オペーレーティング・システム 54－2
（１９９２．３．１３）

高並列計算機ＡＰ１０００のメッセージ通信機構

池坂守夫　堀江健志　清水俊幸　石畑宏明

㈱ 富士通研究所

AP1000は，分散メモリ型の並列計算機であり，細かい粒度の並列性を対象としたデータ並列処理の高速化を目的とする．このため，AP1000では，低レイテンシと高スループットの両方を達成する通信を実現している．AP1000は，１対１通信を実現するトーラスネットワーク等の３種類のネットワークでプロセッサが接続され，ネットワークとプロセッサのインタフェースをとるメッセージコントローラは，通信のレイテンシを小さくする重要な機能を持っている．本論文では，AP1000アーキテクチャについて述べ，その性能評価を行い，AP1000が細粒度のデータ並列処理に適したアーキテクチャであることを示す．

Message Communication of Highly Parallel Computer AP1000

Morio Ikesaka, Takeshi Horie, Toshiyuki Shimizu, and Hiroaki Ishihata

FUJITSU LABORATORIES LTD.

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

The AP1000 is a highly parallel computer with distributed memory. In the AP1000, the newly developed routing scheme is used for a torus topology network to reduce latency and achieve high throughput. The message controller on each processor reduces message handling time such as message assembly/disassembly or data transfer/reception setup for communication. In this paper, we present the AP1000 architecture and some experimental results.

1. はじめに

AP1000は，富士通研究所が開発した分散メモリ型の並列計算機である．16～1024台のプロセッサ（以下セルと呼ぶ）を，3種類のネットワークで結合した構成をとり，フロントエンドのホスト計算機としてワークステーションが接続されている[1]，[2]，[3]，[4]，[5]．

数値計算，CAD，コンピュータグラフィックスなど多くの応用では，データ並列により効率良く問題を解くことができる．AP1000では，粗い粒度の並列性だけではなく，細かい粒度の並列性を対象とした，データ並列の処理を高速化することを目標に，そのアーキテクチャを設計した．特に，低レイテンシと高スループットの両方を達成するメッセージ通信が必要である．AP1000の設計思想は，次の4点に集約できる．

(1)低レイテンシ通信

数値計算におけるデータ並列処理では，計算結果を必要とするプロセッサに，いかに速く結果を届けるか，すなわち，通信の遅延時間をいかに小さくするかが，システムの高性能化の鍵である．これには，通信ネットワークの遅延時間を短縮するだけではなく，メッセージの送信や受信のときに要する遅延時間の短縮も必要である．細かい粒度の並列処理では，特に重要な点である．

(2)高スループット通信

並列計算機の通信ネットワークには，上記の低レイテンシに加えて，ネットワークの輻輳状態による性能低下をいかに小さくするかが要求される．低レイテンシを実現する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などが提案されている．このワームホールルーティングの基本メカニズムとして，一つのメッセージを転送中はその使用チャネルをブロックする性質をもち，通信のスループットを低下させる要因となっている．

(3)高速なデータの分配と収集

典型的なデータ並列処理では，計算をはじめる前にホスト計算機からセルにデータを分配し，計算終了後にセルに分散したデータをホスト計算機に収集する．この分配と収集のために，ホスト計算機は多数セルとの通信を必要とするので，セル台数の増加に従って，通信セットアップ等のオーバーヘッドが増加する．

(4)高速バリア同期

多数セルに渡る同期的なアルゴリズムを実現するためには，バリア同期の機構は必須である．このバリア同期専用のハードウェアを持たせ，同期に要するオーバヘッドを削減すれば，細かい粒度の並列処理を効率良く実行させることができる．

本論文では，AP1000で実現した高速メッセージ通信機構について述べる．まず，第2章で，AP1000のアーキテクチャについて述べる．次に，第3章で，AP1000のメッセージ通信機構について述べる．最後に，第4章では，実現したメッセージ通信の性能評価について述べる．

2. AP1000アーキテクチャ

2.1 システム構成

図1に，AP1000のシステム構成を示す．

AP1000は，16台から最大1024台の，セルと呼ぶ高性能プロセッシングエレメントを，3種類のネットワークで接続した構成をとっている．

全てのセルは，トーラスネットワーク（T-Net）と呼ぶ通信ネットワークで接続されている．T-Netは，任意セル間での1対1通信に用いられ，自動ルーティング機能によって，隣接セルだけではなく遠隔セルとも高速に通信できる．

ホストと全セルは，もう一つの通信ネットワークであるブロードキャストネットワーク（B-Net）で接続されている．B-Netは，ホストとセルを含めた，1対多通信，データの分配と収集に，用いられる．

