1H−4

高並列計算機ＡＰ１０００のＢ−Ｎｅｔアーキテクチャと評価

加藤　定幸　石畑　宏明　清水　俊幸　堀江　健志

㈱富士通研究所

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AP1000を開発している．AP1000のように最大1024台のセル（プロセッサ）から構成される並列計算機ではホスト計算機とセルの間の通信が問題になる．ホスト計算機が個々のセルと通信するとには，通信するセルの切替えが必要になる．AP1000のホスト—セル間ネットワークであるＢネットはこの問題を解決するためにスキャタ，ギャザという転送モードを持っている．

２．Ｂネットの機能

セルから見たAP1000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．同時に複数のセルあるいはホストがＢネットにデータを送りだすこはできない．

Ｂ−ネットでのデータ転送はパケットと呼ばれる単位で転送制御をおこなう．パケットは大別して次の４種類の転送モードをもっている．

①　ブロードキャスト
②　グループキャスト
③　スキャッタ
④　ギャザ

ブロードキャストではネットワークに接続されたすべてのセルがデータを受信し，グループキャストでは１つ以上のセルがデータを受信する．ブロードキャストとグループキャストでは，Ｂネットに送りだされたデータは対象となるセルに全てのデータが取り込まれる．

スキャタ・ギャザは1 つのパケット内のデータをセルごとに分割して転送したり，それぞれのセルが送りだしたデータをあたかも1 つのパケットとしてホストとして取り込んだりする転送モードである．

例えば，図1 の様に2 次元配列を分割してセルに割り当てて通信を行う場合，従来の方式ではセルごとにグループキャストを繰り返す必要があるが，スキャタとギャザを使うと1 回の転送でデータを送ることができる．

スキャタ・ギャザ時のデータ分割は図2 の示すようなバリエーションがあり，アプリケーションの性質に応じて適当な物を選んでしようする．図中で同じ網が掛かっている部分ごとにセルに分割して転送する．

また，Ｂネット上ではホストもセルも同等に扱われるので，例えばセルから他のセルへスキャタをおこなったりギャザを行うことも可能である．

３．Ｂネットの構造

Ｂネットは1024台までの拡張性と高速転送を同時に満たすために図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て実現している．

Ｂネットではホストあるいは任意のセルがデータを送りだすので，送信元の位置に応じてネットの構造を変化させる必要がある．具体的には送信元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．ツリー型の部分ででは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，送信元の存在するツリーでは送信元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．

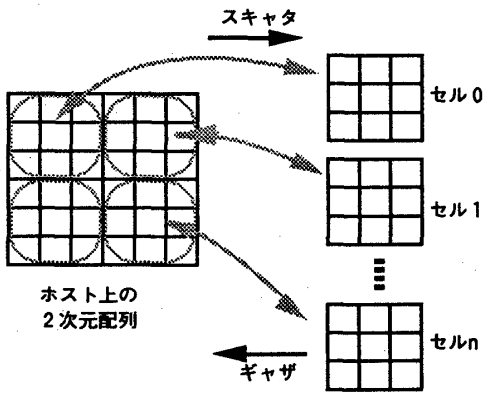

図１スキャタとギャザ

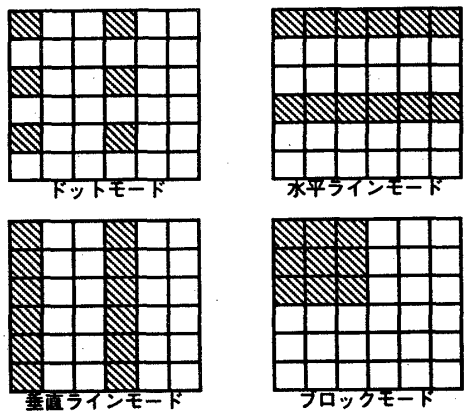

図２スキャタ・ギャザのモード

B-Net of Highly Parallel Processor AP-1000:Architecture & Evaluation
Sadayuki Kato Hiroaki Ishihata Toshikyuki Shimizu Takeshi Horie
FUJITSU LABORATORIES LTD.

