

[大規模科学計算システム]

スーパーコンピュータ SX-ACE のハードウェア

萩原 孝 浜口 博幸 山信田 恒

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本年2月より運用開始される、新スーパーコンピュータシステム SX-ACE のハードウェアを紹介します。

SX-ACE は、増大する科学技術計算需要に応えるためのシステム規模の拡大とシステムの省電力化／省スペース化を両立させることを目標に開発しました。これを実現するために、SX-9 の大規模共有メモリ型ノードを少数接続するマルチノードシステムから、単一 CPU ノードを多数接続するマルチノードシステムにシステム・アーキテクチャを刷新しています。本稿では、SX-ACE の本体ハードウェアのシステム概要、CPU コア、CPU ノード、マルチノードの構成、及び LSI 技術、実装技術を紹介します。

2. システム概要

SX-ACE システムは、4 つの CPU コアで構成される単一 CPU ノードを総計 2560 ノード、総理論演算性能 707TFLOPS (総ベクトル演算性能 655TFLOPS)、総メモリ容量は 160TB を有したシステムです。このシステムは、512 ノードを1つのクラスタとして5クラスタで構成されます。クラスタ内のノード間は、ノードあたり 4GB/秒 × 2 (双方向) でノード間スイッチ (Inter-node X-bar Switch: IXS) で接続されます。システムの主な諸元とハードウェアの階層構成を表 2-1、図 2-1 に示します。

表 2-1 SX-ACE システム主要諸元

CPU ノード	CPU コア数	4
	CPU 数	1
	理論演算性能*1／ベクトル演算性能*2	276GFLOPS／256GFLOPS
	メモリ容量／メモリ帯域	64GB／256GB/秒
	ノード間通信性能	4GB/秒 × 2 (双方向)
クラスタ	CPU コア数	2048
	CPU 数	512
	理論演算性能*1／ベクトル演算性能*2	141TFLOPS／131TFLOPS
	メモリ容量／メモリ帯域	32TB／131TB/秒
	ノード間通信性能	2TB/秒 × 2 (双方向)
システム	CPU コア数	10240
	CPU 数	2560
	理論演算性能*1／ベクトル演算性能*2	707TFLOPS／655TFLOPS
	メモリ容量／メモリ帯域	160TB／655TB/秒
	ノード間通信性能	10TB/秒 × 2 (双方向)

*1) 同時に並列演算することができるすべての演算器(ベクトルユニットの乗算演算器、加算演算器、除算/平方根演算器、ならびにスカラユニットの浮動小数点演算器)における浮動小数点演算処理能力の総和。

