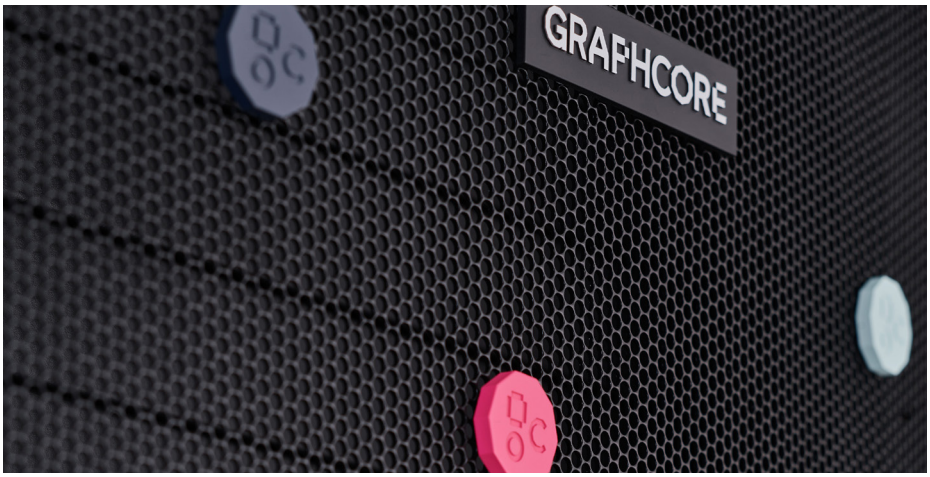
GRAPHCORE

構築する：
IPU-POD$_{64}$



機械知能をスケールアウトするためのビルディングブロック

IPU-POD™システムは、大規模で要求の厳しい機械学習モデルを加速し、柔軟かつ効率的にスケールアウトできるように設計されています。

IPU-POD$_{64}$は、革新的なGC200インテリジェンス処理ユニット（IPU）をベースにした、16枚のIPU-M2000™コンピュートブレードを搭載するシングルラック構成の製品です。IPU-POD$_{64}$は最大で16ペタフロップスのAI演算能力を発揮します。

ハードウェアとソフトウェアを含めたシステム全体が一緒になって設計されているIPU-POD$_{64}$は、標準的なフレームワークやプロトコルをサポートしているので、既存のデータセンター環境にスムーズに統合できます。これによりイノベーターは、最先端の性能の恩恵を受けながら、使い慣れたツールを使ってAI作業負荷を大規模に展開することに集中できます。

分離による演算のカスタマイズ

機械知能の作業負荷の演算需要は実に様々です。実稼働展開では、AIとホスト演算の比率を最適化することで、性能と効率を最大化しながら、一方で総所有コストを軽減できます。IPU-PODシステムでは、必要な数のIPU-M2000プラットフォームにサーバーやスイッチの数を柔軟にマッピングできるので、実稼働のAI作業負荷に合わせて展開を調整できます。IPU-POD$_{64}$は1～4台のサーバー構成に対応しています。

スケーリングのために構築された通信アーキテクチャ

データのアクセスと転送を効率的に行うことで、より高いAI性能を引き出すことができます。IPU-Fabric™はシステム全体のデータ転送のために作られた革新的な通信アーキテクチャで、個々のIPU内やIPU-M2000間、IPU-POD間、そしてデータセンター全体に高速な相互接続を拡張します。データセンターの標準的な通信技術と連携するように作られているIPU-Fabricを利用することで、高性能な低遅延通信でAIアプリケーションの効率を最大化できます。

システムの仕様

IPU	64 x GC200 IPU
IPU-M2000	16 x IPU-M2000
IPU Core	94,208
スレッド	565,440
性能	16ペタフロップスFP16.16 4ペタフロップスFP32
交換メモリ	最大4.15TB（57.6GBのインプロセスメモリと4.1TBのストリーミングメモリを含む）
IPU-Fabric	2.8Tbps
ホストリンク	100 GE RoCEv2
ソフトウェア	Poplar® SDK
システム重量	450 kg + ホストサーバーとスイッチ
システム寸法	16U + ホストサーバーとスイッチ
ホストサーバー	Graphcoreパートナーが提供する承認済みホストサーバーから選択
ストレージ	Graphcoreパートナーが提供する承認済みソリューションから選択
温度制御	空冷式

