

可変構造型並列計算機のPE間通信プロトコル

甲斐康司 森 眞一郎 村上和彰 福田 晃 富田眞治

(九州大学大学院総合理工学研究科)

4 L - 7

1 はじめに

現在我々は、128 台の PE (Processing Element) を 128×128 の多重化クロスバー網 (MC-net) で接続した MIMD 型の並列計算機「可変構造型並列計算機」を開発中である。^[1] 本システムは、可変構造ネットワーク・アーキテクチャおよび可変構造メモリ・アーキテクチャを特徴としている。本稿では、PE間の通信プロトコルについて述べる。

2 PE間通信プロトコルの概要

本システムの各 PE はプロセッサ (PU) と並列に通信処理を行うメッセージ通信ユニット (MCU) を備えており、全ての PE 間の通信は MCU を介して行う。そのための PE 間通信プロトコルは、可変構造ネットワークを十分に活用できるものでなければならない。さらに、PE間の通信手段として、プロセス間のメッセージ交換、および、他のプロセッサのローカル・メモリへのアクセス (リモート・メモリ・アクセス) といった異なる PE 間通信手段を提供する必要がある。

これを実現するために、本システムでは PE 間通信プロトコルを図 1 に示す 3 つの階層にわけて実現している。^[2] 以下、各階層の通信プロトコルを概説する。

- ① メッセージ転送プロトコル (第 1 層): メッセージ転送プロトコルでは、MC-net 上での回線の接続/切断、双方向のパイプライン・データ転送、方向制御の手順を定める。
- ② メッセージ伝送プロトコル (第 2 層): MCU が規定するプロトコルである。MCU は、第 3 層の 2 つのプロトコルを統合し、すべての情報をメッセージ・パケットに変換する。このメッセージ・パケットの伝送により他 PE の MCU との間で情報の交換を行う。
- ③ PE 間メッセージ通信プロトコル (第 3 層): プロセッサが直接関与する階層であり、リモート・メモリ・アクセスのプロトコルとメッセージ交換のプロトコルを規定する。前者は通常のメモリ・アクセスと同様の手順である。一方、後者は OS のメッセージ・ハンドラが規定するプロトコルである。

3 メッセージ伝送プロトコル

メッセージ伝送プロトコルは、具体的には MCU 内のメッ

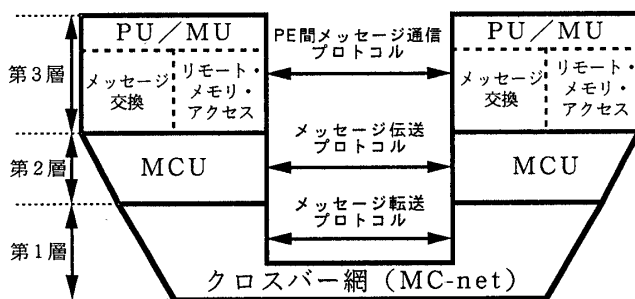


図 1. PE 間通信路の階層モデル

Inter-PE Communication Protocol in the Kyushu University Reconfigurable Parallel Processor
Koji KAI, Shin-ichiro MORI, Kazuaki MURAKAMI,
Akira FUKUDA and Shinji TOMITA
Kyushu University

セージ・センダ (MS) とメッセージ・レシーバ (MR) 間におけるメッセージ・パケット送受の取り決めである。このパケット送受の手順は、伝送するメッセージ・パケットの種類により若干異なる。そこで、まず 3.1 節で generic なプロトコルを示した後、3.2 節および 3.3 節でパケットの種類による相違点を述べる。

3. 1 generic なメッセージ伝送プロトコル

メッセージ伝送プロトコルは以下の 6 つのフェーズで構成する。これらのフェーズを順次遷移して、メッセージ・パケットの伝送を行う (図 2 参照)。

- ① 回線接続: MC-net 上で物理的な回線の接続を行う。
- ② リンク設定: MCU 間の論理的なリンクの設定を行う。
- ③ データ転送: MS から MR への Forward 転送を行う。
- ④ メモリ・アクセス: MR 側でメモリへの READ/WRITE アクセスを行う。
- ⑤ データ転送: MR から MS への Backward 転送を行う。
- ⑥ リンク解放, 回線の切断: 一連のプロトコル処理を終了する。

