インテル® デベロッパー・® クラウドの事例：Supermicro のサーバーで AI コンピューティングを高速化

インテル ® Xeon® プロセッサーとインテル ® Gaudi® 2 AI アクセラレーターを搭載した Supermicro の高度な AI サーバーで、AI の計算、トレーニング、推論を高速化。

Supermicro SYS-820GH-TNR2

インテル ® Gaudi® 2 AI アクセラレーター

業界

- クラウドサービス・テクノロジーのプロバイダー

課題

- 高性能サーバー
- AI トレーニングに必要な消費電力の削減

概要

インテルは、現代のデータセンター、ワークステーション、パーソナルコンピューター市場を支える最先端のハードウェアおよびソフトウェアテクノロジーの世界的リーダー企業です。同社は、高性能、高効率が求められる生成 AI トレーニングと推論の需要に応えるための新世代のアクセラレーターを開発しました。新たに AI ソリューションのポートフォリオに加わったインテル ® Gaudi® 2 AI アクセラレーターは、消費電力を抑えながら優れた AI パフォーマンスを実現することができます。

インテル ® デベロッパー・® クラウドは、柔軟でスケーラブルな AI インフラストラクチャを提供するプラットフォームで、これにより企業と開発者はモデルのトレーニングと推論のワークロードを高速化することができます。このクラウドには、CPU と各種アクセラレーター機能を備えたシステムが備わっています。

一方で、インテル ® Xeon® プロセッサーおよび インテル ® Gaudi® 2 AI アクセラレーターによるオンデマンド・インスタンスを提供する Supermicro のサーバーは、開発者が実稼働環境でモデルとアプリケーションを開発、最適化、テスト、デプロイできるように設計されています。インテル ® デベロッパー・® クラウドの最大の特徴のひとつは、Supermicro のサーバーによる大規模なクラスターによって構成されており、開発者が高性能な AI システムを体験できるということです。

課題

インテル ® デベロッパー・® クラウドの設計には、最も要求の厳しい AI アプリケーションの実行を可能にするサーバーが必要でした。それも、開発者が大規模な AI コンピューティングにアクセスするための大量のサーバーです。また、サーバーには、要求の厳しい生成 AI トレーニングと推論ワークロードを実行し、複数のシステムが連携して最大の AI モデルをトレーニング、デプロイできるような

ソリューション

- **デュアル 第 4 世代 インテル ® Xeon® スケーラブル・プロセッサー**
- **8x インテル ® Gaudi® 2 AI アクセラレーター**
- **2TB DDR4 Memory**
- **6x 400 GbE QSFP Connectors**

拡張性が求められ、さらには最大 8TB のメモリーのサポートと、インテル ® アドバンスト・マトリクス・エクステンション、インテル ® In-Memory Analytics Accelerator など、Xeon® の高度な AI 機能があることが条件でした。

ソリューション

インテルは、デュアル 第 4 世代 インテル ® Xeon® スケーラブル・プロセッサーと 2TB のメモリーを搭載した Supermicro の SYS-820GH-TNR2 サーバーを採用しました。各システムには、最大モデルのトレーニングとデプロイができる 8 つのインテル ® Gaudi® 2 AI アクセラレーターが搭載されています。7nm プロセステクノロジーで製造された各インテル ® Gaudi® 2 AI アクセラレーターは、チップ上に 24 個の 100 Gb イーサネットポートが統合されており、インテル ® Gaudi® 2 AI アクセラレーターのコンピューティングエンジンには、24 個の AI カスタム Tensor コア、デュアルマトリックス乗算エンジン、96 GB の HBM2E メモリー、48 MB の SRAM が備わっています。

