SX-9のハードウェア技術(2) ～ノード間スイッチ～

安藤 憲行・春日 康弘
鈴木 正樹・山本 孝人

要　旨

SX-9のノード間接続装置は、最大512ノードを接続する高いスケーラビリティを持つ専用高速ネットワークです。ノード間データ転送スループット性能が高いだけでなく、RCUとIXSが提供する機能によって短通信レイテンシも実現しています。本稿では、SX-9システムを構成するノード間結合装置の構成、特徴、機能、性能、及びアーキテクチャについて紹介します。

キーワード

●ノード間通信 ●転送性能 ●共有メモリ ●クロスバースイッチ ●RAS

1. まえがき

近年のスーパーコンピュータシステムは、共有メモリノードを高速ネットワークで接続したクラスタ構成システム（マルチノードシステム）が主流となっており、ノード性能の向上に伴って、ノード間接続ネットワーク性能がより重要なファクターとなってきています。スーパーコンピュータSX-9ではこれらの要求に応えるため、従来のノード間接続装置IXS（Internode Crossbar Switch）に対して8倍の転送性能を達成しています。以下、これらの装置を中心に紹介します。

2. マルチノードシステム構成

SX-9シリーズのマルチノードシステムは、共有メモリ型のシングルノードをクラスタ化して、超高速の専用クロスバースイッチであるIXSに接続することにより、最大512ノードを結合したシステムです。

マルチノードシステムは、ノード間のデータ転送帯域幅が非常に広いだけでなく、各ノード内のIXSとの接続装置であるリモートアクセス制御ユニット(Remote Access Control Unit ：RCU)と、IXSが提供する機能によって短通信レイテンシを実現しています。

SX-9マルチノードシステムの構成図を **図1** に示します。各ノードには、RCUが最大16台構成され、ケーブルでIXSに接続されます。各RCUは1レーンを構成しており、RCU当たり4Gバイト/秒×2の接続ポートを2ポート持ち、8Gバイト/秒×2の転送性能を有しています。従って、ノード当たり、最大16レーン、32ポートの接続ポートを持ち、転送性能は最大128Gバイト/秒×2になります。

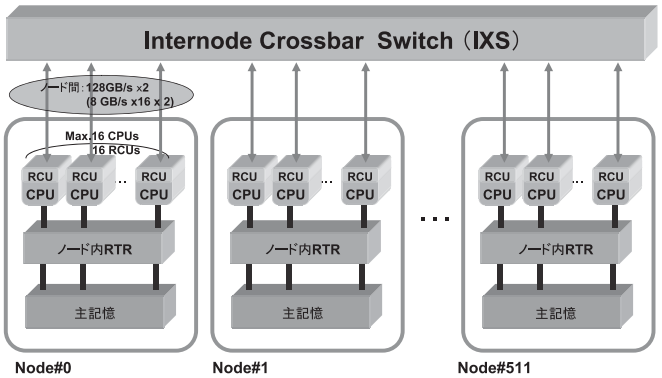

図1　SX-9マルチノードシステム構成

3. リモートアクセス制御ユニット(RCU) の構成と機能

3.1 RCUの構成

RCUは、各CPU LSI内に実装され、ノード間転送制御部、グローバルアドレス変換部、及びデータ送受信部から構成されます。データ送受信部は、CPU内クロスバスを介してノード内RTR（ルータ）、MMU（Main Memory Unit）と、セルフ

ルーティングのクロスバースイッチであるIXSに接続されています。ノード間転送制御部はCPUからの転送要求を受け付けノード間転送リクエストの発行を行い、ノード間データ転送を開始します。また、各RCUは他CPU LSIのCPUとノード内RTRを介して接続されており、ノード間転送制御部において他CPU LSIのCPUからも転送要求を受け付けることが可能です。データ転送に先立ちグローバルアドレス変換部において論理アドレスから物理アドレスへの変換が行われます。データ送信部はMMUからIXSへのデータ転送を行い、RCU当たり8Gバイト/秒、ノード当たり最大128Gバイト/秒の性能を有しています。データ受信部は、IXSからMMUへのデータ転送を行い、共にRCU当たり8Gバイト/秒、ノード当たり最大128Gバイト/秒の性能を有しています。データ受信部とデータ送信部は独立に動作し、ノード当たり最大128Gバイト/秒×2の通信バンド幅を実現しています（**図2** 参照）。

