可変ルーティング機能を付加した相互結合網のスイッチング・ユニット†

坂井修一†† 計宇生†††
田中英彦†††† 元岡達††††

VLSI 技術の発達とともに高並列計算機（数十～数千台）への期待が高まっているが，一般に高並列の計算機システムでは通信系が処理の隘路になる場合が多く，転送性能の高い相互結合網の設計が必要となる．本論文では，相互結合網の構成単位となる蓄積交換スイッチング・ユニット (SU) の設計と評価に関して述べる．当 SU は，可変ルーティング機能をもち，転送性能・信頼性の両面から，マルチパスの多段結合網，格子型結合網，CCC 網などのさまざまな網構成に有利である．市販 TTL を用いた設計では，SU は 4750 個のゲート，メモリおよび結線部より成り，将来 114 ピンの LSI として1チップ化が可能である．さらに，当 SU を $(\log_2 N+1)$ 段の多段結合網であるガンマ網に適用した場合のシミュレーション評価を行った．その結果，高スループット・低遅延の網が実現されることが示され，また，可変ルーティングの効果は，とくにデータ生成率の高い場合に顕著であることが示された．たとえば，10 MHz のクロックを用いた場合，64 ポートのガンマ網のポートあたりのスループット（制御や閉塞によるオーバヘッドを含めた値）は約 8 MB/s，パケット（平均の長さ 9 B）の平均転送時間は約 3.2 μs である．

1. まえがき

VLSI 技術の発達とともに高並列計算機（数十～数千台）への期待が高まっているが，一般に高並列の計算機システムでは通信系が処理の隘路になる場合が多く，転送性能の高い相互結合網の設計が必要となる[1]．

本論文は，結合網を構成する単位となるスイッチング・ユニット (SU) の設計と評価に関して述べたものである．その特徴として，動的にルート変更を行うことが挙げられるが，これは次に述べる背景・理由から SU に必要な機能と考えられる．

（1） 従来多段結合網の主流であったオメガ網[2]・間接キューブ網[3]・デルタ網[4),5]などの $\log N$ 段の結合網は，ハードウェア量の小ささ・制御の容易さの点で優れているが，閉塞による性能低下が大きい，1ルートの網であるため故障に弱いなどの欠点がある．これに対して，閉塞率の低い網であるベネス網 ($2\log N-1$ 段)[6] はルーティング制御に手間がかかり，バイトニックソート網 ($\log N\times(\log N+1)/2$ 段)[7]やクロスバスイッチ網（接点数 N^2）はハードウェア量が大きいなどの難点をもつ．

このような理由から，現在 $\log N+1$ 段から $2\log N-1$ 段程度の多ルート（マルチパス）の多段結合網に関する研究・開発が行われている．Gajski[8] らは，大規模並列マシン Cedar に，マルチパス化した改良型オメガ網を採用している．また，Adams[9),10] らは，主として信性頼の点から，$\log N+1$ 段に拡張した間接キューブ網の検討を行っている．さらに，Chin[11] らは，パケット交換方式の多ルート多段結合網を提案し，これを解析した．

これらの改良によって網の転送性能や信頼性が大幅に改善されることが示されているが，ルーティング制御を含む詳細な網設計が行われているとはいえない．

（2） CCC 網[12]・格子型網[13),14]・超立方体網・superimposed 木網[15] などの結合網は，本来マルチパスの網であり，適切なルートの選択を行うスイッチが，転送性能・信頼性の面から要求される．

本論文で報告する SU は，分散制御・同期・蓄積交換方式であり，ルーティングテーブルを用いた行先制御を行うため汎用性が高く，(1)，(2)の両方の用途に適している．本論文では，まず当 SU の構成と動作を示し（2章），これを用いた相互結合網の特徴を述べ，具体的な網を想定したシミュレーション評価を行う（3章）．さらに，実効的な転送速度・信頼性・LSI 化の問題点などについて考察し（4章），最後に今後の課

† Switching Unit with Variable Routing Control Facility: A Component of Interconnection Networks by SHUICHI SAKAI (Information Engineering Course, Graduate School of Engineering, The University of Tokyo), USEI KEI (Department of Electronics Engineering, The University of Tokyo), HIDEHIKO TANAKA and TOHRU MOTO-OKA (Department of Electrical Engineering, The University of Tokyo).

