計算機アーキテクチャ 83−38
（１９９０．７．１８）

並列計算機ＣＡＰ－ＩＩの ルーティング・コントローラ

堀江 健志　池坂 守夫　石畑 宏明

㈱ 富士通研究所

本論文では，高並列計算機ＣＡＰ－ＩＩの相互結合網について述べる．ＣＡＰ－ＩＩは，数値計算，映像生成の高速実行を目的とした分散メモリ型の並列計算機である．この目的に適した相互結合網のルーティング方式として，ワームホールルーティングに構造化バッファプールを取り入れた方式を提案し，トポロジとして二次元トーラス構造を選択した．本論文では，性能・ハードウェアの規模から判断し，本相互結合網の有効性を示す．次に，相互結合網を構成するＬＳＩ（ルーティング・コントローラ）について述べる．

Routing Controller of Cellular Array Processor CAP-II

Takeshi Horie, Morio Ikesaka, and Hiroyuki Ishihata

FUJITSU LABORATORIES LTD.

This paper presents interconnection networks of Cellular Array Processor CAP-II. CAP-II is the parallel computer for high-speed numerical calculation and image generation. For this purpose we proposed a routing technique which incorporated the structured buffer pool algorithm in wormhole routing and selected two-dimensional torus as a network topology. Also the chip designed for this network is outlined.

1. はじめに

我々は，数値計算と映像生成の高速実行を目的とした，高並列計算機CAP-Ⅱを開発している[1]．CAP-Ⅱは，64～1024台のプロセッサ（セル）がネットワークで結合された分散メモリ型の並列計算機である．本論文では，CAP-Ⅱのセル間結合網（ネットワーク）について述べる．

高並列計算機におけるネットワークの基本的な機能は，セル間のデータ通信を，セルの計算性能を十分に引き出すだけの高いスループットと小さい転送遅延時間で実現することである．また，ハードウェアとして実現可能な規模で最適なネットワークを設計する必要がある．

我々は，ワームホールルーティングに構造化バッファプールのアルゴリズムを取り入れたルーティング方式を開発することにより，高いスループットと小さい遅延時間でセル間通信を行うネットワークを実現した．また，ハードウェアの規模・拡張性を考慮し，ネットワークのトポロジとして，二次元トーラスを選択した．

ここでは，まず，第2章で，CAP-Ⅱのネットワークに必要とされる条件に関して検討を行う．第3章では，この条件をみたすようなネットワークのルーティング方式を提案する．第4章では，シミュレーションにより，本ルーティング方式の有効性を検証するとともに，次元の異なるトーラスアレイ型ネットワークの比較を行い，二次元トーラスネットワークの優位性を示す．第5章では，本ネットワークを構築するLSIを開発したので，LSIの構造・諸元を紹介するとともに，LSIの持つ通信範囲を指定した放送機能について述べる．

2. ネットワークの設計方針

CAP-Ⅱのネットワークの設計にあたり，数値計算とグラフィックスへの応用を考慮に入れ，以下の4項目に留意した．

(1)自動ルーティング　任意セル間通信における中継処理を各セルのＣＰＵの介在なく，ハードウェアで高速に行う．もちろん，デッドロックが発生しないルーティング方式でなければならない．

(2)小レイテンシ　メッセージの送信から受信までの遅延時間（レイテンシ）を小さくする．数値計算の応用では，あるセルの計算結果を他のセルが必要とするとき，その遅延時間が性能に大きく影響する．また，小レイテンシを実現するためには，ネットワークへのメッセージの送受信の時間（ユーザプログラムの送信コマンド発行から，ユーザプログラムの受信コマンドの発行により，実際にメッセージを受信するまでの時間）もネットワークそのもの以上に大きな要因となる．これに関しては，文献2で述べる．

(3)高スループット　ネットワーク全体で通信が行われているとき，ネットワークの一部分の輻輳によりネットワーク全体の性能が低下することがなく，多くの通信路で通信が行われていなければならない．グラフィックスでは，比較的細かいメッセージが不規則にネットワークに送り出されるので，高スループットを必要とする[3]．

(4)規模・拡張性　1セルに物理的に接続されるワイヤの数，1ボードに複数のセルをのせたときのボードに接続されるワイヤ数等のハードウェア規模を小さくし，６４～１０００台規模のセル構成がとれる．実際のシステムを構成する上では，最も基本的なことがらである．

