C-015

DSM を用いたプロセス間通信における優先度制御
A Priority Control for Inter-Process Communication with DSM

田中 晶†
Akira Tanaka

1．はじめに

我々は，物理的に離れたノード間でネットワークワイドメモリ空間を共有する DSM（Distributed Shared Memory）機構を用いて，ノード A の AP（Application（図 1，α）内のプロセス（以下 AP で総称する）が自ノードの DSM にメッセージを書き込むと（図 1，(a)），ハードウェアによりノード B の同一番地の DSM にメッセージが自動的にコピーされる（図 1，(b)(f)(a')），高速通信方式を開発してきた[1][2]．この方式は，DSM 書込みとノード間メッセージ転送をオーバラップして実行するので，小さなレイテンシで転送が行える．ノード間ネットワークではソフトウェア負荷の少ないハードウェア組込みの通信手段も注目され，DSM 通信機構もその小さなレイテンシから有効である．

しかしながら，DSM 書込み処理速度よりネットワーク転送速度が小さいと，ネットワークが隘路になり大規模メッセージが転送し終わるまで次の送信ができず，ノード間制御情報のような実時間性を要する小規模メッセージの遅延を引き起こす可能性がある．そこで，大規模メッセージのフラグメンテーションにより，この課題を解決する優先度制御を提案する．

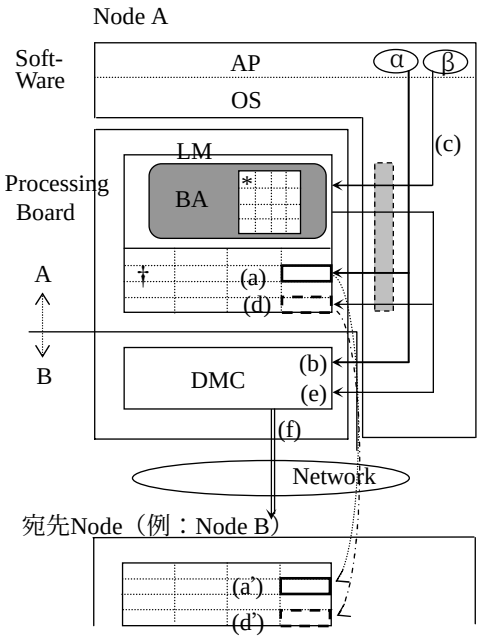
2．基本概念と制御方法

小規模メッセージの実時間通信を確保するために，ルータによるフラグメンテーション技術を参考に，次のようなスケジューリングとバッファリングを組み合わせた方式を検討した．本提案で主に必要な構成要素は図 1 に“提案”として示すローカルバッファとその管理機能である．

＜スケジューリング＞小規模メッセージの優先度を大規模メッセージの優先度より上げ，大規模メッセージの転送中に小規模メッセージ転送要求が発生するとその処理を中断し，小規模メッセージの転送を行った後，再び大規模メッセージの転送を開始する．

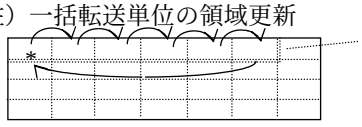
＜バッファリング＞OS は大規模メッセージに対して，メッセージ毎に所定の規模単位（一括転送単位，図 1 の*）に区切ってローカルバッファ（BA: Buffer Area）を割り当て，AP（図 1，β）は BA に大規模メッセージを書き込む（図 1，(c)）．OS は 1 つの一括転送単位が一杯になる毎に一括転送単位を DSM へコピーし，並行して DSM コピーハードウェア（DMC: Distributed Memory Coupler）が相手ノードへ転送する．また，OS は次の転送のため一括転送単位を更新する（図 1，注）．小規模メッセージは，従来通り AP から直接 DSM に書き込まれ，DMC によりノード間転送がすぐに開始される．

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



BA：ローカルバッファ-提案
LM：ローカルメモリ
α：従来のメモリアクセス
β：提案のメモリアクセス
(a)：APからのDSM書込み
(b)(e)：DMCへ自動取込み
(c)：BA書込み
(f)：宛先Nodeへの送信
(d)：BAからのDSM書込み
(a')(d')：宛先Nodeでの DSM書込み
*：一括転送単位
†：MB(Message Buffer)
A, B：並列動作
：BA管理(提案)+DSM管理(BAM: Buffer Manager)

（注）一括転送単位の領域更新



使用される最大領域≦MB領域

図1　DSM通信の動作と提案機能

このため，DSM の領域（MB: Message Buffer）割当てや MB 空塞判定を行う OS の DSM 管理機能を，その対象を BA まで拡張し（以降 BAM（Buffer Manager）と呼称），バッファリングの際の一括転送単位の割当て，一括転送単位空塞判定，及び BA から DSM へのコピーを行う．

3．提案手法による DSM 通信の動作

AP から発生したメッセージが相手ノードへ送信される手順を図 2-2 で示す（〔〕は図 2 の記号）．

i. メッセージ送信開始時，AP は DSM 書込みのために MB 確保要求と共にメッセージ規模を OS（BAM）に送る．

ii. BAM はそのメッセージ規模を元に一括転送単位の規模を定める．優先度の低い大規模メッセージを送信する AP〔Px〕では一括転送単位の規模>0 となり，BAM は DSM の番地の代わりに BA の番地を AP に返す．優先度の高い小規模の実時間メッセージを送信する AP〔Pa〕では一括転送単位の規模は 0 となり，BAM は DSM 内に確保する MB の番地を AP に返す（以下，一括転送単位の規模>0 の送信要求が先に生じた場合を述べる）．

iii. AP がメッセージを ii で番地指定された BA 内の一括転送単位に書き込む〔A〕．BA への転送量が指定された一括転送単位の規模に達すると DMC がそれを検出し〔①〕BAM に通知する．