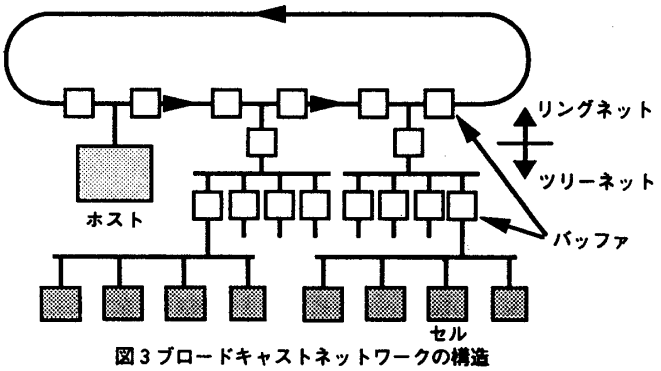

図3 ブロードキャストネットワークの構造

スキャタ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と等しいので，送信側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．

逆にギャザでは，前述のブロードキャストやスキャタとは反対に複数の送信元からホスト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，この制御はデータ転送時間と等しい時間でおこなってるので，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．

4．性能評価

16台のシステムで転送性能の評価をおこなった結果を以下に示す．転送速度は全転送データ数を初期設定の開始から最後のデータの受信までの時間で割ったものである．

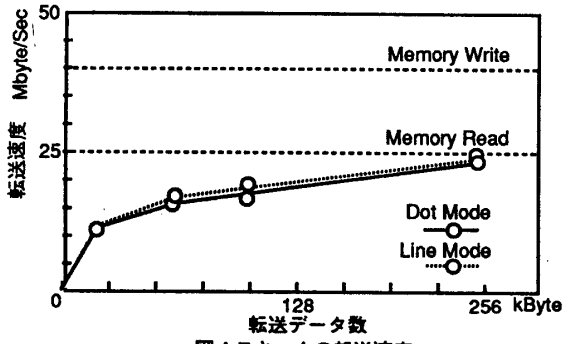

図4 スキャタの転送速度

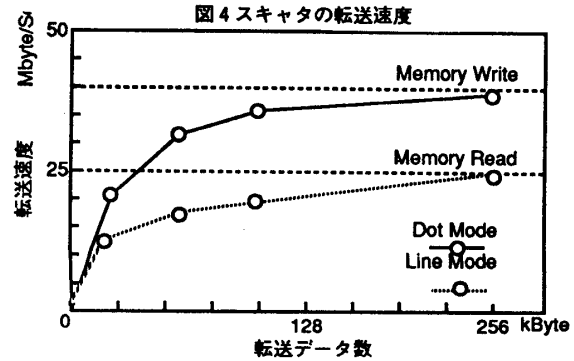

図5 ギャザの転送速度

まず，図4,図5 にラインモード，ドットモードでの転送速度を示す．転送データ数が少ないほど，初期設定のオーバヘッドの影響が大きいく鳴っている．図中の水平な線はセルのローカルバスのメモリリードとメモリライト時の理論性能をしめしている．通常Bネットの転送レートはメモリリードサイクルによって制限されているが，ドットモードではセルでのメモリアクセスが間欠的になるので転送性能はホスト側のライトサイクルまで転送性能が向上する．

図7,8 にラインモードで転送したときに，スキャタ，ギャザと1 つ1 つのセルを切り換えて通信する従来方式との比較をおこなった例についてしめす．従来方式ではセルか切り替わるたびに，初期設定をしているので，このオーバヘッドの為に見かけ上の転送速度が低くなってしまう．特にホストにデータを集める時には，セルが変わる度にBネットの転送方向を切り換えているために見かけ上の転送速度が大きく低下していることがわかる．

5．まとめ

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．

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．情処研報　90-ARC-83-37
2）加藤他，高並列計算機ＣＡＰ－Ⅱのブロードキャストネットワーク．情処研報　90-ARC-83-39

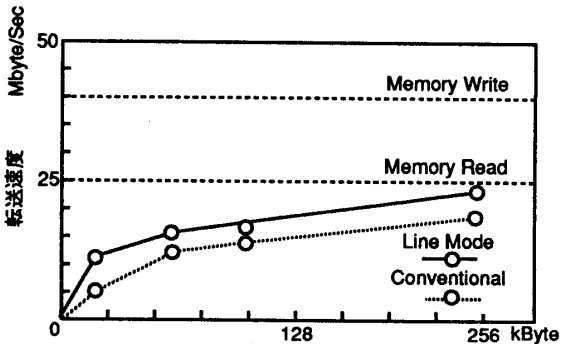

図6 従来方式との比較(スキャタ)

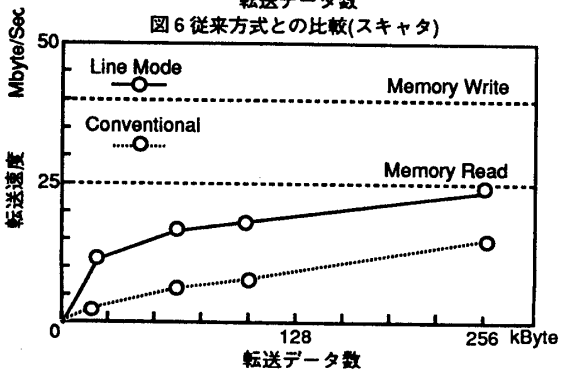

図7 従来方式との比較(ギャザ)