*2) 同時に並列演算することができるベクトルユニット内の乗算演算器、加算演算器のみによる浮動小数点演算処理能力の総和。

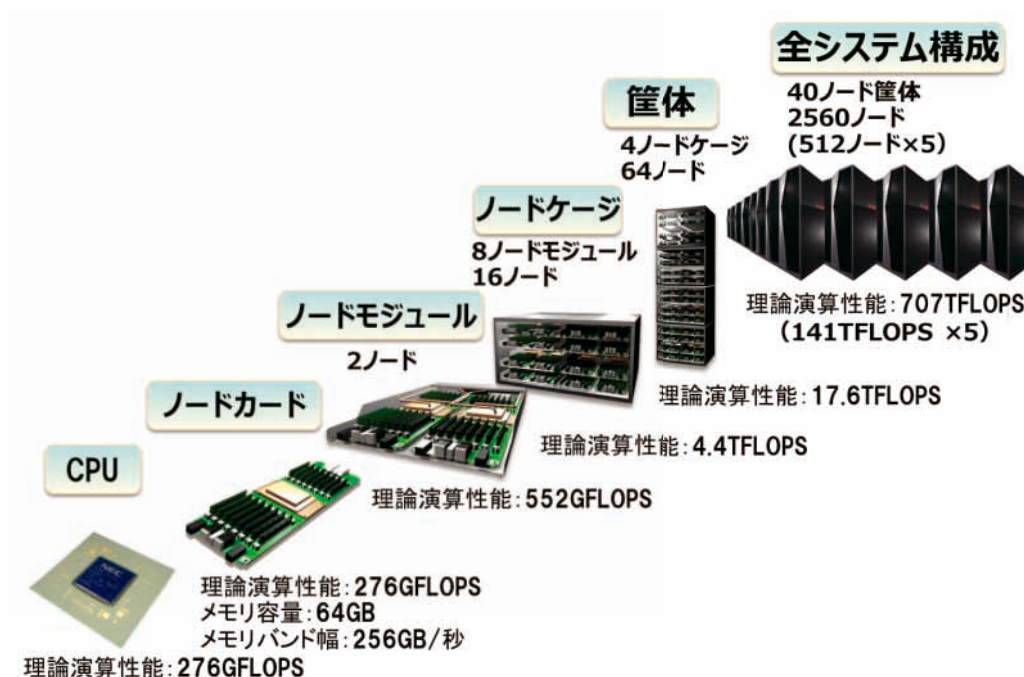


図 2-1 SX-ACE ハードウェア階層とシステム構成

3. ハードウェア・アーキテクチャ

SX-ACE は、これまでの SX シリーズと同様に高メモリバンド幅を必要とする多くの科学技術計算アプリケーションで高い実効性能を提供するスーパーコンピュータシステムです。主な特長としては、

■ビッグ・コア、高メモリバンド幅設計による高実効性能

- 64GFLOPS の世界トップクラスの CPU コア性能
- コアあたり 64GB/秒の高メモリバンド幅

■1CPU/ノード化、先端テクノロジー採用による省電力化／省スペース化

- 同一ピーク性能比較で前機種 SX-9 の 10 分の 1 の低電力化と、5 分の 1 の省スペース化(保守エリア含む)

が挙げられます。以下では、SX-ACE のハードウェア構成と特長を中心に紹介します。

3.1 CPU コア構成

SX-ACE の CPU コアは、従来の SX アーキテクチャを継承しつつ、さらにマイクロ・アーキテクチャレベルで改良を加え、高性能化を図っています。図 3-1 に CPU コア構成を示します。CPU コアは、スカラー処理部 (Scalar Processing Unit: SPU)、ベクトル処理部 (Vector Processing Unit: VPU)、ソフトウェア制御可能な高速バッファ部 (Assignable Data Buffer: ADB)、冗長なメモリアccessを抑制する MSHR (Miss Status Handling Register) で構成されています。SPU は、全ての命令発行制御を行い、ベクトル命令やノード間通信命令を VPU やノード間通信制御部 (Remote access Control Unite: RCU) に発行します。SX-ACE では、SPU から VPU への命令発行専用のパスを設け、SX-9 に比べてベクトル命令の発行をより効率よく行えるようにしています。VPU は、16 本のベクトル・パイプラインから構成され、SX-9 の 2 倍のベクトル・パイプライン数となっています。これにより、1 つのベクトル演算命令を 1 サイクルあたり 16 演算のスループットで処理することで、

CPU コアのベクトル演算性能(ベクトル乗算・加算演算性能)は、64GFLOPS を実現しています。ベクトル長 (VL: Vector Length)は、SX-9 と同じ最大 256 です。各ベクトル・パイプラインの内部構成は、SX-9 と同様、乗算器×2、加算／シフト演算器×2、除算／平方根演算器×1、論理演算器×1 の 6 種のそれぞれ独立に動作可能な演算器と、マスク演算パイプライン、ロード／ストアパイプライン、マスクレジスタ、及びベクトルレジスタにより構成されています。SX-ACE では、乗算器や加算器などの主要演算器間で先行する演算結果を後続の演算器へ直接フォワーディングするパスを設け、ベクトル長の短いアプリケーションの性能向上を図っています。

