

[大規模科学計算システム]

LX 406Rz-2 のハードウェア

横山 淳 乾 重人 水子 陽一
日本電気株式会社 AI プラットフォーム事業部

1. はじめに

東北大学サイバーサイエンスセンターで本年 10 月から運用を開始するスーパーコンピュータ AOBA は、576VE(72VH) のベクトルスーパーコンピュータ SX-Aurora TSUBASA (AOBA-A) と、68 ノードのスカラ処理向け大規模並列コンピュータ LX 406Rz-2 (AOBA-B) が、InfiniBand HDR による高速インターコネクトで接続され、1.76PFlops の総合演算性能を有する最先端の大規模 HPC クラスシステムです。

このうち、サブシステム AOBA-B で採用されている LX 406Rz-2 は、最新の AMD EPYC™ プロセッサを搭載し、4.096TFlops の高いノード性能を実現するとともに、InfiniBand HDR(200Gbps)をサポートした高速な通信性能を併せ持つ高密度実装サーバです。本稿では、LX 406Rz-2 のハードウェア、AMD EPYC™ 7002 シリーズプロセッサならびにインターコネクト技術について紹介します。

2. ハードウェア構成

2.1 LX 406Rz-2 の特長

LX 406Rz-2 は、米 AMD 社の最新プロセッサ EPYC™ 7002 シリーズを搭載し、2U ラックマウント型の筐体に 4 ノードを収容した、高性能、高効率、高密度な HPC クラス型計算サーバです。LX 406Rz-2 は、80PLUS® Titanium Level に対応した最大 2200W の大容量電源ユニットを筐体あたり 2 基搭載し、効率的で信頼性の高いシステムを実現しています。AOBA-B では、コア数が最も多く、電力と性能のバランスがとれた EPYC™ 7702 を採用しました。表 1 に LX 406Rz-2 の主要諸元を示します。本サーバの主な特長は以下のとおりです。

- ・高さ 2U の筐体に 4 ノード搭載し、ノード単位でホットプラグが可能
- ・ノードあたり AMD EPYC™ 7002 シリーズプロセッサを 2 基搭載
- ・DDR4-3200 対応の DIMM を 1 ノードあたり 256GB 搭載
- ・PCI Express Gen4(x16)に対応した拡張 IO スロットを 1 ノードあたり 2 スロット搭載
- ・InfiniBand HDR (200Gbps)をサポート
- ・IPMI2.0(Intelligent Platform Management interface 2.0)に対応し、遠隔地からシステムの環境監視、障害監視、障害ログの収集等が可能
- ・80PLUS Titanium に対応した 2200W の高効率電源を筐体あたり 2 基搭載



写真 1 LX 406Rz-2 正面



写真 2 LX 406Rz-2 背面

表 1 AOBA-B で採用されている LX 406Rz-2 主要諸元

項目		諸元
CPU (1 ノードあたり)	搭載 CPU	AMD EPYC™ 7702
	搭載数	2
	演算性能	4.096TFlops (倍精度浮動小数点演算)
メモリ (1 ノードあたり)	搭載メモリ	DDR4-3200、ECC 対応 RDIMM
	容量	256GB
	メモリバンド幅	409.60GB/s
内蔵ディスク(1 ノードあたり)		480GB SATA SSD
標準インターフェース(1 ノードあたり)		1x IPMI2.0
		2x USB3.0、1x VGA
拡張 IO スロット(1 ノードあたり)		2x PCI Express 4.0 (x16) 1x SIOM card
ノード間接続		InfiniBand HDR 200Gbps x1
筐体	外形寸法(HxWxD)	88mm(2U) x 447mm x 730mm
	質量	38.6kg
	電源	2x 2200W 80PLUS Titanium Level Power Supply

2.2 サーバの内部構成

図 1 に LX 406Rz-2 筐体のハードウェア構成を示します。LX 406Rz-2 は 2U シャーシの中に独立したノードを 4 台搭載しています。各ノードは電源ユニットとファンを、HDD バックプレーンを介して共有しています。各ノードと内蔵ディスク、電源ユニットはホットスワップが可能です。