3. ルーティング方式の原理

2.の設計方針から，レイテンシが小さいことを特徴とするワームホールルーティングに構造化バッファプールのアルゴリズムを取り入れることにより，トーラス状のネットワークを構築してもデッドロックを起こさない，小レイテンシかつ高スループットを実現するルーティング方式を考案した．ここでは，本方式について述べる．

3.1 ワームホールルーティング

ワームホールルーティングでは，メッセージのヘッダが入力チャネルから出力チャネルへ中継ルートをつくりながら，メッセージが送り出されていく．ストア・アンド・フォワード・ルーティングでは，中継ノードがメッセージ全体をストアするのに対し，ワームホールルーティングでは，フリットと呼ぶ数バイト（ビット）のデータのみが中継セルにストアされる．あるセルがメッセージのヘッダを受信すると，中継ルートのチャネルを選択し，フリットをそのチャネルへ転送する．後続のフリットはヘッダのフリットが選択したルートと同じルートに転送されていく．メッセージの先頭は受信セルに到着しているにもかかわらず，メッセージの後部はまだ送信されずにネットワーク上にないということもありえる．

ワームホールルーティングの特徴は，レイテンシが小さい点にある．一方，ワームホールルーティングでは，1つのメッセージが転送されている間，そのメッセージが使用しているチャネルをブロックするので，①デッドロックの発生，②スループットの低下，を起こす可能性がある．①に対しては，バーチャルチャネル[4]と呼ばれる方式により，デッドロックの発生しないルーティング方式が開発されているが，②のスループット性能に対しては，解決されていない．

3.2 提案するルーティング方式

本論文で提案するルーティング方式は，ワームホールルーティングに構造化バッファプール[5]のアルゴリズムを取り入れることにより，デッドロックの発生とスループットの低下を回避している．これは，1つのメッセージが転送されている間，チャネルをブロックすることがないからである．以下，図1の例を用いて，この方式について述べる．

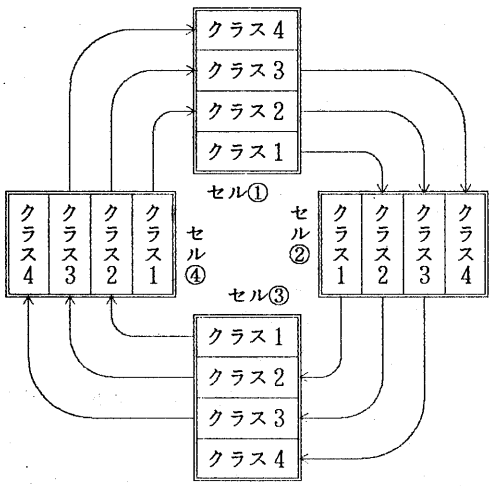

図1　ルーティング方式のバッファ使用方法

各セルは，フリットをストアするためのバッファを用意する．このバッファは，「セル間の最大距離＋1」個のフリットをストアできる大きさである．

4つのセルが単方向のチャネルで接続されているとすると，各セルは，4つのフリットをストアできる大きさのバッファを持つ．

バッファの使用方法は，セル①からセル②へメッセージを送信するときは，「セル①のクラス1⇨セル②のクラス2」，セル①からセル③へメッセージを送信するときは，「セル①のクラス1⇨セル②のクラス2⇨セル③のクラス3」，セル①からセル④へメッセージを送信するときは，「セル①のクラス1⇨セル②のクラス2⇨セル③のクラス3⇨セル④のクラス4」である．セル②，セル③，セル④のときも同じように，「クラス1⇨クラス2⇨クラス3⇨クラス4」，を使用する．

各セルのクラスにフリットがストアされると，次のセルにそのフリットを転送する．転送するとき，データのクラスとデータそのものを転送する．どのクラスのフリットを転送してもよく，フリット転送ごとに転送するクラスをかえることができる．

本アルゴリズムは，まず，ワームホールルーティングにおけるデッドロックを回避している．これは，どのセルからもクラス1⇨クラス2⇨クラス3⇨クラス4という経路が存在し，ループを形成しないからである．

次に，本アルゴリズムにおいて，全てのセルが，同時に，右回りに，①⇨③，②⇨④，③⇨①，④⇨②，へ転送するときを考えると，すべてのチャネルを用いて転送が行われ，スループットの低下を引き起こすことはないことがわかる．本方式は，チャネルが両方向の場合，あるいは，セルの接続チャネル数が多い場合にも適用することができる．

