超並列ネットワーク MDX 用ルータチップの自動設計システム*

2G－2

久々宮守　原田智紀　村上祥基　朴泰祐†
筑波大学 電子・情報工学系‡

1. はじめに

並列計算機のネットワークには, Processing Unit (以下 PU) 同士を直接つなぐ直接網と PU 間にクロスバ・スイッチ (以下 XB) 等をはさんだ間接網がある. 間接網の一つのクラスに MDX (Multi-Dimensional X'bar)[1] がある. これらのネットワークの転送性能の評価は, 我々が開発した汎用ネットワークシミュレータ生成システム INSPIRE[2] を用いることで容易に行なうことができる. ところが INSPIRE では実際の回路を考慮した評価を行なうことができない. 現実の回路を考慮した評価を行なうためには, 回路を設計して, そのハードウェア量, 最大動作周波数などを知ることが必要である. しかし回路設計を行なうには大変な労力を要する.

そこで本研究では, MDX の INSPIRE 上での記述から VHDL 記述への変換を行なうトランスレータを作成する. このトランスレータを用いることにより, INSPIRE における記述を行なうだけで, そのネットワークの要素回路を設計でき, そのネットワークの転送性能と, 実際の回路のハードウェア量と最大動作周波数を知ることによって, 容易に現実的な評価ができるようになる.

2. 背景

本研究の対象となるネットワーク・クラス MDX, ネットワーク・シミュレータ生成系 INSPIRE, ハードウェア記述言語 VHDL について簡単に紹介する.

2.1. MDX

MDX は並列計算機用の間接網の 1 クラスである [1]. MDX は, エクスチェンジャ (以下 EX) と呼ばれるルータ・スイッチと XB により構成され, PU 間に必ず一つの XB を持つという性質を持つ. MDX は, 多段接続網 (MIN) の転送半径が一定という特性と, 直接網の局所性との両方の特性を持ち, 通信の局所性を生かすことができ, 少ないハードウェア量で高いバンド幅と小さいレイテンシを実現できる. 現在提案されている MDX として, ハイパクロスバ (以下 HXB) , MDX-Star, MDX-Baseline, MDX-De Bruijn 等がある.

2.2. INSPIRE

INSPIRE は, 専用言語 NDL でネットワークの諸特性を記述し, それに基づいてネットワーク・シュミレータを生成するシステムである [2]. このシミュレータを用いて, そのネットワークの性能を容易に評価することができる. NDL では, 以下のことを記述することによりネットワークの諸特性を定義する.

- ネットワークの形状
- PU や各種スイッチの個数とそのチャネルのサイズ
- 各ネットワーク資源におけるチャネルの結合
- 各 PU 及びスイッチにおけるルーティング・アルゴリズム

2.3. VHDL

VHDL[3] は, 論理回路の構成と動作を記述するハードウェア記述言語の一つで, 多方面で用いられている. NDL をこの VHDL に変換し, 回路の合成を行ない, ハードウェア量と最大動作周波数を容易に評価可能な環境を作ることが, 本研究の目的である. 変換した VHDL 記述を, シノプシス社の VHDL 論理合成ツールによって論理合成することにより, ネットワークのより現実的な評価を行なうことができる.

3. NDL→VHDL トランスレータ

MDX に属する網の共通部分として, XB の回路全てと, EX の, ルーティング・アルゴリズムによらない部分を, システム・テンプレートとして用意する. このことによりルーティング・アルゴリズムに依存する部分のみを変更することで, 任意の MDX を合成できるようになる.

3.1. NDL 記述の制約

現在のトランスレータでは 3 次元構成の MDX を対象とし, wormhole ルーティングを行なう固定ルーティングのみを変換可能である. そして NDL 記述と VHDL 記述の言語の違いを埋めるために, 実際の NDL で使える関数・命令・演算子, そして構文にもある程度制限が加わる.

3.2. 出力される回路

変換系は EX に関して図 1に示す回路構成を生成することを前提とする. "AR" はアービタ回路を示し, これはバッファ制御回路からの要求を処理する. "SW" はスイッチ回路を示し, AR からの制御信号を受けて複数の入力のうちの一つだけを出力につなぐ. "B" はバッファ回路であり, 容量は 2 flit である. "BC" はバッファ制御回路であり, 各バッファに用意されていて, バッファにメッセージ・ヘッダが到着した時に起動され, その情報を基に経路決定を行なう. この回路要素のうち, AR 部分と SW 部分は, 予めテンプレートとしてシステムが用意し, B 部分の一部と BC 部分をトランスレータにより生成する.

3.3. トランスレータの実装

トランスレータの実装を以下のように行なう.

- 回路記述のテンプレートは, 予め VHDL で記述し用意する.
- NDL 中のルーティング・アルゴリズム記述に注目し, これを先述の EX の B, BC 回路の VHDL 記述に変換する.

---

*Automatic design system for MDX network router chip

†Mamoru Kugumiya, Tomoki Harada, Yoshiki Murakami, and Taisuke Boku

‡Institute of Information Sciences and Electronics, University of Tsukuba

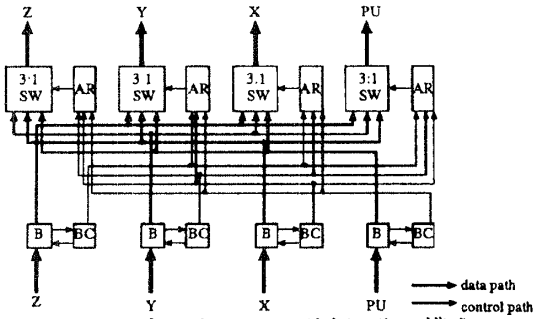

図 1: ネットワーク要素回路の構成

トランスレータは lex 及び yacc を用いて書かれている. その記述量は lex 部約 280 行, yacc 部約 650 行, C 部約 130 行である.