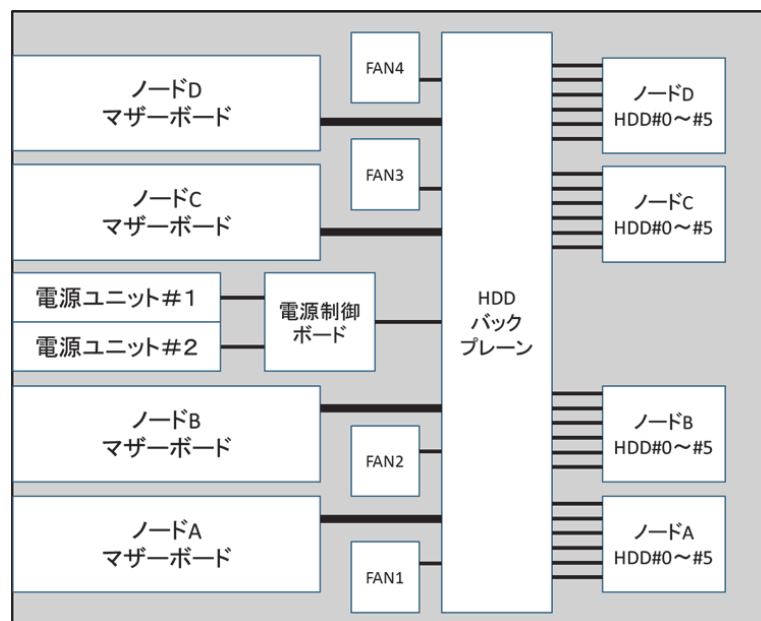


図 1 LX 406Rz-2 ハードウェア構成

図 2 にノードのブロック図を示します。

LX 406Rz-2 は、AMD EPYC™ 7002 シリーズプロセッサを 2 基搭載します。プロセッサ間は、4 チャンネルの xGMI インターフェースで接続されています。また、プロセッサあたり 8 チャンネルの DDR4-3200 対応メモリーインターフェースと、NEC のサーバとして初めて PCIe Gen4 に対応した拡張 PCI スロットを装備しています。初めて PCIe Gen4 に対応したことで、最速のインターコネクトである InfiniBand HDR(200Gbps)に対応することが可能になりました。LX 406Rz-2 には、サーバ管理機能として IPMI2.0 に対応したベースボード管理コントローラ(BMC)を搭載しています。これにより、温度、電力、ファンの動作状態、CPU やメモリのエラー監視、リモートからの電源のオン/オフなどを行うことが可能です。

