press Release

アカマイ・テクノロジーズ合同会社

Akamai、Akamai Cloud Inference により AI 分野での競争力を強化

従来のハイパースケーラーのインフラと比較し、スループットが 3 倍に向上、
レイテンシーを 60% 低減、コストを 86% 削減

※本リリースは 2025 年 3 月 27 日 (現地時間) 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とクラウドコンピューティングの企業、Akamai Technologies（NASDAQ：AKAM）は、Akamai Cloud Inference を発表しました。このサービスにより、予測的な大規模言語モデル（LLM）を実世界で実用化することを目指している組織のイノベーションを、より迅速かつ効率的に推進することができます。Akamai Cloud Inference は Akamai Cloud 上で実行され、世界で最も分散化されたプラットフォームを基盤に集中型クラウドモデルにおける様々な制約に対処します。

Akamai の Cloud Technology Group の Chief Operating Officer 兼 General Manager である Adam Karon は「AI データをユーザーやデバイスに近づけるのは困難であり、従来のクラウドではこれが大きな課題となっています」「LLM のトレーニングという負荷のかかる処理作業は今後も大規模なデータセンターで行われますが、実用的な推論（Inferencing）処理はエッジで行われます。Akamai は過去 25 年間にわたってエッジでプラットフォームを構築しており、エッジは Akamai と市場の他のすべてのクラウドプロバイダーを分ける差別化要素であると同時に、AI の将来に不可欠なものでもあります」と述べています。

Akamai Cloud 上での AI 推論技術

Akamai の新しいソリューションは、プラットフォームエンジニアや開発者が AI アプリケーションやデータ集約型ワークロードをエンドユーザーの近くで構築および実行できるようにするツールを提供し、スループットを 3 倍向上させながら、レイテンシーを最大 60%削減（2.5 倍高速化）します。Akamai のソリューションを活用することで、従来のハイパースケールインフラと比較して、AI 推論およびエージェント型 AI のワークロードを最大 86% 削減できます。Akamai Cloud Inference には次の機能が含まれています。

- コンピューティング：Akamai Cloud は、従来の CPU によるファインチューニングされた推論から、GPU による強力な高速コンピューティングオプション、AI 推論のさまざまな課題に最適な処理能力を提供するためにカスタマイズされた ASIC VPU まで、汎用性の高いコンピューティングを提供します。Akamai は、Triton、TAO Toolkit、TensorRT、および NVFlare を活用して NVIDIA の AI Enterprise エコシステムと統合し、NVIDIA GPU での AI 推論のパフォーマンスを最適化します。

- データ管理：Akamai は、最新の AI ワークロードに特化して構築された最先端のデータファブリックにより、AI 推論の可能性を最大限に引き出すことを可能にします。Akamai は VAST Data と提携し

て、リアルタイムデータへのアクセスを合理化し、推論関連のタスクを高速化します。これは、より関連性の高い結果と応答性の高い体験を提供するうえで欠かせません。これに加えて、スケーラビリティが高いオブジェクトストレージにより、AI アプリケーションに不可欠なデータセットのボリュームと多様性を管理し、Aiven や Milvus などの主要なベクトル・データベース・ベンダーと統合して、検索拡張生成（RAG）を可能にします。このデータ管理スタックを使用して、Akamai はファインチューニングされたモデルデータとトレーニングアーティファクトを安全に保存し、低レイテンシーの AI 推論をグローバル規模で実現します。

- コンテナ化：AI ワークロードをコンテナ化することで、需要ベースのオートスケール、アプリケーションの耐障害性の向上、ハイブリッド／マルチクラウドのポータビリティを実現しながら、パフォーマンスとコストの両方を最適化できます。Kubernetes を使用することで、Akamai はペタバイト規模のパフォーマンスにおいて、より高速で低コストかつ安全な AI 推論を実現します。Linode Kubernetes Engine（LKE）Enterprise（大規模なエンタープライズワークロード向けに設計された Akamai Cloud の Kubernetes オーケストレーションプラットフォーム）と Akamai App Platform に支えられている Akamai Cloud Inference は、KServe、Kubeflow、SpinKube などのオープンソース Kubernetes プロジェクトの AI 対応プラットフォームを迅速に展開し、シームレスに統合して、推論用 AI モデルの展開を合理化することができます。

