計算機アーキテクチャ　83－39
（１９９０．７．１８）

高並列計算機ＣＡＰ－Ⅱの ブロードキャスト・ネットワーク

加藤　定幸　　清水　俊幸　　堀江　健志　　石畑　宏明

㈱富士通研究所

本論文では，高並列計算機ＣＡＰ－Ⅱのブロードキャスト・ネットワークについて述べる．ブロードキャスト・ネットワークはホスト計算機とセル（プロセッサ）を接続するネットワークで，主にホストとセルのインターフェースに用いられる．ブロードキャスト・ネットワークはホスト―セル間で，２次元配列を高速に交換するスキャッタ，ギャザの2つの転送モードを持っている．本論文ではこれらのブロードキャストネットワークの機能と実現方法について述べる．

Broadcast-network of Highly Parallel Processor CAP-Ⅱ

Sadayuki Kato　Toshikyuki Shimizu　Takeshi Horie　Hiroaki Ishihata

FUJITSU LABORATORIES LTD.

1015, kamikodanaka, Nakahara-ku, Kawasaki 211, Japan

This paper presents a Broadcast-network of highly parallel processor CAP- Ⅱ. Broadcast-network is a network which connect between host and cells (processor), and it has "scatter" and "gather" transfar mode of Broadcast-network which excha ange 2-dimensional array between host and cells rapidly. This paper presents fun ction of Broadcast-network and its hardware desigen.

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，分散メモリ型の高並列計算機CAP-Ⅱを開発している．本論文では，CAP-Ⅱのホスト・セル（プロセッサ）間の通信路であるBraodcast network（Bネット）について述べる．

CAP-Ⅱは64～1024台のセルを持っているので，ホスト計算機はこれらのセルと効率よく通信を行うことが出来なくてはならない．ホスト・セルの通信にはホストからセルへのプログラム，データの転送，セルからホストへの計算結果の転送がある．ホストからの転送のうちプログラムや各セル共通のデータはブロードキャストによって高速に転送可能であるが，セル毎に異なるデータを送る場合あるいは，セルからホストに計算結果を吸い上げる場合は，転送するセルの選択あるいは，転送するデータの選択といったオーバヘッドがある．しかも，このオーバヘッドはセルの台数が増加するにつれ大きくなる．

この問題を解決するため，我々はB ネットの転送モードとして，従来からあるブロードキャストに加えて，セルに異なるデータを転送するスキャッタ（分配），セルのデータをホストに効率よく吸い上げるギャザ（収集）を設けることにした，

本論文では，まず，第２章でＢネットの機能，特にスキャッタ，ギャザについてのべる．第３章ではCAP-Ⅱのコマンドバスのハードウェアについて，スキャッタ・ギャザの実現法を中心に述べる．第４章では，スキャッタ，ギャザの両機能について，従来の転送方式との比較を行い，本方式の有効性について検討する．

2. Braodcast network （Ｂネット）の仕様

2.1 Ｂネットの機能

CAP-ⅡのＢネットはホスト計算機と全てのセルを接続した共通バスである．ＢネットはCAP-Ⅱの1000台のセルとホスト間で効率よく通信する事を目的としたネットワークである．並列計算機のセル（プロセッサ）間の通信については広く検討されているが，セルとホスト計算機の通信も以下の理由により高性能化が不可欠である．

ＣＡＰ－Ⅱにおいては，個々のセルの処理能力の向上および，セル台数の増加によって，システム全体で処理可能なデータの量が増大する．ホスト―セル間通信はこのデータを供給するために高いバンド幅が要求される．また，システムを構成するセルの数が増大するにつれて通信時のセル切替えのオーバヘッドが増大する．

また，ブロードキャストで全てのセルに同じデータを送るときはホストは１度のデータ転送でデータを送る事ができるが，ホストのデータを分割してセルに送る場合，あるいはセルの計算結果をホスト計算機に集める場合には，受信するセルあるいはデータを送りだすセルの切替えのオーバヘッドが入る．この切替えの回数はセルの台数に比例して増加し，１回の切替え時間もセルの増加によるハードウエア規模の増大にともなって長くなる傾向がある．そのため，長さがｎのデータをｐ台のセルに等分に分割して送るときの転送時間Ｔは以下のようになる．

$$T = Tc \times p + Tt \times n \quad \text{(1)}$$

Ｔｃ:セル切替え時間
Ｔｔ:1 ワードの転送時間

従ってセル台数の増加にともなって，実際のデータの転送時間よりもセルの切替えのオーバヘッドが大きく影響してくることになる．この問題を解決するためにＣＡＰ－ⅡのＢネットには通信切り換え時のオーバヘッドが少なくなるような転送方式が必要になってくる．

