

SX-9のハードウェア技術(1)

～本体系～

中里 聡・多賀谷 聡・中込 師人
渡井 孝行・澤村 明寛

要 旨

SX-9のCPUは、8セットのベクトルパイプラインを有することにより、100GFLOPSを超える高い単体性能を実現しています。また、最大1Tバイトのメモリを最大16CPUが共有することも可能で、CPU-MMU間データ転送性能は最大4Tバイト/秒、I/O装置-MMU間データ転送性能は最大64Gバイト/秒×2を実現しています。本稿では、この高い性能を実現しているSX-9のCPU、メモリ、I/O装置について説明します。

キーワード

●スーパーコンピュータ ●プロセッサ ●ベクトル処理 ●共有メモリ ●主記憶装置 ●入出力処理

1. まえがき

SX-9は、SX-4からSX-8で実証された高い実効性能とノード内共有メモリの利便性を継承しつつ、更に増大する科学技術計算ニーズに応えるため、システム全体性能の強化はもとより、100GFLOPSを超える単一プロセッサ性能の実現とそれに伴う主記憶データ転送能力の強化を行っています。

以下では、その特長を中心に中央処理装置(CPU：Central Processing Unit)、主記憶装置(MMU：Main Memory Unit)、及び入出力処理装置について紹介します。

2. プロセッサ

SX-9のCPUは、従来のSXアーキテクチャを継承し、更に機能・性能の強化を図っており、前機種であるSX-8の1ノード相当の性能を1CPUで実現しています。

2.1 プロセッサ構成

CPUは大きく分けて、スカラユニットとベクトルユニットで構成され、プロセッサメモリ間ネットワークを介して主記憶装置に接続されています。また、SX-9ではプロセッサメモリ間ネットワーク内にADB (Assignable Data Buffer) を新規機能として搭載しています。図1にCPUの構成を示します。

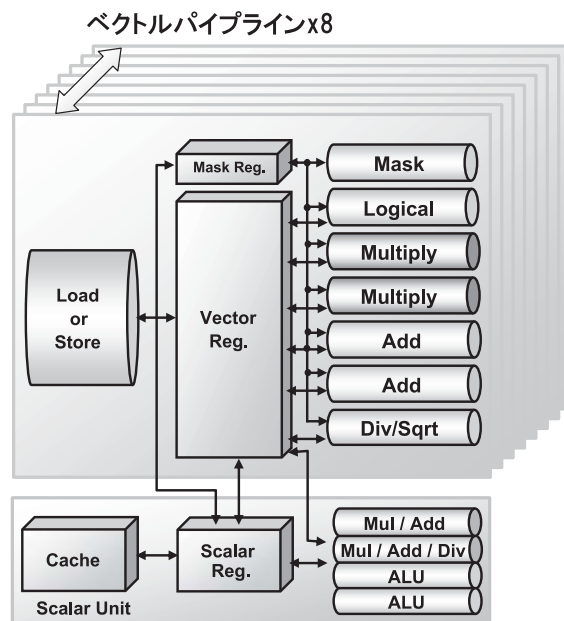


図1 CPUの構成

2.2 スカラユニット

SX-9のスカラユニットは、SXシリーズ互換の64bitRISCアーキテクチャであり、64bit幅の汎用レジスタを128本、命令・データそれぞれ32KバイトのL1キャッシュを搭載しています。スカラユニットの特徴を以下に示します。

(1) 6ウェイスーパースカラ構造

SX-9では、浮動小数点演算系2本、整数演算系2本、メモリ系2本の合計6本の実行パイプラインを持ったスーパースカラ構成になっています。これにより並列度の高いアプリケーションに対する性能向上を実現しています。

