1F−1

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

立川純† 木原大悟† 田中里奈† 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 マイクロ化総合技術センター
††† 九州工業大学 情報科学センター

1 はじめに

新しい並列処理プラットフォームとして，既存PC/WSを，Ethernet Card等の汎用ネットワークインターフェイス(以下，NI)で接続したクラスタが有望視されている．特に，システムがDSM(Distributed Shared Memory)を提供し，プログラマはその上で共有メモリモデルに基づいた並列プログラムの実行を可能とする，DSM型クラスタの実現に向けた研究が盛んに行われている．

汎用NIを用いて構成されるDSM型クラスタでは，プロセッサの持つ仮想記憶管理のためのハードウェアをうまく活用し，共有メモリへのデータ参照を，実行時にNIへの通信に変換することで仮想的な共有メモリ(ソフトウェアDSM)を実現している．この時のNIを介した通信では，参照データが本来存在していたノード(ホームノード)からホストメモリへのデータコピーを行う．また，共有メモリモデルに基づく並列プログラミングに必須となるバリア・ロック等の同期プリミティブもNIによる通信によって実装される．

こうした汎用のNIだけを用いて，DSM環境を構築しようとするアプローチは，コストパフォーマンスに優れたシステムとして今後重要である．しかし，システムの規模が大きくなった場合には，スケーラビリティの点において限界があると考えられる．並列処理が普及するにつれ，スケーラブルな性能を持つ大規模システムの需要は高まるであろう．そこで，システムのスケーラビリティを目的とし，DSM型クラスタに必要となる機能を提供する専用ハードウェアを用いたアプローチとして，現在我々が開発中である“HYPHEN(HighlY scalable Parallel processing system on HiErarchical routing Network)クラスタ”を紹介する．HYPHENクラスタは，ノードPCに接続するNIであるHyphenLink/PCIと，専用スイッチHR-netによって構成されるPCクラスタである．本稿では，HYPHENクラスタの構成と，HyphenLink/PCIの提供する機能及びその設計方針を中心に述べる．

2 HYPHENクラスタ

DSMデータのキャッシュ操作や同期操作は，並列プログラム中で頻繁に実行されるため，これらの操作の性能がシステム全体の処理性能に直接影響を与える．HYPHENクラスタでは，高速にこれらの機能を提供するメカニズムとして，ハードウェアDSM，PBタスクモデルを提案しており，ノードPC間の専用NIであるPCIカードHyphenLink/PCIのサポートによって実現される．

2.1 ハードウェアDSM

ソフトウェアDSMでは，アドレス情報を含んだパケットの送受信を行う，NIによる通信によってメモリ参照のためのキャッシュ操作を実現する．その時，参照される側のプロセッサは，そのパケットの到着をNIからの割り込みによって認識する．従って，他の複数のノードから，同じノードに対する参照が集中した場合などには，一連の割り込み処理を連続して実行する必要があり，ソフトウェア的なオーバヘッドが大きい．

HYPHENクラスタでは，図1に示すように，システムのDSM本体を各ノードのホストメモリでなく，NI上のメモリによって構成する．つまり，DSM本体をノード外部のHyphenLink/PCI上に配置することで，メモリ参照先の割り込み処理によるプロセッサの介入を極力削減する．なお，従来のホストメモリは，システムの3次キャッシュとして捉え直し，プロセッサによるメモリ参照は，ホストメモリへキャッシュして行われる．

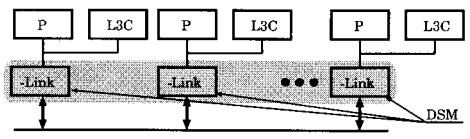

図1: HYPHENクラスタにおけるDSM構成

2.2 PBタスクモデル

PBタスクモデルは，HYPHENクラスタにおける並列プロセスの実行モデルである．このモデルでのプログラミングは，並列プロセスを最小の処理単位であるPBタスク(以下，タスク)の集合として捉え，複数タスクのプロセッサ割り当てを明示的に記述することで行われる．このプロセッサ割り当ては，システムの提供するPB(Parallel Branch)，EXT(EXchange Task)の二つのプリミティブによって記述される．従って，一つのタスクは，PB，EXTを含んだ命令流によって構成される細粒度な処理単位となる．

PBプリミティブは，タスクの実行を要求するノードIDと，要求するタスクIDを引数として，タスクの起動要求を発行する操作を示す．タスクの起動が要求されたノードは，すぐにそのタスク実行を開始するのではなく，各ノードが持つFIFOキューへその要求を投入する．実際の処理開始は，次に述べるEXT実行まで遅延される．EXTプリミティブは，自ノードのFIFO

キューにタスク要求があれば，それを取り出し実行に移し．なければFIFO キューへの投入を待つ．

このPB タスクモデルに基づく並列プログラムは，PB，EXT プリミティブが高速に実行できれば，より柔軟で高性能な同期実行メカニズムとして応用できる可能性がある [1]．