3. 2 リモート・メモリ・アクセス時の伝送プロトコル

基本的な上記のすべての手続きを実行する。リンク設定時には MS はリモート・メモリのアドレスを送出する。Backward 転送ではメモリ・アクセスの結果 (正常/異常終了) およびデータ (READ アクセス時のみ) を転送する。

このプロトコルでは WRITE アクセスの際、リモート・メモリでの WRITE 動作が終了し、そのアクセス結果を MS が受信するのを待ってプロトコル処理を終了する。ところが、PU にとっては WRITE アクセスが正常に終了するという保証が得られさえすれば、アクセス結果を待つ必要はない。そこで、WRITE アクセスに関し、次の 2 種類のモードを設けた。^[3]

- ① 同期 WRITE モード: リモート・メモリでの WRITE アクセスの結果を待ってリンクの解放を行う。

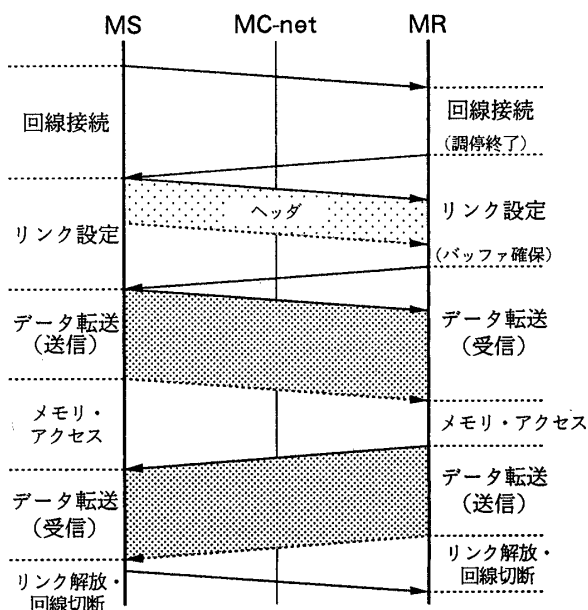


図 2. PE 間メッセージ伝送プロトコル

- ② 非同期 WRITE モード：WRITE アクセス時に Backward 転送フェーズを省略し、アクセス結果を待たずにリンクの解放を行う。実際の WRITE アクセスはリンク解放後に MR の処理として行われる。ただし、このモードでは WRITE アクセスの正常終了を保証するため、4.1 節で後述するような回線接続条件を課している。

3.3 メッセージ交換時の伝送プロトコル

MS は PU よりメッセージ交換の要求があると、まず MC-net 上で回線を接続する。リンク設定では、MS はメッセージのサイズなどを含むヘッダを送信する。MR が受信バッファ領域を確保し、リンク設定条件が満たされると MCU 間にリンクが確立される。次にデータであるメッセージを転送した後、リンクの解放、回線の切断を行う。

4 メッセージ転送プロトコル

MC-net は、8×8 クロスバー LSI を 256 個用いて構成する。プログラム実行時に回線接続要求を調停し回線を接続する単一デマンド・モード、クロスバー LSI 内の制御メモリの内容に従い回線を接続する単一プリセット・モード、および、これらを混合したハイブリッド・モードの 3 種の動作モードを持つ。^[4] そのため、メッセージ転送プロトコルは、これらのモードに応じて回線の接続／切断、データ転送、および、方向制御を行わねばならない。

メッセージ転送プロトコルは、次の 2 種類のプロトコルから成る。

- ① データ転送プロトコル：SEND と ACK の 2 制御信号を、それぞれを Forward 転送時、Backward 転送時のデータ・ストローブとしたパイプライン・データ転送を規定する（図 3 参照）。
- ② MC-net 制御プロトコル：回線の接続／切断、方向制御を行う SYN, SEND, ACK, ARB の 4 制御信号のハンドシェイクの手順を規定する。

この章では、単一デマンド・モード時と単一プリセット・モード時の回線接続手続きについて詳説する。ハイブリッド・モードでは、これらの両者を混合することで実現できるので省略する。

4.1 回線接続条件

同期 ERITE モードでは、MR のすべての処理が終了した時点で 1 回のメッセージ伝送が終了するので、MR は次の接続要求を直ちに受け付けることができる。

ところが、非同期 WRITE モードでは 1 回のメッセージ伝送が終了した後でも、MR は当該メッセージ伝送に関する処理を継続している場合がある。これには、メモリへの WRITE アクセス、および、それにとまうページ・フォルト処理やコンシステンシー制御などがある。したがって、MR がすべての処理を終了していることが次の接続要求を受け取るための条件となる。