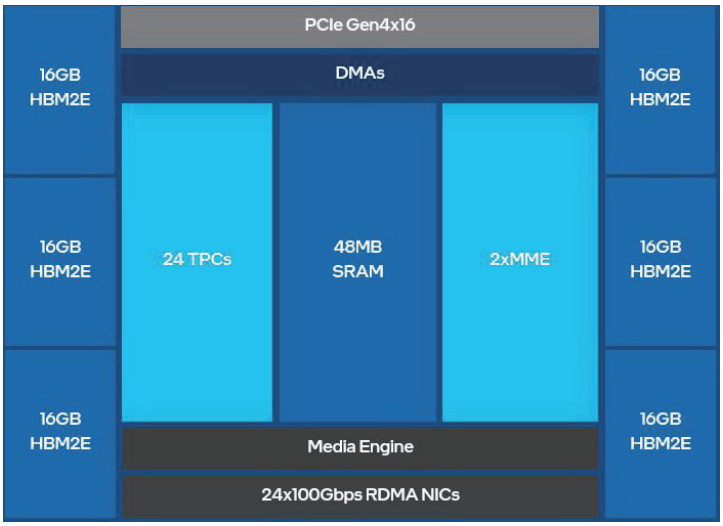
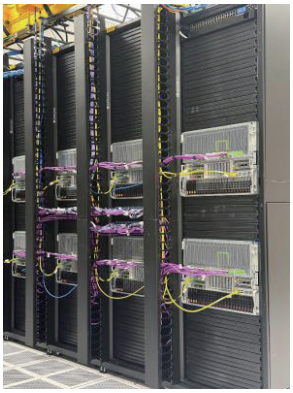


図 1 インテル ® Gaudi® 2 AI アクセラレーターのアーキテクチャ

インテル ® Gaudi® 2 AI アクセラレーターによるスケーリング

業界標準のイーサネットをインテル ® Gaudi® 2 AI アクセラレーターに統合することで、1 ノードから数千ノードまでの柔軟で効率的なスケールアップとスケールアウトが可能になり、近年の生成 AI の需要に応えることができます。インテルは、システム全体でオープンスタンダードの 400 ギガビット・イーサネットスイッチを使用し、新しい Supermicro の AI サーバーをインテル ® デベロッパー・® クラウドに迅速かつ効率的に統合しました。これにより、開発者と運用ユーザーは、シンプルなイーサネット接続で AI トレーニングを高レベルに拡張することができます。

Supermicro はまた、1 ～ 8 台の Intel® Gaudi® 2 ベースのサーバーを使用して MLPerf v3.0 BERT ベンチマークをテストしました。最大 8 台のサーバーを使用した場合でも、パフォーマンスは非常に線形に近く、優れたスケールアーキテクチャを示しました。インテル ® Gaudi® 2 AI アクセラレーターをベースとした Supermicro のサーバーは、MLPerf ベンチマークと推論ベンチマークの両方で強力なパフォーマンスを実証しました。



インテル ® Gaudi® 2 AI アクセラレーターを搭載した Supermicro のサーバー

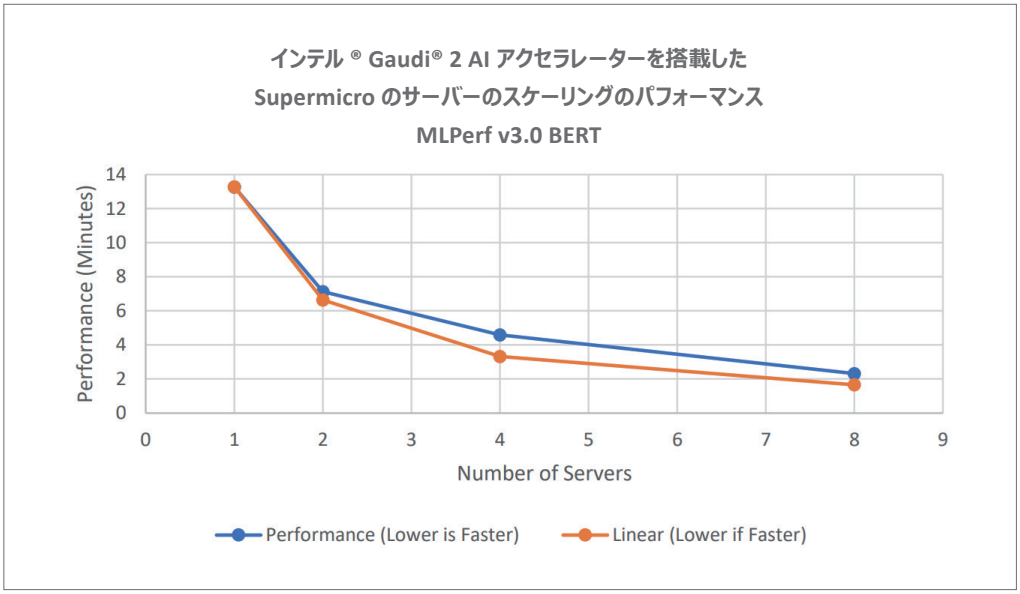


図 2 インテル ® Gaudi® 2 AI アクセラレーターを搭載した複数のサーバーを使用した場合のパフォーマンス

利点

- トレーニングと推論にかかる時間に対する高いパフォーマンス
- 業界標準のネットワーク・ファブリックに基づく効率的な拡張性
- オープンソースソフトウェアでの開発が可能（PyTorch と Hugging Face）

