[新スーパーコンピュータ SX-9]

スーパーコンピュータ SX-9のハードウェア

稲坂 純　　萩原 孝
日本電気株式会社 コンピュータ事業部

1. はじめに

本年 4 月よりセンターに導入された新しいスーパーコンピュータシステムである SX-9 のハードウェアについて紹介します。SX-9 は、旧システムの SX-7 のアーキテクチャを継承し、単一 CPU 性能を約 12 倍（8.8GFLOPS→102.4GFLOPS）、ノード性能で約 6 倍（281.6GFLOPS→1638.4GFLOPS）に引き上げています。また SX-7 向けに開発したソフトウェア資産を継承するシステムです。ノードあたり 16 台の中央処理装置（Central Processing Unit、CPU）を有し、ノード間を専用の超高速のノード間接続装置（Inter-node Crossbar Switch、IXS）により接続し、システム全体で 16 ノード、総 CPU 数 256、総合演算性能 26.2TFLOPS、総メモリ容量 16T バイトを有する、実効演算性能、スケーラビリティ、及び使いやすさに優れたスーパーコンピュータです。

特長は以下の通りです：

① CPU あたり 100GF を超えるベクトル演算性能
② ノードあたり 1.6TFLOPS の演算性能と 1TB の大容量メモリ
③ 専用のノード間接続装置（IXS）により、ノードあたり 128GB/s×2 のノード間データ転送バンド幅
④ 65nm 銅配線技術を用いた超高速、高集積 CMOS LSI による高密度実装
⑤ 省電力設計による消費電力・発熱量の削減、及び高密度実装による設置面積の削減

本稿では、上記特性を備えた SX-9 のシステム構成（アーキテクチャ）、及びテクノロジの概要について説明します。

2. SX-9 システム構成

SX-9 は、CPU と主記憶装置（Main Memory Unit：MMU）を密結合した共有メモリ型のシングルノード 16 台を超高速なノード間接続装置（IXS）によりクラスタ接続することにより、分散並列処理を可能としたシステムです。SX-9 のシステム諸元を表1、システム構成を図 1 に示します。

表 1 SX-9 システム主要諸元

中央処理装置（CPU）	CPU 数	256
	ベクトル性能	26.2TFLOPS
主記憶装置（MMU）	容量	16TB
	データ転送性能	64 TB/s
入出力機構（IOF）	スロット数（チャネル数）	128（512）
	総入出力性能	256 GB/s
ノード間接続装置	データ転送性能	128GB/s×2／ノード

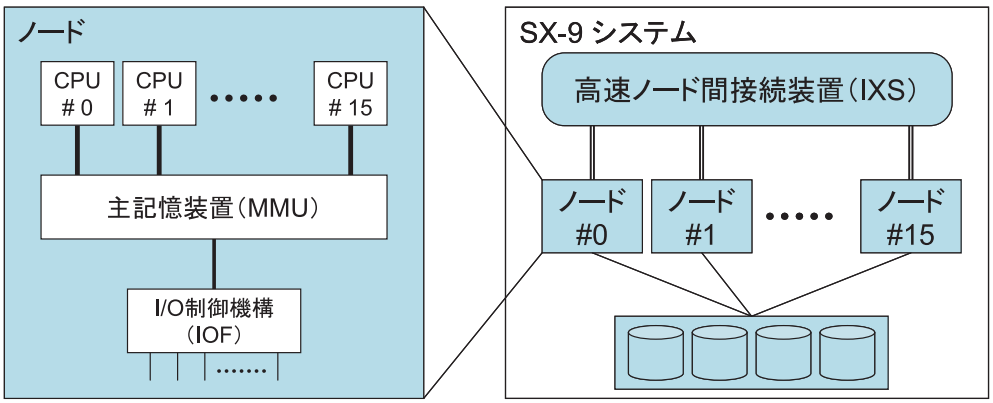

図1　SX-9 システム構成

各ノードは CPU 数 16、総合演算性能 1638.4GFLOPS、主記憶容量 1T バイト、8 スロットの入出力用の PCI スロットを持ち、演算性能、メモリスループット性能、入出力性能などのトータルバランスに優れ、高い実効性能を実現します。また、ノード内の共有メモリを利用した並列処理用に通信レジスタ（Communication Register：CR）を備えており、自動並列処理や、Open MP 指示行による並列処理時の CPU 間同期制御を高速に実行することが可能です。

