[大規模科学計算システム]

ベクトル型スーパーコンピュータ SX-Aurora TSUBASA の最新ハードウェア

上山根慎　　春日康弘　　新井雅裕　　長瀬悟

日本電気株式会社

1. はじめに

東北大学サイバーサイエンスセンターで 2023 年 8 月より新しいスーパーコンピュータ AOBA-S の運用が開始されます。AOBA-S は約 21PFLOPS の総理論演算性能と約 10PB/s の総メモリバンド幅を実現する世界最大のベクトル型スーパーコンピュータシステムになります。注目すべき特徴は、演算性能とメモリ性能のバランスに優れていることであり、実際に使用した時の実効性能や性能の出しやすさの面で優位性が期待できます。

近年、情報社会が急速に発展して世界が大きく変化する中、シミュレーションをはじめとする科学技術計算需要はますます増大しています。一方、より大規模かつ複雑な問題の解析を行うためには、大量かつ多様なデータを扱える高性能計算とそれを支える電力効率に優れた計算基盤が求められます。この要請に応えるため、NEC では長年のスーパーコンピュータ開発で培った LSI 技術と高密度実装技術等を結集した PCI Express カード型の Vector Engine を多数搭載する SX-Aurora TSUBASA シリーズを 2018 年から提供してきました。東北大学サイバーサイエンスセンターでは 2020 年 10 月からシリーズ第二世代製品が AOBA-A として稼働中であり、今回導入するシリーズ第三世代の製品は AOBA-S として稼働します。

AOBA-S は AOBA-A と比較して、2.5 倍の処理性能と 2 倍の電力効率を発揮します。この高性能、省電力化は、コア数の倍増(8コアから16コア)、Level 3キャッシュの新規採用、最先端プロセスの採用等により実現しています。

本稿では、AOBA-S を構成する SX-Aurora TSUBASA システムのアーキテクチャ、システム概要、ハードウェア構成、テクノロジについて紹介します。

2. SX-Aurora TSUBASA アーキテクチャ

AOBA-S のアーキテクチャは AOBA-A と同様の SX-Aurora TSUBASA アーキテクチャを継承します。SX-Aurora TSUBASA は、NEC の 40 年に渡るベクトル型スーパーコンピュータ SX シリーズの流れを汲む製品として 2018 年に第一世代の製品を出荷開始しました。Vector Engine と呼ばれる PCI Express カードにベクトルプロセッサと主記憶を搭載し、これを Vector Host と呼ばれる標準的な x86 サーバに接続することによってシステムが構成されます。

Vector Engine に搭載されるベクトルプロセッサは、従来の SX シリーズのベクトルプロセッサ構成を踏襲しています。通常、サーバに接続される GPU や FPGA のような PCI Express カード型のアクセラレータは、ホスト側で実行されるアプリケーションの一部分を実行し、全体の処理時間の短縮を図ります。一方 Vector Engine は、SX-Aurora TSUBASA アーキテクチャの採用により、コンパイルされたアプリケーション実行ファイルを丸ごとカード上で実行することが可能となります。

この実行モデルにはいくつかのメリットがあります。第一に、ユーザにとって使い慣れた一般的な Linux OS 環境から、SX シリーズの高性能なベクトルプロセッサを利用できます。SX シリーズでは従来、専用の OS 環境を提供していましたが、SX-Aurora TSUBASA は標準的な x86 サーバと Linux OS 環境から演算処理だけにベクトルプロセッサを利用することができます。第二に、プログラム全体を Vector Engine カード上で実行することにより、PCI Express バス上の頻繁なデータ移送を避け、性能ボトルネックを解消することができます。

AOBA-S は現行 SX-Aurora TSUBASA のプログラム資産を継承しており、AOBA-A で利用していたソースプログラムをそのまま流用できます。また原則としてアクセラレータ向けのプログラム修正を行うことなく、SX-Aurora TSUBASA 向けコンパイラを用いてコンパイルしたプログラムをそのまま実行することができ