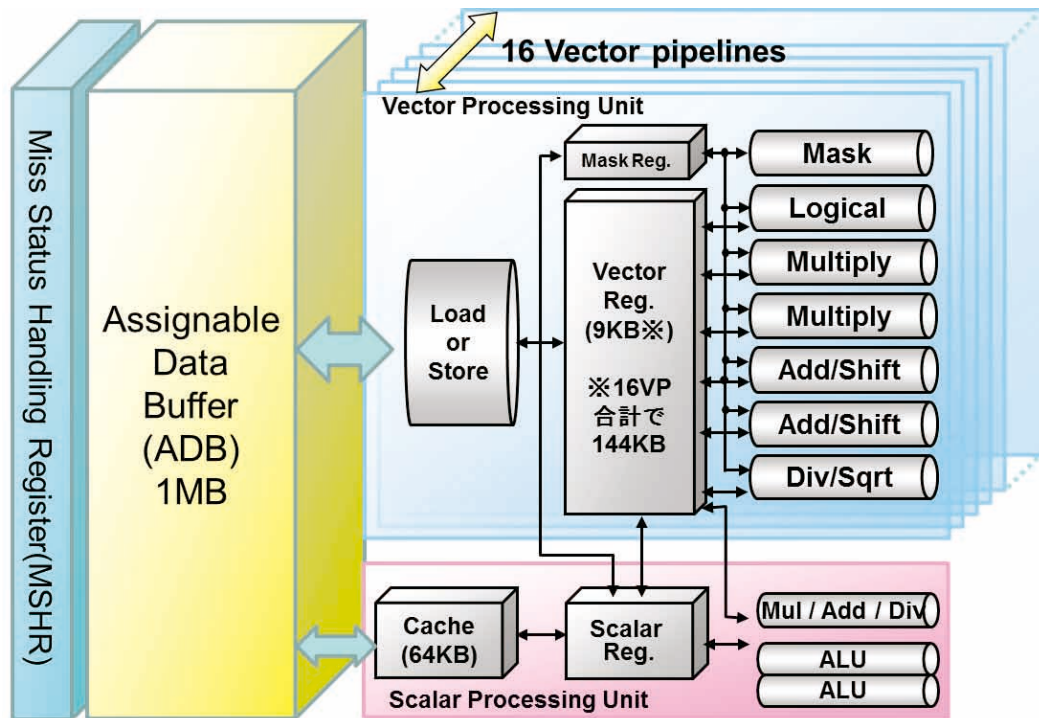
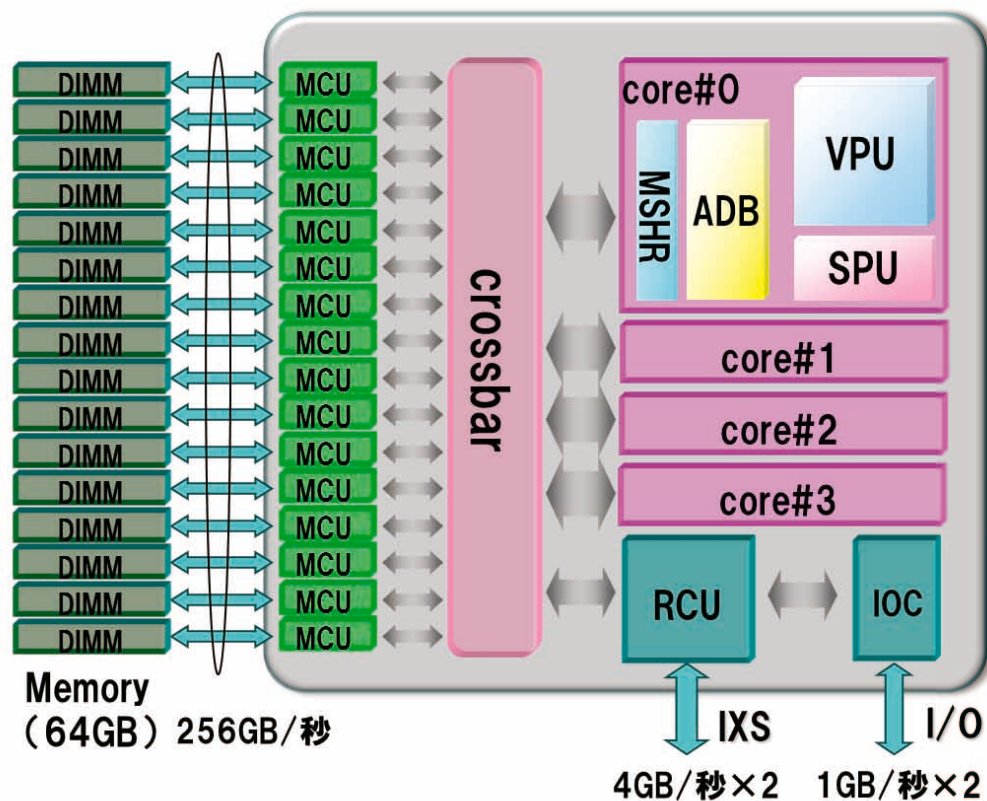


