7H−2

スーパーデータベースコンピュータ SDC2 における データネットワーク系の管理方式

田村孝之 中村 稔 喜連川 優 高木幹雄
東京大学 生産技術研究所

1 はじめに

現在我々が開発している高並列関係データベースサーバ SDC2 では, 処理モジュール間の相互結合網として, 制御通信用のコントロールネットワーク (CNet) と処理データ送信用のデータネットワーク (DNet) との2種類のネットワークを用いている [1]. DNet には間接多段網を採用し, それぞれのスイッチ素子に並列関係演算処理を支援する "バケット平坦化機能" を持たせている.

バケット平坦化は負荷分散の一種であるが, さまざまな動作環境に対応するために, 可変長タブルのサポートや一部処理モジュールの障害についても考慮されている [2]. また, DNet 自体も性能向上のために 2 重化されており, 一系統の障害に備えて 1 つのネットワークを論理的に独立な 2 つの系統に見せる機能を持っている.

このように, ネットワークの高機能化とフォールトトレランスの向上に伴って, スイッチ素子全体を一括して管理する機構が不可欠となる. そこで, 我々はスイッチ素子の動作パラメータの設定や動作状態のモニタリングを行なうための専用の管理プロセッサを DNet の構成要素として導入した.

本論文では, SDC2 の問合わせ処理におけるデータネットワークの位置付けから管理プロセッサに要求される役割を検討し, スイッチ素子とプロセッサ間のインタフェースについて述べる.

2 データネットワーク系の構成と機能

SDC2 の データネットワークは, 図1に示すように, スイッチ素子網およびネットワーク管理プロセッサと, 処理モジュール・ネットワーク間のインタフェースで構成される [3].

スイッチ素子網は, 図1に示すように 8 × 8 のオメガネットワークを変形したものである. スイッチ素子のデバイスとしては, FPGA の一種である LCA XC4010 を用いており, このチップ 1 つで, 2 × 2 のクロスバスイッチとルーティング制御回路の組を構成している.

スイッチ素子には, 宛先指定ルーティングに加えて, バリア同期機構とバケット平坦化機能が必要であり, 一部の処理モジュールや DNet の 1 系統が故障した場合にもこれらが動作することを保証しなければならない. バケット平坦化のために必要となる資源には, 以下のものが挙げられる.

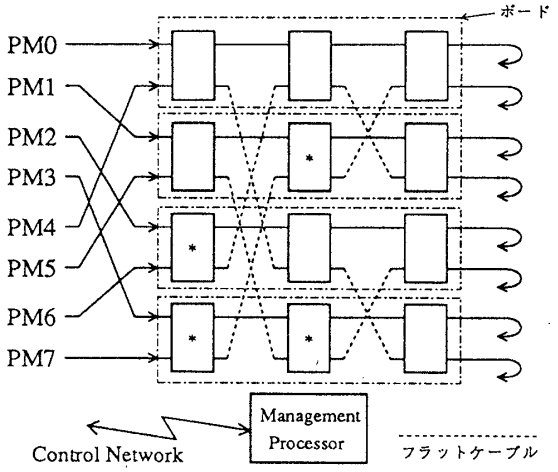


図 1: データネットワーク系の構成

- データの履歴を記憶するメモリ (外付け).
- 履歴に基づいて経路を決定するための比較器.
- 履歴を更新するための加算器.
- データの平坦度とブロック率のトレードオフを設定する閾値レジスタ.
- 一部の処理モジュールが故障した際に, 出力ポート毎に重みを設定するためのレジスタ.

また, ネットワーク管理プロセッサは, バックプレーンバスを通じてスイッチ素子と接続されており, また, 処理モジュールやホストマシンとは CNet を通じて接続されている. 上記のスイッチ素子の内部レジスタ等は, フロントエンドマシンからの指示に基づいて管理プロセッサが設定する.

3 管理プロセッサの処理

管理プロセッサが果たす役割を処理の段階を追って示すと以下のようになる.

3.1 システム起動時

スイッチ素子には FPGA を用いているため, 電源投入時にはデバイスのコンフィギュレーションデータを書き込む必要がある. このデータは, ROM に書き込んで自動的に FPGA に読ませることも可能であるが, デバッグ時の効率や, バージョン管理の容易さ, および動的な再コンフィギュレーションへの応用を考えて, 管理プロセッサからダウン

Data network management scheme for SDC2.
T.Tamura, M.Nakamura, M.Kitsuregawa, and M.Takagi.
Institute of Industrial Science, University of Tokyo.

ロードする方式を採用した.

現在, FPGA の開発はワークステーション上で行なっているが, 管理プロセッサから CNet を通じてリモートファイルアクセスを行なうことができるので, 配置配線が完了したデータを直ちに FPGA にダウンロードしてデバッグを行なうことができる. また, スイッチ素子毎にダウンロードする内容を変えることも容易である.