ます。特殊なプログラミングは必要なく、Fortran、C/C++のプログラムをコンパイルするのみで、プログラムが自動的に最適化され、高速化できます（さらに高い実行性能を得るためには、Vector Engine 向けのコードチューニングを実施することが望ましいです）。図 1 に SX-Aurora TSUBASA アーキテクチャにおける実行モデルを示します。アプリケーション全体を VE 上で実行する OS Offload 実行モデルの他、GPU のように x86 上のアプリケーション中のソルバー等の一部処理を VE にオフロードする VEO 実行モデルや、逆にベクトル化が困難な一部の処理を x86 にオフロードする VH call 実行モデルがあります。

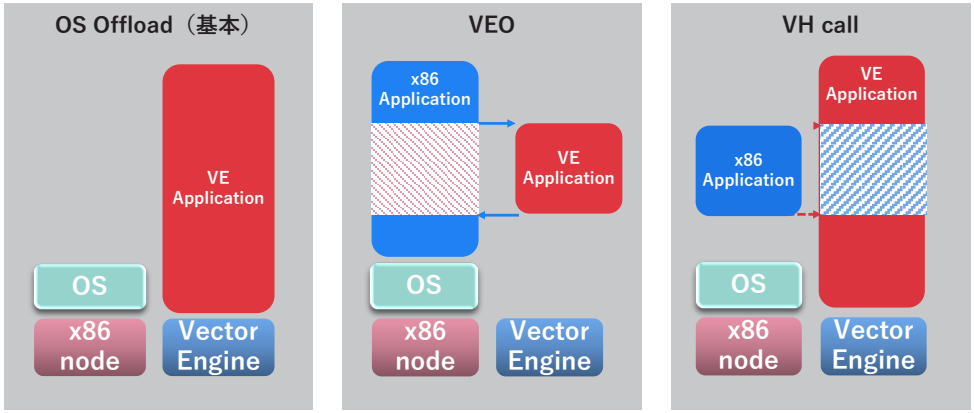

図 1. SX-Aurora TSUBASA アーキテクチャの実行モデル

3. システム概要

AOBA-S は、Vector Engine カードを 8 枚搭載した Vector Host（SX-Aurora TSUBASA C401-8）計 504 台で構成され、各 Vector Host は InfiniBand Network によって接続されます。

システムの主要な諸元を表 1 に示します。

表 1. AOBA-S の諸元

Vector Engine	モデル名	Type 30A
	コア数	16
	理論演算性能（倍精度）	4.91TFlops
	メモリ容量	96GB
	メモリ帯域	2.45TB/s
	インタフェース	PCI Express Gen4 x16
	最大消費電力	370W
Vector Host	モデル名	AMD EPYC 7763
	コア数	64
	メモリ容量	256GB
	理論演算性能（倍精度）	2.50TFlops
	OS	Rocky Linux
	搭載 Vector Engine 数	8
	ノード間ネットワーク	InfiniBand NDR200 x2
システム	Vector Host 数	504
	Vector Engine 数	4,032

	総コア数	32,256(Vector Host)、64,512(Vector Engine)
	総理論演算性能	21.05PFlops
	総メモリ容量	504TB
	総メモリ帯域	9.97PB/s

4.　ハードウェア構成

4.1　Vector Engine Type30A

Vector Engine Type30A は、NEC が新規に開発した第三世代のベクトルエンジンで SX-Aurora TSUBASA の心臓部です。フォームファクタは従来同様 PCI Express(PCIe)カードタイプを踏襲し、PCIe カード上に Vector Engine Type30A のベクトルプロセッサを1つ搭載し、水冷方式により冷却されます。図 2 に Vector Engine Type30A カードの外形写真を、表 2 にカードの実装仕様を示します。