図 3-1 CPU コア構成

SX-9 から導入した ADB は、SX-ACE では SX-9 の 4 倍の 1MB の容量となります。ADB から VPU へは最大 256GB/秒(4Bytes/Flop) のデータ供給性能を備えています。ADB は、ラインサイズ単位(128B)で管理され、ADB を使用するベクトルデータの指定方法としては、コンパイラによる自動最適化、コンパイラ・オプション指定、およびプログラム中に指示行で指定する方法があります。SX-ACE では、ADB の容量増に伴い、コンパイラによる自動最適化や ADB にバッファリングするベクトルデータを指定するコンパイラ・オプションを強化し、ADB をより有効に活用し易くしています。

また、SX-ACE では、新たに冗長なメモリアccessを削減するために MSHR を実装しています。図 3-2 に MSHR の活用例を示します。科学技術計算プログラムでは、格子上の 1 点を計算するために周囲点をアクセス(決まった型のアクセス)するステンシル計算が多くみられます(図 3-2 a))。ベクトル機で周囲点をアクセスする場合、例に示す配列 A の参照は 4 個のベクトルロード命令でアクセスされます。このうち $A(i-1, j)$ と $A(i+1, j)$ の参照は、例えば $i=2\sim 257$ では、それぞれ $A(1:256, j)$ 、 $A(3:258, j)$ をアクセスし、254 要素が重なります(図 3-2 b))。MSHR はこの重なった要素に対する二重のメモリアccessを抑止する機構で、この機構により、 $A(i+1, j)$ のメモリアccessは、 $A(257:258, j)$ の 2 要素となり、CPU の外部メモリ帯域を有効活用する役割を果たします(図 3-2 c))。



SPU: Scalar Processing Unit, VPU: Vector Processing Unit, ADB: Assignable Data Buffer
MSHR: Miss Status Handling Register, RCU: Remote Access Control Unit,
MCU: Memory Control Unit, IOC: Input Output Control unit

図 3-3 CPU ノード構成

流体計算などの科学技術計算アプリケーションで高い実効性能を得るためには、高い演算性能に見合うだけの高いメモリ性能が必要となります。SX-ACE の CPU に搭載された 16 個の MCU は、外部メモリに効率よくアクセスするため、メモリバンク競合で待たされている先行するロード/ストア・リクエストを後続リクエストが追い越す機能や、ストアデータをメモリアクセス単位である 128B のブロックにまとめ、バンク競合を低減して効率よくストア処理するストアデータのマージ機能などを有しています。

RCU は専用開発の IXS を介したノード間通信を制御するユニットです。RCU は 4 つの CPU コアで共有され、各 CPU コアからのノード間通信リクエストの管理機構、論理ノード番号から物理ノード番号へのアドレス変換機構、自 CPU ノード内のメモリアクセスのアドレス変換機構などを備えます。

一般的な分散メモリ型の大規模並列システムでは、IO 機能を一部のノードに集約するか、専用の IO ノードを設ける方式が取られますが、SX-ACE では、各ノードに IOC を設け、SX-ACE の高いプロセッサ性能、及びメモリ性能に見合った高速なデータ入出力転送性能を実現しています。今回のシステムではノードあたり 1GB/秒×2(双方向)の IO 性能を有しています。

