2024 年 9 月 11 日

【報道関係各位】

本リリースは 2024 年 6 月 6 日に本社 G-Core Labs S.A により発表されたリリースの抄訳に 2024 年 9 月 1 日時点での付加情報を加えたものです。

AI ソリューション — Gcore Inference at the Edge — GA リリース（2024 年 9 月 1 日）
ML モデルのグローバルエッジデプロイを高速・セキュア・最適コスト効率で実現

ルクセンブルク：2024 年 6 月 6 日 — AI、クラウド、ネットワーク、セキュリティのグローバルエッジソリューションプロバイダ Gcore は、本日 Gcore Inference at the Edge を発表しました*。Gcore Inference at the Edge は、AI アプリケーションの超低遅延体験を提供する画期的ソリューションです。事前学習済み ML（Machine Learning；機械学習）モデルのエッジ推論ノードへの分散デプロイを可能にし、シームレスかつリアルタイム推論の実行を実現します。
* 2024 年 9 月 1 日付けで GA リリースとなりました。

Gcore Inference at the Edge は、自動車、製造、リテール、テクノロジを含む多種多様な業界の企業組織の AI モデルのデプロイを、コスト、スケーラビリティ、セキュリティの観点で望ましいデプロイを実現します。グローバル規模の活用シーンとしては、生成 AI、オブジェクト認識、リアルタイムビヘイビア分析、ビジュアルアシスタント、プロダクションモニタリング等が挙げられます。

Gcore Inference at the Edge は、低遅延スマートルーティングテクノロジで相互接続された 180 超エッジノードで構成された Gcore グローバルネットワーク上で稼働します。それぞれのハイパフォーマンスノードは、サーバを戦略的にエンドユーザに近い地点に配備する Gcore ネットワークのエッジに配置されています。市場をリードする AI 推論向け設計がなされている NVIDIA L40S GPU をチップとして採用、ユーザリクエスト送信がされると、エッジノードが最低遅延の最至近推論リージョンを選択し、30ms 未満の平均レスポンスタイムを実現しつつルートします。

このソリューションではファンダメンタル ML およびカスタムモデルをサポートしています。Gcore ML Model Hub に含まれるオープンソースのファウンデーションモデルは、LLaMA Pro 8B、Mistral 7B、Stable-Diffusion XL です。モデルが選択可能で、また、Gcore Inference at the Edge グローバルノードへの送信前のモデル事前学習は非依存型であるため、様々な活用シーンへの展開が実現します。加え、開発チームがこれまで直面してきた課題、事前学習を実施した同一サーバ上で AI モデルが動作することに起因するパフォーマンス不足に対する解決策をもたらします。

Gcore Inference at the Edge の特長：

- コスト効率に秀でたデプロイメント：柔軟な価格体系で使用したリソース分のみの支払い
- ビルトイン DDoS プロテクション：Gcore インフラストラクチャで ML エンドポイントを DDoS 攻撃から自動プロテクト
- 傑出したデータプライバシーとセキュリティ：GDPR、PCI DSS、ISO／IEC 27001 基準に準拠したソリューション
- モデルオートスケーリング：オートスケーリングでロードスパイクを処置、ピークデマンドと予測外のサージにモデル側で常時対応可能
- 無制限オブジェクトストレージ：モデル進化に呼応するスケーラブルな S3 互換クラウドストレージ

Gcore CEO　Andre Reitenbach コメント

Gcore Inference at the Edge で目指すのは、顧客の方々に、AI アプリケーションのグローバルデプロイに必要なコスト、スキルやインフラストラクチャではなく、ML モデルを学習済みにするという本来の目的に注力していただくことです。Gcore では、エッジこそが、パフォーマンス体験およびエンドユーザ体験、双方の最大化が可能な領域であると考えており、これまでになかったスケールとパフォーマンスを顧客の方々に享受していただくべく、継続改革を推し進めています。Gcore Inference at the Edge は、障壁なきパワーを具現化し、先進性・効率性・有益性を伴う AI 推論体験を提供するものなのです。

Gcore Inference at the Edge 情報；
[製品 Web ページ] https://gcore.com/inference-at-the-edge
※2024 年 9 月 1 日に GA（General Availability）リリース、価格情報：上記ページに近日掲載予定
[参考記事～ベータ版リリース時]
Web News：https://gcore.com/news/meet-inference-at-the-edge/
Web Learning：https://gcore.com/learning/ai-iate-transforming-industries/

Gcore について ～ Gcore は 2024 年 2 月に 10 周年を迎えました（参考英文記事；Blog, Linkedin Post）。
Gcore はエッジ AI、クラウド、ネットワーク、セキュリティのグローバルソリューションプロバイダです。本社はルクセンブルク、600 超の従業員と世界各地に 10 の営業拠点を擁しています。Gcore の IT インフラストラクチャは自社運用、拠点は 6 大陸にまたがり、グローバル平均レスポンスタイムは 30ms、ヨーロッパ、アフリカ、LATAM における屈指のネットワークパフォーマンスを実現しています。ネットワークは Tier IV と Tier III のデータセンタに配備された世界各地の 180 超の PoP で構成され、200 超 Tbps の帯域を誇ります。
Web サイト
https://gcore.com/
ソーシャルメディア
https://www.linkedin.com/company/g-core/
https://www.youtube.com/@GCoreOfficial
https://www.facebook.com/officialgcore
https://x.com/gcore_official
https://www.instagram.com/gcore.official/

同件に関するお問い合わせ先
Gcore Japan 株式会社
Marketing Manager：白石
tel:03-4567-2817/email: Japan-marketing@gcore.com

共同ピーアール株式会社
担当：栗木、峰松
email: Gcore-pr@kyodo-pr.co.jp