SX-9 システム全体は、ノード間を超高速に接続する専用の IXS により、16 ノードをクラスタ接続したシステムであり、表 1 に示すように総 CPU 数 256、総合主記憶容量 16T バイト、総入出力チャネル数 256 の構成であり、総合演算性能 26.2TFLOPS という高い演算性能を実現します。

また SX-9 の 1 ノードにおける保守エリアを含む設置面積、及び消費電力はそれぞれ約 $2m^2$、及び約 30KVA であり、設置環境及び性能あたりの消費電力は前システム SX-7 と比較して大幅に改善しています。次節以降、SX-9 システムのハードウェアについて説明します。

中央処理装置（CPU）

SX-9 の CPU は従来の SX アーキテクチャを継承しつつ、さらなる機能・性能の強化を図っています。図 2 に CPU の構成を示します。CPU は、スカラユニット部、及びベクトルユニット部により構成され、プロセッサ／メモリ間ネットワークを介して MMU と接続されます。スカラユニットは命令の解読、ベクトルユニットへのベクトル命令の供給・起動、及びスカラ命令の実行を行います。一方、ベクトルユニットはベクトル演算部、及びベクトル制御部から構成されます。ベクトル演算部は 8 セットのベクトルパイプラインを備え、各ベクトルパイプラインは、乗算器×2、加算／シフト演算器×2、及び除算／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パイプライン、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。したがって、科学技術計算で主に利用される乗算と加算は、3.2ns のクロックサイクルあたり 32 個が同時に処理することが可能であり、最大 102.4GFLOPS のベクトル演算性能を実現します。

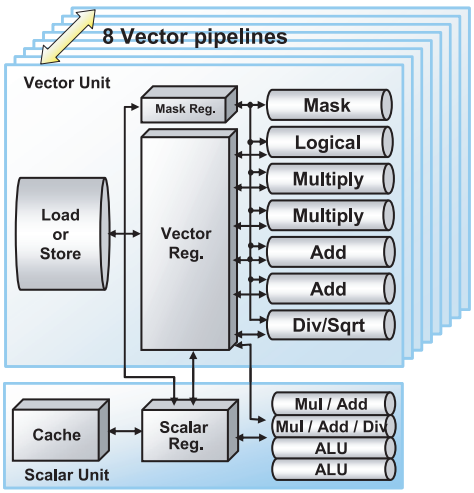

図 2　CPU の内部構成

スカラユニットは 64 ビット RISC アーキテクチャであり、L1 キャッシュ、命令同時デコード数 4、最大命令同時発行数 6 であるスーパスカラアーキテクチャ、アウトオブオーダー実行、及び命令投機実行の採用により短ベクトル時のベクトル命令発行性能、及びスカラ性能を向上させ、3.2GFLOPS の最大演算性能を実現しています。

主記憶装置（MMU）

スーパーコンピュータにおいて高い実効性能を実現するためには、高い演算性能に見合うだけの高いデータ供給性能が必要となります。SX-9 の MMU は、高速、かつ均一にメモリアクセス可能な共有メモリ方式を採用しています。MMU は 32768 個のメモリバンクを 64 バンク毎にメモリバンクを管理する制御部で管理し、その制御部にメモリバンクキャッシュ機構を備えてバンク競合を最小限に抑えつつ、高いメモリスループット性能を実現しています。

これにより SX-9 は表 2 に示すように、シングルノードにおいて、1T バイトのメモリ容量、及び 4T バイト／秒のメモリスループット性能、システム全体では 16T バイトの総メモリ容量、及び 64T バイト／秒の総合メモリスループット性能を実現します。