3.3 マルチノード構成

SX-ACE の 1 クラスタのマルチノード構成図を図 3-3 に示します。クラスタの総理論演算性能は 141TFLOPS、総主記憶容量は 32TB になります。ひとつのクラスタは 512 ノードで構成され、各ノードは IXS を介して接続されます。IXS は専用開発の 32 ポートを持つスイッチ LSI を使って、2 段のフル・ファット・ツリー・

トポロジーを採用しています。RCU のデータ受信部、及び送信部は、それぞれ独立に動作可能であり、ノード当り4GB/秒×2(双方向)の通信バンド幅を実現しています。また、RCUはCPUとは独立に動作するデータムーバ(DMA エンジン)を持つことにより、ノード間のデータ転送をCPU 動作とは独立して行なうことが可能です。512 ノードのクラスタのバイセクション・バンド幅は、2TB/秒になります。SX-ACE の IXS は、SX-9 の IXS 同様に高速なハードウェア・バリア同期機構を持ち、MPI 通信の同期処理などの高速化を実現しています。

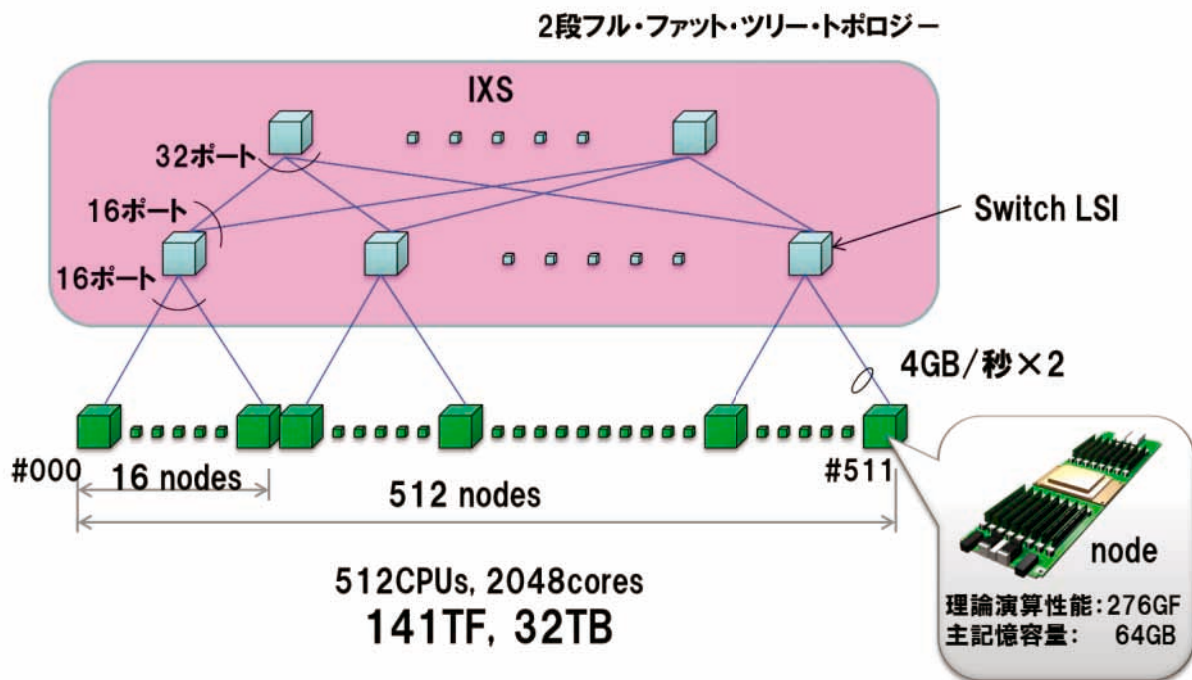


図 3-3 マルチノード構成(接続図)

4. SX-ACE のテクノロジー

4.1 LSI 技術

これまで SX シリーズでは CMOS テクノロジーによる高集積化、及びプロセッサの並列化により高性能化を実現させてきました。SX-ACE では、さらに高い性能を実現するために LSI 技術及び回路技術を発展させています。より高いメモリバンド幅を実現するために、インタフェース回路の低消費電力化、小面積化により LSI への多チャネル搭載を実現しています。また、高速化、高集積化による消費電力の増大を抑えるために様々な工夫を施しています。表 4-1 に SX-ACE の CPU チップ諸元を示します。