†† 東京大学大学院工学系研究科情報工学専門課程
††† 東京大学工学部電子工学科
†††† 東京大学工学部電気工学科

題を列挙する（5章）.

2. 可変ルーティング機能をもつ SU

2.1 設計の目的と指針

われわれは，TTL-IC を用いて，可変ルーティング機能をもつ SU の設計を行った．今回の設計の目的を以下に列挙する．

（1） SU の機能と制御方式・ハードウェア構成を明確化すること．

（2） SU のハードウェア量の見積り・LSI 化の検討を行うこと．

（3） 制御時間を含めた網の転送性能の詳細な見積りを行うこと．とくに可変ルーティングの効果を測定すること．

なお，われわれはすでに固定ルーティング方式のスイッチング・ユニットを設計・試作しており[18]，今回の設計はその改良である．前者からのおもな変更は，

（1） 多ルート化

（2） バッファの高速化

の2点である．

2.2 基本構成

当 SU は，m 個の入力ポートと n 個の出力ポート（$m=n=4\sim5$ 程度を考えている）を各入力ポートに対応する内部バスにより結合した小規模のクロスバスイッチである（**図1**）．制御は各入力ポートと出力ポートに分散し，入力ポート・コントローラでルーティングを行い，出力ポートでスイッチングを行う．各入力ポート内には FIFO 型のバッファ・メモリがあり，可変長パケットを単位とする蓄積交換を行う（同期式）．転送は1パケット内でパイプライン化される．すなわち，パケットの末尾の語の到着を待たずに，先頭の語から次々に次段の SU に転送される．

ルーティングは，入力ポート・コントローラ内にルーティング・テーブル（RT）を置き，これを用いる表引き方式である．他に，組合せ論理回路のみで制御を行う方式が考えられる．後者はハードウェア量の点で有利であるが，汎用性に乏しい欠点があるため採用しなかった．

入力ポートの FIFO メモリは，容量の点・拡張性の点から，今回の設計では RAM を用いて実現した（**図2**）．RAM は2面設け，パケットをインタリーブして格納することにより，1語／クロックの高速転送を行う．RAM の読出しアドレス・書込みアドレスは，それぞれ読出しカウンタ（R-CNTR）・書込みカ

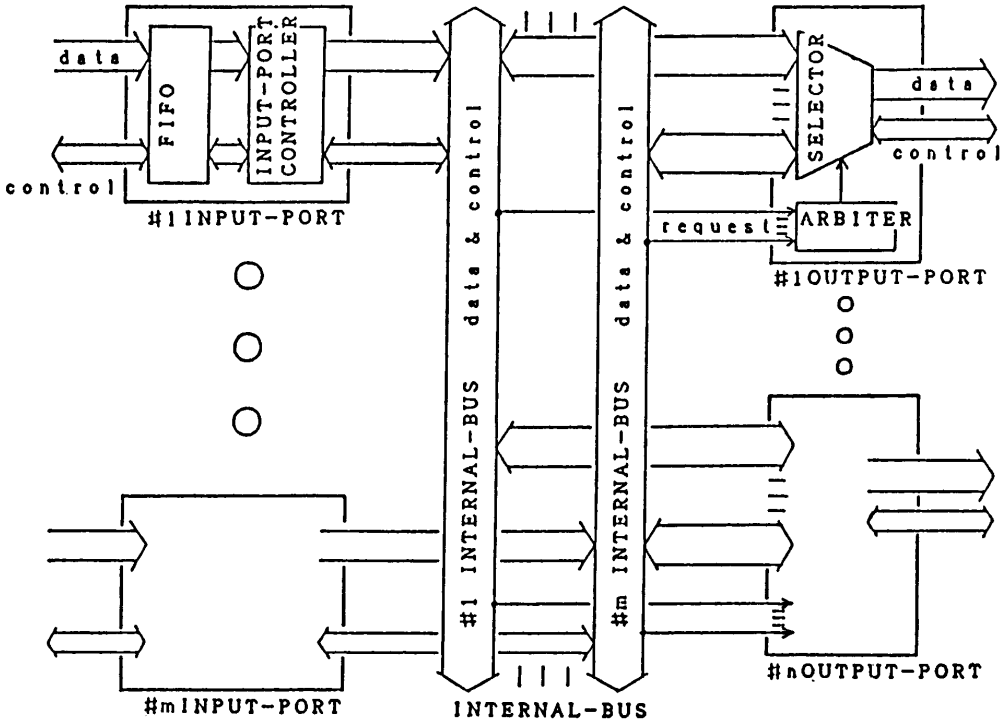

図1 SU の全体構成
Fig. 1 Overview of SU.