4. 評価

ここでは，3.2で述べたルーティング方式を評価する．本ルーティング方式は基本的にワームホールルーティングに基づいており，小レイテンシな通信を実現することができる．そこで，評価として，ネットワーク上に多くのメッセージが転送されている状態での通信の性能，すなわち，スループット性能を評価する．

まず，本ルーティング方式の特性を調べ，ネットワークの負荷が大きいときでもスループット性能が低下しないことを示す．次に，ワームホールルーティングを使った一ルーティング方式であるバーチャルチャネルとの性能を比較し，本ルーティング方式がスループット性能向上に有効であることを示す．

本ルーティング方式は，アレイ型ネットワークに適用することができる．ネットワークのトポロジを検討するため，次元の異なるアレイ型ネットワークのスループット性能を

求め，その性能を比較する．なお，対象とするネットワークアレイの端どううしはトーラス状に接続されているものとする．

4.1 シミュレーションの条件

性能評価を目的としたソフトウェアによる「ネットワーク・シミュレータ」を開発した．ここでのシミュレーションの条件を以下に示す．

●ルーティング方式　n次元のネットワークのとき，まず一次元内のセルでルーティングを行い，次に，二次元内のセルでルーティング，最後にn次元内のセルでルーティングというようにルーティングしていく．次元内のルーティングは，3.で述べた方式により，静的な最短ルーティングを行う．

●チャネル　1チャネルを双方向に使用する．1データ（1ワード）が1単位時間に転送される．

●受信　データがあるセルに到着すると，直ちに，ネットワークから除かれる（受信される）．

●送信　セルからデータを送信するかどうかを，乱数で決定している．送信する割合を，ここでは，「ネットワークへの負荷」と定義する．

●通信の形態　不規則で局所性のない通信を基本とする．

●メッセージ　8ワードの固定長とする．また，メッセージは，

宛先ヘッダ＋データ＋・・・＋データ

とういう形式をしている．

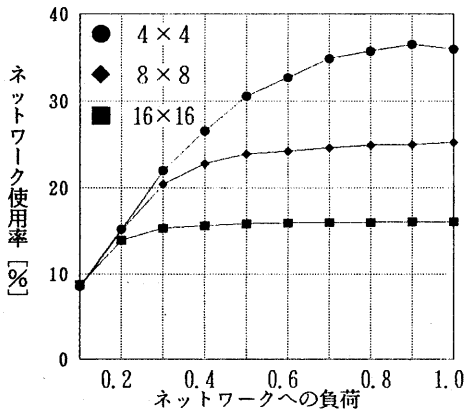

図2　ネットワークへの負荷とネットワーク使用率

以上の条件でシミュレーションを行い，その結果として，ネットワーク使用率をスループット性能として得る．ここでは，ネットワーク使用率を，

単位時間内（送信しないときも含む）にネットワークへデータを送信することができた確率

と定義する．ネットワークが輻輳状態になるとネットワークにデータを送信できなくなり，使用率は低下する．ネットワーク使用率は，全セルの平均の値とする．いつも送信し，いつも送信できたとすると，ネットワーク使用率は１００%となる．

4.2 ルーティング方式の特性

本ルーティングを二次元トーラスネットワークへ適用したときの，ネットワークへの負荷とネットワーク使用率との関係を図2に示す．図からネットワークへの負荷が大きくなってもネットワーク使用率が低下することがないのがわかる．なお，一次元，三次元，四次元等の場合でも，同じ特性を示した．また，メッセージの長さが8以上のときも，ネットワーク使用率が全体的に下がるが，同じ特性を示した．実際の転送では，ネットワークへの負荷は，0.1～0.4の範囲になると考えられるので，ネットワークの輻輳状態によりメッセージが送信できなくなる場合は，あまり多くなく，ネットワークが有効に利用できるものと思われる．

4.3 バーチャルチャネルとの比較

バーチャルチャネルは，ワームホールルーティングにおけるデッドロック問題を解決するルーティング方式である．提案したルーティング方式がチャネルを仮想的に多重化しているのに対し，バーチャルチャネルは，チャネルを仮想的に二重化している．ここでは，このバーチャルチャネルのアルゴリズムをネットワーク・シミュレータに組み込み，本方式との比較を行った．