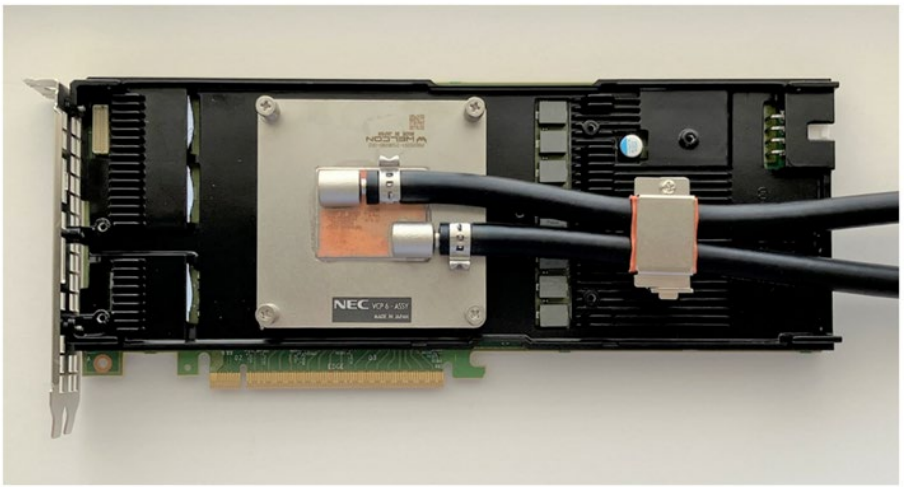

図 2. Vector Engine カード(水冷タイプ)

表 2. Vector Engine カード実装仕様(水冷ホースを除く)

カード長	266.46mm
カード高	111.15mm
カード幅	2 スロット
補助電源コネクタ	8-pin EPS 12V

続いて Vector Engine Type30A(VE30A)プロセッサについて説明します。図 3 に VE30A プロセッサの主要な諸元および概略ブロック構成を示しています。VE30A プロセッサは、前機種(VE20B)の 2 倍となる 16 個のコアを搭載し、16 コアへのデータ供給能力を向上させるためにメモリサブシステムを一新しています。主記憶には HBM2e メモリを 6 個搭載し最大で 2.45 テラバイト/秒という非常に高いメモリ帯域を実現した他、コアと主記憶の経路上にあるキャッシュ階層の見直しも図りました。

具体的には、従来から踏襲のレベル 1 キャッシュ、レベル 2 キャッシュ、ラストレベルキャッシュ(LLC)については、レベル 1、レベル 2 キャッシュはそれぞれ容量を倍増、ラストレベルキャッシュについては容量を従来の 16MB から 64MB に大幅増強しています。さらに VE30A では、レベル 3 キャッシュ(L3C)と呼ばれるコアあたり 2MB の大容量を持つコア専用のキャッシュを新設し、コアへのベクトルデータ供給能力を大幅に改善しています。また、このレベル 3 キャッシュは出来る限り多くの有効なベクトルデータを格納で

きるよう、ソフトウェア制御機能を有しています。ソフトウェア制御機能とはソフトウェアがキャッシュ上にベクトルデータを搭載する/しないを命令単位に指定することができる機能で、この機能により空間的・時間的局所性のある再利用可能なベクトルデータのみを選択的にキャッシュに残すことが可能です。本機能を有効に活用することでアプリケーションの性能を高める効果が期待できます。

VE30A プロセッサと外部のインタフェースは PCI Express Gen4 を採用し、従来の VE20B の 2 倍となる最大 64GB/s の帯域を実現しました。プロセッサに内蔵する DMA エンジンを活用し、コアの処理とは独立して高速にデータ通信をおこなうことが可能です。

	VE30A (AOBA-S)	VE20B (AOBA-A)
コア数	16	8
コア理論演算性能	307GF(倍精度) 614GF(単精度)	←
プロセッサ理論演算性能	4.91TF(倍精度) 9.83TF(単精度)	2.45TF(倍精度) 4.91TF(単精度)
L1 cache for scalar	64KB per core (命令/オペランド別)	32KB per core (命令/オペランド別)
L2 cache for scalar	512KB per core	256KB per core
L3 cache	2MB per core SW controllable	N/A
Last level cache	64MB shared	16MB shared
メモリバンド幅	2.45TB/s	1.53TB/s
メモリ容量	96GB	48GB
PCIe	Gen4	Gen3
消費電力(TDP)	< 370W	300W