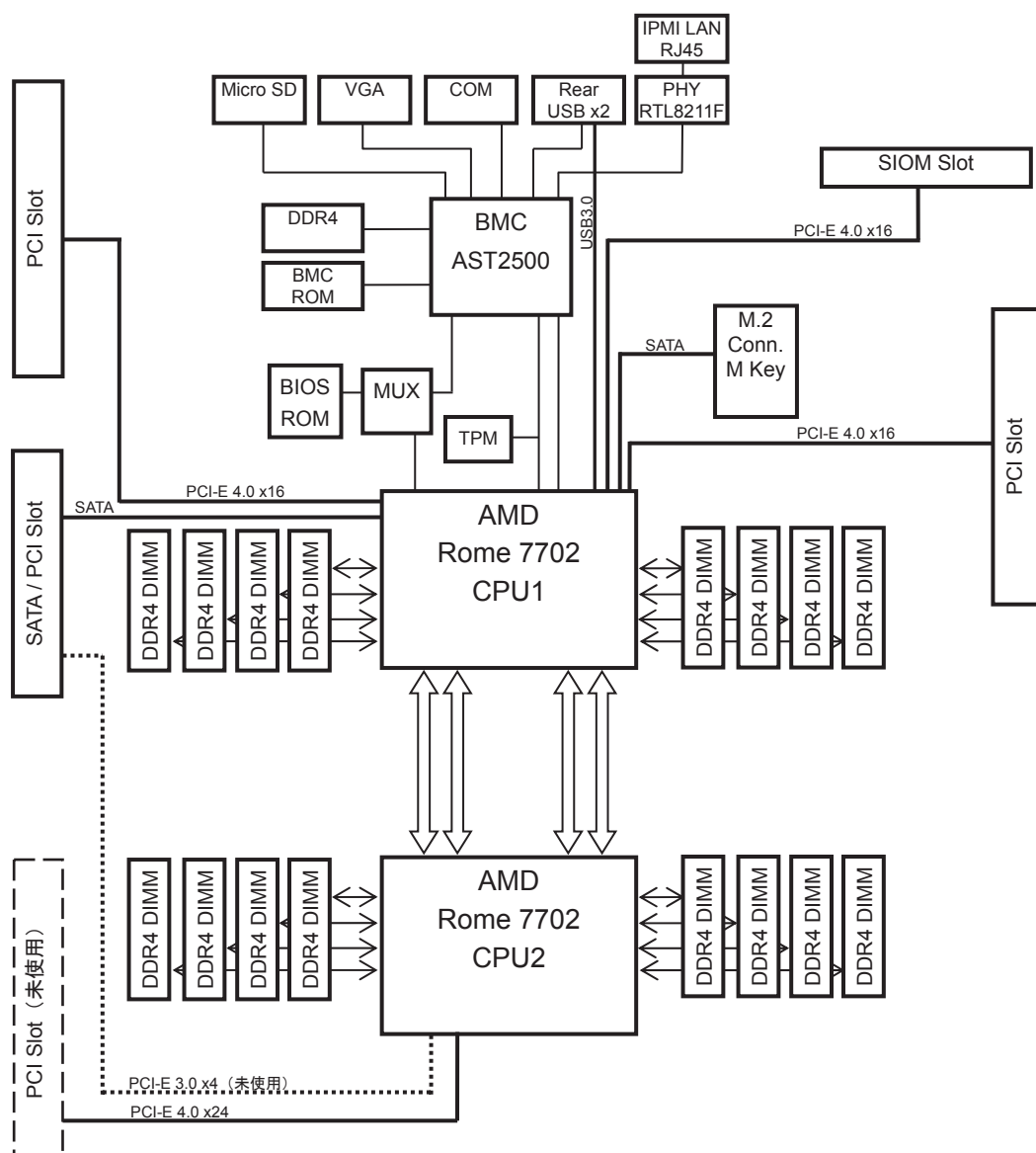


図 2 LX 406Rz-2 ノードブロック図

3. AMD EPYC™ 7002 シリーズ

3.1 AMD EPYC™ 7002 シリーズの特長

AMD EPYC™ 7002 シリーズ(コードネーム:Rome)は、2017 年に発売された第一世代である EPYC™ 7001 シリーズ(コードネーム:Naples)プロセッサの後継としてリリースされた製品です。第二世代である EPYC™ 7002 シリーズは、7nm プロセスで製造される CPU ダイ 8 個と 14nm プロセスで製造される I/O ダイ 1 つがパッケージ上に実装され、それらが高速インターコネクトで接続されています。CPU ダイには、最大 8 個の Zen2 コアとキャッシュメモリが搭載されています。I/O ダイには、8 チャンネルのメモリコントローラと 128 レーンの PCI Express Gen4、SATA 等を制御する I/O コントローラが搭載されています。これにより、1 ソケットあたり最大 64 コアの CPU、8 チャンネルの DDR4-3200 メモリ、最大 128 レーンの PCI Express Gen4 を備えています。

EPYC™ 7002 シリーズの Zen2 コアは、浮動小数点演算ユニットの強化、L3 キャッシュの容量拡大、CPU 間接続速度の向上などにより、第一世代の Zen コアと比較し IPC(Instruction Per Clock cycle)がおおよそ 15% 向上しています。

AMD EPYC™ 7002 シリーズプロセッサは、サーバ向けのプロセッサとして初めて PCI Express Gen4 をサポートしています。これにより、第一世代の EPYC™ 7001 シリーズと比較して、2 倍の I/O 性能を実現し、HPC クラスタシステムを接続するインターコネクトの帯域幅を 2 倍に向上しています。EPYC™ 7002 シリーズの主な特長は以下の通りです。

- 7nm プロセスの「Zen2」アーキテクチャを採用
- ソケットあたり最大 64 コア 128 スレッド
- DDR4-3200 メモリに対応し、ソケットあたり最大 4TB のメモリを搭載可能
- サーバプロセッサとして初めて PCI Express Gen4 をサポート
- 第一世代と比較しソケット当たりの理論ピーク性能(FLOPS)をおおよそ 4 倍に向上

3.2 AMD EPYC™ 7702 諸元

AOBA-B の LX 406Rz-2 では、64 個のコアをもち、性能と電力のバランスが取れた EPYC™ 7702 を採用しました。表 2 に AMD EPYC™ 7702 の諸元を示します。

表 2 AMD EPYC™ 7702 諸元

項目	諸元
コア数/スレッド数	64/128
周波数	2.0GHz
消費電力(TDP)	200W
製造プロセスルール	7nm
メモリーインターフェース	DDR4-3200
メモリーチャンネル数	8
最大メモリー容量	4TB
最大メモリー帯域幅	204.8GB/s
PCI Express	Gen4 128 レーン

4. インターコネクト

4.1 インターコネクトの特長

スーパーコンピュータ AOBA のインターコネクトは、InfiniBand HDR による 2 段 Fat-Tree ネットワークを採用し、フルバイセクションバンド幅、ノンブロッキング構成によりすべての演算サーバ、ストレージ、フロントエンドサーバで高速なデータ通信を可能にしています。

