# お茶の水５号のハードウェア分散共有メモリ機構

4N－8

田中 清史† 松本 尚† 対木 潤‡ 平木 敬†

†東京大学大学院理学系研究科情報科学専攻　‡富士通株式会社

## 1 はじめに

並列／分散システムにおいて汎用性を維持し、容易なプログラミング環境を提供するためには分散共有メモリが有効である。その際、遠隔データアクセスのレイテンシを小さくするためにキャッシュを行なうことが主流となっている。また、プロトコル制御をハードウェアで実現することにより、キャッシュのコンシステンシ管理に付随するオーバヘッドを削減することが可能である。

本稿では効率の良いコヒーレンス制御とスケーラブルなディレクトリを備えた分散共有メモリを、タグ情報のための特別なSRAMチップを使用せず、メモリコントローラに簡単な回路のみを付加することによって低コストで実現する方式を提案する。すなわち、DRAMで構成される主記憶上にタグ情報を確保し、メモリコントローラがプロセッサからのメモリアクセスリクエストに対してタグの参照、更新およびリプライを一括して管理する。これにより、基本的な分散メモリ型システムに新たな構成要素を付加することなしにCC-NUMAが実現可能となる。更に並列計算機プロトタイプお茶の水５号に本機構を実装し、プログラムを実行した結果を示す。

## 2 お茶の水５号の構成

お茶の水５号(OCHANOMIZ-5:Omnipotent Concurrency-Handling Architecture with Novel OptiMIZers-5)[1]は東京大学平木研究室における並列処理プロジェクトの第５号並列計算機プロトタイプである。

お茶の水５号は２個のPE(SuperSPARC+:動作クロック60MHz、マルチキャッシュコントローラ内蔵モジュール)とメモリを共有バス(MBus:SPARC専用バス)結合したものをクラスタとしている。DRAMモジュールによる主記憶はクラスタ毎に32Mbyteあり、その他にメモリコントローラ、バス使用権を調停するアービター、ネットワークへのインターフェースのそれぞれにXilinx社のSRAMベースFPGAを使用している。なお、共有メモリのキャッシュ管理はメモリコントローラ内部のhardwired回路によって高速に行なう。

全体としては４つのクラスタを２進木階層構造の相互結合網で接続した構成をとる。ネットワークの各内部スイッチングノードをFPGAおよびSRAMで構築することにより、一般化されたコンバイニング[2]をサポートする高機能ネットワークを実現している。

---

Hardware Distributed Shared Memory of OCHANOMIZ-5

†Department of Information Science, Faculty of Science, University of Tokyo　‡FUJITSU Ltd.

Kiyofumi TANAKA, Takashi MATSUMOTO, Jun TSUIKI, Kei HIRAKI

tanaka@is.s.u-tokyo.ac.jp

## 3 コヒーレンス制御方式の設計方針

コヒーレンス制御のオーバヘッドを削減し、多数のクラスタ間の共有を適切なコストで扱うことを目標とするために、以下のことが挙げられる。

- スケーラブルなディレクトリ方式
  ディレクトリのビット数を節約するために、共有情報を階層距離で管理する。この方法は疑似フルマップディレクトリ[3]のサブセットである。この方法はStanford大学のDASH[4]のフルマップディレクトリに比べて大幅にメモリ使用量が少ない。
- マルチキャスト／コンバイニング機構の利用
  クラスタ間ネットワーク内の各階層でマルチキャスト／コンバイニング[3]を行なう。これにより一回のコヒーレンス制御あたりのネットワークの負荷が軽減され、チェインドディレクトリ方式よりも効率が良い。

また、ハードウェア制御を容易にするために、以下のように簡単化する。

- キャッシュライン単位の管理
  共有データのクラスタレベルキャッシュのラインサイズをプロセッサキャッシュと同サイズにし、ライン単位で管理を行なう。またこれにより、false sharingを削減することが可能である。
- ホームの固定
  ディレクトリを持つクラスタは各キャッシュブロックに対して静的に固定される(Home)。
- OwnerID／ディレクトリのHomeでの一括管理
  Invalidateプロトコルで一時的にOwnerがHome以外の他のクラスタに移る場合、OwnerクラスタIDをHomeに記録する。
- リクエストはHomeへ転送
  Invalidな場合のRead、Sharedな場合のInvalidateリクエストはHomeへ転送する。Readリクエストに関して、HomeがOwnerならばReply、InvalidならばOwnerクラスタへリクエストをフォワードする。Invalidateに関して、HomeはShared領域へのinvalidateを行ない、その後Acknowledge(以下Ack)を返送する。

## 4 実装

クラスタキャッシュの状態遷移を図1に示す。(中間状態Pendingは省かれている。)ブロックがshared状態になるときに、Homeを必ずclean状態にすることにより、shared-dirtyを回避する。これによりディレクトリはHomeにのみ用意すればよい。

共有最大距離によるディレクトリは、疑似フルマップディレクトリ[3]の簡易版である。ディレクトリとしてブロックを

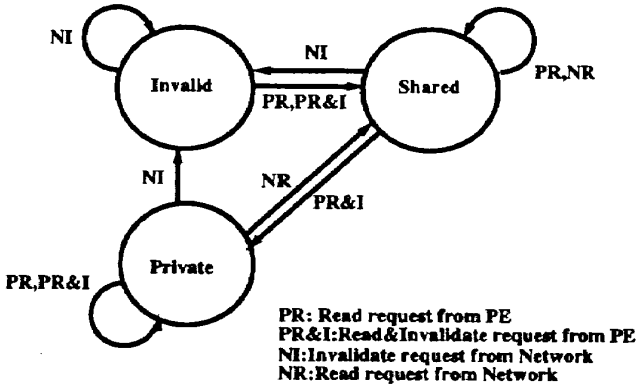


図 1: クラスタキャッシュの状態遷移

共有するクラスタの Home からの最大距離を記録する。ここで、距離とは 2 つのクラスタ間でパケットを転送する際に通過するネットワークの階層の段数である。Invalidate パケットは共有最大距離内の全てのクラスタへマルチキャストされる。(共有最大距離だけ階層ネットワークをさかのぼったノードを root とする部分木へのブロードキャスト。) また、ネットワーク上でのコンバイニングにより Ack メッセージの収集を効率良く行なう。