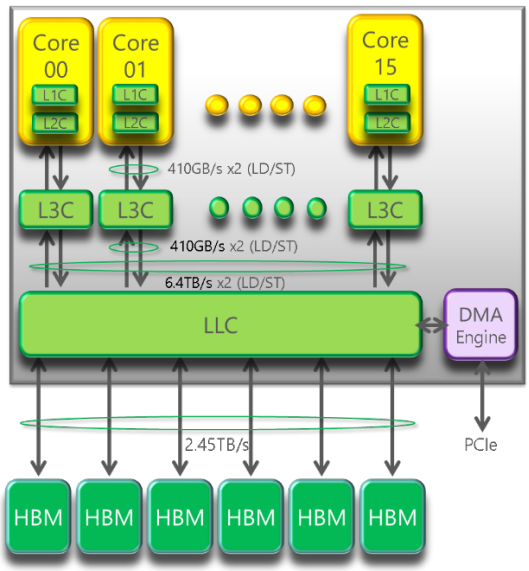

図 3. VE30A プロセッサの主要スペックと概略ブロック図

図 4 はコアの概略的なブロック構成を示したものです。コアは、スカラ処理部(Scalar Processing Unit: SPU)、ベクトル処理部(Vector Processing Unit: VPU)の他、主記憶へのロード/ストアを制御するアドレス生成部、およびリクエスト/リプライクロスバ部から構成されます。VE30A で新設の L3 キャッシュを経由してコア外部のメモリネットワークへデータ転送の送受信を行います。

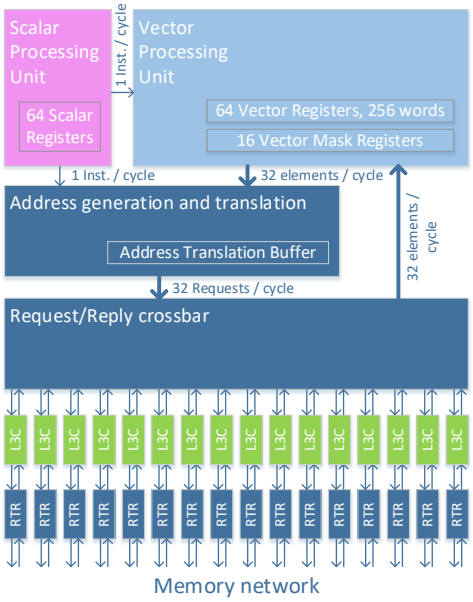

図 4. コア ブロック図

図 5 に SPU の概略ブロック構成を示します。SPU は x86 プロセッサ同様の汎用プロセッサの機能を持ち、全ての命令の解釈を行った上でスカラ命令を処理する他、ベクトル命令やノード間通信命令をそれぞれ VPU や DMA エンジンへ発行します。SPU は 1 クロックサイクルに最大 4 命令のフェッチ/デコード処理が可能で、VPU への専用パスを用いて 1 クロックサイクルあたり1つのベクトル命令を発行することができます。アプリケーションの実効性能を高めるには全体制御をおこなう SPU の性能向上が重要な要素であり、VE30A では L1 キャッシュ(命令・オペランド別)、L2 キャッシュの容量をそれぞれ 64KB、512KB に倍増した他、Unified Scheduler と呼ばれる命令発行制御の強化により最大仕掛命令数を VE20B 比 1.33 倍に増強して並列性を向上させる等、全体的な性能向上を図っています。