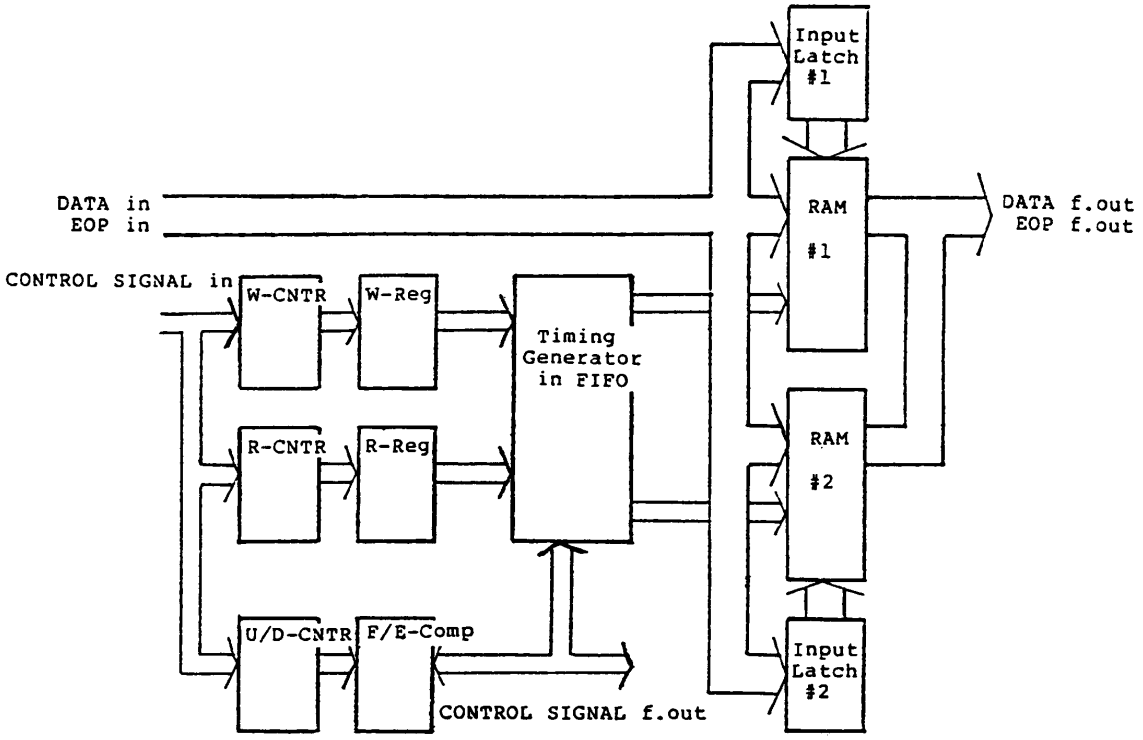

図 2 入力ポート内の FIFO の構成
Fig. 2 FIFO in input port.

ウンタ (W-CNTR) によって指定される．両アドレスの差分はアップダウンカウンタ (U/D-CNTR) に保持され，これによって FIFO の Full, Empty が検出される．

なお，パケット整形・デッドロック検出などの機能は，今回の設計には含まれていない．

2.3 可変ルーティング

可変ルーティング機能は，一つの行先に対して複数個 ($\leqq n$) の出力ポートを指定できるように RT を構成することで実現される (**図 3**)．入力ポート・コントローラは，パケットの到着のたびに，テーブルに記された出力ポートに，第 0 エントリから順番に接続要求を出し，受理された時点から当該パケットの転送が行われる．

今回は，適応型のルーティング（行先の交通量によってＲＴを書き換えるルーティング）を採らなかった．したがって，出力ポート・アドレスをＲＴに格納する順が経路の優先順位であり，これを最適化することが課題となる (3.1 節参照)．

