4B－4

超並列 *Teraflops* マシン *TS/1*

～分散共有メモリアーキテクチャ～

鈴木 真樹　田邊 昇　菅野 伸一　小柳 滋
RWCP† 超並列東芝研究室‡

1　はじめに

超並列 Teraflops マシン TS/1[1] では分散共有アクセス機構により共有メモリモデルによる並列プログラムを容易に作成することができ、また細粒度通信におけるオーバヘッドを低減することができる。

PE に汎用マイクロプロセッサを用いる時、汎用マイクロプロセッサの物理アドレス空間よりも大きな分散共有メモリを高速にアクセスするためには工夫が必要である。また、分散共有メモリを常にワード単位でアクセスしていたのではメモリアクセスのたびに結合網遅延が発生し、効率が低下する。

本稿では PE に汎用マイクロプロセッサを用いる TS/1 における分散共有メモリを効率良くアクセスするためのアーキテクチャを発表する。

2　TS/1 における分散共有メモリ

分散共有メモリの導入の目的は、プロセッサ間チェイニング機構ではカバーできない不規則細粒度通信の通信オーバヘッドの削減とプログラマに対して陽に送受信の記述をしなくても良い簡便な通信手段を提供することである。

2.1　マルチユーザ環境のサポート

図1は、TS/1 における分散共有メモリのアドレス変移を示しており、仮想記憶の管理を容易に行なうために2レベルアドレス変換方式を採用している。しかし、単に2レベルアドレス変換方式をサポートしただけではマルチユーザ環境で使用する場合にアドレス空間の断片化やプロテクション違反の問題が発生する。そこで、TS/1 ではユーザ単位に独立したグローバル仮想アドレスを用意する (複数の分散共有メモリ空間を用意する) ことで、マルチユーザ環境のサポートを行なう。2 レベル変換において複数のグローバル仮想アドレスを処理するためには、2 レベル目の変換の際にユーザ毎に異なるアドレス変換テーブルを用いるだけで済むが、1 レベル変換でユーザ毎に異なる分散共有メモリ空間を処理する為には、リモート PE でアクセスに対する正当性をチェックするための機構が新たに必要となる。

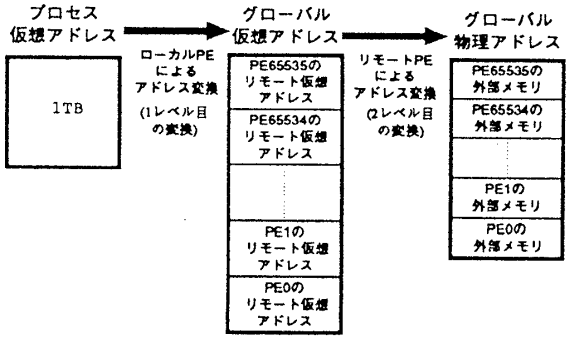

図 1: TS/1 の分散共有メモリのアドレス変移

2.2　分散共有メモリアクセス

汎用プロセッサでは TS/1 に実装されるようなテラバイトクラスの分散共有メモリをアクセスするには物理アドレスが不足する。このため従来 [2] は、分散共有メモリのアクセスのとき、分散共有メモリに対するアクセスが発生する度にプロセッサに割込みをかけてソフトウェアにより処理を行なっていた。しかしこれでは低オーバヘッドの通信は不可能である。

TS/1 では、上記の問題を解決するために、プロセッサ物理アドレスをセグメント方式を採用してグローバル仮想アドレスに拡張する機構を用いる。図2のように R4400PC から出力される物理アドレス (プロセッサ物理アドレス) は TSC1 の BANK レジスタの値を上位に連結されてグローバル仮想アドレスとなる。BANK レジスタを更新しないでプロセス仮想アドレスをプロセッサ物理アドレスに変換できるエントリのみを R4400PC の TLB に格納しておくことで、BANK レジスタを書き換えるときは R4400PC の TLB 例外で発見できる。従って、BANK レジスタの管理は OS だけで処理され、ユーザから BANK

Massively Parallel Teraflops Machine "TS/1", – Distributed Shared-Memory Architecture –
Masaki SUZUKI, Noboru TANABE, Shin-ichi KANNO, Shigeru OYANAGI

†Real World Computing Partnership(新情報処理開発機構)
‡(株) 東芝 研究開発センター 内

レジスタの存在を意識すること無く分散共有メモリをアクセスすることが可能である。

また、ユーザからは分散共有メモリのアクセスもローカルメモリのアクセスと変わらず、また、BANKレジスタの使用に関して若干のオーバヘッドが発生するが、従来のように分散共有メモリのアクセスの度にプロセッサに割込みを発生する方式に比べてオーバヘッドは遥かに少ないと思われるので低オーバヘッドの通信を行なうことができる。本方式において小規模な構成でありアドレスが足りる場合は、BANKレジスタの値は変化しないのでオーバヘッドがほとんどなく、常に割込みがかかる方式より有利である。

