〈SX-8のハードウェア〉

SX-8のハードウェア技術(2)～ノード間接続・入出力装置～

Internode Crossbar Switching Unit and I/O Processing Unit of SX-8

安藤憲行* 井川康宏* 山本孝人**
Noriyuki Ando Yasuhiro Ikawa Takahito Yamamoto
長野知明* 萩原 孝*
Tomoaki Nagano Takashi Hagiwara

要 旨

SX-8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。また，SX-8の入出力処理装置は，SX-7までの専用IOP方式からHBAにCPUが直接起動を指示するダイレクトIO方式に変更し，高速ファイルアクセス，通信ネットワークを提供します。本稿では，SX-8システムを構成するノード間接続装置および入出力処理装置の構成，特長，機能，性能，およびアーキテクチャについて紹介します。

The internode crossbar switch of SX-8 has a high-speed network with a high scalability that connects 512 nodes. I/O processing unit of SX-8 is changed to a direct IO method that CPU issues IO instructions to HBA (Host Bus Adaptor) directly from the IOP method of SX-6/7.

This paper describes the configuration, characteristics and architecture of internode crossbar switch and the I/O processing unit of the SX-8 system.

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。また，システム規模の拡大につれて，スーパーコンピュータが扱うデータ量は飛躍的に拡大しており，強力な入出力機構のニーズも高まっています。スーパーコンピュータSX-8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（Internode Crossbar Switch：IXS）に対して2倍の性能を達成しています。また，従来の専用IOP方式からCPUが直接起動を指示するダイレクトIO方式に変更し，即応性を高めた入出力機構を実現しています。以下，これらの装置を中心に紹介します。

2．システム構成

2.1 シングルノードシステム

SX-8のシングルノードシステムの構成，および構成要素のシステム内での位置づけを**図1**に示します。

シングルノードシステムは，大規模な主記憶装置（Main Memory Unit：MMU）に内部ネットワークによって最大8台のCPUが接続された共有メモリシステムを構成しています。さらにMMUに同様に接続される最大4台のIOユニットが存在します。

2.2 マルチノードシステム

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スイッチであるIXSによって最大512ノードを結合したシステムです（**図2**）。特に転送帯域幅が非常に広いだけでなく，各ノードのIXS接続ポートであるリモートアクセス制御ユニット（Remote Access Control Unit：RCU）とIXSが提供する高速同期機能によって低い通信レイテンシを実現しています。

3．RCUとIXS

SX-8マルチノードシステムは，共有メモリの範囲を超えて

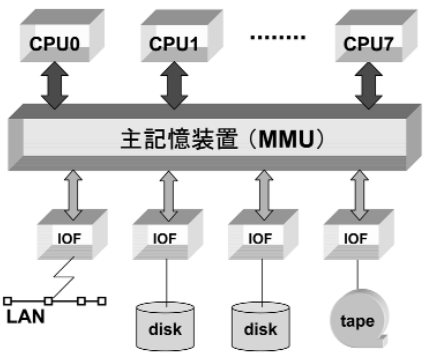

図1 SX-8シングルノード構成
Fig.1 Configuration of SX-8 single-node system.

＊ コンピュータ事業部
Computers Division

＊＊ NECコンピュータテクノ
NEC Computertechno, Ltd.

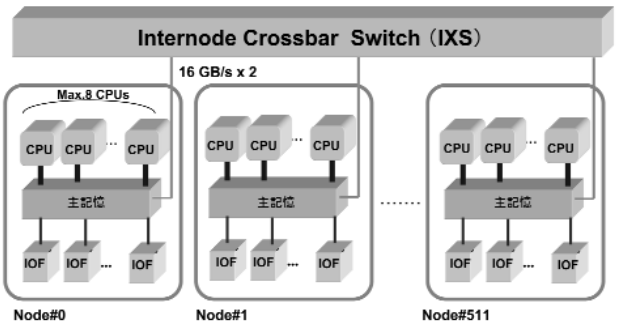

図2　SX-8マルチノードシステム構成
Fig.2　Configuration of SX-8 multi-node system.