iv. 小規模メッセージの送信要求が無ければ，BAM は BA から DSM へコピーを行った後，一括転送単位毎に番地を更新しながら，AP に残りのメッセージを BA に書き込むよう指示し，iii と iv を繰返す．

v. iii において，小規模メッセージの送信，即ち一括転送単位=0 の優先度の高い送信要求があれば〔Pa〕，BAM は一括転送単位一杯の際に大規模メッセージの BA から DSM へのコピーを保留し，小規模メッセージに DSM の MB を割当て，書込みを指示する〔②〕．小規模メッセージ送信 AP はメッセージを DSM に書き込み，DMC が並行してノード間転送する〔B〕．DSM 書込み終了後〔③〕，BA に保留していた大規模メッセージを，BAM が BA から該当 DSM 領域へ転送する〔④, C〕．DSM 領域への転送終了後〔⑤〕，Px のメッセージの残り，或いは，Px の送信が完了した場合は他のメッセージの送信を引き続き行う〔⑥, D〕．

vi. AP が送信を終了すると AP が BAM に通知し，BAM が BA 内の一括転送単位を解放する．

即ち，小規模メッセージ送信要求の際，大規模メッセージの送信中でなければ，Pa への MB 割当て，DSM 書込み〔B〕とノード間転送のみが直ちに行われ，大規模メッセージ送信中であれば一括転送単位一杯を契機にその処理を中断して，小規模メッセージを送信することになる．

尚，iv での書込み番地の更新は，DSM 転送が完了する前に一括転送単位が上書きされるのを防ぐためである．従って，1 ノード向けにバッファリング用に確保されるメモリ規模は，最大で，本来の DSM 通信機構が 1 回で送れる規模，即ちせいぜい 1 ノード向けの MB[1]領域の規模程度を，図 1 の注に示すように繰返し利用する．

4．本優先度制御の効果とまとめ

図 2-2 に示すように，大規模メッセージ Px をバッファリングし，さらに，一括転送単位毎にスケジューリングにより分割することで，実時間性の高い Pa が割込む形で送信される．その結果としてネットワーク側の待ちが回避され，従来図 2-1 の*1 であった実時間メッセージ Pa の送信を，図 2-2 の*2 のように大規模メッセージに先行して行える．そして次の利点を持つ．(a)一括転送単位の規模は AP が通知するメッセージ規模に基づき BAM が定めるため，間接的ながら AP 自身がその実時間性により優先度制御が可能である．これらは，タスク管理に類する複雑な制御無しに比較的簡易に実現できる．(b)AP は単に BAM の返す番地にメッセージを書き込むだけであり，実時間性や規模によらず AP インタフェースは共通で良い．(c)送信の優先度，即ち，転送保留の有無及び送信順序を，唯一のパラメータである一括転送単位の規模のみにより設定できる．

本提案は，送信処理速度と比べネットワーク速度が遅い場合や実時間性の低いメッセージ規模が大きい場合に効果が高く，一方，ネットワークでの遅延が殆ど無い場合は効果が得にくい．このような特徴を活かせる AP，例えば，システムバックアップ情報の送信と混在して制御情報を即時に送る場合等に有効である．

DSM による高速通信方式において，既存の主要構成要素である DMC を改造することなく，OS の機能を少しばかり拡張することで，大規模メッセージが混在しても，制御情報等の小規模メッセージの実時間性を確保できる技術を明らかにした．

【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] 山田茂樹, 田中 聡, 田中 晶, 向井良: メモリ結合型分散処理ノードアーキテクチャ MESCAR のインプリメンテーションと性能評価, 電子情報通信学会論文誌 D-I, Vol.J83-D-I, No.4, pp.418-429 (2000).

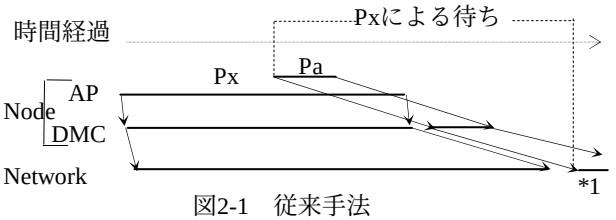
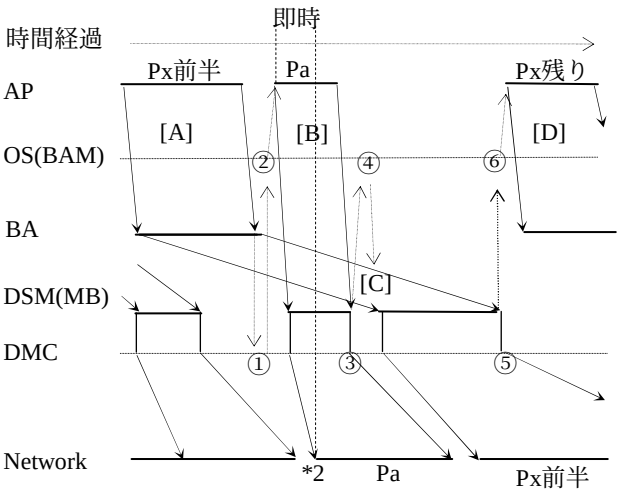


図2-1　従来手法



①：BA書込みが一括転送単位に達したことを検出.
②：高優先度メッセージがあれば書込み開始.
③：PaのDSM書込み終了を検出.
④：Px前半のBA->MB転送を開始.
⑤：Px前半のBA->MB転送終了を検出.
⑥：Px残り或いは次メッセージ転送を開始.

図2-2　スケジューリングとバッファリング

Pa：実時間メッセージ　　Px：大規模メッセージ

図2　提案手法導入の効果