また、SX-9は、従来SXシリーズ同様に3次元実装構造のMMUカード実装を引き続き採用しています。これにより、RAM とメモリ制御部間の物理的距離を短くし、RAM の高速動作、及び RAMとCPU 間の高速信号伝送を実現しています。一方、メモリ信頼性確保のための ECC（Error Check and Correct）符号、タイミング、パリティ、2 重化回路などの採用による高いメモリ故障検出率の実現、擬似障害によるチェック回路の診断機能、エラー内容から即座にエラー箇所を指摘するビルトイン機能などにより、RAS（Reliability, Availability, Serviceability）機能の充実を図り信頼性を高めています。

表 2　主記憶装置諸元

	諸元
主記憶装置容量	1Tバイト
インターリーブ数	32768
CPU あたりのデータ供給能力	256G バイト／秒
最大データ供給能力	4Tバイト／秒

入出力機構（Input/Output Features：IOF）

入出力機構はシステムのスループットを高く保つために、SX-9 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転送性能を備えており、ノードあたり 8 スロット（但し、1 スロットはシステム制御用）、16G バイト／秒、SX-9 システム全体では 112 スロット、256G バイト／秒の性能を有します。入出力インタフェースは、ファイバチャネル（Fiber Channel: FC）、シリアル SCSI（Serial Attached SCSI: SAS）などの汎用インタフェースをサポートしており、様々な周辺機器を接続することが可能です。また、ネットワークインタフェースとして、ジャンボフレームに対応したギガビット・イーサネットを備えています。

I/O 処理において、CPU は全ての I/O 装置に対して対等にアクセスすることが可能であり、実行負荷の低い CPU を I/O 制御に割り当てるなど、CPU の効率的利用を可能としています。

ノード間接続装置（IXS）

SX-9 は図 1 で示したように、共有メモリ型のシングルノード 16 台を超高速専用クロスバスイッチを介して結合することにより、分散並列処理を可能としています。各ノードは、ノード間通信制御部（Remote Control Unit: RCU）を介して IXS とケーブル接続され、分散並列処理において低通信レイテンシ、及び高通信スループットを実現します。

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、ノードあたり 128G バイト／秒×2（双方向）の通信バンド幅を実現します。また、RCU は CPU とは独立に動作するデータムーバを持つことにより、異なるノードのメモリ間でデータ転送を行なうリモートメモリアクセスを CPU 動作とは完全に独立して行なうことが可能です。SX-9 の IXS は、従来の SX シリーズの IXS で採用していた回線交換型から、パケット交換型にデータ交換方式を変更しました。これにより、従来の回線制御オーバーヘッドを削減し、小データサイズの転送性能の向上を図っています。

3. SX-9 のテクノロジ

LSI 技術

これまで SX シリーズでは CMOS テクノロジによる高集積化，及びプロセッサのパラレル化により高性能化を実現しつつ、コストパフォーマンスを向上させてきました。

SX-9 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。また、システムの性能向上のためには、LSI 間信号伝送の高速化も非常に重要です。SX-9 では、新規のマルチチャネル・シリアル・インタフェースを開発し、LSI 間の高速データ転送を実現しています。また、インタフェース回路の低消費電力化、小面積化により LSI への多チャネルの搭載を実現しています。表 3 及び図 3 に SX-9 の CPU LSI 諸元、及び CPU チップ外観をそれぞれ示します。

表3　CPUチップ諸元

テクノロジノード	65nm
搭載トランジスタ数	3億5千万トランジスタ
電源電圧	1.0V
ピン数（内信号ピン）	8,960（1,791）
配線層構成	銅11層
I/Oインタフェース	CML（Current Mode Logic）
実装形態	ベアチップ実装