最大4,096台のCPUによる分散並列処理を可能としました。

SX-8マルチノードシステムのハードウェア構成を**図3**に示します。SX-8の各シングルノードシステムには，RCUが最大2台実装され，ケーブルでIXSに接続されます。

3.1　RCUの構成と機能

（1）構　成

RCUは，ノード内に最大2台実装され，データ送受信部と送受信制御部から構成されます。データ送受信部は，MMUとセルフルーティングのデータ系クロスバスイッチであるXSWに接続されています。データ送信部はMMUからXSW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16Gバイト/秒の性能を有しています。データ受信部は，XSW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16G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16Gバイト/秒×2の通信バンド幅を実現しています（図3参照）。送受信制御部は，データ送信部，データ受信部とIXS制御部であるXCTに接続され，XCTとの制御情報に基づき，データ送信部とデータ受信部の制御を行います。

（2）リモートアクセス機能

IXSとRCUによって，SX-8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。

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，分散された多次元配列の境界データ転送の再配置に柔軟に対応するため，2ディスタンス転送，および，3ディスタンス転送機能をサポートしています（**図4**参照）。

（3）非同期リングバッファ

SX-8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。

（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

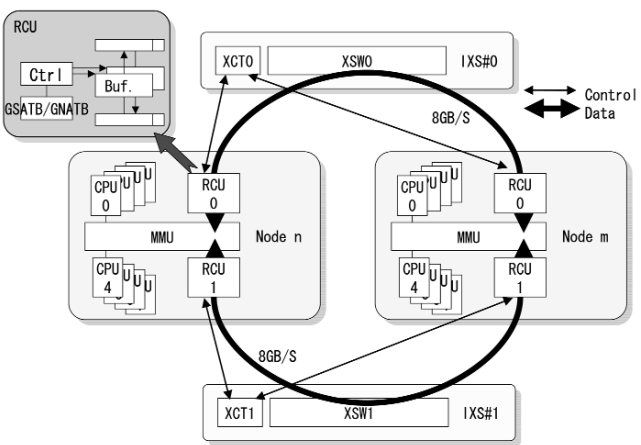

図3　ノード間接続ハードウェア構成
Fig.3　Configuration of internode communication.

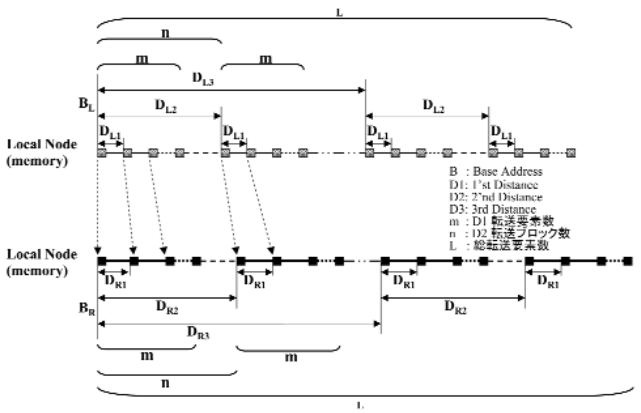

図4　3ディスタンス転送
Fig.4　3 distance data transfer.

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。

3.2 IXSの構成と機能

IXSは，データ系クロスバスイッチXSWとコントロール系クロスバスイッチXCTで構成され，各ノードのRCUとのインタフェースを持っています。XSWは転送データをルーティングするフルクロスバ・スイッチであり，ノード当たり16Gバイト/秒のスループット性能を有しています。XCTはコントロール系のノード間通信用クロスバ・スイッチ，ノード間転送のパス管理機構，GCR（Global Communication Register）を備えています。GCRは複数ノード間の同期・排他制御などを効率よく行うために，システム内の任意のCPUが，Test & Set，Fetch & Incrementなどを実行できる共有レジスタです。