2.4 動作速度

当 SU の動作速度をクロック数で示す．

（1） 行先アドレスを知って出力ポートを決め，

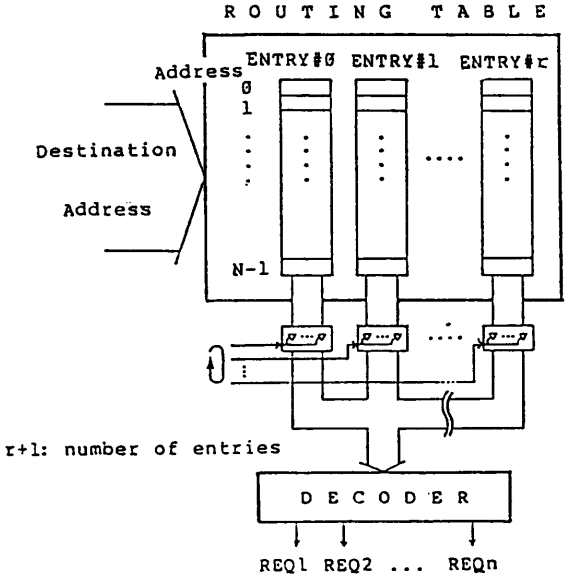

図 3 可変ルーティングの実現
Fig. 3 Variable routing control.

アービトレーションを行って経路を設定し，実際の転送を始めるまでの制御時間：3 クロック

（2） 1 回のルート変更に要する時間：2 クロック

（3） FIFO 間の 1 語転送時間：1 クロック

表 1 当 SU のハードウェア量
Table 1 Hardware complexity of the SU.

	1 Port	5 Ports
Gates in IP	840	4200
FIFO RAM	2.30 kbits	11.5 kbits
Routing Table	768 (r+1) bits	3.84 (r+1) kbits
Gates in OP	110	550

r+1: number of entries of RT

クロック周波数を決める要因として，(ⅰ)アービタの遅延，(ⅱ)アドレスコンバータ・メモリ (RT を格納するメモリ) のアクセス時間，(ⅲ)FIFO メモリのアクセス時間が挙げられる．今回の設計では，(ⅲ)が最も支配的であり，クロックの1周期は，アクセス時間の2倍以上に設定されねばならない．

2.5 ハードウェア量

表 1 に当 SU のハードウェア量を示す．SU は5入力5出力*，バス幅を 9 bit (8 bit がデータ，1 bit がパケットの終了信号) とした．

当 SU では，総ゲート数が 4,750，総入出力線数が 114，内部バスの総配線数が 80，内部バスと出力ポートの間の総接続線数が 300 である．

3. 当 SU を用いた相互結合網

3.1 網の構成法

当 SU の特徴は，マルチパスの結合網に有効であることと，汎用性・拡張性が高いことである．多ルートの多段結合網[8)~11)]や，CCC 網[12)] などへの適用が考えられる．**図 4**，**図 5** に，冗長構成のオメガ網と格子型網に当 SU を適用した例を示す．

マルチパスの網では，適切な経路選択を行って転送効率を高く保つことが課題となる．当 SU では出力ポート・アドレスを RT に登録する順序を決める際，この点について配慮せねばならない．具体的には，

(1) 網全体のトラヒックのバランス
(2) SU 内の閉塞
(3) 転送時の中継回数
(4) 以後の経路数
(5) ストアアンドフォワード・デッドロック[18)]

などを考慮する必要がある．

* SU を入出力5ポートずつの構成にしたのは，
①4×4 のスイッチとして多段結合網の構成単位とし，1ポートは故障時のための予備とする．②格子型網などでは，隣接の4ノードと自ノードのプロセッサのために計5ポートが必要となる．③LSI 化 (4.3 節参照) の際，ピン数の上からも妥当である．などの理由による．

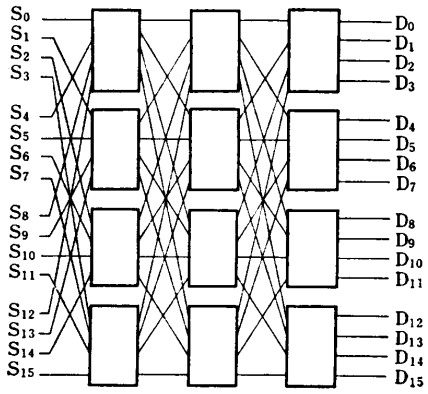

図 4 冗長構成のオメガ網 (4×4 SU)
Fig. 4 Omega network with an extra stage (4×4 SU).