Ｂネットは論理的にはセルとホストを接続した均一な共通バスに見える．

セルとホストはＢネット上では対等であり，セルからセルへのブロードキャストも可能になっている．そこで，ホストとセルをまとめてノードと呼ぶことにする．Ｂネットには同時に複数のノードよりデータを送りだすことは出来ない．ブロードキャストの権利を獲得したノードをマスタ，その他をスレーブと呼ぶ．また，各ノードにはそれぞれＩＤがふられている．このＩＤは各セルで異なっているが複数のノードに共通のＩＤを与えることも可能である．共通のＩＤをもったノードを同じグループに属しているノードと呼んでいる．

表１に，Ｂネット上での転送モードについて説明する．マスタはデータを送りだす際に。１まとまりのデータ（パケットと呼ぶ）のまえにヘッダをつける．ヘッダには転送のモードを示すビットならびにこのデータを受信するグループのＩＤを示すビットが含まれている．

表１　Ｂネットの機能

ブロードキャスト	そのパケットを全てのスレーブが受信する．ホストから全てのセルに共通のデータを送ったりプログラムを送るのに使用する
グループキャスト	ヘッダのＩＤビットと各スレーブのＩＤを比較して等しいスレーブだけがパケットを受信する
スキャッタ	２次元配列を分割し，異なるセルに送る転送モード．
ギャザ	セル上の２次元配列をホスト上に集めて，１つの２次元配列を作る転送モード．
その他	①ブートストラップコードをセルに送る． ②セルにＩＤを付ける．

2.2 スキャタとギャザ

ＣＡＰ—Ⅱは主に数値計算を高速実行するために，作られた並列計算機であるが，このような用途では２次元配列のデータを扱うことが多い．また，この配列をセルに送る場合，すべてのセルに同じ配列を送る場合だけでなく，図１のように，配列を分割してセルに割り当てることも多い．

従来のネットワークでこの様な事を行う場合，まずデータをホスト計算機上でブロックに切り分けて，それぞれのブロックを送り先のセルを指定して転送していた．したがって，前述のようにデータ転送に要する時間には切りわけの時間と転送先のセルの切替え時間がオーバーヘッドとして加わっていた．しかし，Ｂネットでは受信側に送られてきたデータが自分が受信するブロックに入っているかどうかを判定するハードウエアをもうけブロックの切りわけを受信側で行っている．この判定の時間がデータの転送速度より早ければ，送信側（マスタ）はブロードキャストと同等の速度でデータを転送することができる．この機能をスキャタとよんでいる．

逆にセルの２次元配列をデータをホストに収集する転送方式としてギャザがある．このモードでは前述のブロードキャストやスキャタとは反対にスレーブからマスタに向かってデータが送られる．各セルは他のセルとホストの通信状況を監視しホスト上の配列のどの部分が転送されているか判定して，自分がデータを送りだすタイミングを決めている．同時に多数のセルがデータを送りだすためにセル間の順序の制御が必要になるが，この制御に要する時間はマスタを決定する調停時間と異なり，データ転送時間と等しいので，ブロードキャストと等しい転送速度でデータを収集することが出来る．

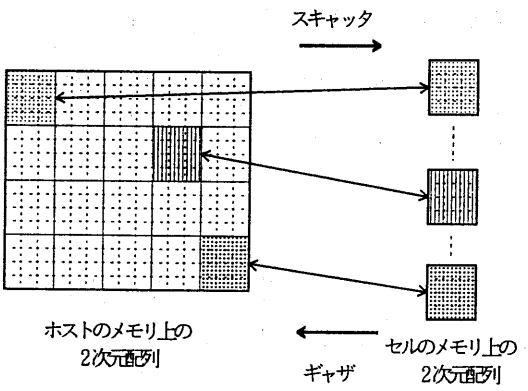

図１．２次元配列の交換

スキャタ，ギャザ時のブロックのパターンには図２に示すようにＸ方向Ｙ方向のストライプ状のパターンの組み合わせである．このパターンはそれぞれのスレーブでごとに設定し，そのパラメータはＸ，Ｙ方向それぞれの幅と間隔と原点からのオフセットの６種類である．このパターンはスレーブごとに独立しているので，１部のスレーブにはブロック状に他のスレーブにはストライプ状に配列を切り分けることも可能である．また，スキャタ時には隣あったセルにオーバラップしたブロックに切り分けることもできる．

