HYPHEN クラスタにおける基本プリミティブの設計と実装

木原大悟† 立川純† 福田健一郎† 大西淑雅†† Bernady O. Apduhan† 佐藤寿倫†††† 有田五次郎†

† 九州工業大学 情報工学部 知能情報工学科
†† 九州工業大学 情報科学センター
††† 九州工業大学 マイクロ化総合技術センター

5K−1

1 はじめに

分散共有メモリ型並列処理環境では，ユーザに対するメモリの位置透過性は確保されるが，データ依存関係を保つための同期操作や，それに付随するメモリ一貫性を効率よく行えるかどうかがシステム全体の処理性能を大きく左右する．

一般的なソフトウェアDSM の場合，キャッシュ一貫性制御等のメモリ管理はプロセッサによって実行する必要がある．つまり，プログラムの本質ではなく，システムの制御にプロセッサ資源を消費してしまう．そのため，システム全体におけるオーバヘッドが無視出来なくなる．

そこで我々は，高速な同期手段をもち，キャッシュの一貫性の制御が不必要なプロセス実行モデルにもとづく分散共有メモリ型クラスタHYPHEN(以下，HYPHEN クラスタ) を設計・実装中である．HYPHEN クラスタは，分散共有メモリ，実行管理機構，メモリアクセス機構をHyphenLink/PCI[1] としてハードウェアで実装し，より細粒度な並列処理を行おうとするものである．

本稿では，HYPHEN クラスタにおける並列プロセス実行モデルと，実行を支援する基本プリミティブの中のメモリ管理に関する基本プリミティブについて述べる．

2 HYPHEN のプロセス実行モデル

HYPHEN クラスタにおける並列プログラムはPB タスク (以下，タスク) と呼ぶ複数の実行主体に分割され，分散共有メモリに格納・実行される．

タスクの実行は各プロセッサが持つFIFO キューによって制御され，後述するPB 操作 (en-queue) によって実行が指定され，EXT 操作 (de-queue) で実行を終了 (次のタスクの実行を開始) する．このとき並列実行されるタスク間にデータの依存関係がなければ，このようなプログラムは自然に同期が取れ，共有データ間の干渉も起こらない．このような実行モデルをPB タスクモデルと呼ぶ．

2.1 PB タスクモデル

PB タスクモデルとは，HYPHEN クラスタにおける並列プロセス実行モデルである．

PB タスクモデルは，PB 操作，EXT 操作によるFIFO キューへの操作によって実現される．PB 操作は，静的に割り当てが決定されたプロセッサのFIFO キューに対して，次に処理すべきタスクID を登録する操作である．EXT 操作はFIFO キューに登録されている先頭タスクID を取り出し，処理を開始する操作である．もしFIFO キューにタスクID が登録されていなければ，プロセッサは次のタスクID が登録されるまで待機する．また，EXT 操作は必ずPB タスクプログラム内の終了時に記述し，実行される．

PB 操作及びEXT 操作が高速に実行される条件下では，PB タスクモデルを効率よい同期操作として応用できる可能性がある [2]．例として，barrier 同期をPB タスクモデルで実現する方法を図 1 に示す．下図の例では依存関係のないB，C，D タスクが完了した後に同期を取り，依存関係を有するE が実行されるとする．ここで，FIFO キューの実行順序を利用し，C，D タスクの後にS1，S2 を同期の為のタスクとして登録することで同期がとれる．そして同期を司る最後のPB タスクS2 が次のタスクを登録することで次の処理が開始される．このようにしてバリア同期を実現できる．

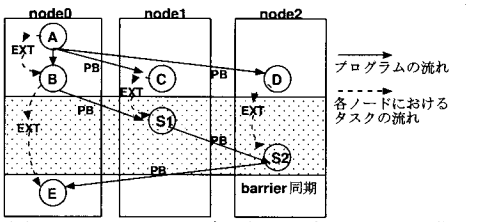

図 1: PB タスクモデルを用いたbarrier 同期

2.2 HYPHENクラスタのキャッシュ管理

HYPHEN クラスタにおける実行モデルはPB タスクモデルである為，タスク実行中にデータ共有をシステムが意識する必要はない．しかし，SVM(Shared Virtual Memory) を用いてページ単位でキャッシュ管理をしている為，false sharing が発生する．そこでHYPHEN クラスタではMultiple-Writer-Protocol を利用し，複数タスクからメモリ中の同一ページに書き込めるようにして，メモリ中のページの一貫性を保証する．

これを行う方法としては，Write-Through とDifference-Write-Back(以下，WB) が考えられるが，外部メモリのアクセス速度，ネットワークの輻輳を考慮してWB を採用する．Difference-Write-Back とは，キャッシュ時にキャッシュページのコピーを保持し，書き戻す際に，保持しておいたページコピーとページに書き込まれた差分 (Difference) のみを転送するWrite-Back である．

他プロセッサ上のタスクを起動するPB はタスクの任意の場所で発行可能である．従って，WB 操作は，EXT 操作と共にPB 操作にも付随して実行しなければならない．

