GPGPUソリューション

VCAE-2U04LA001-G

JCS
Japan Computing Systems

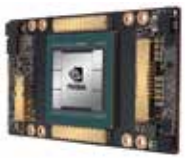
NVIDIA HGX A100 4基搭載 2Uラックマウント GPGPUモデル

2Uラックマウントシャーシに、4基のHGX A100を搭載可能な高密度ラックマウントGPGPUモデルです。GPU間を相互接続するNVLink、NVSwtichをサポートし、4基のA100 GPUは高速に相互接続されます。これによりボトルネックであった帯域幅の遅延を減らしディープラーニングのパフォーマンスを劇的に向上させます。

- NVIDIA HGX A100 4基搭載
- AMD 第3世代 EPYC プロセッサー 2基搭載
- DDR4 3200 ECC メモリ搭載
- 最大8TBまで拡張可能な大容量メモリ
- NVIDIA CUDA プリインストール

NVIDIA A100 搭載

NVIDIA A100 Tensor Core GPUはAIとHPC用途のためにデータセンター向けGPUとして開発されたVolta世代の後継として製造プロセスルール7nmのAmpereアーキテクチャを採用。第3世代のTensorコアを実装し、BFloat16、Tensor Float 32を新たにサポート、前世代のVoltaと比較しFP32演算、INT8のディープラーニング推論処理において20倍の性能を実現し、トレーニングと推論の両方が強化されています。
また、Multi-instance GPU(MIG)機能により単一のA100を最大7個のGPUとして扱うことが可能で複数のユーザーが独立したインスタンスとして利用できます。

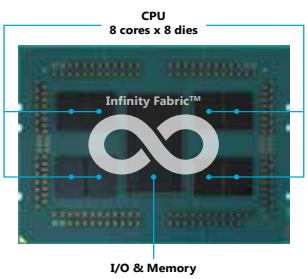


■NVIDIA HGX A100	
GPU アーキテクチャ	NVIDIA Ampere
フォームファクタ	SXM
FP64 CUDA コア	3,456
FP32 CUDA コア	6,912
Tensorコア	432
GPUメモリ	40GB HBM2 / 80GB HBM2
GPU間接続帯域	最大600 GB/sec
マルチインスタンスGPU	最大7分割
最大消費電力	400 W
主要アプリケーション実行性能	100%
冷却方式	パッシブ（冷却ファンなし）

AMD 第3世代 EPYC™ プロセッサー搭載

第3世代 EPYCは前世代より製造プロセス、最大コア数などに大きな変化はないものの、新しいアーキテクチャ「ZEN3」に刷新されており、第2世代 EPYCに採用されていた「Zen2」と比較すると、IPC（Instruction Per Clock-cycle）が最大19%ほど改善されています。
アーキテクチャが新しくなったことにより、8つのCPUコアが32MBのL3キャッシュを共有し1コアあたりが使えるL3キャッシュが増加しているほか、6チャンネルのメモリチャネルをサポート。これによりメモリ構成が柔軟に選択でき、導入におけるコスト削減に貢献しています。

このほか、セキュリティ面も強化されており「AMD Infinity Guard」と呼ばれるセキュリティ機能を搭載、新たに「SEV-SNP」「Shadow Stack」機能が追加され仮想マシン環境の安全性を引き上げています。



統合型開発環境CUDAプリインストール

NVIDIA社より提供される開発環境「CUDA」をインストールしています。最新のCUDAドライバのインストール、お客様ご指定のコンパイラ、数値演算ライブラリー式のインストール、CUDA SDKサンプルコードの実行検証を行うことで、直ちに並列アプリケーション等の開発が始められます。「CUDA」は、C言語等の標準のプログラミング言語に慣れたプログラマが、簡単に利用できるように設計されたNVIDIA社の並列プログラミングモデル及びソフトウェアです。従来のコンピュータ・グラフィックス用のシェーディング言語によるプログラミングに比べて、格段に効率良く開発することができます。

■仕様	
プロセッサー	AMD 第3世代 / 2世代 EPYC™ プロセッサー x2基
チップセット	System on Chip
メモリ	DDR4-3200 SDRAM ECC
メモリスロット	32スロット（最大8TB）
ストレージ	2.5インチ
ドライブベイ	4（ホットスワップ対応）SATA/NVMe/SAS
オプティカルドライブ	USB接続 外付けDVDマルチドライブ（オプション）
グラフィックス	onboard
GPU	NVIDIA HGX A100 40GB / 80GB x4基
ネットワークI/F	10GBase-T 2port (RJ45)
管理インターフェース	IPMI 2.0, KVM-over-LAN
I/Oポート	VGA x1（背面） USB 3.0 x2（背面）
拡張スロット	PCI-Express 4.0 x16（4スロット Low Profile） PCI-Express 4.0 x8（1スロット Low Profile）
ケース	2U ラックマウント　(外形寸法：W437mm x H89mm x D823mm)
電源	200V 1800W リダンダント電源（80-plus Titanium 高効率電源）
対応OS（別売）	● Ubuntu 64bit ※その他Linuxディストリビューションについてはご相談ください。
ソフトウェア	NVIDIA CUDA（デバイスドライバ、ツールキット、CUDA SDK開発環境、コンパイラ等）。 * Deep Learning各種フレームワークのインストールにつきましてはお問い合わせください。
Deep Learning用フレームワーク	DIGITS、TensorFlow、Chainer、Caffe、Pytorch、Docker
保守	1年間全国出張オンサイトサービス 3年間全国出張オンサイト（オプション）

株式会社日本コンピューティングシステム TEL.03（3821）3200 FAX.03（3821）3241　http://www.jcsn.co.jp