RC-004

分散共有メモリを用いた同報通信についての一検討
A Study on Multicast by Distributed Shared Memory

田中 晶† 寺元 光生†
Akira Tanaka Mitsuo Teramoto

1. まえがき

インターネットの普及により、物理的に離れていてもネットワークを介した分散システムの構築が容易になってきた。これまでに、我々は、AO(Application Object)が直接共有メモリ空間にメッセージを書込むことでバッファ間コピーのオーバヘッドをなくした、分散共有メモリ（Distributed Shared Memory: DSM)による高速通信方式を実現した[1]（以下、DSM 通信と記述）。分散メモリカップラ（Distributed Memory Coupler: DMC）が、ネットワークへのデータ送出／データ受信と、ノード内のメモリ書込みを並行して行うので、通信リンク速度に追随可能な高速性を有している[1]。

ところで、内部処理速度がネットワークに対して高速（同等）な場合、同報通信等により多量のデータを送出すると、ネットワークにその処理能力を上回る負荷を与え、輻輳や転送速度の低下が生じるため、DSM 通信の高速性を活かせない。既に、リアルタイム性が高いデータは輻輳を回避して送信する優先度制御を提案しているが[2]、本稿では、本 DSM 通信システムの、(a)高速性、(b)ノード間通信とプロセス間通信処理が共通、(c)メモリアドレスと送信先のネットワークアドレスがリンク、等の特徴に基づき、複数の転送処理を共通化し、冗長なデータ送出を避け、通信処理機能やネットワークに負荷をかけずに容易に双方向と片方向の同報通信を実現する方式を提案する。尚、本提案は、共有メモリを用いたシステムであれば、例えば、ネットワークではなくバス等による蜜結合のプロセッサ等にも適用範囲を拡張できる。

2. 分散共有メモリを用いた通信

対象とする通信システムの各ノード或いはデバイス（以下、ノードで総称）は、ホストプロセッサ（プロセッサモジュール(Processor Module: PM)）、ローカルメモリ、DMC 等から構成される（図 1）。ノードのローカルメモリが、アプリケーション空間(Application Space: AS)、共有空間(DSM space）、OS 等の空間(Kernel space）に分離され、アプリケーションオブジェクト(Application Object: AO)即ちプロセスが、DSM 上のメッセージバッファ(Message Buffer: MB)にデータを書込み／読み出すことで通信を行う（図 2）。このため、AO が同一ノード内でも、異なるノード上にあっても、同一の処理で通信できる。尚、OS や AO が存在する PM をも送信元／受信先として扱えるので、PM 内で AO を区別する場合を除き、送信元／受信先は、PM で代表して記述する。図 1 で通信処理手順の概略を述べる。各ノードのローカルメモリ内の DSM 領域は、送信元 PM と受信先 PM の組み合せ毎に区分され（この区分が MB である）、各 MB は受信先 PM のネットワークアドレス（イーサネット、ATM、IP 等のアドレス）とリンクされている。例えば、ノード A において、PM が自ノード内 DSM のノード B 向けにリンクされた MB にデータを書込むと、データと共にメモリアドレスが内部バス上を流れる（これらのデータ類は、バストランザクションと呼ばれる）[①]。DMC がこのバストランザクションを取込み[②]、書込み先 MB にリンクされたネットワークアドレスとメモリアドレスを基に、受信先ノード内の、送信元ノードでデータが書込まれた MB と同じアドレスに、ネットワーク経由でデータを書込む[③]。この一連の動作中、送信側メモリ書込み／送信、受信／受信側メモリ書込みは、並行して行われる。データのメモリ書込みを検出した受信先 PM は、データを読み取り、MB 解放通知を送信元に送り、MB を再利用可能にする。ノード A から C への送信も同様である[④]。MB と受信先のネットワークアドレスのリンクは、連想記憶メモリ(Content Addressable Memory: CAM)を用いて、バストランザクションのメモリアドレスから簡単な演算で受信先ネットワークアドレスに変換し DMC が下位レイヤの転送処理（インタフェース）に送る。尚、メモリの冗長領域は縮退して実装しており、必要なサイズは、32bit 実アドレスの場合、携帯電話相互間のメールを全て保存するとしても約 5 年は上書せずに使用可能な程度である。