4. MDX 各網の評価

提案されている MDX の代表として HXB , MDX-Star, MDX-Baseline, MDX-De Bruijn をとりあげ, INSPIRE での評価と, 今回作成したトランスレータで変換し合成した結果についての評価を行なう.

4.1. 評価環境

3次元で 512PU (8×8×8) のネットワークを対象とする. INSPIRE では全 PU が独立してランダムに行き先 PU を決定し, 毎クロック一定の確率でメッセージを転送するという条件で, シュミレーション時間 20000clock, メッセージ長 20 flit として評価を行なった. 評価指標として, ネットワーク・スループットとメッセージの平均レイテンシを用いる. またハードウェア量や最大動作周波数の評価環境としてシノプシス社の VHDL 論理合成ツールを利用する. 本研究における回路合成では, 配線及び供給電源等は考慮せず, セルについてのみ考慮した合成を行なっている. これは特定のテクノロジを意識した設計を行なわず, トポロジの違いによるハードウェア量及び転送性能の比較を目的としているためである. 従って後の評価も各セルに設定されている値のみの評価である. INSPIRE では clock 単位の評価しか行なうことができないので, 回路合成による最大動作周波数の評価により, シミュレーション結果を clock 単位から nsec 単位に修正する. 現在の実装ではデータ線の幅は, 16bit に固定されている. つまり 1flit = 2Byte である.

4.2. 評価

表 1に回路合成結果, 図 2にランダム転送による転送性能の評価結果を示す. 合成した回路それぞれの信号ピン数は, システムの仕様により, 外部との入出力信号数が全ての回路で同じため, どの回路も 154 となる. 図 2 のスループット (横軸) は, PU が 1 nsec あたりに受信できるデータ量を表す. また平均レイテンシ (縦軸) は, 送信 PU でメッセージが生成されてから, 受信 PU に到達するまでの時間である. また, 今回評価した全てのネットワークに共通する XB のゲート数と最大動作周波数は, それぞれ 14890 と 29.34MHz となった. 全てのネットワーク要素回路は同期していると仮定しているので, HXB は EX 自身の最大動作周波数が 29.4MHz であるにも関わらず, 網全体として最大動作周波数が 29.34MHz となる. このことを踏まえても局所性の面で有利な HXB が他の網に比べて有利になっている. 結果として対称性が良く, シンプルである HXB が優れていることが分かる. このように今回作成したシステムで変換を行なうと, 容易に現実的なネットワークの評価を行なうことができる.

表 1: 回路の合成結果

トポロジ名	EX のゲート数	最大動作周波数 (MHz)	
		EX	網全体
HXB	6641	29.4	29.34
MDX-Star	6622	24.13	24.13
MDX-Baseline	6792	25.72	25.72
MDX-De Bruijn	6781	25.51	25.51

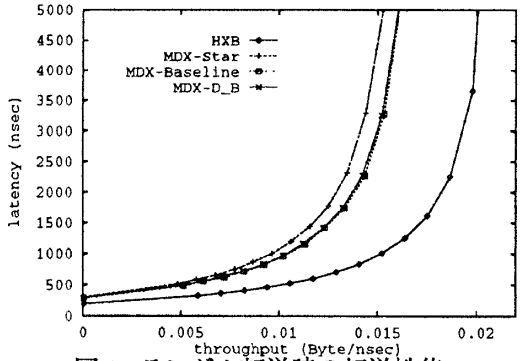

図 2: ランダム転送時の転送性能

5. おわりに

本研究では, MDX 網に対し NDL 記述を VHDL 記述に変換するトランスレータを作成した. これにより, 計算機シミュレーションだけでは知ることのできなかった, より現実的な評価を容易に行なうことができるようになった. また実際にいくつかの MDX 網を NDL で記述し, シミュレーションと回路合成を行いシステムの有効性を示した. 実際の変換にかかる時間も 0.2 sec 程度で, 十分高速である.

今後の課題として, virtual channel への対応と, MDX だけでなく, さまざまなネットワークについて, 計算機シュミレーション用の記述から, 容易にハードウェア量及び最大動作周波数を評価することを可能にすることが挙げられる.

謝辞

本研究に関し貴重な御意見を頂いた, 筑波大学西川博昭助教授ならびに坂井修一助教授, アーキテクチャ研究室諸氏に深く感謝します. なお, 本研究の一部は創成的基礎研究費 (08P0401) の補助によるものである.

参考文献

[1] 村田淳 他, "大規模並列計算機用結合網クラス MDX の提案と評価" JSPP'96 予稿集,pp.137-114,1996.

[2] 原田智紀 他, "並列処理ネットワークのための性能評価用シュミレータ生成系 INSPIRE" 情報処理研報 Vol.95, No.80, pp.65-72

[3] R. Lipsett, C. Schaefer and C. Ussery, "VHDL:Hardware Description and Design" Kluwer Academic Publishers,1989

[4] 村上祥基 他, "VHDL によるハイパクロスバ網用ルータ・チップの設計" 情報処理研報 Vol.96, No.121, pp.17-24