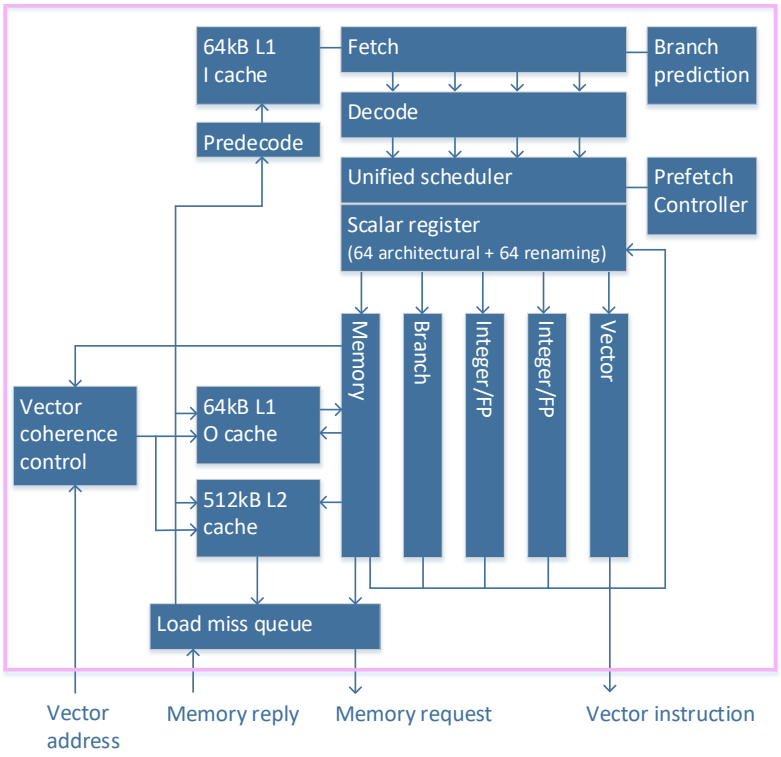

図 5. SPU 概略ブロック図

図 6 に VPU の概略ブロック構成を示します。VPU は 32 本のベクトルパイプライン(Vector PiPeline: VPP)から構成され、各 VPP は FMA0～2 の 3 セットの FMA 演算器(浮動小数点積和演算器)を有し、最大で 3 つの FMA 演算命令を並列に実行することが可能です。よって、理論ベクトル演算性能は 32VPP×FMA3 セット×2(乗算＋加算)×1.6(GHz)=307.2GFlops という高い性能を発揮することができます。また、FMA 演算器は単精度データ×2 要素の単精度 Packed 演算を行うことも可能であり、その場合の最大演算性能は 2 倍の 614.4GFlops となります。各 VPP は 1 クロックサイクルに 64bit 倍精度データのロードする可能であるため、コア辺りのメモリバンド幅は 8 バイト(64 ビット) ×32 パイプライン×1.6GHz = 409.6GB/s(x2=ロード+ストア)となります。

これらの演算器にデータを供給するために、SX-Aurora TSUBASA アーキテクチャ上は 64 本のベクトルレジスタを有し、1 本のベクトルレジスタは倍精度データであれば最大 256 要素、単精度 Packed データであれば最大 512 要素のベクトルデータを格納することが可能です。また、SX-Aurora TSUBASA アーキテクチャ上の 64 本のベクトルレジスタに対し、物理的には 256 本のベクトルレジスタを備えており、ベクトルレジスタのリネーミング機構により命令列上の論理ベクトルレジスタ間の依存関係を解消し、ベクトル演算命令・ベクトルメモリアクセス命令のアウトオブオーダー実行を実現しています。