スイッチは NVIDIA Mellanox QM8700 シリーズを採用し、6 台のコアスイッチと 13 台のエッジスイッチでノード間接続ネットワークを構成しています。これらのスイッチにはマネージドスイッチを採用しており、うち 2 台のスイッチはサブネットマネージャとして動作し、ファブリック内のホストや通信経路の検索、監視、ならびに障害検出・回復などを行います。

2 台のサブネットマネージャは冗長化されており、1 台がプライマリサブネットマネージャ、もう 1 台がセカンダリサブネットマネージャとして動作します。プライマリサブネットマネージャで障害が発生すると、経路情報などをセカンダリサブネットマネージャが引き継ぎ動作を継続します。このように、スーパーコンピュータ AOBA のインターコネクトは、信頼性の高い高速ネットワークを実現しています。

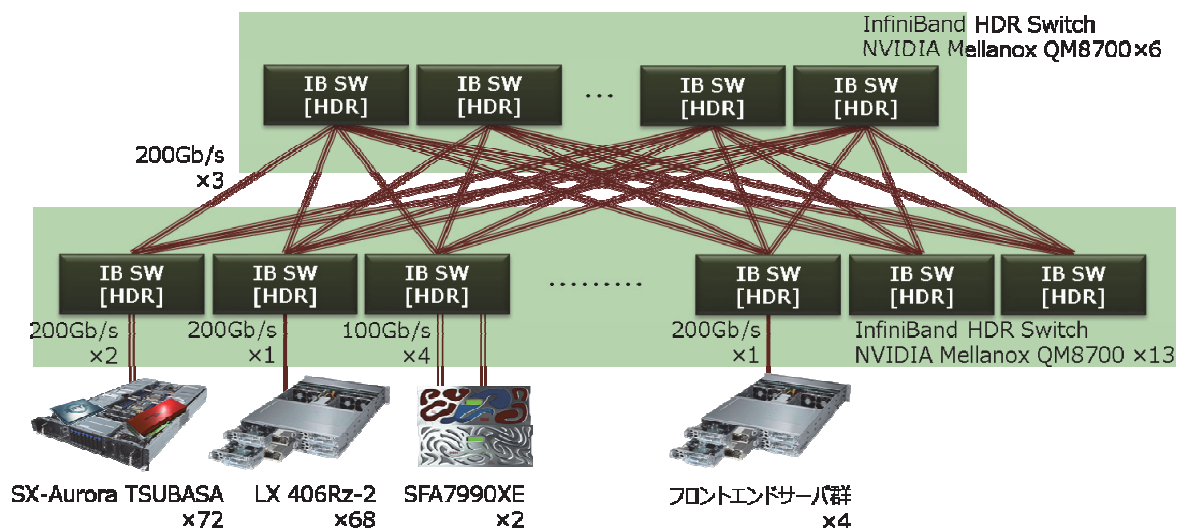


図 3 InfiniBand ネットワーク図

4.2 InfiniBand HDR スイッチ(QM8700 シリーズ)

QM8700 シリーズは、1 筐体に 40 ポート、16Tbps のノンブロッキング帯域幅と、90ns 以下のポート間レイテンシを実現します。LX 406Rz-2 に搭載する InfiniBand HDR(200Gbps) HCA カードと組み合わせることで最高のファブリック性能を実現します。表 3 に QM8700 シリーズの主要諸元を示します。

表 3 QM8700 シリーズ 主要諸元

項目		諸元
筐体		19 インチラック対応 1U 筐体
ポート数		HDR (200Gbps) 40 ポート HDR100 (100Gbps) 80 ポート (スプリッタケーブル使用時)
ファブリック管理機能		On-board Subnet Manager (SM) Unified Fabric Manager (UFM) agent
運用管理	管理ポート	100/1000 RJ45 Ethernet ポート RS232 コンソールポート USB ポート
	管理インターフェース	CLI、WebUI、SNMP v1/v2/v3
電源		2x ホットスワップ可能 冗長電源モジュール
FAN		ホットスワップ可能ファンモジュール

5. むすび

本稿では、AOBA-B で採用されている HPC クラスタ型計算サーバ LX 406Rz-2 の主要な機能、特長についてご紹介しました。

科学技術計算など高い性能を追求する HPC 領域では、設置スペースや消費電力等の様々な問題に向き合いながら、高い処理能力が求められます。NEC の HPC クラスタソリューションは、さまざまなワークロードに適用した高い電力性能比を有する CPU と高密度実装サーバにより、お客様の課題に応じて参ります。