3.2 スイッチ素子の初期設定

問合せ処理の実行に先立って, 各処理モジュールや DNet が動作可能であるかどうか調べ, システム全体の構成を決定する必要がある.

正常なモジュールに対する重みを 1, 使用不能なモジュールに対する重みを 0 として, 各スイッチ素子の出力ポートから到達できる処理モジュールの重みの総和を該当するスイッチ素子の内部レジスタに設定する. 以後, バケット平坦化時にこの重みを用いることにより, 正常な処理モジュールだけにデータを送りながら負荷分散を行なうことが可能になる.

また, バリア同期の際に, 動作していない処理モジュールからは同期信号が出力されないため, その処理モジュールに対する同期信号はマスクする (同期信号を受け取ったことにする) 必要がある.

さらに, 今回実装した DNet ではネットワークのトポロジーを変形しているため, 図 1で * 印の付いたスイッチ素子については, straight 接続と crossed 接続の状態を反転するようにフラグを設定する.

バケット平坦化機能を用いて関係演算処理を行なう場合には, スイッチ素子の外部メモリに記憶されている履歴情報もあらかじめクリアしておく必要がある.

3.3 動作状態のモニタリング

初期設定等が完了して, ネットワークをデータが通過している間は, 管理プロセッサはスイッチ素子の動作状況をモニタリングする.

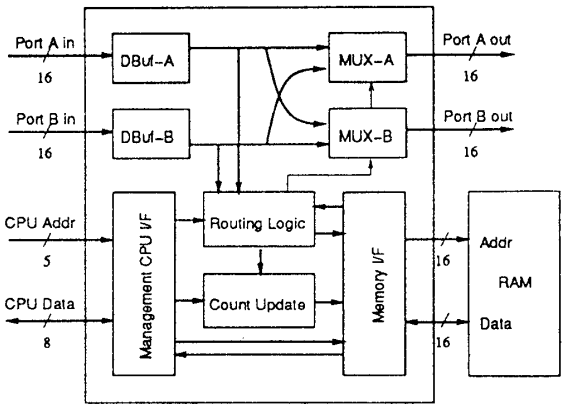
監視する内容としては, 各スイッチ素子における輻輳率などのネットワーク性能に関するものと, 履歴情報のオーバーフローや動作不能なモジュール宛のタプルの到着などの例外状態とが考えられる.

管理プロセッサは, これらのモニタ結果をまとめてフロントエンドに報告する.

4 スイッチ素子のプロセッサインタフェース

図 2にスイッチ素子のブロック図を示す. 図で CPU I/F となっている部分には, 16 個の 8 bit レジスタが含まれるが, これらは以下のように分類できる.

1. 内部ステート制御レジスタ
2. 内部ステート監視レジスタ
3. バケット平坦化のパラメータ (閾値, 重み)
4. メモリ制御レジスタ
   (アドレス, データ, カウント, 操作指定)



(Xilinx LCA XC4010-5 +64K SRAM)

図 2: スイッチ素子のブロック図

内部ステートの制御レジスタの機能には, バリア信号をマスクすることや, ルーティングをプロセッサから直接制御することなどが挙げられる. また, 内部ステート監視レジスタには, スイッチ素子内のいくつかの信号の状態を直接読み取るものに加えて, 特定の信号がアクティブなサイクル数をカウントするレジスタも含まれる.

メモリ制御レジスタは, プロセッサからスイッチ素子を経由して外部メモリをアクセスするためのものである. ルーティング動作を行なっている間にも, メモリアクセスが互いに干渉しないように, スイッチ素子内部でアービトレーションを行なっている. メモリ操作としては, 指定したアドレスの読み出し, 指定したアドレスへのデータの書き込みに加え, 指定した領域を特定データで塗りつぶすことが可能になっている.

5 まとめ

本論文では, SDC2 のデータネットワーク系の管理方式について述べた. FPGA の採用と管理プロセッサの導入により, データネットワークの柔軟性は大きく増した. 今後は, さらにモニタリング機能の強化について検討していきたい.

参考文献

[1] 中村, 田村, 喜連川, 高木. スーパーデータベースコンピュータ第二版 SDC2 におけるシステムソフトウェアの構成. 情処 第 47回全国大会 発表予定, 1993.

[2] 田村, 中村, 喜連川, 高木. スーパーデータベースコンピュータ (SDC) のバケット平坦化ネットワークにおける縮退動作支援アルゴリズムとその評価 *SWoPP '92*. 情処計算機アーキテクチャ研究会, 1992.

[3] 田村, 中村, 喜連川, 高木. スーパーデータベースコンピュータ (SDC2) におけるデータネットワーク系の実装. *SWoPP '93*. 信学技報, 1993.