表4-1 CPUチップ諸元

テクノロジーノード	28nm
搭載トランジスタ数	20億個
電源電圧	0.9V
ピン数(内信号ピン)	4,344(2,586)
配線層構成	銅11層
実装形態	BGA (Bump pitch:0.8mm)

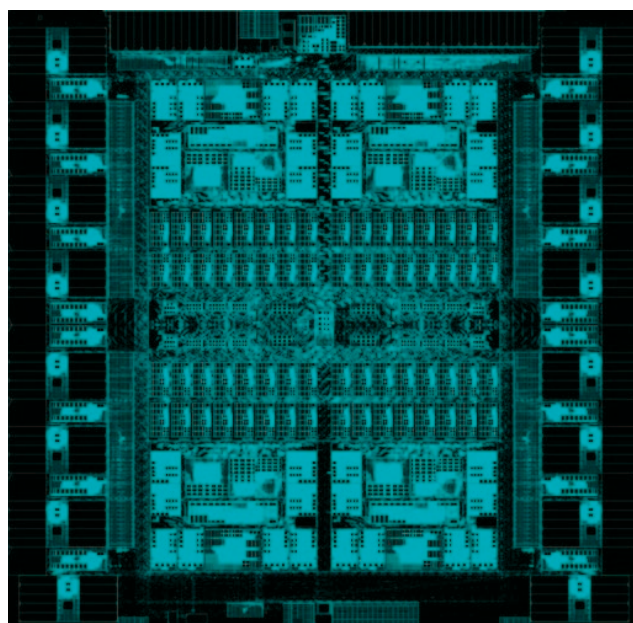


図 4-1 CPU チップレイアウト

SX-ACEに使われるLSIは、28nm CMOSプロセス、11層銅配線テクノロジーの採用による高速・高集積化、低電源電圧による低消費電力化、及び大容量オンチップキャパシタの搭載による高いノイズ耐力を実現しています。さらに、低消費電力実現のための設計上の工夫として、Multi-Vthトランジスタの最適な使い分け、レジスタ単位およびCPUコア単位でのクロックゲーティング、チップ内各種センサを用いた電源電圧の最適制御

など、最先端の低消費電力化技術を盛り込んでいます。

システムの処理能力の向上には、LSI 内信号伝送の高速化とともに、LSI 間信号伝送の高速化と並列化が必要となります。信号伝送の高速化や並列化を妨げる要因となる電源ノイズ対策も重要です。SX-ACE では高速、かつ安定した信号伝送を実現するために、伝送信号の波形を改善するエンファシス、イコライズ機能を備えた回路、及び波形ひずみの少ないコネクタなどを採用しています。また隣接する信号の配線間隔を始めた信号種に応じた配線ルールを定め、配線間のクロストークを低減しています。さらにトランジスタが高速化し、電源電流の時間変化が大きくなることにより電源ノイズが増加するため、デカップリング用コンデンサの搭載数や LSI パッケージの層構成を最適化するなどにより電源ノイズの低減を実現しています。

4.2 実装技術

SX-ACE では、実アプリケーションの高い実効性能と省スペース化のために、高密度接続技術、高密度配線基板技術、および高効率冷却技術により高いメモリバンド幅と高密度実装を実現しています。

高いメモリバンド幅を実現するために 4,000 ピン超の CPU パッケージと多ピン化対応と微細化を追求した高多層+HDI (High Density Interconnect) 基板を採用しています。また、省スペース化を実現するためにコンパクトな水冷構造と高密度ノードモジュールを採用しています。

表 4-2 に CPU パッケージ、ノード基板、およびノードモジュールの諸元を示します。また、図 4-2 から図 4-5 に CPU パッケージ、ノード基板、ノードモジュール、およびノード筐体を示します。