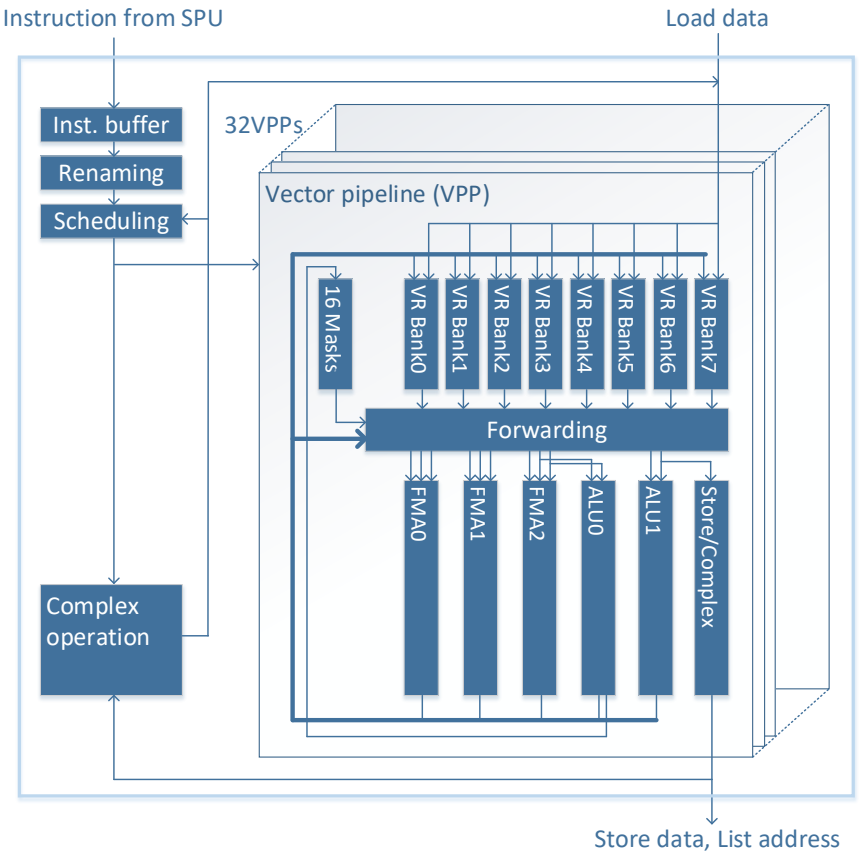

図 6. VPU 概略ブロック図

4.2 SX-Aurora TSUBASA C401-8

AOBA-S の構成要素（ノード）である SX-Aurora TSUBASA C401-8 の外観とブロック図をそれぞれ図 7 と図 8 に示します。C401-8 は Vector Host と 8 枚の Vector Engine で構成されます。Vector Host のプロセッサは AMD EPYC 7763 で、Vector Host あたり 1socket、64 コアを備えています。プロセッサの基本動作クロックは 2.45GHz で、シングルコアであれば最大で 3.5GHz までのブースト動作が可能です。プロセッサは 4 つの PCI Express switch と 2 つの InfiniBand NDR200 カードと直接接続されています。4 つの PCI Express switch はそれぞれ 2 つの Vector Engine カードと接続されています。それらの接続は PCI Express Gen4 x16 です。SX-Aurora TSUBASA C401-8 は Vector Host あたり 256GB の主記憶と 1.92TB の SSD を備えています。AOBA-S は AOBA-A の SX-Aurora TSUBASA B401-8 に対して Vector Host のプロセッサ性能を増強、SSD 容量を倍増しています。また、Vector Engine の性能向上に合わせて、PCI Express を Gen4 に、電源容量も増やしています。AOBA-S と AOBA-A の Vector Host の差分を表 3 に示します。

