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

6Ｌ－3　秋山幸広*、小林耕三*、岡野信保**、浅間知之***、森山建三*

*(株)日立製作所　ソフトウェア開発本部　**(株)日立製作所　公共情報事業部
***(株)日立マイクロソフトウェアシステムズ

1.はじめに

高速ノード間通信の実現のためには、ハードウェアの高速化に加え、ソフトウェア処理の高速化も重要である。そのためには、可能な限りユーザレベルで処理を行う必要がある。日立の並列計算SR2000シリーズには高性能な通信を実現するために、リモートノードのメモリを直接操作するリモートＤＭＡ転送機能を装備している[1][2][3]。我々の目標は、このリモートＤＭＡ転送機能の性能を十分に引き出すことのできる通信ＡＰＩの開発にあった。

ハイエンドモデルのSR2201においては、専用のカーネルコードを開発することでソフトウェアオーバヘッドの低減を図った。その結果、１対１通信において、ハードウェアのオーバヘッドを含めて通信遅延が３.５μsec未満、スループットが２８０MB以上と高い性能を得ることが出来た。本稿ではリモートＤＭＡ転送機能によるアプリケーションインタフェース(API)の実現方式と評価について述べる。

2.従来のノード間通信における問題

並列計算機のノード間通信におけるソフトウェアオーバヘッドの支配的な要因には以下のものがある。

(1)メモリコピー

送受信プロセス間での非同期な通信を可能とするため、バッファリングが必要となる。このため、送信側と受信側の双方でメモリコピーが発生する。メモリコピーと通信は逐次的に実行されるため、通信性能はメモリコピー性能に抑さえられる。

(2)コンテキストスイッチ

１ノードマルチプロセス環境を前提にした設計では受信プロセスのブロックとデータ到着後の再起動に伴い、コンテキストスイッチが生じる。

(3)割り込み

送信および受信の完了を割り込みにて検出する場合、割り込みハンドラでのマシン状態の退避/回復が行われる。

(4)システムコール

送受信を行うのに,カーネルのサービスを受けると、システムコールを使用することとなる。システムコールはトラップ割り込みにより実現されており(3)と同じオーバヘッドを要する。

3.リモートＤＭＡ転送機能

上述した問題点を解決するネットワークアーキテクチャとしてSR2000シリーズでは、リモートＤＭＡ転送機能を提供する。本機能は図１で示すように送受信側とも連続物理メモリ領域をユーザプロセスの仮想アドレス空間に固定的に割り付けておき、送信側が受信側のデータ格納領域を指定して直接データを書き込む、送信者主導のノード間メモリコピーである。送信側プロセスの仮想空間上のデータを直接受信側プロセスの仮想空間に書き込むことにより、送受信の双方でメモリコピーが発生しない[1][3]。

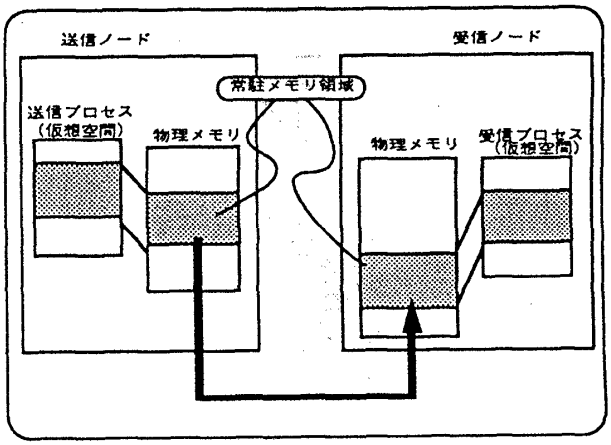

図１.リモートＤＭＡ転送機能

4.通信ＡＰＩの実現

高速な通信ＡＰＩの実現において、カーネルのサービスを極力避け、通信処理のほとんど全てをユーザレベルで行うことが課題となる。

リモートＤＭＡ転送機能によりメモリコピー処理のオーバヘッドを削減すると共に、データ到着の検出をユーザプロセス上にてスピンしてテストするインタフェースを提供することによりコンテキストスイッチおよび割り込みのオーバヘッドを削減した[3]。

SR2201では、システムコールのオーバヘッドも削減するため、直接起動インタフェースを提供し、より低オーバヘッドの通信を実現する。

The implementation and evaluation of high performance inter node communication API on SR2201

Yukihiro Akiyama,Kouzou Kobayashi,Nobuyasu Okano,Tomoyuki Asama,Kenzo Moriyama
*Software Development Center,Hitachi,Ltd. **Goverment & Public Corporation Information Systems Division,Hitachi,Ltd.
***Hitachi Microsoftware Systems,Inc.