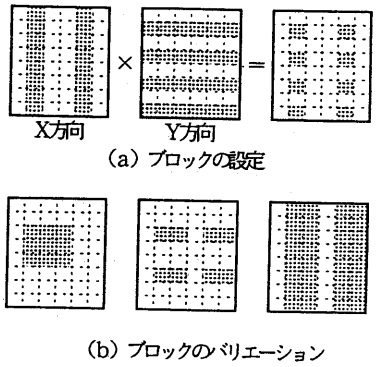

図２．配列の切りわけ例

3．Ｂネットのハードウエア

3.1 ネットワークの構造

Ｂネットは1024台までのセル構成に対応可能でなくてはならない．そこで我々は，Ｂネットの構成として図３に示すようなリング型のネットワークと階層型のネットワークを組み合わせた構成を取ることにした．バックプレーン間の接続はリング型のネットワークを用い配線を簡略化している．またバックプレーン内ではツリー型のネットワークを用いている．

リング型のバスを用いるとセル台数の増大に比例して伝搬遅延（レイテンシィ）が増加する欠点があるが，ツリー状のネットワークと組み合わせて伝搬遅延の増加を抑えている．コマンドバスではブロードキャスト，スキャッタ，ギャザなどまとまったデータが流れる事が多く，転送性能はスループットによって決まるので，リング型のバスを用いても性能はそれほと低下しない．

マスタとなるノードを決めるために．ノード間で，コマンドバスの転送権のアービトレーションを行っている．アービトレーションの回路もバスの構造と同様に階層構造となっており，調停に時間を要するが，前にも述べたようにコマンドバスは大きなデータの転送を主に考えているので，アービトレーションの高速化は特に考慮しなかった．

Ｂネットではホストあるいは任意のセルがマスタになることが出来るので，マスタの位置に応じてネットの構造を変化させる必要がある．具体的にはマスタのあるツリー型ネットへのリングの分岐点の入力側でリングネットを切断する．ツリー型の部分ではでは基本的にはリング

からセルへのブロードキャストになるようにバッファの向きを決めているが，マスタの存在するツリーではマスタからリングとの接続点に向かう経路上のバッファの転送方向をを反転してデータをリングへおくりだす．これによってＢネット全体はマスタをルートにしたツリー状のネットワークになる．このネットワークはマスタとスレーブの距離が不均一であり，また，MIMD型計算機であるので各スレーブが同期して動作している保証もない．そこで，データの受信が遅れたスレーブにあってもデータの取りこぼしがないようにハンドシェークを行っている．ただし全スレーブでハンドシェークを行うと1 ワードの転送に時間が掛かるので各バッファ内に２段のFIFOをもうけツリーの各枝毎にハンドシェークを行っている．したがって，ネットのサイズによらず転送速度は一定である．

ギャザの時はマスタに向かってスレーブからデータが送られるので，ツリー型ネットでは転送方向を全て反転し，リングの部分ではマスターのあるツリー型ネットへのリングの分岐点の出力側でリングネットを切断する．しかし，複数のバッファが同時にデータを出力することは出来ないので，あとで述べるようにバッファの出力制御を行っている．

また，ネットワーク上のバッファにデータが残っているままで，マスタを切り換えてしまうとデータが正しくスレーブに伝わらない．そこで，データが全てのスレーブに送られた事を検出するために，データ転送終了後，マスタはコマンドバスに転送の終了を示す特殊なデータを送りだす．このデータが全てのセルに行き渡った事をCAP-ⅡのＳネット（同期機構）で検出して，マスタを開放する．ギャザ時はマスタがあらかじめ決められた数のデータを受けたことで終了を判定する．

表２．Ｂネットのハードウエア諸元

トポロジ	リング＋ツリー
セル構成	64～1024台
伝播遅延	0.8 μs(64台) ～5.6 μs(1024台)
転送速度	50ＭＢ／秒　（32ビット幅）
専用ＬＳＩ	ＣＭＯＳ　1.2 μm ゲートアレー 33000 ゲート

3.2 専用ＬＳＩ

ＢネットとセルあるいはホストとのインターフェースにはBIF と呼ばれる専用のLSI をもちいる．BIF の内部構成は図４の様になっている．また，階層バスの各バッファもこのLSI をデータのバッファとして使用している．

●ヘッダデコーダ：ヘッダのデコードを行い，送られてきたデータが自分がこのパケットを受信するのかどうか，また，スキャッタで受信するのかどうかを判定する．また，グループキャストのためにホストから送られてきたIDを記憶する機能をもつ．