CPU パッケージは、他社比 2 倍以上の 4,344 (信号数:2,596) の IO 端子数を有しています。また、CPU LSI を保護する Lid を無くし、CPU LSI を直接冷却することで高い冷却効率を実現しています。

ノード基板は、スルーホールピッチが 0.8mm、34 層という多層化により信号配線収容性を高めるとともに、表裏面 1 層に Line/Space が 65/70 μm と微細な配線層を形成することによって多ピン CPU パッケージの搭載を可能としています。また、カスタム化された Mini-DIMM を搭載して高メモリバンド幅と高密度化を実現しています。

ノードモジュールには、2 つのノード基板、CPU を冷却するコールドプレート、および外部と信号接続する光モジュールが搭載されています。また、CPU 冷却にコンパクトな水冷構造を採用することでノードモジュールを高密度化しています。

ノード筐体には、ノードモジュールが 8 個で 1 ユニットとして 4 ユニット、つまり 32 ノードモジュール (64 ノード) 搭載できるようになっており、筐体あたりのベクトル演算性能は SX-9 の 10 倍の 16.4TFLOPS を実現しています。

表 4-2 SX-ACE 実装技術の諸元

項目		諸元
CPU パッケージ	材料	セラミック
	パッケージ外形	55mm×55mm×2.0mm
	IO 端子ピッチ	0.8mm ピッチ
	IO 端子数	4,344
ノード基板	構造	HDI(High Density Interconnect)
	基板サイズ	370mm×110mm t=4.9mm(層数:1-34-1)
	配線仕様	スルーホールピッチ:0.8mm、L/S = 65/70 μ m
	実装物	表:CPU,DIMM,高速伝送用コネクタ 裏:電源部品
ノードモジュール	シャーシサイズ	253mm×420mm t=60mm
	実装物	表:CPU×2,DIMM×32(RAM×320),光 mod×4 裏:DC-DC コン,高速伝送用コネクタ×4
ノード筐体	筐体サイズ	1200mm×700mm
	実装物	Node モジュール×32(64 ノード),電源
	性能	16.4TFlops/筐体



図 4-2 CPU パッケージ(TOP 面)