図 7. SX-Aurora TSUBASA C401-8 外観

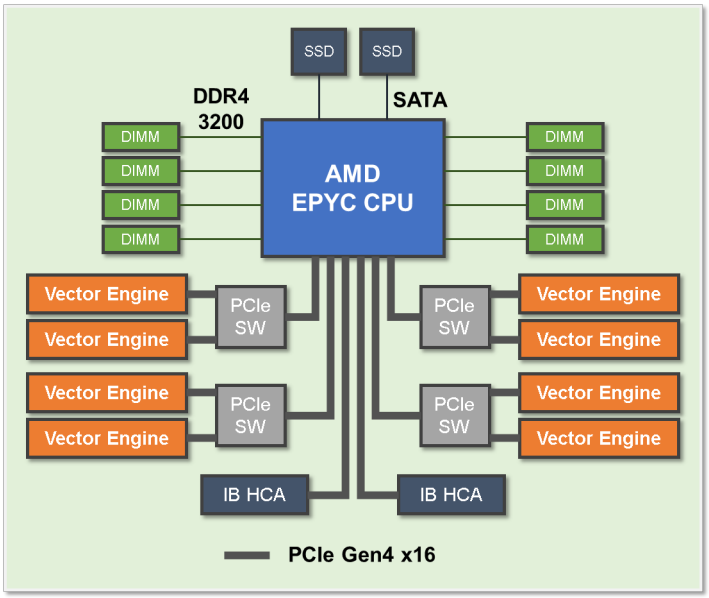

図 8. SX-Aurora TSUBASA C401-8 ブロック図

表 3. AOBA-S と AOBA-A の Vector Host の差分

	AOBA-S	AOBA-A
プロセッサ	AMD EPYC 7763 64 コア, 2.45GHz	AMD EPYC 7402P 24 コア, 2.8GHz
主記憶	256GB, DDR4/3200	256GB, DDR4/3200
デバイス	1.92TB SSD	960GB SSD
Vector Engine I/F	PCIe Gen4 x16	PCIe Gen3 x16
ノード間 I/F	InfiniBand NDR200 (200Gbps)	InfiniBand HDR (200Gbps)
電源	2200W PSU x2	2000W PSU x2

4.3 ノード間ネットワーク

AOBA-S のシステムは最新の InfiniBand NDR ネットワークにより構成されています。図 9 に InfiniBand ネットワーク構成図を示します。InfiniBand NDR スイッチとして NVIDIA QM9700 シリーズを利用します。QM9700 は、スイッチ 1 台あたり NDR(400Gbps)64 ポートまたは、NDR200(200Gbps) 128 ポートを接続可能であり、従来の HDR スイッチと比べてポート密度が高く、より大規模な InfiniBand ネットワークを構成できます。

AOBA-S は 504 台の Vector Host から構成されており、Vector Host 1 台あたり 2 枚の NDR200 カードを介してエッジスイッチに接続します。InfiniBand NDR ネットワークを使用することにより、これら Vector Host 504 台を、シンプルかつ高性能な、2 段 Fat-Tree ネットワークで構成しています。

これにより、ノード間接続ネットワークは、フロントエンドサーバなどの各種サーバやストレージを含めたシステム全体を、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接続し、高速なデータ通信を可能とします。

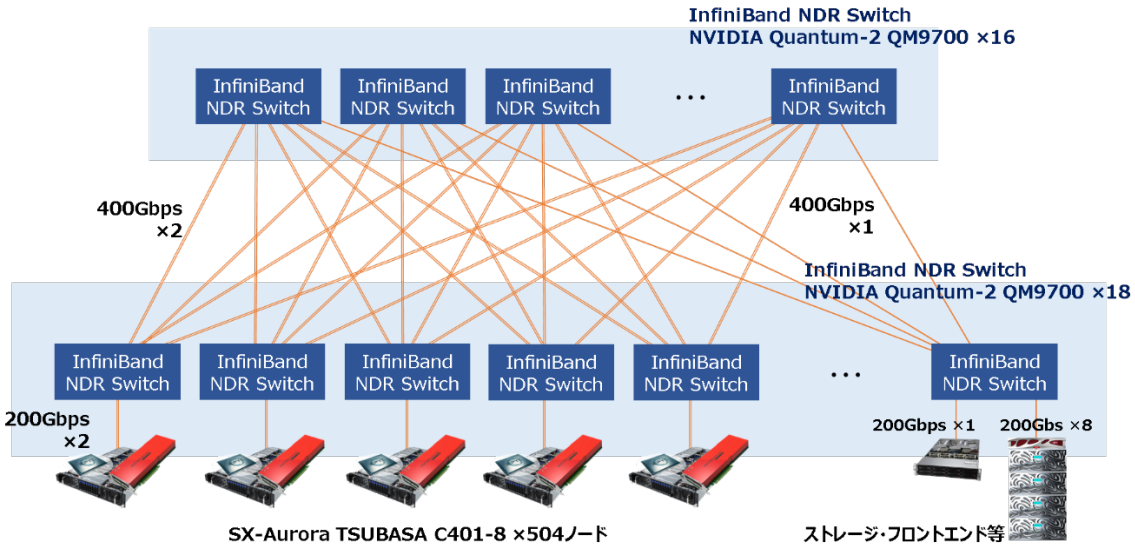

図 9. InfiniBand ネットワーク構成図

5. SX-Aurora TSUBASA のテクノロジ

Vector Engine VE30A は、従来の VE20B と比べて 2 倍の演算性能、1.6 倍のメモリ帯域と性能の向上を図っています。カードは、従来機種と同様、PCIe アドインカードのフォームファクタを採用し、同じサイズで実現しています。