ホストと全セルは，ハードウェアでバリア同期を実現する専用の同期ネットワーク（S-Net）で接続されている．S-Netは，木構造をとっており，バリア同期の要求から成立までの時間は1.6μsである．

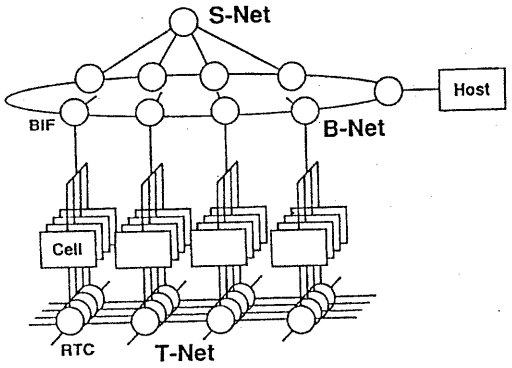

図1：AP1000のシステム構成

2.2 セル構成

図2に，セル構成を示す．

各セルは，SPARC プロセッサ(IU, FPU)，16 MB のDRAM，128 KBのキャッシュメモリ，メッセージコントローラ(MSC)とネットワークデバイス(RTC, BIF) で構成されている．RTC は，ルーティングコントローラと呼び，T-Netを制御する．BIF は，B-Netインターフェースと呼び，B-Netの制御を中心に，S-Netの機構も備えている．これらは，32 ビットの内部バスで接続されている．

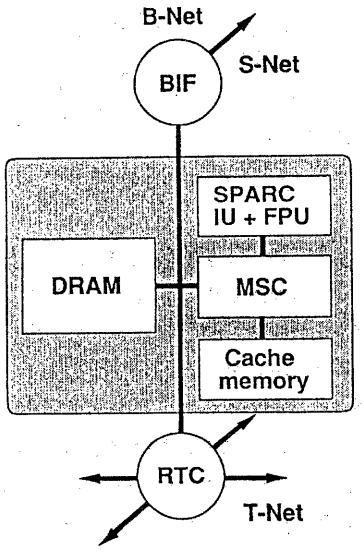

図2：セルの構成

3. メッセージ通信機構

3.1 MSC

メッセージコントローラ（MSC）は，メッセージの送信と受信におけるセットアップオーバーヘッドを削減する特別な機構（ラインセンドとバッファレシーブ）を実現している．この機構によって，小さなメッセージを効率良く通信できる．

MSCには，この機構に加えて，通常のDMAコントローラの他，ストライド，リスト，ベクタ転送の機能も備えている．

(1)ラインセンド

MSCは，小さなメッセージを最小のオーバーヘッドで送信するラインセンドと呼ぶ機能を持つ（図3）．

プログラムで通信データ（メッセージ）を作った直後は，特に小さなメッセージの場合，そのメッセージがキャッシュ上にある確率が高い．このメッセージを送信する最も速い方法は，キャッシュからネットワークに直接転送することである．MSCでは，この機能をラインセンドと呼び，実現している．メッセージがキャッシュ上にあるときは，メッセージがネットワークに直接送信される（図3の①）．メッセージがキャッシュ上にないときは，MSCに備えたDMA機能が自動的に起動され，メモリ(DRAM)上のメッセージがネットワークに送信される（図3の②）．通常の書き込み操作でラインセンドを実行でき，一回でキャッシュのラインサイズ分のデータ（16バイト）をネットワークに送信できる．

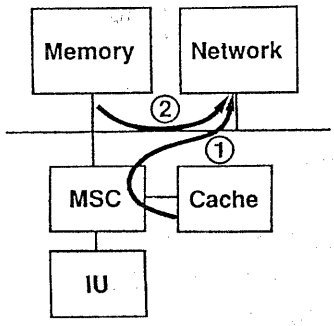

図3：ラインセンド

ラインセンドを用いると，まず，メモリ転送が不要であること，次に，DMAスタートのための設定やI/O チャネルへのアクセスが不要であるという利点がある．従って，メッセージ送信時の遅延を短縮することができる．

(2)バッファレシーブ

MSCは，ネットワークからメモリ上のリングバッファに，メッセージを直接受信するバッファレシーブと呼ぶ機能を持つ（図4）．

MSCでは，READアドレスとWRITE アドレスを示すポインタレジスタを持ち，これらを使って，メモリ(DRAM)上にリングバッファを定義できる．ネットワークからリングバッファへのメッセージ受信に対応してWRITE ポインタが自動的に制御され，リングバッファにあるメッセージのIUによる読み込みに対応してREADポインタが自動的に制御される．