(2) 命令実行制御

アウトオブオーダー実行機構によってプログラム順によらず実行可能となった命令から実行することで高性能を実現しています。しかし、更なる性能向上のためには複数の分岐命令をまたいだより大きな領域に対する実行可能命令の検索が必須となっており、SX-9では、最大8つの分岐命令をまたいで検索を可能にしています。また、実際に分岐命令が実行されるまでの間、その後続命令の仮実行を行い、分岐予測が失敗した場合には正しい状態から実行を再開する投機命令実行機構を採用しています。なお、アウトオブオーダー実行機構における命令ウィンドウを広げるべく、SX-9ではリオーダーバッファが64エントリに強化されています。

2.3 ベクトルユニット

ベクトルユニットは、ベクトル演算部、ベクトル制御部から構成されます。

(1) ベクトル演算部

ベクトル演算部は、おのおのに並列動作可能な論理演算、乗算、加算/シフト演算、除算の基本演算パイプライン（乗算器及び加算/シフト演算器はそれぞれ各2パイプライン）に加え、マスク演算、ロード/ストアの各パイプライン、及び64ビットの容量を持つ16個のマスクレジスタ、おのおの512バイトの容量を持つ72個のベクトルレジスタから構成されるベクトルパイプラインを1セットとし、プロセッサ当たり8セットのベクトルパイプラインを有することにより、合計144kバイトのベクトルレジスタを備え、48本の基本演算パイプラインを同時に動作させることにより、強力なベクトル処理を行います。

1) ベクトル演算パイプライン

おのおのの演算パイプラインは、IEEE倍精度浮動小数点データ形式及びIEEE単精度浮動小数点データ形式をサポートしており、それぞれのデータ形式の切り替えは命令ごとに可能です。

また、通常、先行する演算の結果を後続の演算で使用する場合には、先行の演算結果をレジスタを介して使用しますが、先行演算の演算器出力のデータをそのまま後続演算の演算器に転送する機能を強化しました。この機能により短ベクトル長の命令の処理を従来機種以上に高速化しました。ベクトルパイプ間相互のデータ転送においても、100Gバイト/秒の転送性能により、ベクトルデータのベクトルパイプ間移送、マスクビットによるベクトルデータの圧縮・伸張も高速に処理することができます。

2) ベクトルレジスタ

各ベクトル演算パイプライン及びロード/ストアパイプラインへのデータの供給/格納を同時に行えるようにし、機能リソースを有効に使用できるように構成してあります。

3) ベクトルマスクレジスタ

マスクビットの生成は、論理演算パイプラインを使用して並列に行うようにし、また、他のマスク付き演算とベクトルマスクレジスタ間の演算、あるいは、ベクトルマスクレジスタ間の演算と他のマスク付き演算とのチェイニング機構をより強化することで性能の向上を図っています。

(2) ベクトル制御部

ベクトル制御部においては、SX-8では2系統であった命令発行ステージを、SX-9では24系統まで強化し、最大24命令を追い越しての発行を可能にし、かつ、7命令の同時発行を可能にしました。さらに、演算実行時のマスク制御を司るベクトル制御レジスタを多重化することにより、先行のマスク読み出し命令の実行終了を待たずに後続のマスク更新命令の追い越し発行を可能にするとともに、更にその後のマスク読み出し命令の追い越し発行も可能にしています。これにより、特に短ベクトル長の命令において、マスクビット確定待ちのタイムロスを大幅に削減しています。

2.4 ADB (Assignable Data Buffer)

ADBはプロセッサメモリ間ネットワーク内にあり、主にベクトルデータを選択的にバッファリングする機能を有しています。ADBとベクトルユニット間のデータ転送は、プロセッサメモリ間より高いデータ転送性能を、より短いレイテンシで実現できるため、利用頻度の高いデータをADBに配置することで、更に実効性能を高めることが可能です。また、これによりメモリトラフィックの削減とメモリバンク競合の