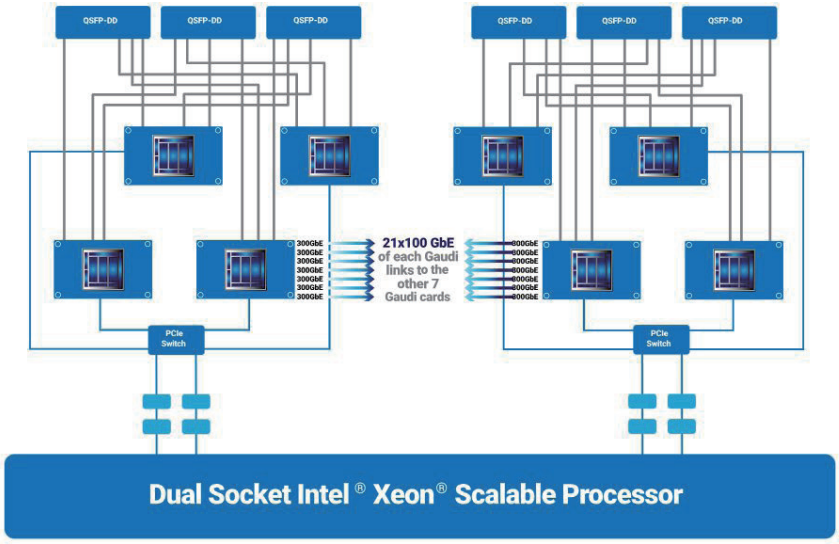


図 3 インテル ® Gaudi® 2 AI アクセラレーターを搭載したサーバーの仕組み

「当社はインテル ® デベロッパー・® クラウドの設計に Xeon® プロセッサーとインテル ® Gaudi® 2 AI アクセラレーターを搭載した Supermicro のサーバーを選択しました。その理由は、Supermicro のサーバーの AI の利用における強力なパフォーマンス、効率性、拡張性です。当社のユーザーは、この最先端のプラットフォームを活用することで、業界をリードするオープンソースモデルと柔軟なスケールアウトでの開発を体験できるでしょう。当社はこれからも Supermicro と協力し、さらに大規模なスケーリングオプションを提供できるよう、インテル ® デベロッパー・® クラウドの次世代サーバーに期待しています。」

- Markus Flierl, Corporate Vice President of Intel® Developer Cloud

Supermicro の GPU サーバー：
https://www.supermicro.com/en/products/system/ai/8u/sys-820gh-tnr2

インテル ® デベロッパー・® クラウド：
https://www.intel.com/content/www/us/en/developer/tools/devcloud/overview.html

Supermicro について

Supermicro は、高性能なグリーンコンピューティング・サーバーを提供する世界的リーダー企業です。当社は、ブレード、ストレージ、GPU ソリューションでカスタマイズし、アプリケーション用に最適化されたサーバーとワークステーションを世界中の顧客に提供しています。当社の製品は、実証済みの信頼性、優れた設計、業界で最も幅広い製品構成を提供し、あらゆるコンピューティングのニーズに応えています。詳しくは https://www.supermicro.com/ja/ をご覧ください。

インテルについて

インテルは業界のリーダーとして、世界中の進歩を促すとともに生活を豊かにする、世界を変えるテクノロジーを創出しています。ムーアの法則に着想を得て、顧客企業が抱える大きな課題を解決する半導体製品を設計・製造し、その進化に向けて日々取り組んでいます。クラウド、ネットワーク、エッジ、あらゆるコンピューティング機器のインテリジェント化によりデータの価値を最大化し、ビジネスと社会をより良く変革します。インテルのイノベーションについては、https://intel.co.jp をご覧ください。

お問い合わせ：スーパーマイクロ株式会社
〒 150-0031 東京都渋谷区桜丘町 20-1 渋谷インフォスタワー 21 階
電話：03-5728-5196　FAX：03-5728-5197　Email：Sales_Inquiry_JP@Supermicro.com