メモリコントローラはバスリクエストを受け取り、共有アドレスに対するデータ／Invalidate 要求ならば対応するタグ情報を読み出す。キャッシュの状態によって要求元にデータ供給、あるいはリトライで返答する。リトライの場合、必要ならば同時にネットワークインターフェースに通知してネットワークトランザクションを起動させる。その後、必要ならばタグの更新を行なう。

共有アドレス以外への通常の MBus トランザクションと比較すると終了応答までにタグの読み出し／デコードの時間が付加されるが、これは高々 4 バスクロックサイクルである。一度プロセッサのキャッシュに読み込まれ、その後高い確率でヒットすればこのタグ参照のコストは無視できるほど小さくなる。

## 5 性能評価

評価アプリケーションとして、SPLASH-2 の LU factorization(512 × 512) を採用し、トータル実行時間を計測した。比較として、同じくお茶の水 5 号上での non-cacheable の DSM の実行時間を併せて示す。実行台数は 1、2、4 台である。1、2 台に関しては 1 クラスタ内での実行、4 台に関しては 2 クラスタを使用した。図 2が実行結果である (縦軸は sec、横軸は実行台数)。CC が本方式、NC が non-cacheable の場合であり、下の図は、CC に関して拡大したものである。CC は NC に対して大幅に高速化していることがわかる。また、NC は 4 台でネットワークが飽和しているが、CC は 4 台まで台数効果が得られた。

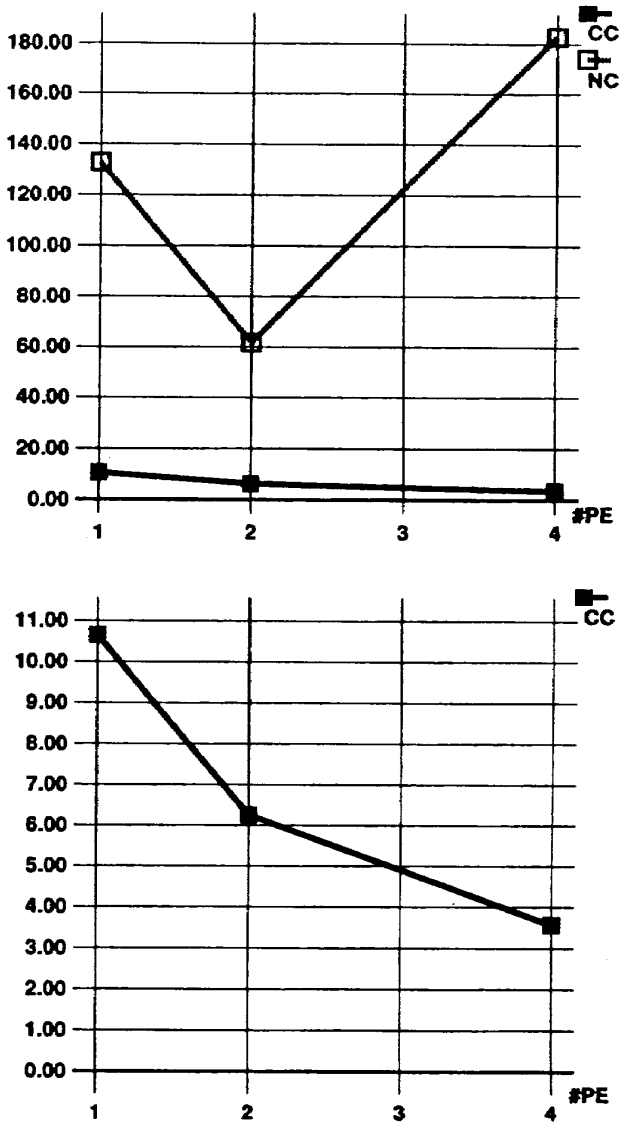


図 2: SPLASH-2 LU (512 by 512 Matrix)

## 6 おわりに

本研究において低コストでハードウェア分散共有メモリ機構を実現する方式を示した。更に本機構をお茶の水 5 号に実装し、性能評価を行なった。

謝辞

本研究の一部は通商産業省 RWC プロジェクトの一環として行なった。

## 参考文献

[1] 田中 清史, 対木 潤, 松本 尚, 平木 敬: 並列計算機プロトタイプお茶の水 5 号. 第 3 回 FPGA/PLD Design Conference & Exhibit 応用技術論文集, pp.505-514 (Jul. 1995).

[2] 田中 清史, 対木 潤, 松本 尚, 平木 敬: 一般化されたコンバイニング機構. 情報処理学会研究報告, ARC Vol.96,No.13, pp.31-36 (Jan. 1996)

[3] 松本尚, 平木敬 : 超並列計算機上の共有メモリアーキテクチャ. 電子情報通信学会技術研究報告, CPSY92-26. pp.47-55 (Aug. 1992)

[4] D.Lenoski, et al.: Design of Stanford DASH Multiprocessor. Technical Report CSL-TR-89-403, Stanford Univ. (Dec. 1989)