回避も可能となり、マルチプロセッサでの並列処理における性能低下を抑制することができます。

ADBに格納するデータは、ベクトルデータのみ、スカラデータのみ、ベクトルデータとスカラデータの両方、のいずれかを設定により切り替えることが可能です。このため、実行するアプリケーションに応じて、最適な利用方法を選択することができます。また、ベクトルデータの格納に対しては、それぞれのベクトルロード/ストア命令ごとにADBを利用するかどうかを制御できるため、必要なデータが意図せずADBから追い出されることを防ぐことが可能です。更に、よりADBを積極的に利用できるように、新規命令としてプリフェッチ命令を追加して、ソフトウェア制御によるプリフェッチ機能をサポートしています。

3. 主記憶装置

MMUは、各CPUが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。記憶素子としてはDDR3-SDRAM(Double Data Rate-SDRAM)を使用し、大容量メモリを提供しています。

3.1 主記憶装置の諸元

MMUは最大構成時、512枚のMMUカードから構成されます。MMUカード当たりのメモリ容量は2Gバイト、システムとしては256Gバイトから1Tバイトまでサポートしています(表)。

おのおのMMUカードはCPUからのメモリアクセス要求に対して同時に並列して動作可能であり、メモリスループット性能は、MMUカード当たり8Gバイト/秒、システム当たり最大4Tバイト/秒という極めて高いデータ転送能力を有しています。

表 主記憶装置の諸元

項目	諸元
記憶容量	256Gバイト～1Tバイト
インタリーブ数	8192 way～32768 way
記憶素子	1Gビット DDR3-SDRAM
論理素子	CMOS LSI
データ供給能力	1Tバイト/秒～4Tバイト/秒

加えて、メモリを最大32,768wayに分割してインタリーブすることにより、DDR3-SDRAMの動作率を上げ、実効性能を向上させています。

3.2 主記憶装置の特長

MMUは理論的なピーク性能値だけではなく、高い実効性能を実現するために、制御LSIの機能強化を行っています。

主記憶装置内にはメモリバンクキャッシュをCPUごとに備えています。各CPUからのロードリクエストに対しキャッシュヒットした場合はキャッシュからデータを読み出すため、後続リクエストのメモリアクセス待ち時間の短縮が可能となり、スループット性能を向上させています。

また、MMUカードはシステム当たりの使用個数が多いため、高い信頼性が必要とされます。そのため、SX-9では以下のような各種施策を実現しています。

制御LSI-記憶素子間のタイミングバジェットは、記憶素子のデータ転送速度の高速化のため年々厳しくなっており、DDR3-SDRAMにおいてはWrite LevelingとMulti Purpose Register (MPR) の2つのキャリブレーション機能が追加されています。この2つの機能を活用し、高速化が進む記憶素子インタフェースの最適化を実現することにより、タイミングマージンを拡大し、更なる信頼性向上を可能にしています。

更に、SX-9では高速CMOS LSI 1個とBGA (Ball Grid Array) 封止のDDR3-SDRAM 24個とを同一基板に実装する方式を選びました。LSIとRAMの配線長を短くすることにより、高速信号伝送の信頼性を高めています。

上記の施策以外にも、ECC (Error Correcting Code) 符号の採用による信頼性向上、更にタイミング、パリティ、2重化回路などのチェックにより高い故障検出率を実現しています。また、擬似障害発生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できるビルトイン診断機能などを有しており、RAS (Reliability Availability Serviceability) 機能の充実を図っています。

4. 入出力処理装置

SX-9の入出力処理装置は、SX-8から採用したDirect I/O方式を引き続きサポートしています。

4.1 入出力処理装置の特長

- (1) システムに実装されるすべてのプロセッサは、すべてのI/O装置に対して対等にアクセスすることが可能です。また終了通知の割り込み先CPUもソフトウェアから自由に設定することが可能であり、アプリケーションによる実行負荷の低いCPUをI/O制御に割り当てるなど、効率的なCPUの使用を可能にしています。
- (2) SX-9ではI/O制御方式としてCPUからのI/Oアクセス命令により、Host Bus Adapter (HBA) 内のメモリに直接アクセスするDirect I/O方式を採用しています。
- (3) 搭載するチャンネルカードの種類に応じて入出力インターフェースを選択できます。チャンネルカードの増設、削除、種類変更によるハードウェアの設定が不要なため、拡張が容易です。