3.2 リモートアクセス機能

IXSとRCUによって、SX-9のCPUは他ノードのメモリと自ノードのメモリ間でデータ転送を行うことが可能になります。これはリモートメモリアクセスと呼ばれ、マルチノードシステムの基本動作となります。RCUは、CPU動作とは独立に動作するデータムーバを備えているため、ノード間のデータ転送をCPUでの演算動作とメモリアクセスとはまったく並列に行うことができます。

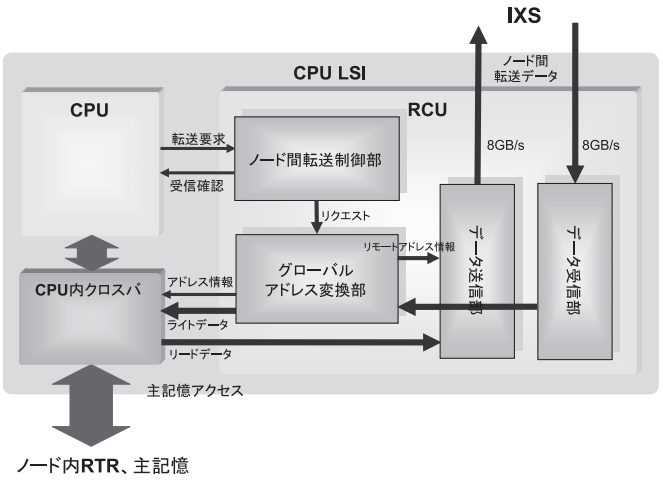

図2　RCUハードウェア構成

RCU内のデータムーバは、2種類のデータ転送命令（総称してINA命令と呼ぶ）をサポートしています。1つがCPUリソースの有効活用を優先する非同期転送命令で、もう1つがCPUと同期して転送を行う同期転送命令です。両タイプの命令ともデータ転送時のシステムコールによるオーバーヘッドを最小に抑えるために、非特権ユーザモードによる実行を可能としています。

また、SX-9では、非同期転送命令に短レイテンシを実現するための転送機能もサポートしています。

3.3 非同期リングバッファ

SX-9では、RCU内に設けられたポインタで主記憶上にジョブごとに配置された非同期リングバッファ内の非同期コマンドエリアを制御することで、ジョブごとに約15,000個の非同期命令をキューイングすることが可能です。これにより、CPUが非同期命令をキューイングできずに後続処理を停止させてしまうことを最小限に抑えています。この非同期リングバッファのポインタ制御は、RCU構成数によりハードウェアで制御しています。これにより、ソフトウェアはRCU構成を意識しないキューイングが可能となります。

3.4 メモリ保護機構

複数のプログラムが互いに干渉することなく、不用意なメモリ破壊が起きないよう、RCUでは、GSATB（Global Storage Address Translation Buffer）を備え、メモリ保護を実現しています。

また、論理ノードという概念を導入しています。分散マルチノードプログラムを実行する物理ノード番号や、物理ノード数に変更があっても処理が行えるように、GNATB（Global Node Address Translation Buffer）を備えることにより、論理ノードと物理ノードのマッピングをハードウェアで実現しています。

3.5 高速ショートメッセージ通信

通常のリモートアクセスは、ローカルノードの主記憶とリモートノードの主記憶を介して転送が行われるため、これまでのシステムでは、P2Pショートメッセージ転送のように、