5.1 プロセッサ技術

Vector Engine の要となるのがベクトルプロセッサです。その製造プロセスは、従来の 16nm から微細化した 7nm を採用しました。これにより、ゲート数や SRAM ビット数は、それぞれ 5 倍超の実装を実現しました。

HBM は、HBM2 から高速な HBM2e を採用し、プロセッサあたりのメモリバンド幅 2.45TB/s を実現しています。

LSI の実装形態は、シリコンインタポーザを用いた 2.5 次元実装です。ベクトルプロセッサも HBM もサイズが大きくなりましたが、65x65mm の大きさに収めることができました。

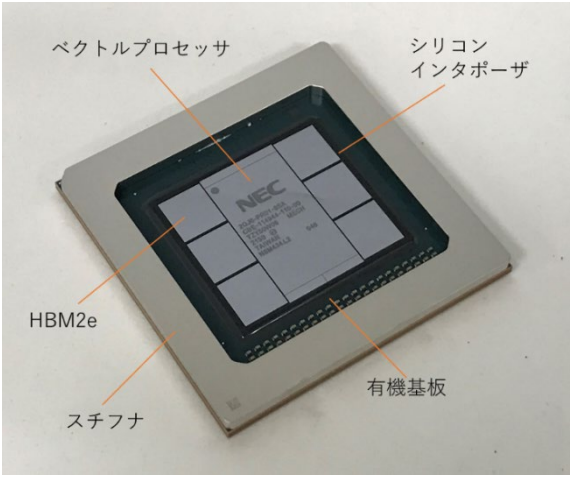

図 10. Vector Engine プロセッサ LSI 外観

5.2 冷却技術

プロセッサの演算性能に対する消費電力比率は従来機種に比べて向上しています。しかし、プロセッサとして見た場合は、従来機種に比べて消費電力は上がっています。プロセッササイズも大きくなっていますが、消費電力をプロセッサのサイズで割って算出した単位面積当たりの電力(電力密度)は上がっています。このため、従来よりも冷却性能の高いコールドプレートが必要になります。プロセッサのフロアプランにジョブ実行時の電力をマッピングすることにより、動作時にプロセッサの温度が最も上がる部分(ホットスポット)を見積り、そのホットスポットが冷却できるよう、コールドプレートの構造見直しと高精度な冷却シミュレーションを繰り返し実施することにより、最適なコールドプレートを設計しました。

5.3 電源の安定供給

微細化された大電力プロセッサを安定に動作させ、その性能を最大限に引き出すためには、急激な電流変動に対しても電圧変動を抑えた電源が必要不可欠です。電源の安定供給のためには、プリント基板上にコンデンサを実装する手法が一般的です。コンデンサへの配線にはインダクタンス成分や抵抗成分があるため、プロセッサの電源ピン、グランドピンにできるだけ近づけて配置することが鉄則です。プロセッサの大電力化に伴い、要求されるコンデンサの数が多くなってきています。コンデンサの数が多くなると、プロセッサの近傍に配置できなくなり、特性改善のために更にコンデンサの数を増やすといった悪循環に陥ります。

これを解決するために、VE30A では、プリント基板の電源層とグランド層の間に薄型のキャパシタ材料を挟み込み、プリント基板内にコンデンサを形成しました。これによりプリント基板の給電系の電気特性が従来比 4 倍改善しました。

6. おわりに

NEC のベクトル型スーパーコンピュータ SX シリーズは、従来から東北大学サイバーサイエンスセンターで採用されてきました。このたび AOBA-S として稼働する最新の SX-Aurora TSUBASA システムは、航空機や発電タービンなどのものづくり分野で求められる大規模数値流体シミュレーションや津波浸水や河川氾濫の被害予測などの防災減災などの気候変動への適応策に役立つシミュレーションにおいて、多くの方々の研究を後押しする役割を担います。NEC は今後も社会・市場の課題解決に向けて積極的に貢献していきます。