図3に，セル台数とネットワーク使用率との関係を示す．3.2で述べたように，本ルーティング方式を用いることにより，高いスループット性能を得られることがわかる．これは，バーチャルチャネルが，チャネルをブロックする確率が高く，実際の通信チャンルの稼働率が低くなるためである．

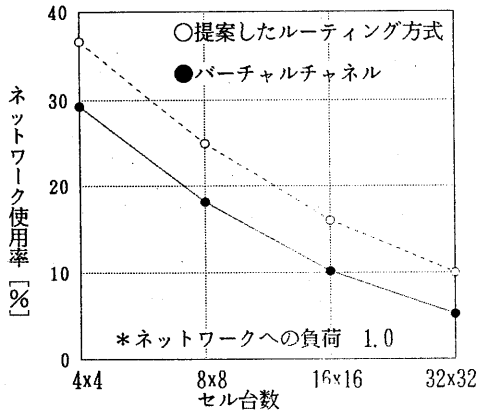

図3　バーチャルチャネルとの比較

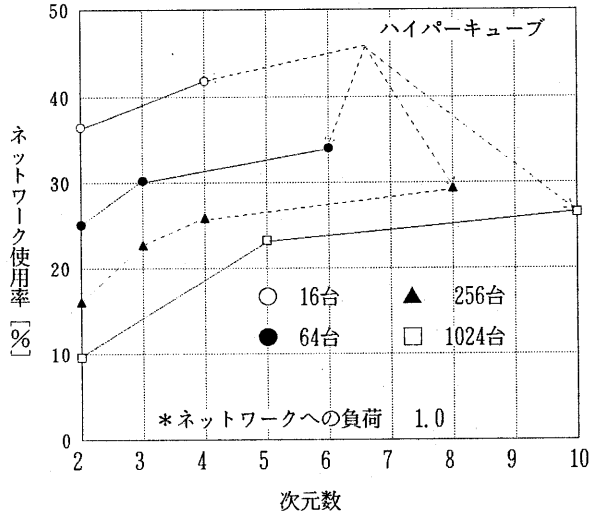

図4　次元数とネットワーク使用率

表1　セル間距離

セル台数	二次元	三次元	ハイパーキューブ
64	8 (4)	6 (6)	6 (6)
256	16 (4)	10 (6)	8 (8)
1024	32 (4)	16 (6)	10 (10)

＊ () 内は，1セルに接続されるチャネル数

4.4　トポロジの評価

本ルーティング方式は，アレイ型ネットワークへ適用することができる．ここでは，次元数の異なるネットワークのスループット性能を評価し，ハードウェアの規模を考慮にいれた最適なトポロジを決める．

図4に，台数一定のもとで，次元を変えたときのネットワーク使用率を示す．また，表1にセル台数と平均セル間距離との関係を示す．この表は，ネットワークの静的な特性を示す．なお，64台構成の6次元ネットワーク，256台構成の8次元ネットワーク等は，バイナリハイパーキューブネットワークになり，このときは，本ルーティングアルゴリズムを用いなくても，デッドロックは発生しない．

図から例えば１０２４台構成のときの二次元アレイ構成とハイパーキューブとの性能比は，1対2.5であり，これは，平均セル間距離の比5対16よりも小さい値になっているのがわかる．

1チャネルのビット幅が同じであれば，ハイパーキューブが二次元あるいは三次元アレイよりも優れているのは明らかである．しかし，ハイパーキューブはハードウェア量が多く，結果として1チャネルのビット幅を小さくする必要がある．例えば，1台のセルに接続されるワイヤ数を等しいとすると，１０２４台構成のとき二次元アレイとハイパーキューブの1チャネルのビット幅の比は，10対4になり，提案したルーティング方式を用いれば，スループットにおいてほぼ同程度の性能が得られることになる．なお，レイテンシと拡張性においては，二次元アレイの方が優れているのは明らかである．

5.　ルーティングＬＳＩ

ここでは，CAP-Ⅱのネットワークを構成するLSI について述べる．本LSI は，3.2で述べたルーティング方式を用いた二次元トーラスネットワークを実現するチップである．さらに，任意セル間の1対1の通信だけではなく，放送通信の機能も盛り込んでいる．

5.1　LSI の構成

本LSI は，各プロセッサに1個接続され，プロセッサ間通信を実現する．チップのピン数を削減するために，通信チャネルは，双方向で使用する．また，通信チャネルには，パリティビットを付け，エラーを検出するようにしている．