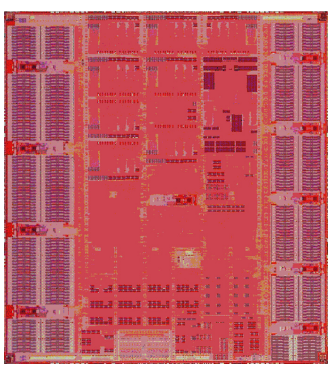

図 3　CPU チップ外観

SX-9 に使われる LSI の共通仕様として、65nm CMOS プロセス、11 層銅配線、及び低誘電率層間絶縁膜などの採用による配線遅延の改善、MIM（Metal-insulator-metal）プロセスの開発による大容量オンチップキャパシタの実現、ゲート酸化膜の薄膜化による高性能な低電圧電源の実現などを行なっています。さらに、新規のマルチチャネル・シリアル・インタフェースの開発により、CPU-MMU 間の低転送レイテンシを実現し、同時に LSI 上の電源分離、アナログ回路の削減、制御信号のデジタル化などによりノイズ耐力及びエラーレートの格段の向上を実現しています。

LSI 内部の RAM 回路は、デバイス性能を最大限引き出すために専用設計されています。また、

LSIの低消費電力化のために、読み出し回路のダイナミックパワーの削減、及びリーク電流の少ないトランジスタの使い分けによるスタティックパワーの削減を実現しています。また、高速なクロック動作を実現するために、LSI外部からクロックを逓倍するAPLL（Analog Phase-Locked Loop）回路を採用しています。

高速システムにおける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が必要となります。同様に、信号伝送の高速化を妨げる要因となる電源ノイズ対策も重要です。SX-9では高速、かつ安定した信号伝送を実現するために、信号伝送時の減衰が小さい低損失材料を使用した基板、伝送信号の波形を改善するイコライズ機能を備えた回路、及び波形ひずみの少ないソケットやコネクタなどを採用しています。また、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最適化などにより電源ノイズの低減を実現しています。

実装技術

SX-9は、世界最高性能、高コストパフォーマンス、及び優れた設置性を実現するために、高密度LSI実装技術、高密度接続技術、高効率冷却技術、及び高性能電源モジュール技術により、従来のワンチッププロセッサをさらに進化させました。

CPU、及びMMUモジュールの諸元を表4に、CPU、及びMMUモジュールの外観を図4にそれぞれ示します。超高速動作が要求されるCPU／MMUモジュールは、高密度実装により大型で多ピンのLSIを搭載可能としています。CPUモジュールは、CPU LSIをビルドアップ基板表面にベアチップ実装し、裏面には高速シリアル信号を伝送するケーブルを接続するための高密度コネクタを搭載しています。MMUモジュールは、メモリ制御用のHUB LSIとSDRAMをプリント配線基板に実装しています。

次にシステム実装技術について述べます。SX-9はルータ（RTR）モジュールと呼ばれるメインボードならびに高周波多芯ケーブルを介して、CPUモジュールとMMUモジュールを相互接続するケーブルインタフェース接続構造を採用しています。これにより、メモリユニットの高密度化とCPUモジュールとMMUモジュール間の効率的な接続を実現することができました。RTRモジュールは図5に示すように32枚のMMUモジュールと2個のルーティングスイッチ LSIを搭載しています。MMUモジュールから送信されたデータは ルーティングスイッチLSIでルーティングされ、高速シリアルインタフェース信号に変換された後、RTRモジュール裏面に搭載されたコネクタを経由し、高周波信号伝送用に開発した多芯ケーブルを介してCPUモジュールに伝送されます。

表 4　CPU／MMU モジュール諸元

項目		CPUモジュール	MMUモジュール
搭載 LSI（形態）		CPU LSI×1 （ベアチップ）	HUB LSI×1 （FBGA）
	ピン数	8960	840
	IO ピッチ（μm）	168	168
搭載 RAM		―	μBGA×24
配線基板	種類	ビルドアップ プリント配線基板	プリント配線基板
	基板サイズ（mm）	140×112.5	110×65
	基板厚（mm）	1.6	1.56
	基板層数	4ビルド－8コア－4ビルド	12
	配線密度（μm）	配線幅／間隙 ＝18／20	配線幅／間隙 ＝80／80
モジュール	入力端子数	2628	170
	冷却	空冷	
	消費電力（W）	240	41