GRAPHCORE

AI開発者向けに構築

IPU-PODシステムは業界標準のソフトウェアツールをサポートしています。開発者はTensorFlow、PyTorch、PyTorch Lightning、Kerasなどのフレームワークや、ONNXなどのオープンスタンダード、Hugging Faceなどのモデルライブラリを使用できます。

より踏み込んだ制御と最高の性能を実現するために、PoplarのフレームワークではpythonとC++での直接IPUプログラミングが可能になっています。Poplarがあれば、開発を複雑にせずに多くのIPU全体でモデルを簡単にスケールアップできるので、開発者はアプリケーションの精度と性能に集中できます。

Graphcoreの製品は、イノベーションにつながる力をAI開発者に提供することを念頭に開発されています。私たちのソフトウェアスタックは業界のオープンスタンダードをサポートしており、そのうえオープンソースです。

AI専門家へのアクセス

Graphcoreは、IPU-PODの設置やアプリケーション開発から実稼働展開までをトータルにサポートするパートナーのネットワークを世界規模で展開しています。ドキュメンテーションなどの資料については、当社ウェブサイトをご覧ください。

簡単な展開

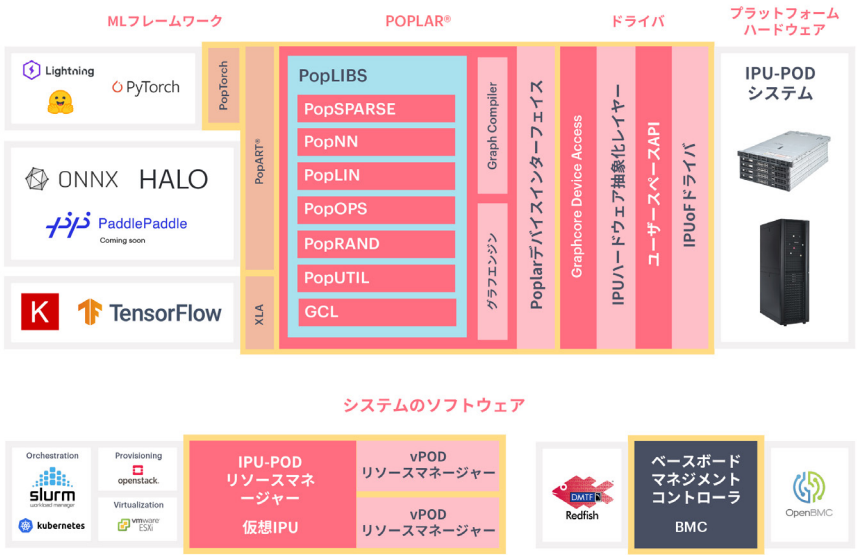
IPU-PODの設計で最優先に検討されたのは、展開のしやすさです。その結果、標準的なハードウェアやソフトウェアのインターフェースとプロトコルをサポートし、既存のデータセンターのインフラストラクチャと効果的に統合するソリューションが生まれました。

IPU-PODは、OpenBMCやRedfish DTMF、IPMI over LAN、Prometheus、Grafanaなど、業界標準のオープンソースソフトウェアやオープンAPIをベースにした、管理や可視化のためのソフトウェアやツールの多くをサポートしています。

業界で実績のある管理ツール

DockerとKubernetesをサポートしているので、アプリケーションの展開やスケーリング、IPU-PODの管理を簡単に自動化できます。またVirtual-IPU™技術により、IPUを様々なテナントや作業負荷に安全にプロビジョニングできます。これにより開発者は、複数のIPU-PODの内部と全体の両方でモデルレプリカを構築し、大規模なモデルの多くのIPU-POD全体でIPUをプロビジョニングできます。

IPU-PODには、IPUリソースの管理を簡素化するための使いやすい直感的なウェブGUIがあります。これによりエンジニアは、ステータスの管理やシステムテストの実行、作業負荷に対するIPUのプロビジョニングを行えます。またIPU-PODは、VMWareのRadiumをはじめとする様々なクラウドプロビジョニングおよび管理スタックと統合できます。



機械知能で次のレベルを体験する用意はできていますか？

AIインフラストラクチャの要件とソリューションの適合性を評価するため、以下の当社パートナーとつながってください。
ご不明な点がまだありますか？Graphcoreまで直接ご連絡ください。info@graphcore.ai

GRAPHCORE.AI