- エッジコンピューティング：開発者による AI 搭載アプリケーションの構築を簡素化するために、Akamai AI 推論には WebAssembly（Wasm）機能が含まれています。Akamai は、Fermyon などの Wasm プロバイダーと協力し、開発者が LLM 向けの推論をサーバーレスアプリケーションから直接実行できるようにします。これにより、顧客はエッジで軽量コードを実行し、レイテンシーの影響を受けやすいアプリケーションで有効化することができます。

これらのツールを組み合わせることで、低レイテンシーの AI 搭載アプリケーション向けの強力なプラットフォームが構築され、企業はユーザーが求める体験を提供できます。Akamai Cloud Inference は、データ集約型ワークロードに 1 ペタバイト／秒以上のスループットを一貫して提供できる、Akamai の大規模な分散型プラットフォーム上で動作します。世界 130 か国以上における 1,200 以上のネットワークにまたがる 4,200 以上の接続点で構成される Akamai Cloud は、クラウドからエッジまでのコンピューティングリソースを提供し、アプリケーションのパフォーマンスを加速し、スケーラビリティを高めます。

トレーニングから推論への移行

AI の導入が進むにつれて、LLM に関する過剰な宣伝によって、具体的なビジネス上の課題を解決するための実践的な AI ソリューションという点に焦点が当たっていないことを企業は認識しつつあります。LLM は、要約、翻訳、カスタマーサービスなどの汎用タスクに優れていますが、これらのモデルは非常に大規模で、学習には高額なコストと時間を要します。多くの企業は、データセンターや計算能力、しっかり構築された安全でスケーラブルなデータシステム、そして場所やセキュリティの要件により課される意思決定の遅延に関する課題など、アーキテクチャやコストの要件による制約を受けています。具体的なビジネス上の課題に対処するように設計された軽量な AI モデルは、個々の業界向けに最適化でき、独自のデータを使用して測定可能な成果を生み出し、今日の企業にとって優れた投資利益率を実現できます。

AI 推論には、さらに分散化されたクラウドが必要

データセンターやクラウドの一元的な領域の外でデータが生成されることが増えています。このような変化により、発生源に近い場所でのデータ生成を活用する AI ソリューションへの需要が高まっています。企業は LLM の構築やトレーニングを超えて、データを利用したより迅速でスマートな意思決定の実現や、よりパーソナライズされた体験への投資に向かい、それに伴ってインフラのニーズが根本的に変化します。企業は、AI を活用してビジネスの運営とプロセスを管理および改善することで、より多くの価値を生み出すことができると認識しています。運用インテリジェンスのユースケースにおける望ましい選択肢として、分散型クラウドやエッジアーキテクチャが登場しています。リモート環境でも分散されたアセット全体にわたってリアルタイムで実用的な知見を提供できるからです。Akamai Cloud の顧客事例には、車載音声アシスタンス、AI を活用した作物管理、コンシューマー向けマーケットプレイスの画像最適化、衣服を視覚化するバーチャルショッピング体験、自動化された製品説明ジェネレーター、顧客フィードバックの心理分析ツールなどがあります。

Karon は「LLM のトレーニングは地図をつくる作業に似ており、データの収集や地形の分析、道の描画が必要です。そのためには時間とリソースがかかりますが、一度構築してしまえば非常に便利に使えます。AI 推論は GPS の利用に似ており、その構築済みの知識を利用して、リアルタイムに再計算し、変化に応じて目的地に到達できます」「推論は AI の次の最先端領域です」と説明します。

# # #

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。