超並列計算機SR2201のネットワーク インタフェース・アーキテクチャ

1F－1

岩嵜正明*，薗田浩二*，樋口達夫**，中越順二***，森山建三****

*（株）日立製作所システム開発研究所　**（株）日立製作所中央研究所

（株）日立製作所汎用コンピュータ事業部　*（株）日立製作所ソフトウェア開発本部

1．はじめに

分散メモリ型並列計算機SR2201は、3次元ハイパークロスバー・ネットワーク（以下HXB）によってノード間を接続する。このHXBの最大サイズは、X,Y,Z各軸方向に8×17×16であり、2176ノードを接続できる。HXBは、高々3回のルーティングによって任意のノード間を接続でき、レイテンシ数μsec以下、転送経路当たりのピークスループット300MB/secのハードウェア性能を有する。

本稿では、このHXBと各ノードを接続するネットワーク・インタフェース・アダプタ（NIA）のアーキテクチャについて述べる。NIAの設計目標は、メモリコピー処理や割込み処理のソフトウェア・オーバヘッドを低減し、ハードウェア性能に見合ったEnd-to-Endの通信性能をアプリケーションに提供することである。

2．NIAの概要

NIAは、物理的に1系統のHXBに対して、論理的に2種類のインタフェース、Remote DMAインタフェース（RDMA I/F）とRingBufインタフェース（RingBuf I/F）とを提供する。前者はRemote Memory Write機能を提供し、後者はメッセージ・パッシング機能の効率的実装を可能とする。

NIAの制御レジスタ群は、RDMA I/FとRingBuf I/Fのそれぞれに対応して存在する。OSは、RDMA I/FとRingBuf I/Fのそれぞれに対応する2系統のデバイスドライバを備える。この2系統のデバイスドライバから発行された送信要求は、送信側NIAによって多重化され、また、HXBから受信したパケットは、受信側NIAによって分離され、各デバイスドライバに渡される。

2.1. パケット通信

NIAインタフェースではデータをパケット化して送受信する。送信側でのパケットへの分割、受信側でのパケットの組立は、ソフトウェアが行なう。OSは、パケット・サイズの上限をハードウェア上限値（64KB）以下に制限し、ネットワーク上でのパケット衝突による転送遅延時間の増加を低減する。1パケットに収まらない大量データを転送する場合には、TCW（Transfer Control Word）チェイン機能を利用して、複数パケットを一括処理する。

2.2. 可変長ヘッダのスプリッティング

NIAが可変長ヘッダのGather/Split機能を備えることによって、ヘッダ付加、削除のためのメモリコピー・オーバヘッドを排除している。ヘッダ長はソフトウェアによってパケット毎に設定できる。送信側NIAは、TCWと転送対象のデータを結合（Gather Read）してパケットを作成する。受信側NIAは、ヘッダとデータをそれぞれ別々のメモリ領域に分割して書込む（Split Write）。

2.3. 送信者主導の受信完了割込み有無制御

NIAは、受信側でパケット到着時に割込みを発生させるか否かを、送信側のソフトウェアにより制御できる機能を備える。各TCW内の受信完了割込み制御ビットによって、パケット毎に受信時の割込み有無を指定できる。このビットがOnの場合のみ、受信データをメモリに書き込んだ後、NIAから受信ノードのCPUに対して「受信完了割込み」が発生する。

この機能により、低オーバヘッドのSpin Wait、あるいは、受信処理の一括化が実現できる。大量データをTCWチェインで連続的に転送する場合、例えば数十パケット毎に1回割込みを発生するといった一括受信処理を行ない、受信側の割込み処理オーバヘッドを低減できる。

2.4. TCWとTCWチェイン

パケットを送出するには、宛先ノード等を指定するTCWをソフトウェア側で作成し、そのTCWの先頭アドレスをNIAの起動レジスタに格納する。NIAはTCW内の情報からパケット・ヘッダを生成する。複数のTCWをチェインすることにより、1パケットに収まらない大量データを転送できる。

TCW内の送信完了割込み制御ビットによって、チェインされた各TCWの実行を完了した時点で、NIAが送信ノードのCPUに対して「送信完了割込み」を発生させるか否かを制御できる。2ノード間で複数パケットにまたがる大量データ転送を行なう場合、TCWチェイン最後尾のTCWのみこのビットをOnに設定する。これによって、CPUの介在なしに連続的にパケットの送出が可能となり、高スループット化が実現できる。

3．Remote DMAインタフェース

RDMA I/Fは、送信側／受信側ともにユーザの仮想アドレス空間の一部をリアルメモリ上に固定的にマッピングしておき、それらのメモリ間でデータ転送を行う高速通信方式である。異なったノード上のユーザ仮想アドレス空間々で直接データ転送を行うため、カーネル空間とユーザ空間々でのデータコピーが発生しない。このリアルメモリ上に固定的に割り当てた領域のことをCombuf（direct mapped COMmunication BUFfer）と呼ぶ。

RDMA I/Fは、送信者主導（Sender Initiative）のノード間直接メモリ転送機能を提供する。送信側のソフトウェアは、送信領域先頭アドレス，転送データ・サイズ，

Network Interface Architecture of Parallel Computer SR2201,
Masaaki IWASAKI, Kouji SONODA, Tatsuo Higuchi,
Junji NAKAGOSHI, Kenzou MORIYAMA
Hitachi, Ltd.