IXSは回線交換型のデータ交換網であるため，データ転送に先立ちXCTに対して転送経路獲得の要求を行い，XCTから許可が得られるのを待ちます。許可が得られたら，送信ノードは，XSWに転送パス設定コマンドと転送データを送出し，転送データ送出後に転送パス解放コマンドをXSWに送出します。転送データ受信ノードは，転送データを受け取った後，XCTに転送経路解放を行い，1つのデータ転送が完了します。

3.3 マルチノードシステムのRAS機能

SX-8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（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65TFLOPS（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-8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IXSの障害に対応してお客様の運用形態に合わせて選択可能な2つのシステム構成を用意してシステムとしての可用性を強化しています。

（1）性能重視型マルチノードシステム

図5に示す性能重視型マルチノードシステムはノード当たり2本有する8Gバイト/秒×2のノード間転送パスを2つのIXSにそれぞれ接続します。この2本のパスを用い並列転送を行うことにより，ノード当たり最大16G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，IXSの障害発生時に，自動的に障害発生IXSを切り離し，残ったIXSでの8Gバイト/秒×2の転送性能による縮退運転に切り替えます。縮退運転時にはジョブを再実行（再投入）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。

（2）運用継続型マルチノードシステム

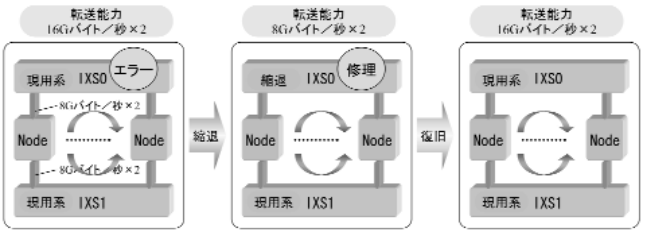

図5 性能重視型マルチノードシステム
Fig.5 High performance model.

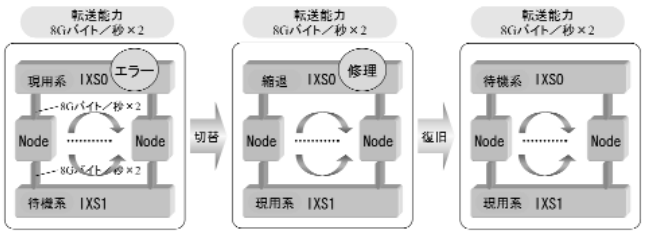

図6 運用継続型マルチノードシステム
Fig.6 Hot stand-by model.

図6に示す運用継続型マルチノードシステムは，2本の8Gバイト/秒×2のデータ転送パスを2つのIXSにそれぞれ接続して片方のIXSを現用系，他方を待機系として8Gバイト/秒×2のノード間転送を行う冗長システムです。この冗長構成により，現用系IXSで障害が発生した場合，転送パスを自動的に待機系に切り替えることが可能です。それにより，システムを停止することなく，運用を継続する高い可用性を実現するマルチノードシステムを提供しています。

4．入出力処理装置

SX-8の入出力処理装置は，SX-7までのIO Processor方式を廃止し，新たにDirect IO方式をサポートしています。また，I/O専用のTLBをI/O Control Chip内に持ち，CPUに負荷をかけることなくI/O動作上のセキュリティを高めるユニークな構造を備えています。

4.1 入出力処理装置の特長

（1）ノード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IO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-8ではIO制御方式としてCPUからのI/Oアクセス命令により，HBA（Host Bus Adapter）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。従来のIOP方式で存在したChannel Program生成過程でのオーバーヘッドを排し，より短TATで高スループットなI/O制御を実現しています。また本方式の採用により，従来のChannel Program方式では実現不可能であったZero Copy転送のサポ