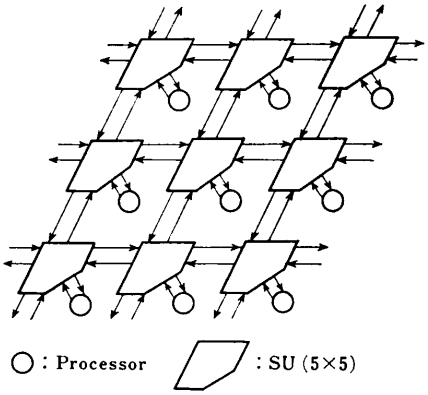

図 5 格子型網
Fig. 5 Mesh-connected network.

3.2 シミュレータ

SU の動作確認，当 SU を用いた相互結合網の転送性能の評価，RT 構成法の具体的な検討などを目的とし，機能レベルシミュレータを作成した．シミュレータは，動作速度の点から FORTRAN 77 で書かれている．

相互結合網の転送性能の解析・シミュレーション評価に関する研究は多い．たとえば，デルタ網について Patel[5)] (回線交換)，Dias[4)] (蓄積交換) らの解析が知られており，また本論文に近いものとしては，Chin[11)] らの報告がある．ただし，相互結合網では計算機網などと異なり，転送単位が小さく，低い転送遅延が要求されるため，ルーティングなどの制御時間が重要なパラメータとなる．現状では，この点を考慮した評価が十分になされているとはいえない．

ここで述べるシミュレーションでは，当 SU のクロックを単位として，次の二つの値を求めた．

(1) 正規化スループット：1クロックの間に網の

出力ポートから出る語数の平均値をポートあたりで計算したもの．なお，網が定常状態を保つ正規化スループットの最大値を，限界スループットと呼ぶ．

（2） 遅延：パケット転送遅延の平均値．

シミュレーションのパラメータとしては，①網の形状，②網の大きさ（N），③データ生成の分布と生成率（r），④平均パケット長（p），⑤FIFO 長（l），⑥行先の決定方式，⑦RT の構成法などが挙げられる．今回は，①ガンマ網[16]，③ポアソン分布，⑥ランダムとしたものに関して報告する．なお，オメガ網とその冗長構成（図4）に関する評価も行ったが，これについては文献19）を参照されたい．

3.3 ガンマ網のシミュレーション評価

当 SU をガンマ網（**図6**）[16]に適用した場合のシミュレーション評価について述べる．ガンマ網は，ほとんどすべての段で可変ルーティングの可能な（$\log_2 N+1$）段の網である．

ここでは，以下の3種のルーティング方式を比較した．

（1） 三つの出力ポートのうち，下の二つのみを用いる固定式ルーティング

（2） 三つの出力ポートを均等に用いる固定式ルーティング

（3） 可変ルーティング

シミュレーション結果を**図7**（正規化スループット）**図8**（遅延）に示す．

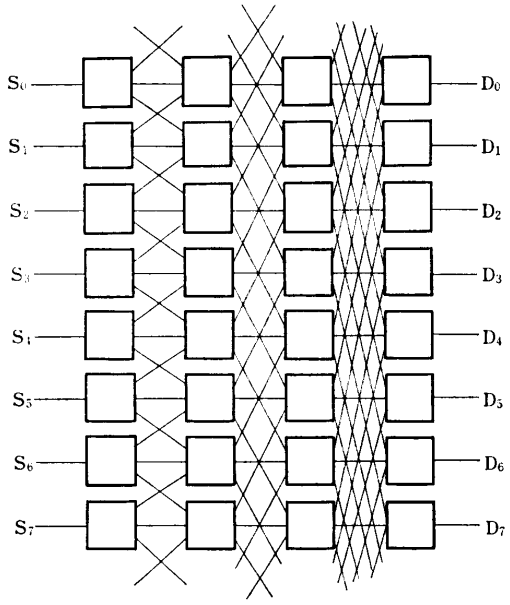

図6 ガンマ網（N=8）
Fig. 6 Gamma network (N=8).

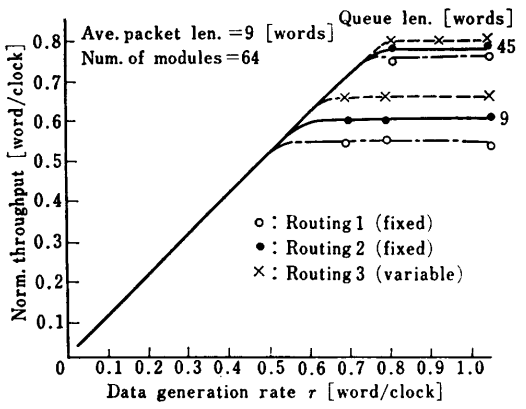

図7 ガンマ網の正規化スループット
Fig. 7 Normalized throughput of gamma network.