受信ノード番号，受信領域先頭アドレスの4パラメータを指定して、NIAにメモリ転送の起動を指示する。通常のDMA転送と同様に、実際のメモリ転送はCPUの介在なしにNIAによって実行される。

3.1. Combufの初期化とLCT

Combufは、物理メモリ常駐を保証されたユーザ仮想アドレス空間上の連続領域である。更に、ひとつのCombufは物理アドレス空間上でも連続アドレスである。OSは、ユーザ仮想アドレス空間上の領域を、物理アドレス連続な物理メモリ上に常駐化させる初期化機能を提供する。

Combuf間でデータ転送を行なうには、データ転送の開始に先立って、送信側ノード及び受信側ノードのそれぞれで初期化処理が完了していなければならない。また、データ転送の開始に先立って、Combufの先頭アドレス，サイズ，アクセスコードをLCT（Local Combuf Table）に登録し、受信側のアクセスコードを送信側に渡す前処理を行う。

LCTは、ノード内のCombuf領域の先頭アドレスやサイズを管理するテーブルで、NIAとOSによって共有する。LCTは、Combuf先頭物理アドレス、Combufサイズ、アクセスコード、有効ビット等のフィールドから構成されるエントリをCombufの個数分保持する。

3.2. アクセスコード

アクセスコードは、リモート・メモリ書込みに関するメモリ保護機能を提供する。即ち、アクセスコードは、パケットの送信者が、受信Combufに対する正当なアクセス権を有するか否かをチェックするために使用する。NIAは、パケット・ヘッダに含まれるアクセスコードと、LCT上のアクセスコードとを比較し、これらが不一致ならば、受信データのメモリへの書込みを抑止し、受信側CPUに割込みを発生する。

3.3. フラグ付き転送とスピン・ウェイト

数百バイト以下の細粒度データ転送を頻繁に行う並列アプリケーションでは、ノード間データ転送のレイテンシがシステム性能に大きく影響する。RDMA I/Fは、このレイテンシの低減を目的として、Spin Waitによってデータ到着を検出できるインタフェースを提供している。

Spin Wait方式では、受信側のCombuf上に「受信フラグ書込み領域」を設ける。ここにNIAが受信フラグを書込むのを、アプリケーションがループして待つことでSpin Waitを実現する。受信フラグはユーザ空間内に書き込まれるので、システムコールのオーバヘッドなしで受信処理を行なうことができる。

4．RingBufインタフェース

RingBuf I/Fは、非同期なメッセージ・パッシング機能を提供することを目的とする。RDMA I/Fがデータ書込みアドレスを送信側で指定するのに対し、RingBuf I/Fはデータ書込みアドレスを受信側で指定する。即ち、受信側で受信バッファ領域を確保し、到着したパケットのデータをこのバッファ領域に書込む。

HXBでは、パケットを受信する毎に割込みを発生させると、CPUの処理速度が追従できず、実効的な転送スループットが低下する。この問題を低減するために、RingBuf I/Fでは、リングバッファを採用している。RingBuf I/Fは、リングバッファの一方の側からNIAがパケットを連続的に書込み、もう一方の側からOSがパケットを一括して取り出す並行動作を可能とする。

4.1. 間接リングバッファ

RingBuf I/Fでは、リングバッファ上に不連続に空きエントリが残存しない様に、リング上にバッファへのポインタを配置する方式としている。受信パケットに含まれるデータの寿命、即ち、そのデータをユーザ・プロセスに渡して、バッファを解放できるまでの時間はデータ毎にばらつきが生じる。このため、リングバッファ上にデータを直接書込むと、空きエントリがリングバッファ上に不連続に残存してしまう。

NIAでは、リングバッファの各エントリにバッファへのポインタを格納する間接アドレス構造とすることで、デバイスドライバがリングバッファ上の使用済みエントリをFirst In - First Outで解放することを可能としている。また、リングバッファの各エントリは、ヘッダ書込みバッファへのポインタとデータ書込みバッファへのポインタを分離しており、ヘッダSplitが可能な構造としている。

4.2. スケーラブル・リング

リングバッファの総エントリ数や、各エントリ毎のバッファ・サイズは、ソフトウェアで設定可能な柔軟性を持った設計としている。これによって、ノード数の増加、あるいは、将来のハードウェア性能向上によって、より多くのエントリ数が必要になった場合、ハードウェア仕様を変更することなく、システムの構成パラメータを変更するだけで対応することができる。

4.3. ハイ・ウォータ／オーバフロー割込み機能

RingBuf I/Fは、リングバッファ・オーバフローを低減するために、リングバッファの空きエントリ数が設定された下限値を下回った時点で、NIAがCPUに対してハイ・ウォータ割込みを発生する。また、リングバッファ・オーバフローが発生した場合、NIAはCPUに割込み、受信動作を一時停止する。

5．まとめ

SR2201のNIAは、以上に述べた機能により、並列アプリケーションが必要とする低レイテンシの細粒度転送から、高スループットの大容量転送に至るまでを幅広くサポートすることができる。

[1] 千葉，岩寄他：分散並列OS「Orion」の試作・高速通信機能の検討，情報処理学会第45回全国大会予稿集，1992

[2] 秋山他：並列計算機SR2201における高速ノード間通信APIの実現と評価，情報処理学会第52回全国大会予稿集，1996

[3] 薗田他：超並列計算機SR2201におけるソフトバリア同期機能の実装と評価，情報処理学会第54回全国大会予稿集，1997