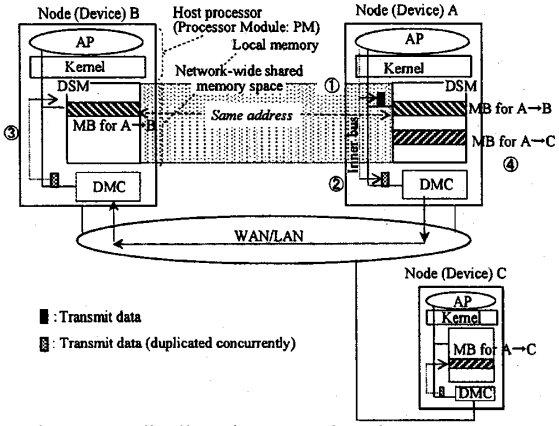

Fig. 1 Inter-distributed-process shared memory

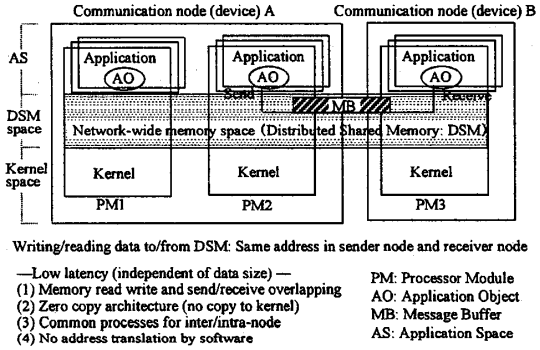

Fig. 2 Network-wide shared memory space

3. 分散共有メモリを用いた同報通信

3.1 同報通信機能の構成方法

同報通信（双方向）を分散共有メモリで実現する方法は、大きく次の二通りに分類される。

(1) 特定の MB を同報用に用いることを他のノードに通知する或いは予め決定しておく：(1-1) MB にリンクする受信先 PM を複数設定する方法（MB は、同報用として他の MB とは区別される）。(1-2) MB にリンクする受信先 PM のネットワークアドレスを、例えば、IP のマルチキャストアドレスのように、下位レイヤの複数の宛先を指示するアドレスとする方法。

(2) 通常の DSM 構成（MB と受信先 PM のネットワークアドレスは一対一にリンク）を用いる：PM が宛先の異なる複数の MB にデータを書込む。

† 日本電信電話（株）NTT 未来ねっと研究所
NTT Network Innovation Laboratories, NTT Corporation

これら何れも、転送インタフェース或いは、ネットワーク内で同一データが大量に複製されるため、ネットワークに負荷をかけ輻輳／遅延を生じ、或いは PM 等の上位層での処理が増大し、DSM 通信のリアルタイム性が阻害される可能性が高い。また、(1-2)は下位レイヤの同報機能の性能に依存し、DSM 通信システム自体で輻輳等の制御を行いにくい。そこで、これらの課題を解決し、同報を行っても DSM 通信の高速性が維持される方式を提案する。

3.2 同報通信に適用する DSM 通信手順

データ複製による転送量の増大を極力抑え、DSM の上位或いは下位レイヤの手順増加を避け、かつ、本来の DSM 通信システム構成に影響しない方法として、DSM 通信手順に最小限の変更を加えた同報用の手順を考案した。直接のデータ送信先は、多くの場合、隣接ノードのみで良いため、転送データ量を抑制できる。尚、ネットワーク上で、他のノードに中継されることを前提としている。図 3 は、3.1 節(1)に示したように、同報データの送信前に（同報データと一体として／データ送信直後でも同様）、該データが同報データであることを送信元 PM が他のノードに通知して、送信データが書込まれた MB が同報データ格納用であることを通知した上で、同報通信を行う場合の、通信手順の概略を示している。通知を行わず、予め同報用 MB を決定しておくと、手順は少ないが DSM の利用に制限が生じるので、本稿は通知手順を含めた手順を述べる。

(n) 通知処理：送信元ノード(nA)の PM-A は、同報通信開始通知と同報対象の PM のメモリアドレスを DSM 管理機能(DSMM-A)に送る。

(d-1) DSMM-A は、同報通信開始通知を契機に同報開始識別子(ntfy)を生成し、宛先 PM が存在するノードの DSMM 向けに DSM 通信により送信する（以下、ノード間送信は全て DSM 通信による）。PM-A は同報対象の送信データ(data)を生成し同報宛先 PM 向けのアドレスに書込むが、1 個のノードに複数の宛先 PM が存在する場合、その内一つの PM 宛のメモリアドレスに書込む。

(d-2) 書込まれた data を中継ノード(nB)或いは宛先 PM が存在するノード(nC)に各 1 個ずつ送信する。DSMM-A は、送信データに関連する識別データ(mbid)を全ての宛先 PM 向けに生成し、宛先 PM へ送信する。さらに、宛先 PM が存在するノードに各 1 個の送信要求(rqst)を生成し、同報対象の data が全て送信された後に送信する。また、送信宛先のノードアドレスを記録する。

(r-1, t) ntfy を受信した中継ノード(nB)或いは宛先ノード(nC)では、受信した mbid が自ノード宛アドレスを示せば、mbid で指定されるメモリアドレスの MB に受信した data を書込む。rqst 到着後、送信元ノード或いは中継ノード(nB)にメッセージ解放通知(rels)を送信する。

(r-2) ntfy を受信した中継ノード(nB)は、mbid が他ノード宛アドレスを示せば宛先ノード毎に(d-1,2)を行う。rqst が到着する迄は、mbid を自ノードの mbid 領域に格納する。