4.2 デマンド・モード時の回線接続手続き

デマンド・モードでは、プログラム実行時に回線の接続を行わねばならない。回線接続の手続きを図 4 に示す。

MS は、回線を接続するために、まず SYN と送信先 PE のアドレスを、MC-net 上の送信先 PE が接続するクロスバー LSI に送出する。デマンド・モードでは、SYN, ARB, ACK の 3 信号の

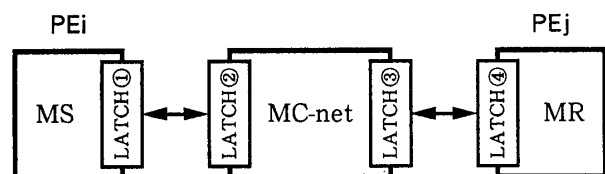


図 3. 通信経路の概略図

ハンドシェイクにより、次の 2 段階の調停を行い回線の接続を行う。

- ① クロスバー LSI 内部調停：クロスバー LSI は、他 MS と競合がある場合は調停を行いスイッチを接続し、さらに送信先 PE に接続要求信号 ARB を送出する。
- ② クロスバー LSI 外部調停：接続先 PE の MR で ARB と SYN を受信すると、他クロスバー LSI 間との調停を行った後 1 つの MS を選択し、かつ回線接続条件が満たされていれば ACK を返す。

MS は ACK の受信で回線接続が成功したことを確認し、SEND をストローブとしてデータのパイプライン転送を開始する。

4.3 プリセット・モード時の回線接続手続き

プリセット・モードでは、電気的には常に MS-MR 間に回線が接続されているためメッセージ伝送ごとに回線接続を行う必要はない。したがって、同期 WRITE モードでは MS はいつでもデータの転送を開始できる。

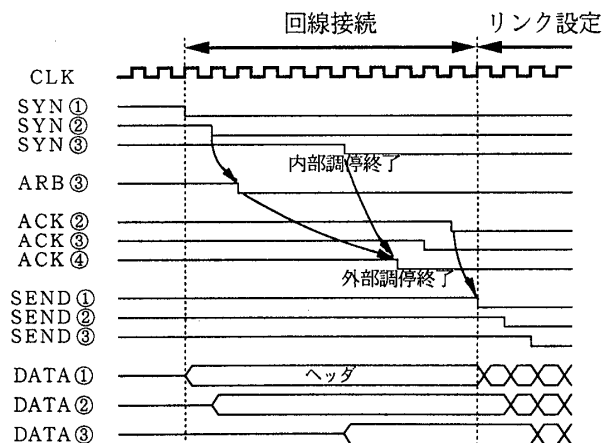
しかし、非同期モードでは、電気的に回線が接続されているにもかかわらず、回線接続条件が満たされているとは限らない。この状況を MS は把握しておく必要がある。そこで、通常はデータの転送方向を Backward として MR の内部状態を MS に送出し、MS 自身が回線接続条件の判定を行う。その結果、条件が満たされている時に限り、MS は転送方向を Forward に切替えデータの転送を開始することができる。

5 おわりに

以上、MC-net 上でメッセージ交換、および、リモート・メモリ・アクセスを実現するための PE 間通信プロトコルについて述べた。今後、このプロトコルの動的な性能の評価を行う予定である。

参考文献

- [1] K.Murakami et al. : "The Kyushu University Reconfigurable Parallel Processor - Design of Memory and Intercommunication Architectures -", Proc. 1989 Int'l. Conf. Supercomputing, pp351-360, June 1989.
- [2] 森ほか：「可変構造型並列計算機の PE 間メッセージ通信機構」, 情報処理学会論文誌, vol.30, no.12 (1989 年 12 月)。
- [3] 甲斐ほか：「可変構造型並列計算機のローカル／リモート・メモリ・アーキテクチャ」, 情処研報, 90-ARC-80-11 (1990 年 1 月)。
- [4] 蒲池ほか：「可変構造型並列計算機のネットワーク制御方式」, 信学技法, CPSY89-16, pp. 7-12 (1989 年 8 月)。



○ 信号線の数字は図 3 のラッチの出力を表す。

図 4. デマンド・モード時の回線接続手続き