本LSI のブロック図を図5に示す．

●ルーティングコントローラ　各クラスの制御，入力データの送信方向判断等を行う．X方向とY方向のコントローラは，同じ制御回路である．

●バッファメモリ　バッファメモリは2ポートのメモリで構成されており，各クラスに対して，最大2データを格納する．実際の転送では，転送が輻輳してこない限り，バッ

ファを使用しないで転送が行われ，レイテンシを小さくしている．1データを中継する遅延時間は，1セルあたり，160 nsである．

●送受信FIFO　ネットワークとローカルバスの転送レートの差によるデータ転送の低下を防ぐため，送信，受信側とも8ワードのFIFOを持っている．

●ローカルバスコントローラ　ローカルバスに対するコントロールを行う．DMAへの要求信号，CPUへの割り込み信号等を発生する．また，メッセージヘッダを自動的に剝ぐ（ストリップする）機能も持つ．

また，表2に本LSI の諸元を示す．

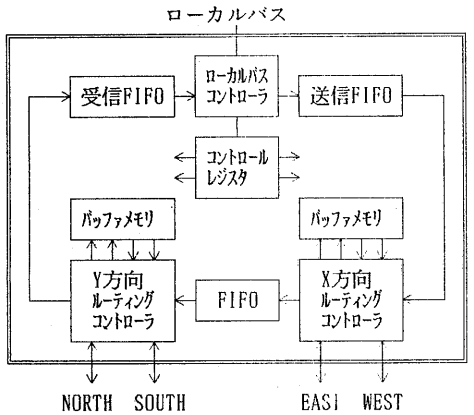

図5　ルーティングLSI のブロック図

表2　ルーティングLSI の諸元

トポロジ	二次元トーラスネットワーク
最大構成	32×32の1024台構成
転送チャネル	データ幅17ビット パリティ3ビット
転送速度	25MB/S　（1チャネル）
転送時間	160 ＋ 160×セル間距離＋ 160×ワード数　(ns)
パケット	任意長ワード（32ビット/1ワード）
回路規模	49,000 ゲート
テクノロジ	1.2 μm CMOSゲートアレイ

5.2　放送機能

ルーティングLSI は，1対1のセル間通信のみならず，1対多の放送通信機能も持っている．放送通信は，例えばLU分解におけるピボット情報の通信など，数値計算では重要な機能と考えられる．図6に放送通信の送信形態を示す．放送通信では，送信するセルの範囲を指定することができる．図では，X方向に1，Y方向に2の範囲に放送している．放送の範囲として，一方向（例えばXの－方向）だけの放送も行うことができる．放送の範囲が同一方向で重ならない限り，複数のセルが同時に放送通信を行うことができる．なお，放送メッセージは，通常の1対1のメッセージと混在することができる．

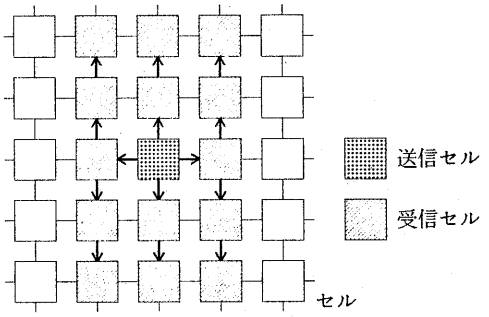

図6　放送通信

6. おわりに

高並列計算機CAP-Ⅱのネットワークについて述べた．まず，ネットワークのルーティング方式として，ワームホールルーティングに構造化バッファプールのアルゴリズムを取り入れた方式を提案し，その有効性を示すとともに，トポロジとして二次元トーラスネットワークの優れていることを示した．

また，本方式を実現するネットワークLSI を開発した．本LSI は，任意セル間の通信とともに放送通信を行うことが可能である．

今後，CAP-Ⅱシステムを完成させ，実機上でネットワークの評価を行うとともに，数値計算に対するシステム全体の性能評価を行う予定である．

参考文献

1）石畑他：高並列計算機CAP-Ⅱの構成とメモリシステム

2）清水他：高並列計算機CAP-Ⅱのメッセージコントローラ

3）佐藤他：高並列計算機CAP-Ⅱによる三次元グラフィックス

4）Dally, W.J.:A VLSI Architecture for Concurrent Data Structures, Kluwer, Hingham, MA, 1987.

5）Merlin, P.M.:Deadlock avoidance in store-and-forword networks-I, IEEE Trans. commun., COM-28