3 HyphenLink/PCI の設計

HYPHEN クラスタでは，2 節で述べたのメカニズムが密接に組み合わさって，並列処理システムとしての機能を果たす．そのためには，各ノード上で動作するOS との連携が必要であるが，ここでは特に2 節での機能実現のためのHyphenLink/PCI の機能に限定して述べる．HyphenLink/PCI は，本システムでのメインメモリとなるDSM モジュールであり，PB，EXT プリミティブのための通信機構である．また，HyphenLink/PCI 間は，専用スイッチHR-net[2] を介して相互に接続される．以降では，HyphenLink/PCI 上のメモリを単にDSM と表記する．

3.1 基本機能

プロセッサからみた共有メモリ空間は基本的にSVM 方式に準拠するため，x86 系プロセッサでのページサイズ(4KB) をDSM の管理単位とする．DSM 関連の機能としては，DSM からホストメモリへのページコピー操作(CACHE) と，ホストメモリからDSM へのページコピー操作(WB) が必要となる．これら操作は，SVM でのページミス時のトラップハンドラによって使用される機能であり，ソフトウェアオーバヘッドを考慮してゼロコピー通信によって実現する．HyphenLink/PCI によるゼロコピー通信は，参照先DSM がローカルOS による仮想記憶管理下になく，DSM へ直接アクセスできるため，容易に実現することが可能である．また，HR-net への通信量を削減する試みとして，CACHE 時にコピーしたページをHyphenLink/PCI 上に別に一時保存しておき，WB 時に更新ページとの差分のみを転送するDIFF 機能も持つ．

また，PB，EXT プリミティブのための機能として，FIFO キューを持つ．キューの深さは現在検討中で，最適な深さは対象とするアプリケーションに強く依存する．PB は指示されたノードのキューへのタスクID の投入で，EXT はキューからの取り出し操作である．

3.2 構成

HyphenLink/PCI は，図2 に示すように，PI，MI，NI モジュールによって構成される．本節では，その設計として各モジュールの動作について述べる．

PI(Processor I/F) : プロセッサからの要求を受け(a)，CACHE・WB 等であれば，DMA 機能を用いてホストメモリ-MI 間とのページ転送を実行する(b)．PB では直接NI に要求を発行し(e)，EXT はFIFO キューから直接データを取得する(f)．

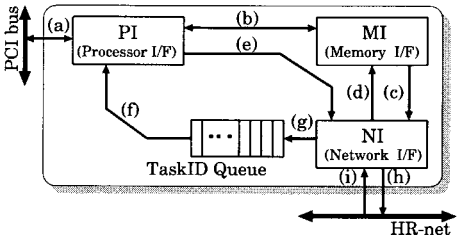

図 2: HyphenLink/PCI のブロック図

MI(Memory I/F) : DSM 本体へのアクセスや，DIFF 機能のためのデータコピーの保存を行う．PI から指示されたノードID が自ノードでなければ，ページ転送要求をそのままNI へバイパスする(c)．他ノードからのDSM 参照(d) にも対応する．

NI(Network I/F) : HR-net 間の通信インターフェイスである．PB 操作では，指示されたノードのFIFO キューにタスクID を投入し(g)，DSM 参照時には，HR-net に対するメモリアクセスサイクルを起動する(h)．また，PB 要求はこの段階で，ある特殊なアドレスへのメモリライト操作にによって変換される．従って，外部からの要求(i) はすべて，リード/ライトとアドレスによって識別される．

3.3 現状と基本性能の見積もり

HyphenLink/PCI の実装として，当研究室が開発したFPGA/DIMM を実装したPCI カードSHOKE2000[3] を用いている．現状では，プロセッサとSHOKE2000 上のFPGA 間の通信が実現されており，シミュレーションの結果，プロセッサによるPCI サイクル起動からのサイクルで，32bit データのSHOKE2000 上のDIMM へのシングルリード/ライトはそれぞれ14，12 サイクルで動作する．また，PB/EXT 操作のための自ノードへのFIFO キューへのシングルリード/ライトはそれぞれ，6,7 サイクルで動作している．

4 まとめ

本稿で述べたHyphenLink/PCI はDSM 型クラスタを想定した拡張カードであり，バリアのための専用機能を持たず，PCI 等の汎用I/O バスをベースにする等，非常にシンプルな構造を持ち，近い将来汎用化可能となる要素を多く含んでいる．今後はさらに具体的なHYPHEN クラスタの設計とともに，HyphenLink/PCI の実装を進める．

参考文献

[1] 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).

[2] 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).

[3] 田中康一郎, 有田五次郎:SHOKE2000:PCI-Based FPGA Card の開発とその評価, 電子情報通信学会論文誌 D-I Vol. J84-D-I No.6, pp. 540 – 547 (2001).