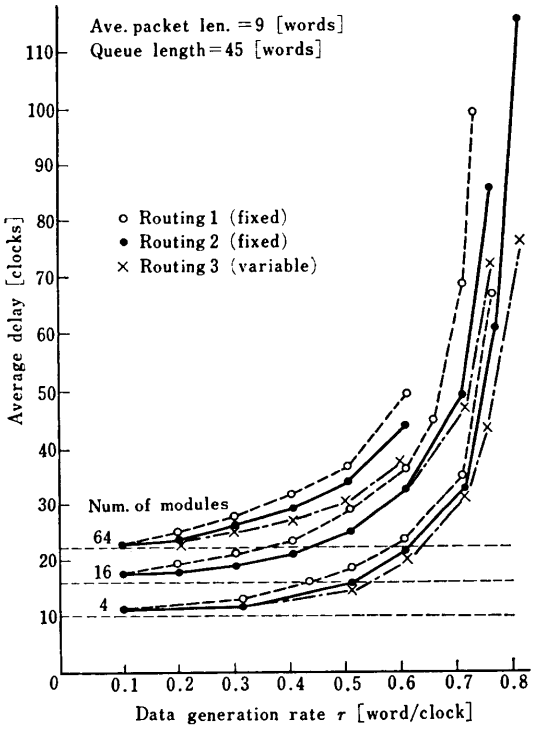

図8 ガンマ網における遅延
Fig. 8 Delay on gamma network.

図より，可変ルーティングの効果が観察される．とくに，データ生成率の高いときの遅延の差が顕著であり，N=64，r=0.6 のとき，（1）と比較して（3）は約23% の転送遅延の短縮が見られる．

4. 考　　察

4.1 実効的な転送性能

相互結合網の転送速度は，①実際にデータが網上を流れる速度，②閉塞によるオーバヘッド，③制御によ

るオーバヘッドの3点から決められる.

前章で述べたシミュレーション結果から,以上3点を含めたガンマ網の実効的な転送性能を示す.なお,〔 〕内の値は,10 MHz のクロック (FIFO のアクセス時間 50 ns) を仮定した場合の数値である.

4.1.1 限界スループット

$N=64$ のガンマ網 (可変ルーティング) では,前章の条件下で約 0.8 語/クロック〔8 MB/s〕の限界スループットを示す.これは,閉塞率 0 の場合の約 90% の性能である.また,網が大きくなったときの限界スループットの低下は小さく,$N=1,024$ の網でも 0.8 語/クロックに近い値が予想される[19].

4.1.2 転送時間

4.1.1 項のガンマ網で,$r=0.2$ のときの遅延は約 23 クロック [2.3 μs],パケットの平均転送時間は約 32 クロック [3.2 μs] である.r が 0.6〜0.7 を超えたところで,転送遅延は急激に増加する.

4.2 網の信頼性

相互結合網の対故障設計には,二つの方法が考えられる.

(1) SU 内の論理回路の多重化

(2) 網の多ルート化による障害部分の迂回

本論文で述べた可変ルーティングは,(2)の観点から有効であり,冗長構成のオメガ網 (図4) やガンマ網 (図6) では,障害をもつSUや結線部を迂回するルートを設定していくことで,網の信頼性を向上させることができる.

Adams[9),10)] らは,$\log N+1$ 段に拡張した間接キューブ網の信頼性を確率論的に検討している.今後,故障発生時の信頼性低下に関する一般論とともに,フォールバック状態での網の性能解析が必要である.これに関しては,上述の (1) によるSUの改良とともに,別途報告する予定である.

4.3 LSI 化の検討

2.5 節で,TTL によって構成される当 SU のハードウェア量を示した.実際に1チップの LSI として実装する場合,次の2点が問題となる.

(1) 配線数が多く,線およびドライバの占める面積が大きく必要になる[17].また,クロック周期は,線の遅延で決まることが考えられる.

(2) アドレスコンバータ・メモリの容量が大きい.

このうち,(2)に関しては,各入力ポートでアドレスコンバータ・メモリを共有するなどの対処が考えられる.

