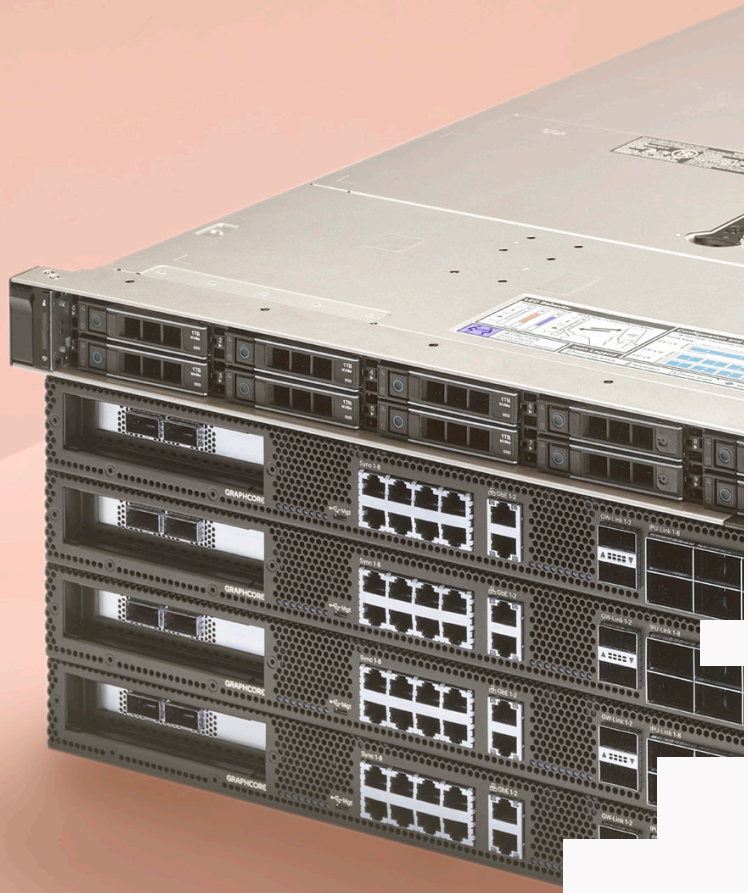
GRAPHCORE

BOW POD$_{16}$



探求 | 構築 | 成長

Bow Podシステムは、機械知能を大規模に展開するための高い性能と効率を実現します。このシステムは、今日の大規模かつ複雑なモデルを加速させるとともに、イノベーターが未来のソリューションを探求し、発明するためのプラットフォームを提供するために設計されています。

Bow Pod$_{16}$システムは探求に最適なシステムです。プロトタイピングの段階で必要なパワー、パフォーマンス、柔軟性をすべて提供し、迅速に実稼働に移行できます。言語や画像認識などのあらゆる機械学習分野において、より優れた革新的なAIソリューションをIPU上で構築したり、GNNやLSTMを探求したり、全く新しいものを創造したりするうえで、Bow Pod$_{16}$は理想的な出発点です。

最新世代のIPU

Bow Pod$_{16}$システムは4台のBow-2000マシンに、先駆的な新型Bow IPUプロセッサをそれぞれ4基搭載しています。この革新的なIPUは、WoW（Wafer-on-Wafer）技術を使って製造された世界初のプロセッサで、実績あるIPU技術の利点を次のレベルへと高めています。

性能と効率

最大5.6ペタフロップスのAI演算と業界トップの効率性を実現するBow Pod$_{16}$の根底には、革新的なシリコン技術の使用や、効率とスケールアウトを重視したコンピュート機能とメモリアーキテクチャ、ソフトウェアとアプリケーションを第一にしたソリューション展開があります。

スムーズな展開と短期間での市場投入

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Bow Pod$_{64}$は、標準的なフレームワークとプロトコルをすべてサポートしており、既存のデータセンター環境はもちろん、プライベートクラウドやパブリッククラウドにも簡単に統合できます。

システムアグリゲーターが構成と展開に要する時間を短縮できるうえ、AI向けに設計された、市場をリードするサーバープラットフォームと高性能ストレージアプライアンスを幅広くテスト・検証することで選択肢を用意しています。