(v) 中継ノード(nB)は、記録した送信元ノードを参照して全ての後段ノード（通過するノード及び宛先ノード）からの rels が到着した後、前段ノードへ rels を送信する。

(f) 送信元ノード(nA)は、記録した送信先ノードを参照して全ての後段ノードからの rels が到着した後、この同報で使用した全ての MB を解放する。尚、rqst は mbid を介して data との対応付けが可能なので、受信ノード或いは中継ノードでは同報終了をも示すこととなり、MB は同報通信用以外に（通常通りに）利用可能となる。

4. 考察と結論

提案技術（3.2 節）を、クロック換算及び転送データ量の積算に基づき、DSM 自体への修正が不要な 3.1 節(2)の方法と比較する。送信データ長を 40B（適用対象の DSM 通信システム [1]のレイテンシはデータ長比依存のため、データ長は評価に影響しない）、同報先 PM 数を 8 とすると、レイテンシ（最終到着ノード宛）は 17～46%に削減される。また、ネットワークが DSM 処理速度と同程度の高速性を持つ前提で、同報先 PM 数を 2 から 8 に増やすと、平均転送速度は 62%になるが、低下率は 3.1 節(2)に対し 38%に抑えられる。同報数が増えると提案技術の有効性は増す。

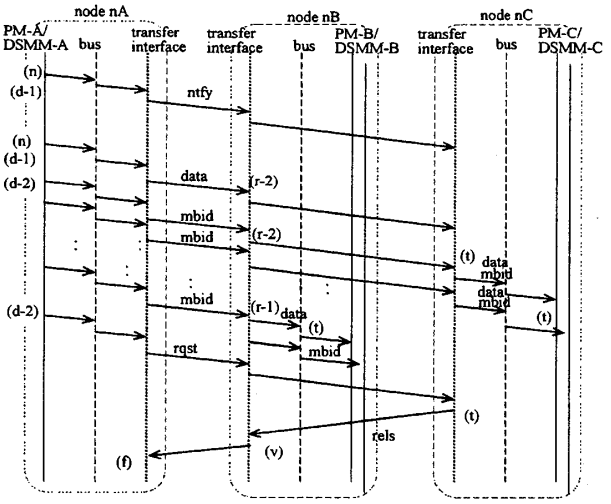

Fig. 3 DSM multicast communication sequence

クラスタ間ネットワークの再帰的方法[3]や、独自のスイッチによる多段ネットワーク[4]、を用いるマルチキャストが提案されているが、ネットワークの輻輳を軽減する技術の例は少ない。本 DSM 通信システム[1]は、WAN/LAN 等の公衆（汎用）通信網を対象とするため、本稿で提案する転送量が小さな同報方式の有効性は大きい。クラスタ構成に適用する DSM 間通信[5]では、グループ間は宛先グループに割当てられた波長を指定し、グループ内はブロードキャストする。グループ間／内のネットワーク負荷とノード負荷は夫々高／低、低／高となるが、1 対 1 通信自体で循環的に波長を割当てることで波長数を抑制している。一概に比較は難しいが、ノードの階層化が不要で、ネットワーク・ノードとも負荷を低減する我々の方式の効果は高い。

DSM により同報を行うと、ネットワーク側がボトルネックになって転送速度が低下する。提案方式は、実装による評価が必要ではあるが、既存の DSM 通信構成を殆ど変更しないで、即ち、同報の開始を示す通知、及び、受信した識別データが示す複数のメモリアドレスにデータを書込む機能を追加するだけで、DSM の並列性や高速性を損なうことなく同報通信を実現する可能性を有する。

【参考文献】

[1] 田中 晶, 山田茂樹, 田中 聡: ネットワーク分散処理ノードアーキテクチャ MESCAR のメモリ間複製機構の設計と評価, 電子情報通信学会論文誌(B), Vol.J86-B, No.2, pp.148-161 (2003).

[2] 田中 晶, 山口 正泰: 分散システム間通信における優先度制御の一検討, 2004 年電子情報通信学会総合大会, B-6-204, (2004).

[3] 額田匡則, 鈴木紀章, 天野英晴, 西村克信, 田村友紀, 長名保範, 小西将人, 五島正裕, 富田眞治: 超並列計算機 JUMP-1 のマルチキャスト機構による性能向上, 電子情報通信学会技術研究報告, CPSY2001-49, Vol.101, No.217, pp.37-44 (2001).

[4] 細見岳生, 加納 健, 中村真章, 広瀬哲也,中田登志之: 並列計算機 Cenju-4 の分散共有メモリ機構, 情報処理学会論文誌, Vol.41, No.5, pp.1400-1409, (2000).

[5] A.K. Kodi and A. Louri: A Scalable Architecture for Distributed Shared Memory Multiprocessors Using Optical Interconnects, Proc. *18th Int'l Parallel and Distributed Processing Symp. (IPDPS 04)*, IEEE Press, pp.11-20 (2004).