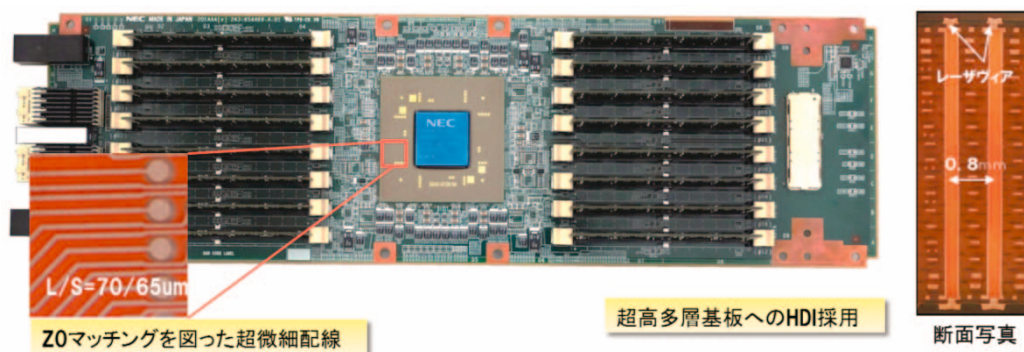


図 4-3 ノード基板

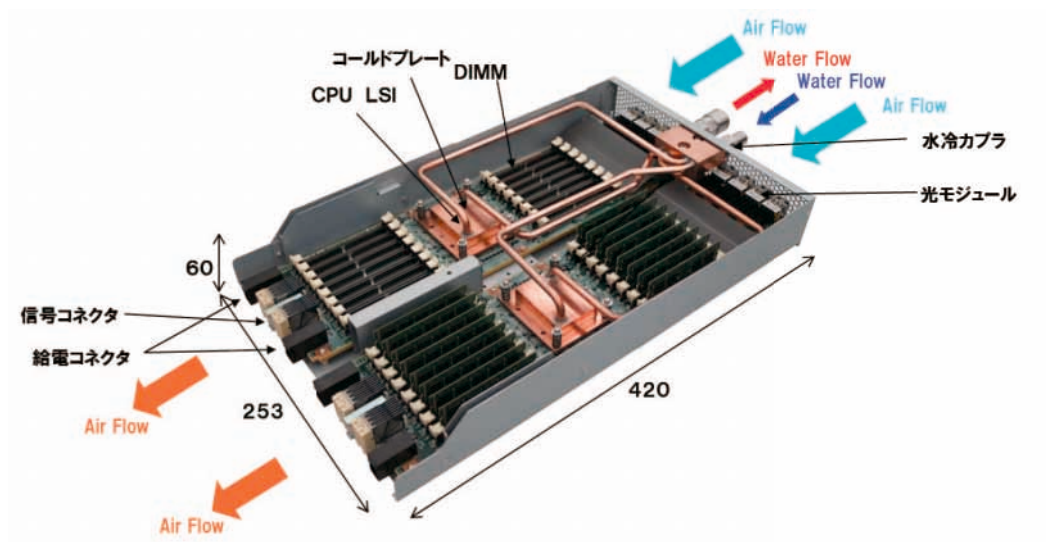


図 4-4 ノードモジュール

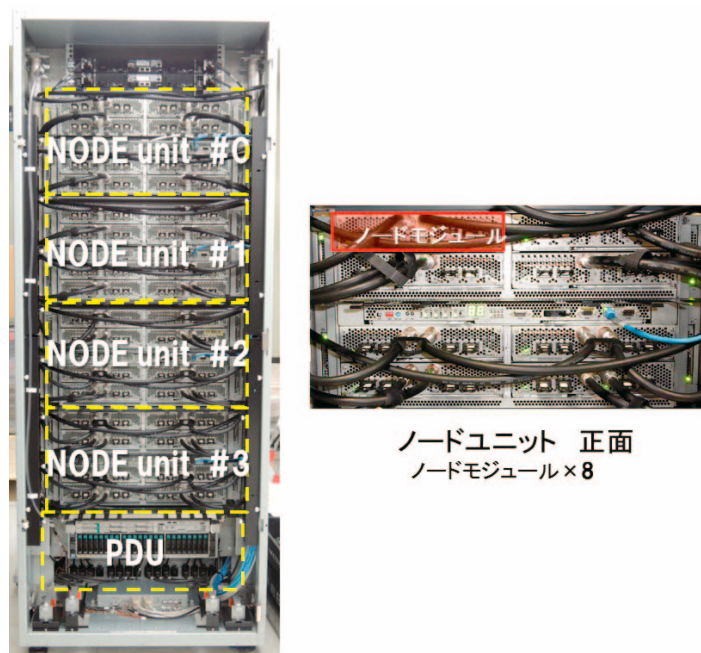


図 4-5 ノード筐体

5. おわりに

SX-ACE は、従来機種比で同一演算性能あたり、10 分の 1 の低消費電力化、5 分の 1 の省設置スペース化を達成しつつ、スーパーコンピュータの要件である高い演算性能と、それに見合うメモリからのデータ供給性能のバランスを重視した分散メモリ型のスーパーコンピュータ・システムとして開発しました。NEC は、今後も様々な研究分野の発展を支える強力なツールであるスーパーコンピュータを開発していきます。

参考文献

- [1] Shintaro Momose, Takashi Hagiwara, Yoko Isobe, Hiroshi Takahara, “The Brand-New Vector Supercomputer, SX-ACE”, Proceedings of ISC14, 2014
- [2] Shintaro Momose, “SX-ACE Processor: NEC’s Brand-New Vector Processor”, HOTCHIPS26, 2014
- [3] Akihiro Musa, Yoshiei Sato, Takashi Soga, Ryusuke Egawa, Hiroyuki Takizawa, Koki Okabe, Hiroaki Kobayashi, “Effects of MSHR Prefetch Mechanisms on an On-Chip Cache of the Vector Architecture”, Proceedings of International Symposium on Parallel and Distributed Processing with Applications, 2008.