またイノベーターであれば、最先端の性能と効率を引き出しながら、使い慣れたツールやフレームワークを使ってAI作業負荷を大規模に展開することに集中できます。

システムの仕様

プロセッサ	Bow IPU x 16
1Uブレードユニット	Bow-2000マシン × 4
分割コア	23,552
スレッド	141,312
性能	5.6ペタフロップスFP16.16 1.4ペタフロップスFP32
メモリ	14.4 GB In-Processor-Memory™ 最大1 TBのStreaming Memory™
ソフトウェア	Poplar® SDK

ホストリンク	100 GE RoCEv2
システム重量	66 kg + ホストサーバー
システム寸法	4U + 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システムから選択
ストレージ	Graphcore®パートナーが提供する承認済みシステムから選択
温度制御	空冷式

BOW POD16

分離による演算のカスタマイズ

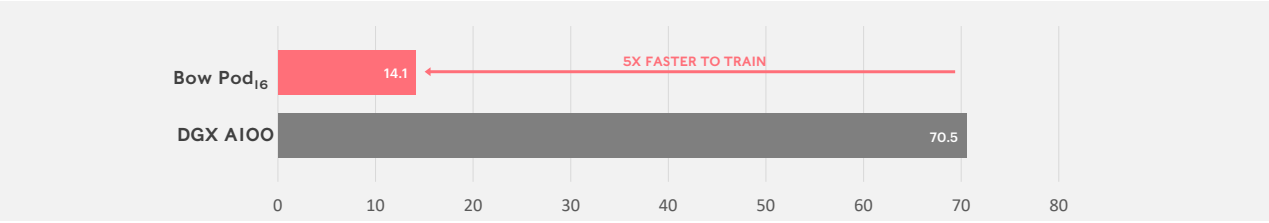
機械知能の作業負荷の演算需要は実に様々です。実稼働展開では、AIとホスト演算の比率を最適化することで、性能と効率を最大化しながら、一方で総所有コストを軽減できます。Bow Podシステムでは、必要な数のBow-2000マシンにサーバーやスイッチの数を柔軟にマッピングできるので、実稼働のAI作業負荷に合わせて展開を調整できます。Bow Pod$_{16}$は複数のサーバー構成に対応しています。

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Bow IPU内やBow-2000間、Bow 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。

AI開発者向けのプラットフォーム

TensorFlowやPyTorch、PaddlePaddle、その他の人気の高いMLフレームワークは、コミュニティ推進型の連携および革新のために、包括的なPopLibs™ライブラリとともに、オープンソースとしてサポートされ、利用可能になっています。完全制御して最大性能を引き出したい開発者のために、GraphcoreのPoplar SDKはC++での直接IPUプログラミングを可能にします。



EfficientNet-B4 Training Time To Train Performance
IPU-POD Platforms | Preliminary Results (Pre-SDK2.5) | G16-EfficientNet-B4 Training
DGX A100 (A100-SXM4-80GB) | TensorFlow | Mixed Precision | https://developer.nvidia.com/deep-learning-performance-training-inference

大規模展開を前提にした設計

Poplar SDKツールとフレームワークイメージを備えたプリビルドDockerコンテナなので、イノベーターはすばやく起動し、実行することができます。また、SlurmやKubernetes、OpenStackなど、コンテナのオーケストレーション、プラットフォームの可視化、プロビジョニングのための一般的な各種フレームワークがサポートされています。

ソフトウェア第一

完全統合で、IPUの最適化されたPoplarソフトウェアは、IPUアーキテクチャ固有の特性を活用して、比類なき性能と柔軟性を備えたAIアプリケーションを構築します。Poplarがあれば、開発を複雑にせずにモデルのスケールを1から数千へとアップできるので、イノベーターはアプリケーションの精度と性能に集中することができます。

AI専門家へのアクセス

インストール、実稼働、アプリケーション開発についての豊富な経験とサポートは、GraphcoreのAI専門家とエリートパートナーネットワークからグローバルに利用できます。

ご用意はよろしいですか？

グラフコアの専門技術者と連絡を取り、お客様のAIインフラストラクチャに関する要求とソリューションの適合性を評価してみましょう。info@graphcore.aiまでお問い合わせください。