3 キャッシュ基本プリミティブ

以下に，HYPHEN クラスタに必要となるキャッシュ基本プリミティブについて述べる．

3.1 ハードウェアのプリミティブ

メモリ更新操作の為のハードウェアプリミティブとして，CACHE プリミティブ，WB プリミティブを HyphenLink/PCI に持たせる．これらのプリミティブはプロセッサによって起動される．

CACHE プリミティブは，HYPHEN クラスタにおけるキャッシュメモリへ，メインメモリから転送するプリミティブである．その具体的動作を以下 (図2) に述べる．

1. プロセッサからキャッシュすべきページの存在する，HyphenLink/PCI のノード番号とページ先頭アドレスを受け取る
2. 受け取ったノード番号が他ノードの場合，HR-net[3] を介して対応する HyphenLink/PCI からキャッシュすべきページを転送する
3. WB 発行時に diff 生成するため，HyphenLink/PCI のキャッシュ管理用に確保したメモリ領域に，キャッシュすべきページデータを書き込む
4. 実際にキャッシュメモリへ書き込む

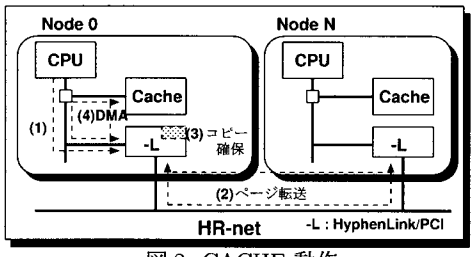

図 2: CACHE 動作

WB プリミティブは一貫性を保つために，HYPHEN クラスタにおけるキャッシュメモリからメインメモリにページデータを書き戻すプリミティブである．その具体的動作を以下 (図3) に述べる．

1. キャッシュメモリから書き戻すべきページデータを転送する
2. 転送したページデータと CACHE 時に確保したページとの diff を取る
3. 他ノードからのキャッシュであれば，HR-net を介して対応する HyphenLink/PCI にページデータを転送する
4. キャッシュ管理用に確保したメモリ領域を解放する

以上の二つのプリミティブを HyphenLink/PCI に持たせ，HR-net を介したメモリ更新操作を行うことで，非常に効率の良いキャッシュ管理を行うことができる．

3.2 ソフトウェアのプリミティブ

3.1 で述べたプリミティブを利用するために，ソフトウェア側で CACHE プリミティブと WB プリミティブを用意する．

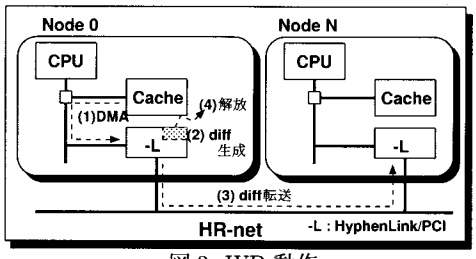

図 3: WB 動作

CACHE プリミティブは共有メモリアクセス保護違反が発生した時に呼ばれるプリミティブで，HyphenLink/PCI に対して違反を起こしたページの先頭アドレスを送信する．

WB プリミティブは，タスクが終了する時にキャッシュブロックとメモリブロックの一貫性を保つために呼ばれる．言い換えれば EXT 操作を行う時に呼ばれる操作である．WB プリミティブは HyphenLink/PCI の持つハードウェア WB プリミティブを起動するだけで良い．

4 実装

3.1 で述べたハードウェアプリミティブは，FPGA と DIMM を搭載した HyphenLink/PCI 上に実装する．キャッシュ制御を行う回路を FPGA 上に搭載し，キャッシュの diff 生成に必要なページデータを DIMM 上に確保する．

3.2 で述べたソフトウェアプリミティブは実際に HyphenLink/PCI を動作させる為のデバイスドライバを作成し，システムに対して公開する CACHE プリミティブと WB プリミティブを実装する．

5 まとめ

本稿では HYPHEN クラスタにおける PB，EXT を用いたプロセス実行モデルと，キャッシュ管理について述べた．PB タスクモデルでは，データの依存関係によるタスクの分割，PB，EXT による実行制御を静的に決めてやる必要がある．これをユーザが行えば，効率の良い細粒度の並列処理が行えるが，一般にはユーザの負担が大きすぎる．

OpenMP 等のような汎用の共有メモリプログラミング言語・環境の上で PB タスクモデルを動作させる配慮が必要である．

参考文献

[1] 立川純, 木原大悟, 田中里奈, 他 : HYPHEN クラスタにおけるノード PC 間接続機構の設計とその評価, 第 63 回全国大会情報処理学会 (2001).

[2] 有田五次郎 : FIFO キューを同期手段とする並列プログラムについて (I) : 待ちなし並列プログラム, 情報処理学会論文誌, 第 24 巻, pp. 221 – 229 (1983).

[3] 立川純, 木原大悟, 福澤毅, 他 : クラスタ計算機 HYPHEN におけるメモリアクセス機構 : HR-net, 並列処理シンポジウム JSPP2001, pp. 111 – 112 情報処理学会 (2001).