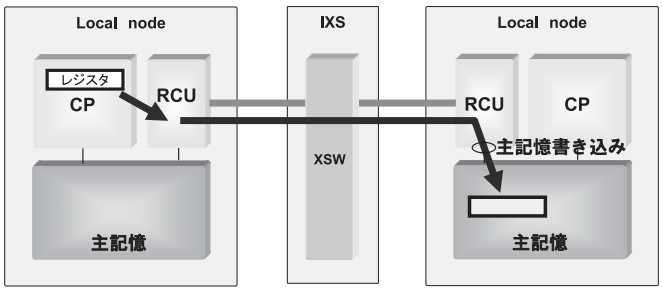

図3　P2Pショートメッセージ転送

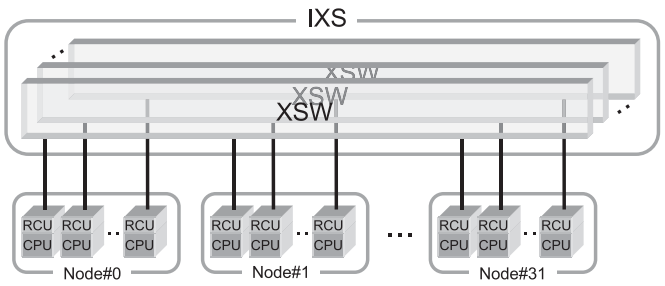

図4　32ノード・フルクロスバ構成

メッセージ長が短いノード間転送では、レイテンシが大きく十分な性能を提供することができませんでした。

SX-9では、このようなショートメッセージにおいて、ノード間通信制御を行うRCUがCPU内スカラレジスタの値を直接リモートノードの主記憶に書き込みを行い、CPUとRCUのハンドシェイク処理を簡素化することにより、レイテンシ削減を行い、性能向上を実現しています（**図3** 参照）。

4. IXSの構成と機能

SX-9のIXSは、専用クロスバースイッチであるXSWによって構成され、最大512ノードまでの16マルチレーンネットワーク接続を可能にしています。また、RCU当たり8Gバイト/秒×2の転送性能を有しており、ノード当たりでは最大128Gバイト/秒×2の転送性能を実現しています。

4.1 IXSの構成

IXSを構成するXSWは転送性能4Gバイト/秒×2のポートを32本装備しており、各ポートがRCUの1ポートと1対1で接続されます。32ノードまでは1つのXSWが各ノードのRCUと直接接続されることによって、XSW単段によるフルクロスバー接続でのマルチノードシステムを構成します。ノード内の各RCUはそれぞれ異なるレーンに接続され、複数レーンを同時動作させることによってスループット性能を向上させています。ノード数の少ない構成では、物理的に1つのXSWが複数のレーンを受け持つことになりますが、XSW内では異なるレーン間の接続は論理的に切り離されています（**図4**）。

33ノードを超える場合はXSWを2段接続したFat Tree（ファットツリー）構成を採用しています。16ノード単位で1段目XSWによりまとめられ、それを2段目XSWが束ねる構造になっています。これにより最大512ノードまでのマルチノードシステムを構築することができます。32ノード以下の単段構成と同様に、複数レーンによる同時動作を行うことでスループット性能の向上が可能で、多段構成の場合はレーン単位で物理的に独立した構成となります（**図5**）。

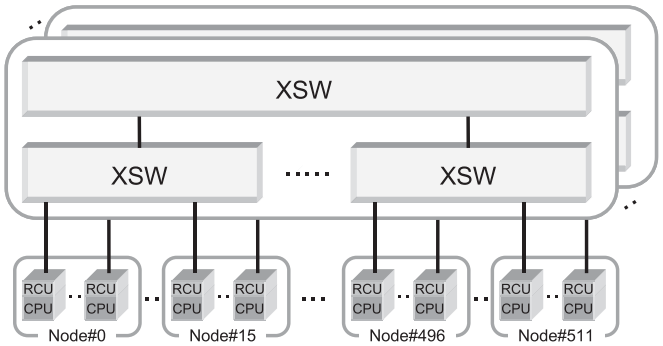

図5　512ノード・ファットツリー構成

4.2 データ転送方式

前機種のSX-8が回線交換方式であったのに対して、SX-9のIXSではパケット交換方式を採用しています。クロスバースイッチであるXSWは、ノードからのパケットを蓄える入力バッファ、パケットを宛先ノードへと振り分けるクロスバースイッチ、ノードへ転送するパケットを保持する出力バッファ、及びノード間の同期・排他制御を効率よく行うための共有通信レジスタ（GCR：Global Communication Register）を備えています。ノードからのデータパケットはヘッダ部に記載された宛先情報に従って目的のノードへルーティングされま