4.2 入出力処理装置の構成

入出力処理装置の構成を 図2 に示します。SX-9の入出力処理装置は、IOF(IO Feature : ホストブリッジユニット)とPCI Express制御ユニットから構成されています。

IOFはシステム当たり最大16個搭載可能であり、IOF当たり2個のPCI Express制御ユニットを接続可能です。IOFに搭載されるすべての機能の設定は、IOFを仮想的に1つのチャンネル装

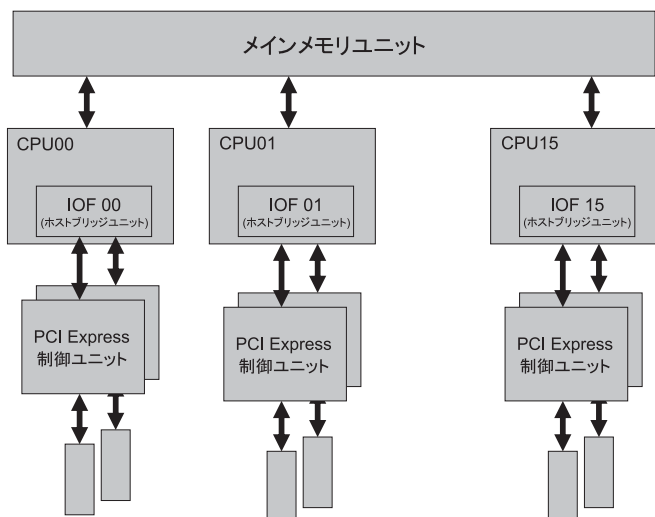


図2 入出力制御装置の構成

置に見せることにより、汎用のチャンネルカードと同一のアクセス方法でソフトウェアから任意のタイミングで設定、変更が可能です。

PCI Express制御ユニットは、PCI Express(x8)スロットを1本搭載するコントローラで、PCI Express制御ユニット当たり最大2Gバイト/秒×2 (双方向) のデータ転送性能を提供します。従って、1システム当たりの総合最大転送性能は64Gバイト/秒×2となります。

また、IOFとPCI Express制御ユニットは、それぞれ独立したデータバッファを搭載し、チャンネルごとにデータの流れを管理する構造としているため、チャンネル間でのデータ待ちなどによる転送処理の干渉を排除し、常に高いスループットがすべてのチャンネルで実現できるように構成しています。

更に、IOFは障害処理パスをノーマルリクエストパスとは別系統で持つことにより、IOF及び配下のPCI Express制御ユニットの初期化・設定・障害ログの採取を、ノーマルリクエストパスの障害時にも実行することを可能としています。

5. むすび

以上、SX-9のプロセッサ、主記憶、I/O装置を中心に紹介しました。SX-9は、高い実効性能で定評のあるSXアーキテクチャを継承しながら、命令実行制御の強化、メモリスループット性能の強化などを実施することにより、更に実効効率の高い、コストパフォーマンスに優れたスーパーコンピュータ製品として開発しました。

今後もユーザーニーズを取り入れ、最先端技術を駆使して、より優れたスーパーコンピュータを開発していく所存です。

執筆者プロフィール

中里 聡
第一コンピュータ事業本部
コンピュータ事業部
技術エキスパート

多賀谷 聡
第一コンピュータ事業本部
コンピュータ事業部
技術エキスパート

中込 師人
NECコンピュータテクノ
コンピュータ第二技術部
技術エキスパート

渡井 孝行
NECコンピュータテクノ
コンピュータ第二技術部
主任

澤村 明寛
第一コンピュータ事業本部
コンピュータ事業部
技術エキスパート