ピン数・ハードウェア量の2点から,本論文で述べた5入力5出力の SU は,LSI 化に適した規模といえる.今後,データのスキューや線の遅延の見積りが課題である.

5. むすび

可変ルーティング機能をもつ汎用 SU の設計を行い,これを用いた相互結合網の転送性能の評価をシミュレーションによって示した.今後の課題として,4章に述べたものの他に,ストアアンドフォワード・デッドロックへの対処法[18]の検討,格子型網などの静的[1]な結合網への適用と転送シミュレーション,非同期方式の検討,放送・パケット整形などの機能の付加,適応型ルーティングの検討などがあり,さらに当 SU の LSI による試作,大規模な網のシミュレーション評価などが挙げられる.

参考文献

1) Feng, T.: A Survey of Interconnection Networks, *IEEE Comput.*, Vol. 14, No. 12, pp. 12-27 (1981).
2) Lawrie, D. H.: Access and Alignment of Data in an Array Processor, *IEEE Trans. Comput.*, Vol. C-24, No. 12, pp. 1145-1155 (1975).
3) Pease, M. C., Ⅲ: The Indirect Binary *n*-Cube Microprocessor Array, *IEEE Trans. Comput.*, Vol. C-26, No. 5, pp. 458-473 (1977).
4) Dias, D. M. and Jump, J. R.: Analysis and Simulation of Buffered Delta Networks, *IEEE Trans. Comput.*, Vol. C-30, No. 4, pp. 273-282 (1981).
5) Patel, J. H.: Performance of Processor-Memory Interconnection for Multiprocessors, *IEEE Trans. Comput.*, Vol. C-30, No. 10, pp. 771-780 (1981).
6) Benes, V.: *Mathematical Theory of Connecting Networks*, Academic Press, New York (1965).
7) Batcher, K. E.: Sorting Networks and Their Applications, Proc. Spring Joint Comput. Conf., pp. 307-314 (1968).
8) Gajski, D., Kuck, D., Lawrie, D. and Sameh, A.: Cedar—A Large Scale Multiprocessor, Proc. of the 1983 Int'l Conf. on Parallel Processing, pp. 524-529 (1983).
9) Adams, G. B., Ⅲ and Siegel, H. J.: The Extra Stage Cube: A Fault-Tolerant Interconnection Network for Supersystems, *IEEE Trans. Comput.*, Vol. C-31, No. 5, pp. 443-454 (1982).

10) Adams, G. B., Ⅲ and Siegel, H. J.: Modifications to Improve the Fault Tolerance of the Extra Stage Cube Interconnection Network, Proc. of the 1984 Int'l Conf. on Parallel Processing, pp. 169-173 (1984).
11) Chin, C. Y. and Hwang, K.: Connection Principles for Multipath Packet Switching Networks, The 11th Ann. Symp. on Comput. Arch., pp. 99-108 (1984).
12) Preparata, F. P. and Vuillemin, J.: The Cube Connected Cycles: A Versatile Network for Parallel Computation, *Comm. ACM*, Vol. 24, No. 5, pp. 300-309 (1981).
13) 成瀬, 雨宮: 科学技術計算向きデータフロー計算機に用いる相互結合ネットワークの性能評価, 信学技報, EC82-36 (1982).
14) 白川, 影山, 阿部, 星野: 並列計算機 PAX-128, 信学論 (D), Vol. 67-D, No. 8, pp. 853-860 (1984).
15) Maekawa, M.: Optimal Processor Interconnection Topologies, The 8th Ann. Symp. on Comput. Arch., pp. 171-186 (1981).
16) Parker, D. S. and Raghavenda, C. S.: The Gamma Network: A Multiprocessor Interconnection Network with Redundant Paths, The 9th Ann. Symp. on Comput. Arch., pp. 73-80 (1982).
17) Franklin, M. A.: VLSI Performance Comparison of Banyan and Crossbar Communications Networks, *IEEE Trans. Comput.*, Vol. C-30, No. 4, pp. 283-291 (1981).
18) 南, 服部, 坂井, 田中, 元岡: プロセッサ間結合網に於けるスイッチング・ユニットの試作と評価, 第 27 回情処全大, 5N-1 (1983).
19) 坂井, 計, 田中, 元岡: 汎用スイッチング・ユニットを用いた高並列計算機の相互結合網, 信学技報, EC84-18 (1984).

(昭和 59 年 10 月 15 日受付)
(昭和 60 年 1 月 17 日採録)