バッファレシーブを用いると，従来行われていたメッセージ受信に対する割り込み処理が不要であるという利点がある．従って，プログラム実行とメッセージ受信を並列に行うことができる．

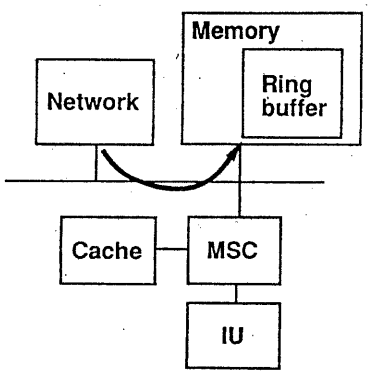

図4：バッファレシーブ

3.2 T-Net

T-Netは，二次元トーラス構造のネットワークである．メッセージのルーティング方式として，ワームホールルーティングを採用している．

ワームホールルーティングでは，中継ノードは，フリットと呼ぶ数バイトのデータのみストアする．あるノードがメッセージヘッダを受信すると，そのノードは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に転送する．後続フリットは全て，ヘッダフリットが選択したルートと同じルートで，転送される．このように，ワームホールルーティングを採用すると，低レイテンシが実現できる．しかし，一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っているチャネルはブロックされるので，デッドロックの発生とスループットの低下を招く可能性がある．

そこで，デッドロックの回避とスループット低下の最小化を実現するために，ワームホールルーティングに構造化バッファプールアルゴリズムを取り入れたルーティング方式を採用した．これを用いると，一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しない．図5にこのルーティング方式の一例を示す．

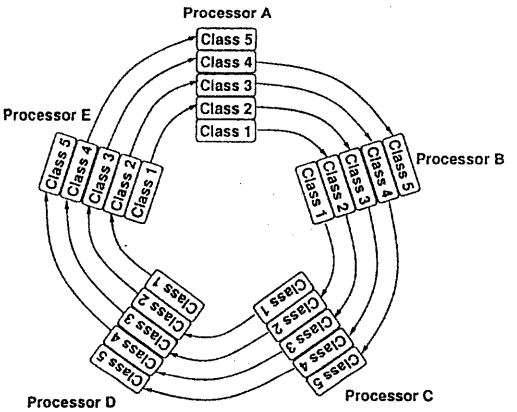

図5：ルーティング方式

各ノードは，フリットをストアするためのバッファを持ち，「ノード間の最大距離＋1」に等しい数のフリットをストアできる．例えば，単方向チャネルに5個のノードが接続されている場合，各ノードは5個のフリットをストアできるバッファを持つ．

フリットがバッファにストアされると，次のノードに転送される．データクラスは，データそのものと一緒に転送される．任意のデータクラスのフリットを転送でき，各フリット転送毎にデータクラスを変えることができる．複数のフリットが転送できるなら，その一つを選択する．

このアルゴリズムは，デッドロックを回避している．すなはち，どのノードからも，クラス1⇨クラス2⇨クラス3⇨クラス4⇨クラス5の経路が存在し，ループを作らない．

また，全てのプロセッサが，A⇨C，B⇨D，C⇨E，D⇨A，E⇨B，というように，時計回りの方向に同時に転送するとき，全てのチャネルが用いられ，スループットの低下を招かないことがわかる．

ＲＴＣは，以上に述べた，低レイテンシ，高スループット，デッドロック回避を満たすメッセージ通信機構を実現するＬＳＩである．ＲＴＣは，二次元トーラス接続された任意セル間の1対1通信を行う自動ルーティング機能を持つ．このとき，まずX方向にルーティングし，次にY方向にルーティングする固定ルートを採用した（図6）．さらに，1対1通信に加えて，放送通信の機能も持つ（図6）．この放送通信では，メッセージヘッダでX方向とY方向の範囲を指定でき，指定した範囲での放送を実現できる．この機能は，数値計算における行列演算等で頻繁に使われ，ハードウェアでの実現は効果的である．

ＲＴＣの各チャネルは，１６ビット幅で，最大転送レートは，1チャネル当たり２５MB/sである．中継ノードでの遅延時間は，1ノード当たり 80ns である．1ワードを32ビットと定義して，ネットワークの輻輳がない場合の転送時間（遅延時間）は，

160 ＋ 160×セル間距離 ＋ 160×ワード数 (ns)