プロセッサ物理アドレス、グローバル仮想アドレスやBANKレジスタの値は、PE番号(X,Y,Z)やリモート仮想アドレスの項目より構成されており、プロセッサ物理アドレスを拡張するときに項目のビット境界を静的に変更できる。そのため、実装されているPE構成に見合った最適なアドレス拡張を行なうことができる。

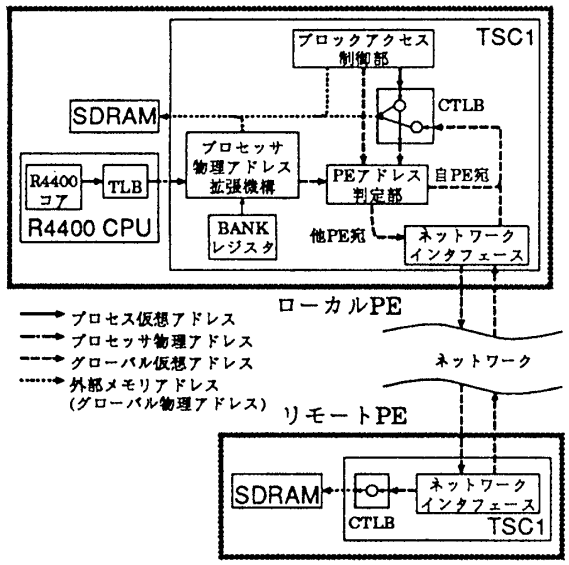

図 2: プロセッサ物理アドレス拡張機構

2.3 ブロックアクセス

前述の分散共有メモリアクセス機構を用いることにより分散共有メモリのアクセスにかかる通信オーバヘッドは削減されるが、リモートの分散共有メモリのアクセスにはキャッシングしない空間をアクセスするため常にネットワークによる遅延が生じるので、ローカルメモリのアクセスに比べるとアクセス時間が遅い。また、分散共有メモリをプロセッサから出力されるプロセッサ物理アドレスでアクセスする場合、リード命令のときにプロセッサがロックされるのでプロセッサの効率が悪くなる。ネットワークの観点で見ても、分散共有メモリのアクセスはネットワークに細かいメッセージが多く発生するのでアドレス等のヘッダ情報の転送に通信バンド幅がとられるため、ネットワークの効率も悪くなる。

以上の問題を解決するためにTS/1ではTSC1にDMAコントローラを内蔵し、分散共有メモリのブロック単位のアクセスをサポートする。これにより、プロセッサをロックすることなく分散共有メモリのアクセスが可能となることからプロセッサの効率が向上する。また、ネットワークにはブロック状のメッセージが発生するので、ワード単位のアクセスに比べてメッセージヘッダの情報が少なくなり、ネットワークの効率も向上する。

ブロックアクセスの起動は、通信オーバヘッドの削減のため、ユーザモードからブロックアクセスのコマンドをメモリに記憶しておき、システムプログラムにより予めプロセス仮想アドレスにマッピングされたDMAコントローラ起動レジスタに対してブロックアクセスのコマンドの先頭アドレスを書き込むことにより行なわれる。これにより、ブロックアクセスの起動は、システムプログラムの介在なしにユーザモードだけで処理を行なうことができ、低オーバヘッドの通信が行なわれる。

また、通常のメッセージパッシングでは、受信側でメッセージの書き込みアドレスを決定するために割込みが発生し、オーバヘッドの一要因となっているが、我々のブロックアクセスでは、書き込みアドレスもメッセージ中に指定されており、受信側での割込みが不要であるためオーバヘッドを少なくすることができる。

3 おわりに

本稿では、低通信オーバヘッドを目指した超並列TeraFlopsマシンTS/1における分散共有メモリアーキテクチャについて述べた。市販の汎用プロセッサを用いた場合でも、テラバイトクラスの分散共有メモリを低通信オーバヘッドでアクセスすることができる方法について述べた。

今後は、詳細設計を進めながら評価を行なう予定である。

参考文献

[1] 田邊, 菅野, 鈴木, 小柳: 「超並列テラフロップスマシン TS/1 の構想」, 情報処理学会研究報告 93-ARC-101-6, pp.41-48, 1993.

[2] 加納, 中田, 奥村, 大竹, 中村: 「並列マシン Cenju2 上の有限要素法による非線形変換解析」, JSPP '93, pp.379-386, 1993.