4.1直接起動インタフェース

データ送信の起動は、メモリマップドされた送信起動レジスタにコマンドワードのアドレスを書き込むことで行われる。送信起動レジスタを保護するために、カーネルモード(特権モード)でのみアクセスできるようにしている。

モード移行のために、従来はシステムコールを使用していた。システムコールは、トラップ割り込みにより実装されており、レジスタやPSW(Proccessor Status Word)等の全てのマシン状態の退避が行なわれる。

起動オーバヘッドを低減する為には、システムコールを発行せずに、直接ネットワークを起動することが望ましい。

SR2201に搭載されるRISCチップには、トラップ割り込みを発生させずに、モードを変更する命令がある。この命令を使用することで、図2のように、直接カーネルモードへの切り換えが可能となる。

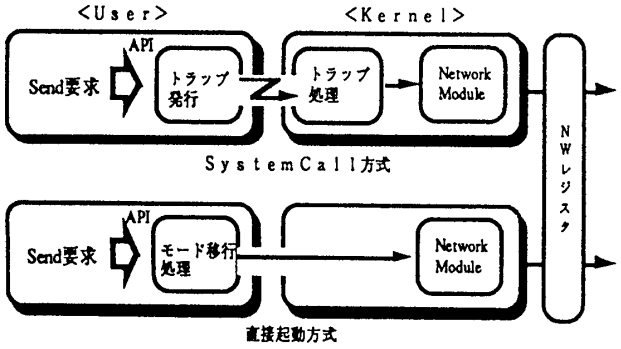

図2.直接起動インタフェース

直接起動インタフェースでは、送信の前準備段階で、あらかじめコマンドワードを作成しておき、一度作成したものを繰り返し再利用する方式を採る。これにより、送信起動のオーバヘッドは、コマンドワードのアドレスをネットワークレジスタに書き込むだけの低オーバヘッドを実現している。

4.2メモリ保護

コマンドワードには、受信ノードのアドレス、書き込み先のアドレス等が含まれる。リモートDMA転送機能では、受信領域にアクセスキーを設定し、受信側のネットワークハードウェアがコマンドワードで指定されたアクセスキーと比較することで不当なメモリ書き込みを防止する。加えて、コマンドワードをカーネル空間に保持することで、ユーザのバグによるコマンドワード破壊を防でいる。

5.評価

直接起動インタフェースの性能評価について述べる。評価プログラムは、隣接する2ノード間でデータ転送を往復させる操作を約100繰り返し、その片道に要した平均時間を通信所要時間とした。
比較対象として、従来のシステムコール版についても同じ測定を行った。

5.1通信遅延

送信プロセスのｓｅｎｄ要求発行から受信プロセスでデータ到着を確認するまでの時間(8B転送時)を表1に示す。直接起動インタフェースの通信遅延は、約3.45μsと従来方式の21.50μsに比べ16%のオーバヘッドに低減された。

表1.直接起動の性能比較

種別	通信遅延(μs)	性能比
直接起動方式	3.45	0.16
システムコール方式	21.50	1.00

5.2スループット

転送サイズ8Bから1MBまでのスループットを図3に示す。直接起動インタフェースは、転送サイズ20KB～30KBでピークのネットワーク性能を引き出している。

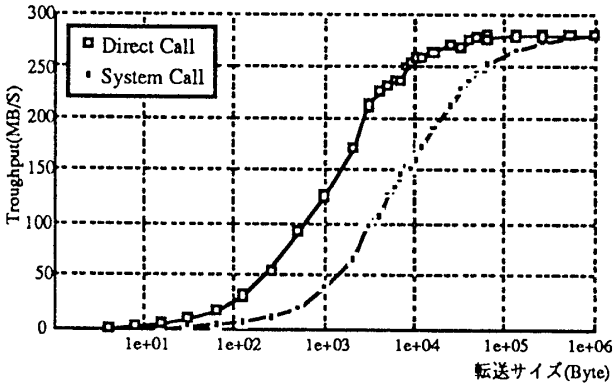

図3.スループット比較

6.おわりに

本稿では、SR2201におけるノード間通信APIの実現と評価について述べた。直接起動インタフェースでは、ソフトウェアオーバヘッドを低減し高性能な通信ＡＰＩを実現することが出来た。このインタフェースは、1ノード・シングルプロセス環境を前提に開発してきたが、マルチプロセス環境においても応用できると考える。

参考文献

[1] 千葉、他：分散OS「Orion」の試作　情報処理学会第45回全国大会(1990)
[2] 西門、他：SR2001 OSの開発コンセプト　情報処理学会第50回全国大会(1995)
[3] 薗田、他：SR2001における高速プロセッサ間通信機能　情報処理学会第50回全国大会(1995)