で与えられる．

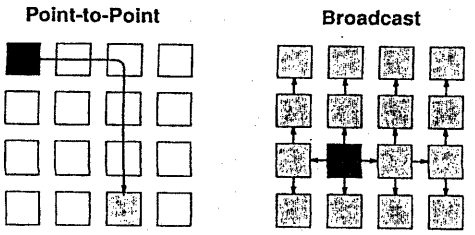

図6： T-Netの通信

3.3 B-Net

B-Netは，階層バスとリングを組み合わせた構造のネットワークである．B-Netは，ホストからセルへあるいはセルからセルへの1対多の放送通信，ホストから特定セルへの1対1通信の基本的な通信に加えて，データの分配と収集の機能を持つ（図7）．

多くの並列プログラムでは，ホストにある初期データをセルに分配し（Scatter），各セルでの計算結果をホストに収集する(Gather)必要がある（図8）．従来では，ホストは多数セルとの通信を必要とするため，セル台数が増加するにつれ，ホストの通信セットアップに要する時間は増加する一方であった．ＢＩＦは，1対多と1対1の基本的な通信を実現するだけでなく，この分配と収集の機構を実現するＬＳＩである．

B-Netは，共通バスと同じように，一つのプロセッサ（ホストあるいはセル）だけが送信できる．但し，収集のときは，一つのみが受信し，他の全てが送信する．

B-Netは，３２ビット幅で，最大転送レートは５０MB/sである．

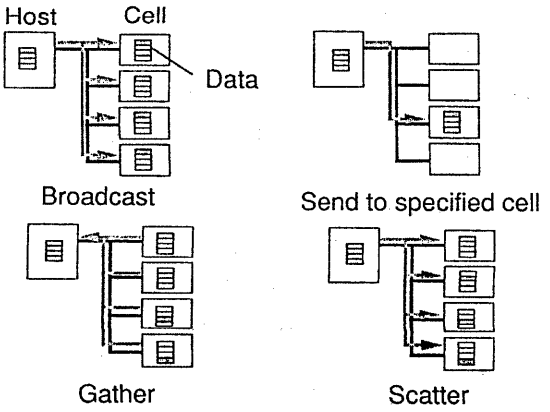

図7： B-Netの通信

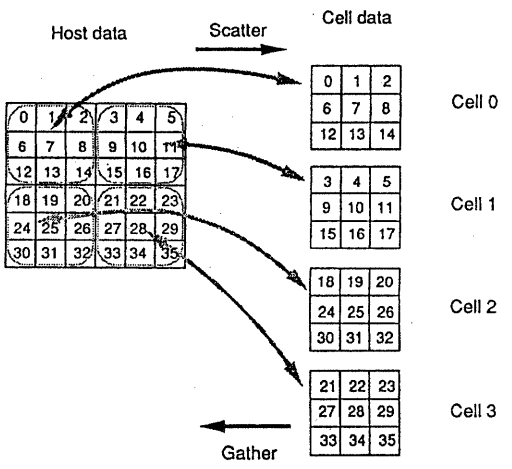

図8：データの分配と収集

4. 性能評価

現在，最大５１２（１６×３２）台のAP1000システムが完成している．

AP1000のメッセージ通信機構（ＭＳＣのラインセンドとバッファレシーブ，T-Netのルーティング方式）の有効性を示すために，２種類のベンチマークプログラムを実行した．

第一のPingpongベンチマーク[6]では，メッセージ通信のレイテンシに関する基本性能を評価する．

第二のLINPACKベンチマーク[7]は，ＬＵ分解によって連立一次方程式を解くものであり，これを用いて，より大きなアプリケーションでの通信の実効性能を評価する．

(1)Pingpongベンチマーク

このベンチマークでは，マスタセルからスレーブセルにメッセージを送り，スレーブセルはマスタセルにそのメッセージを直ちに送り返す．距離３だけ離れたセル間で行ったメッセージ交換の半分の時間を，データサイズを変えてAP1000で図った結果を，図９に示す．図９で，□と○は従来のDMAを用いた結果であり，■はラインセンドとバッファレシーブを用いた結果である．ラインセンドとバッファレシーブの機構が，小さなメッセージの通信に，非常に有効であることが示されている．