●送受信バッファ：Ｂネットとのデータの転送を行うバッファ，この部分はＢネットの各バッファでも使用できるようになっている．

●Ｂネットコントローラ：コマンドバスの獲得，および，スキャタ・ギャザ時のデータの転送タイミングの制御を行う．スキャタ，ギャザのパラメータはセルのCPU によって設定される．スキャタ時は送られてきたデータが自分が受信するデータで無いときは受信バッファをクリアし，受信するときはFIFOに書き込む．ギャザ時はデータを送るときはFIFOのデータを送信バッファにおくり，それ以外の期間はダミーを送る．

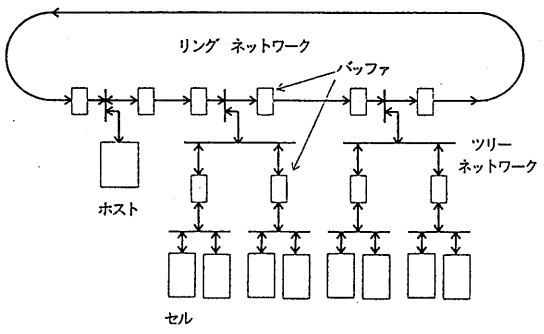

図３　ブロードネットの構造

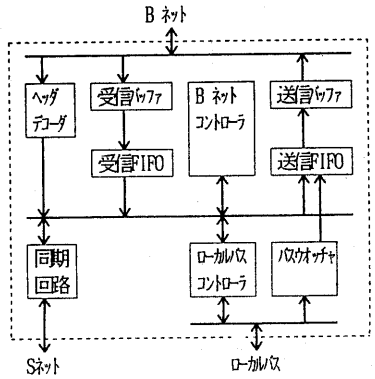

図４　ブローロネット専用ＬＳＩ

●送受信FIFO：Ｂネットとセルのローカルバスの転送速度の差を吸収する8 ワードのFIFO
●ローカルバスコントローラ：セルのローカルバスとBIF 内部のレジスタあるは，FIFOとのデータ転送の制御をおこなう．
●同期回路：ＣＡＰ—ⅡのＳネットの制御を行う．
●バスウォッチャ：セルのローカルバスのパフォーマンス，エラーの監視と記録を行う．

3.3 ギャザの実現

スキャッタについては４．１のブロードキャスト機能と４．２のＢＩＦ内の選択回路によって実現できる．しかし，ギャザではデータを出力するスレーブを高速で切り換える必要がある，また，ＭＩＭＤ型の計算機では各セルの動作にばらつきがあり，同じタイミングでデータを送りだす保証がないためにあるセルが遅れた時でも，マスタに届くデータの順序が正しい事を保証する必要がある．その為に全スレーブのＢＩＦを同期させる必要があるが１データ毎に全セルで同期を取っていたのではデータ転送速度が同期の速度で制限されてしまう．そこで，各スレーブの同期にはデータの転送プロトコルを同期に利用し，各合流点で独立に同期をとることで，ブロード時の転送速度でギャザを出来るようにした．

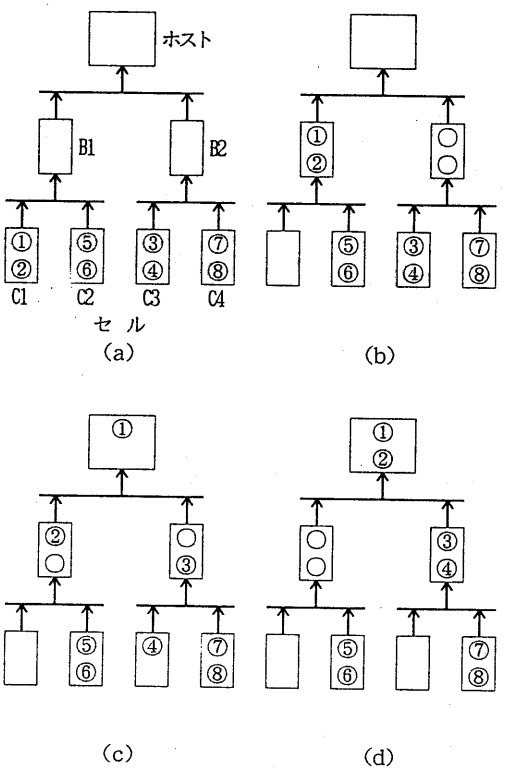

図５　ギャザ時の同期

各セルはパラメータに従ってデータを送りだすが，各スレーブが送りだすデータはホスト上の配列のイメージの中に自分が送りだすデータを埋め込んだデータの列で，自分が送りださない部分のデータはダミーのデータを埋めておく．そのために，データ線にはそのデータがダミーである事を表すフラグをもうける．スレーブから送られてきたデータは各合流点でマージされて送られていく，この時ダミーのデータ同志はダミーのデータを，ダミーのデータでないデータが来たときはそのデータを送るようにする．また，各分岐点で複数のスレーブあるいはバッファの出力を制御は，自分がダミーでないデータを送りだすときに出力をオンにすることによって行う．．またフラグのビットは他のデータとは別にオアをとって上位のバッファにわたしている．