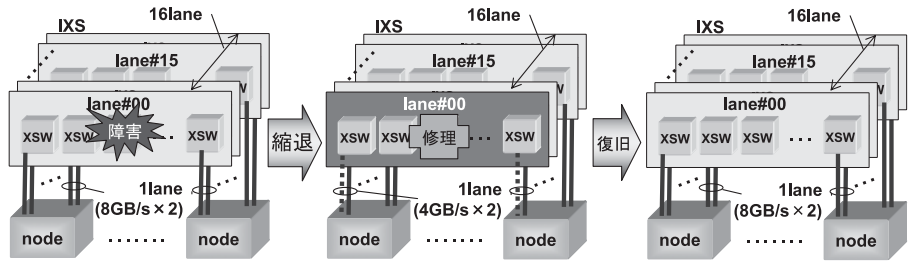

図6　マルチノードシステムのレーン縮退

すが、入力バッファ及び出力バッファを装備することによってスループット向上を図っています。入力バッファのワード数は、最大512ノード構成時のノード筐体配置を考慮して、ノードーIXS間ケーブル長40mでも十分なスループットを確保できるだけの容量を確保しています。また、XSW内部のクロスバースイッチを2分割構成とすることでパケット間の競合を減らし、スループット性能の向上を図っています。

5. マルチノードシステムのRAS機能

SX-9マルチノードシステムは、複数ノードをノード間接続装置であるIXSにより接続したスケーラビリティに優れた大規模システムであり、システム性能の高速化要求に応えるために、最大0.8PFLOPS（1.6TFLOPS×512ノード）のシステム性能を提供しています。

システム規模が大きくなると使用部品数が増加するため、高い可用性を確保するためには各部品の万が一の故障に対する十分な備えが必要になります。

SX-9では従来システムからの優れたRAS機能を踏襲しつつ、新たにマルチレーンで接続されたIXSの障害に対応し、レーン当たりの性能を縮退することでシステムとしての可用性を強化しています。

図6 に示すマルチノードシステムではレーン当たり8Gバイト/秒×2のノード間転送パスを最大16レーンでIXSに接続します。各レーンは2本のパスに分割（1パス当たり4Gバイト/秒×2）され、この32本のパス（2パス×16レーン）を用い並列転送を行うことで、ノード当たり最大128Gバイト/秒×2の高速なノード間転送性能を実現します。

本システムは、ノード間ネットワークを構成するIXSの複数のレーンのうち、一部で障害が発生した場合、障害レーン内の障害パスを自動的に切離し、レーン転送性能を8Gバイト/秒×2から4Gバイト/秒×2に縮退した運転に切り替えます。

このとき、障害が発生していないレーンでは転送パスの縮退は行わないため、縮退運転での転送性能低下は1/32（約3%減）であり、最小限の性能低下を実現しています。

レーン転送性能の縮退運転時にはマルチノードジョブの再投入を行い、また、レーン転送性能の復旧時はノード内の運用を継続したまま組み込みを可能とすることで、可用性の確保が可能なシステムを提供しています。

6. むすび

以上、SX-9マルチノードシステムの高いスケーラビリティと高性能を支えるノード間接続装置について紹介しました。今後もより増大するHPCへのニーズを満たすため、機能・性能を強化した製品を開発していく予定です。

執筆者プロフィール

安藤 憲行
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

春日 康弘
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

鈴木 正樹
NECコンピュータテクノ
コンピュータ第二技術部
技術マネージャー

山本 孝人
NECコンピュータテクノ
コンピュータ第二技術部
主任