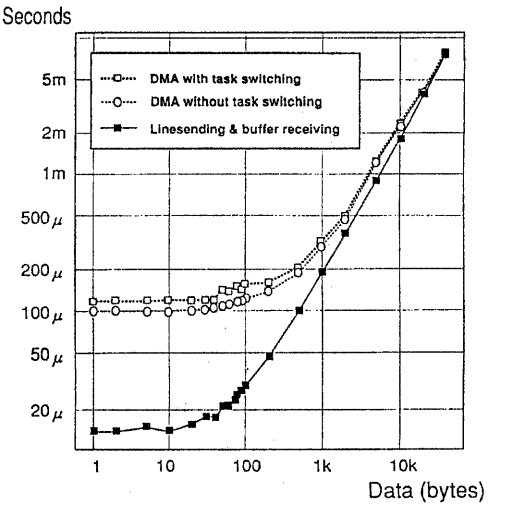

図９：Pingpongベンチマーク

(2)LINPACK ベンチマーク

図10に，AP1000を含めた分散メモリ型並列計算機での結果を示す．レポートとして公表されているこの結果は，問題規模として1000×1000のマトリクスを解くものであり，AP1000では，ＭＳＣのラインセンドとバッファレシーブの機構，T-Netの放送機能を使っている．AP1000では，細かな粒度の通信を高速化したことにより，問題サイズが比較的小さいとき，あるいは，台数が増えたときにも，高い性能が得られている．

マトリクス演算に代表される数値計算では，通信のレイテンシが，並列計算機の性能を支配する．AP1000では，ネットワークとメッセージ送受信時のハンドリングで，通信のレイテンシを小さくしており，細かなメッセージ転送を必要とする応用に対しても，並列性を抽出することができる．また，メッセージ送受信時のハンドリングでは，メモリ対メモリの通信を高速化する方法を採用した．これによって，プログラムを柔軟に記述できる．従って，AP1000は粒度の粗い並列性から，粒度の細かい並列性まで，広い応用に適用できる．

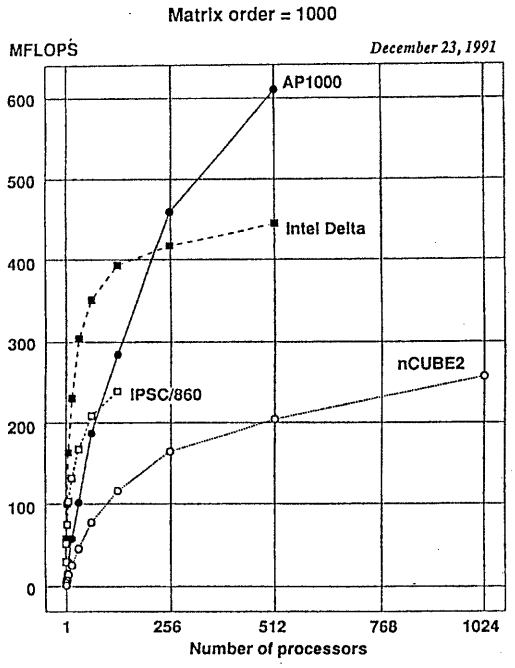

図10: LINPACK ベンチマーク

5. おわりに

AP1000のアーキテクチャについて，そのメッセージ通信機構を述べ，その性能を評価した．この結果，粒度の細かい並列処理を効率よく実行できることが示された．従って，AP1000では，プロセッサ台数が増えても，問題の並列性を十分引き出せ，性能をあげることができる．

今後も，種々の応用問題にAP1000を適用し，アーキテクチャ等の評価を続けていく．

謝辞

日頃，御指導や御助言を頂く，情報処理研究部門石井部門長代理，アーキテクチャ研究部白石部長，第二研究室佐藤室長に深謝いたします．

参考文献

[1] 石畑他，"高並列計算機AP-Ⅱの構成とメモリシステム"，情報処理学会計算機アーキテクチャ研究会，83-37，1990，pp. 217-222.

[2] 堀江他，"高並列計算機AP-Ⅱのルーティングコントローラ"，情報処理学会計算機アーキテクチャ研究会，83-38，1990，pp. 223-228.

[3] 加藤他，"高並列計算機AP-Ⅱのブロードキャストネットワーク"，情報処理学会計算機アーキテクチャ研究会，83-39，1990，pp. 229-234.

[4] 清水他，"高並列計算機AP-Ⅱのメッセージコントローラ"，情報処理学会計算機アーキテクチャ研究会，83-40，1990，pp. 235-240.

[5] 堀江他，"高並列計算機AP1000のアーキテクチャと性能評価"，電子情報通信学会研究会，CPSY91-26，1991，pp. 173-180.

[6] R. Hockney, "Performance parameters and benchmarking of supercomputers", Parallel computing, Vol 17, 10&11, 1991, pp. 1111-1130.

[7] J. J. Dongarra, "Performance of Various Computers Using Standard Linear Equations Software", Technical Report, 1991.