以上の動作を図４で説明する．この例では４台のセルがスレーブとなってマスタであるホストにデータをおくる，各セルが送るデータは（ａ）のセルの箱に入っているデータで，それぞれ番号の順番でホストに転送される．初めにC1以外のセルはダミーのデータを送りだす（ｂ）は，バッファB1,B2 に 2つのデータが送られた状態である．B2はC3,C4 ともにダミーを送りだしているのでダミーのデータ（ 図中の○）が入っているが，B1にはC1のデータが入っている．このあとB1のデータがホストに送られると，B2のなかのダミーも消えるのでその開いた分にC3からのデータが送られる（ｃ)(ｄ)．また，データを送り終わったC1もダミーのデータを送り続ける．

４．性能予測

Ｂネットのスキャッタ，ギャザの機能について，従来方式のブロードキャスト（グループキャスト）で同様の事を行った時との性能比較を行った．

ブロードキャストで同様の処理を行う場合は，ブロックの切りだしや統合をホスト計算機で行わなければならない．しかし，ＣＡＰ－ⅡではＭＳＣ（ＤＭＡＣ）にメモリ上のとびとびの領域のデータを転送する機能があり，データの送受信にこの機能を利用するとホスト上でのブロックの切りだしを行わなくてもデータを転送できる．ここでは，ブロックの切りだしをＭＳＣで転送時に行った場合（ブロードキャスト①）と，ブロックの切りだしをホスト上でソフトで行った場合（ブロードキャスト②）について検討を行う．

また，転送時の性能にはＢネットよりもセルのローカルバスの転送速度（40ＭＢ／秒）が支配的であるので，式（１）での転送速度としては，ローカルバスの転送速度をもちいる．また，スキャタ，ギャザ時のBIF の設定時間はレジスタへのロード時間のみ，として10メモリサイクル1.6 マイクロ秒，MSC の設定時間も同様に1.6 マイクロ秒とした，また，ソフトでのブロックの切りだしの時間も1 ワード当たり，2 メモリサイクル0.32マイクロ秒とした．

また，スキャタ，ギャザ時の転送時は初期設定を最初に１度だけやれば良いので，転送時間は，

$$T = Tc + Tt \times n \quad \text{——————} \quad (2)$$

また，ブロードキャスト②ではソフトで切りだしを行う時間が加わるので，

$$T = Tc \times p + (Tt + Tw) \times n \quad \text{——} \quad (3)$$

Tw:1 ワードあたりのデータ切りだし時間

100 ワードのデータを転送したときの転送時間について性能予測をした結果を，表 3 にスキャタについて，表 4 にギャザについてしめす．

表 3

セル台数	スキャタ新方式	ブロードキャスト①	ブロードキャスト②
64	1.04	1.13	4.40
1024	1.04	2.66	5.93

表 4

セル台数	ギャザ新方式	ブロードキャスト①	ブロードキャスト②
64	1.04	1.13	4.40
1024	1.04	12.49	15.76

（単位ミリ秒）

本方式では，セル切替えのオーバヘッドがないので，台数が増えても転送時間は増加しないが，従来方式では台数が増加するにつれて，転送時間が増大しているのがわかる．特に，ギャザを従来のブロードキャストで実現した場合，転送の切替えをホストの指示によっておこなっている．したがって，セルの切替え時間はホストの切替え命令がセルに届く時間と，その命令を受けてからデータがホストへ届く時間の和なので，Bネットの伝播遅延が支配的になっている．そのため，台数が増えたときの性能の低下が著しい．

5．おわりに

高並列計算機ＣＡＰ－Ⅱのブロードキャストネットワークについて述べた．本ネットワークによって，ホストとセルの間で多量のデータを高速で交換することが可能になる．

現在，我々はＣＡＰ－Ⅱのハードウエアの開発をおこなっている．今後は，システムを完成させ実機上で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．本研究会投稿予定

2）清水他，高並列計算機ＣＡＰ－Ⅱのメッセージコントローラ．本研究会投稿予定

3）堀江他　高並列計算機ＣＡＰ－Ⅱのルーティングコントローラ．本研究会投稿予定