ートを可能とし，高速なユーザレベル通信機能のサポートも新たに実現しています。

（3）一般的なMapped IO方式を採用するシステムでは，IO用TLBはMain Memory用TLBと共用する構造を採りますが，SX-8ではI/O専用TLBをI/O装置内に独立して持つユニークな構造を採用しています。これにより，IO要因によるTLBミスが発生するケースでも，CPUは後続命令の発行を妨げられることがないため，命令処理効率の向上が可能です。また，Direct IO方式でありながらMapped IO方式と同様のIO Memory空間の仮想化を可能にしたことにより，HPCで必要とされるresume/restart時のChannel構成の変更に柔軟に対応するとともに，IOリクエストの発行権限管理を行うことが可能となり，高度なセキュリティ性を提供します。

4.2 入出力処理装置の構成

入出力処理装置の構成図を**図7**に示します。SX-8の入出力処理装置は，IOF（IO Feature：ホストブリッジユニット）とバス制御ユニットから構成されています。

IOFは1ノード当たり最大4つ搭載可能であり，IOF当たり配下に4つのバス制御ユニットを接続可能です。IOF内にはIO専用のTLB機構を搭載し，配下のチャネルに対するアクセス権のチェック，IO空間アドレスの仮想マッピング，物理アドレス変換後のアドレスによる命令の振り分けを行っています。IOFに搭載されるすべての機能の設定は，IOFを仮想的に1つのチャネル装置に見せることにより，汎用のチャネルカードと同一のアクセス方法によりソフトウェアから任意のタイミングで設定，変更を可能としています。さらに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，万一の障害発生時に速やかな障害情報収集と，復旧手段の確保を可能としています。

バス制御ユニットは，配下にPCI-X BUSを2本搭載するバスコントローラで，バス制御ユニット当たり最大2Gバイト/秒のデータ転送性能を提供します。1ノード当たりのバス制御ユニットの搭載可能台数は16台であり，総合最大転送能力は30Gバイト/秒を達成しています。

バス配下に搭載されるチャネルはMulti Functionをサポートし，すべてのFunctionに対して独立したリクエスト調停を行うことにより，すべてのチャネルの効率的な使用を可能としています。

また，IOFとバス制御ユニットは，それぞれ独立したデータバッファを搭載し，チャネルごとにデータの流れを管理する構造となっており，チャ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ネルで実現できるように設計されています。

さらに，IOFは専用の診断パスを搭載し，IOFおよび配下のバス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が可能となっています。

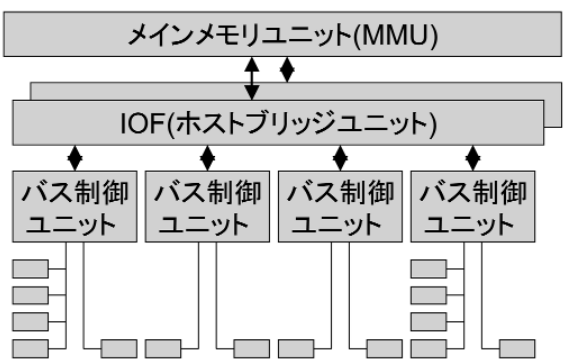

図7 入出力制御装置の構成
Fig.7 Configuration of IO processing unit.

5．むすび

以上，SX-8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とSX-8で新たに開発したDirect IO方式の入出力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。

筆者紹介

Noriyuki Ando
安藤 憲行（あんどう のりゆき） 1989年，NEC入社。現在，第一コンピュータ事業本部コンピュータ事業部第四技術部技術エキスパート。情報処理学会会員。

Yasuhiro Ikawa
井川 康宏（いかわ やすひろ） 1987年，NEC甲府（現NECコンピュータテクノ）入社。現在，NEC 第一コンピュータ事業本部コンピュータ事業部第四技術部技術エキスパート。

Takahito Yamamoto
山本 孝人（やまもと たかひと） 1988年，NECコンピュータテクノ入社。現在，コンピュータ第二技術部主任。

Tomoaki Nagano
長野 知明（ながの ともあき） 1992年，NEC入社。現在，第一コンピュータ事業本部コンピュータ事業部第四技術部主任。

Takashi Hagiwara
萩原 孝（はぎわら たかし） 1989年，NEC入社。現在，第一コンピュータ事業本部コンピュータ事業部第四技術部技術エキスパート。