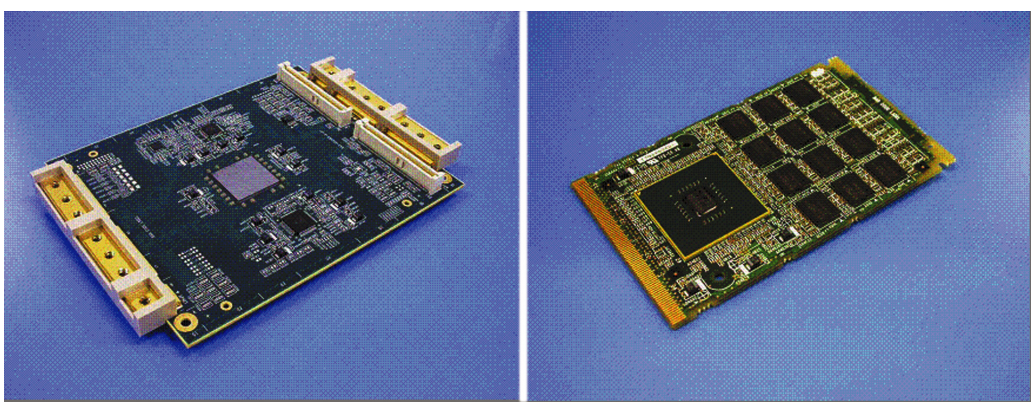

図 4　CPU モジュール、MMU モジュールの外観

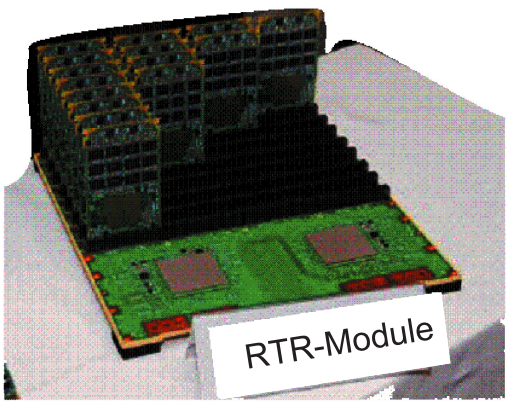

図 5 RTR モジュールの外観

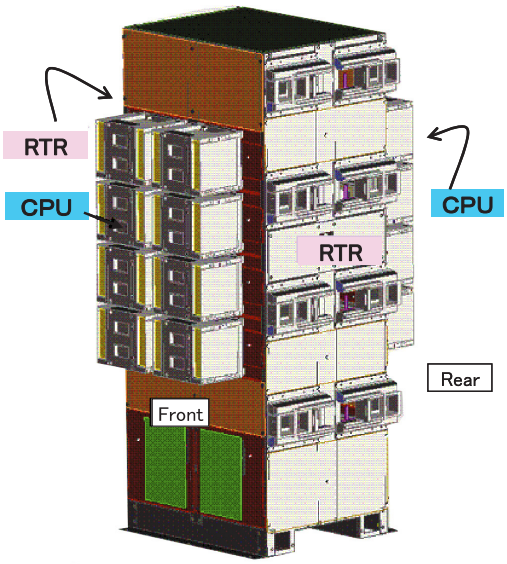

図 6　布線ボックスの外観図

CPU モジュールと RTR モジュール間の高密度多芯ケーブルは図 6 に示す布線ボックスと呼ばれるユニットに収納されています。このケーブルはノード内に搭載される CPU モジュール 16 ユニットと RTR モジュール 16 ユニット間を 1 対 1 で相互に接続することで、CPU チップと MMU モジュール間は独立した専用の高速信号伝送媒体を有することが可能となり、高性能 CPU と大容量のメモリモジュール間の広帯域のデータ転送バンド幅を実現し、SX-9 システムの高性能化に寄与しました。

4. おわりに

本稿では SX-9 のハードウェア概要について説明しました。SX-9 は、スーパーコンピュータの要件である CPU の高い演算性能と、それに見合う主記憶からのデータ供給能力のバランスを重視し、ユーザに使いやすい大規模な分散共有メモリ型スーパーコンピュータ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。