計算機アーキテクチャ　83－37
（１９９０．７．１８）

高並列計算機ＣＡＰ－Ⅱの構成とメモリシステム

石畑　宏明　　稲野　聡　　堀江　健志

清水　俊幸　　加藤　定幸

㈱富士通研究所

本論文は，高並列計算機ＣＡＰ－Ⅱの設計目標とアーキテクチャ，ホストとのインターフェース，各プロセッサエレメントのメモリシステムの構成について述べる．ＣＡＰ－Ⅱは，64～1024台のプロセッサ・エレメントからなる分散メモリ型の並列計算機である．並列計算機をさまざまなアプリケーションに適用するためには，高い単体プロセッサの演算性能と，高速なプロセッサ間通信が要求される．ＣＡＰ－Ⅱでは，プロセッサに32ビット RISCマイクロプロセッサを採用し，用途別の３種類のネットワークを持つことにより高性能を実現する．

An Architecture of Highly Parallel Processor CAP-Ⅱ

Hiroaki Ishihata, Satoshi Inano, Takeshi Horie,

Toshiyuki Shimizu, and Sadayuki Kato

FUJITSU LABORATORIES LTD.

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

In this paper, we present an architecture of CAP-Ⅱ. CAP-Ⅱ is a highly parallel processor consists of 64 to 1024 processing elements (PEs). Highly parallel processor must have both powerful PEs and high speed communication network to achieve high performance in variety of problems. We use 32bit RISC micro-processor with cache memory for PE. And the system has 3 different type networks for broadcast, local communication, and synchronization.

1. はじめに

数百台以上のマイクロプロセッサから構成される高並列計算機の開発が盛んである．Ncube，iPSC のように，すでに商品化されているものもある．我々は，分散メモリ型の高並列計算機CAP-Ⅱ（Cellular Array Processor -Ⅱ）の開発を進めている．CAP-Ⅱは，グラフィクスや図形処理，回路シミュレーション，分子動力学，モンテカルロシミュレーションなどの分野へ適用し，実用的性能を実現することを目的としている．

並列計算機で高性能を実現するためには，問題に内在する並列性を十分引き出すことが必要である．数値計算では，データを分割し多数のプロセッサ・エレメント（セル）に分散させて並列化を行うことにより高い並列度が生まれる．このような，データ分散を基本とした並列処理では，ホスト計算機上のまとまったデータをセル群に分配したり，セル上の計算結果をホストに集めることを高速に行えることが必要となる．

並列処理では，あるセルが計算した結果を別のセルが利用して計算が進むのが一般的である．このような場合は，データ送信から受信までの通信遅延時間（レイテンシ）が小さいことが必要である．グラフィクス[1]や分子動力学[2]では，細かなデータ通信が頻繁に起こる．データの通信順序がランダムで，通信のスケジューリングができないので，小レイテンシという通信特性が重要になる．

CAP-Ⅱでは，データ分散による並列化や少量データのランダムな通信といったアプリケーションの特性を考慮した総合的な通信性能の向上を設計目標の最重点としている．本文では，第2章で，CAP-Ⅱの具体的設計目標とアーキテクチャについて，第3章で，CAP-Ⅱのホストインターフェースについて，第4章で，セルの構成について，第5章では，セルのメモリシステムの構成について述べる．

2．ＣＡＰ－Ⅱのアーキテクチャ

2.1. ＣＡＰ－Ⅱの設計目標

並列計算機で並列処理の効果を出すためには，演算性能と通信性能の両方を高めることが重要である．我々は，CAP-Ⅱの設計にあたり，以下の目標を設定した．

(1)小型高性能なセルプロセッサ

セルプロセッサは，1システムあたり数百以上も使用する．コンパクトに実装できることと信頼性が高いことが重要である．演算能力は，実装とのトレードオフになる．メモリ容量・演算速度がソフトウェア開発上の大きな制限とならないことが最低限必要である．

(2)高速なグローバル通信

ホストとセル群の通信のように，1対多の通信をグローバル通信と呼ぶ．グローバル通信では，ブロードキャスト以外に，データの分散・収集を考慮したデータ転送が効率良く行えることが必要である．

(3)高速なローカル通信

セル同士の1対1の通信をローカル通信と呼ぶ．ローカル通信は，任意のセル間で自由に行える必要がある．ネットワークには，デッドロックの起きないルーティングの機能と，データ転送能力が大きく，レイテンシが小さいことが要求される．

(4)全セルの同期とステータス

セル全体で同期待ちをする機能や，全セルの状態を知るための機能が必要である．並列処理の台数効果を高いものにするためには，これらの機能のオーバーヘッドが小さいことが必要である．

(5)実現性・拡張性

アーキテクチャは，64セルから1024セルまでの広い範囲のシステム構成に対応できる拡張性を備え，現在の実装技術で実現可能でなければならない．また，各セルには，種々の専用ハードウェアを付加できるのが望ましい．

2.2. グローバル通信

グローバル通信は，1対多の通信であり並列化できない．セル側の並列処理によって，並列化可能な部分の処理時間が短縮されるに従い，並列化されないこの部分の占める割合が大きくなる．これを高速化することは，トータルの処理時間を短縮する上で重要である．

数値計算では，データを適当に分割し多数のセルに分散させ，そのデータの分割に基づいて処理を行うことにより高い並列度を引き出すことができる（図1）．このような，データ分散を基本とした並列処理では，ホスト計算機上のまとまったデータをセル群に分配したり，セル側の計算結果をホストに集めることが必要となる．

ホストは，全セルからのデータを受信したり全セルにデータを送信したりする．セル台数分のデータ転送セットアップが必要になり，ホスト側でデータの分配・収集にかかる時間Tは，(1)式のようになる．

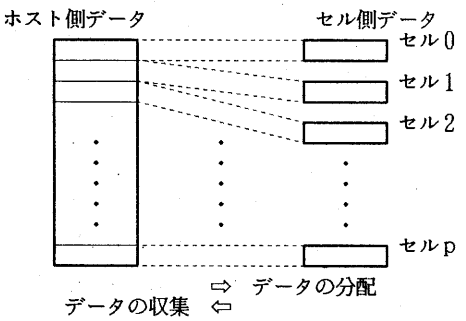


図1　データの分配と収集（スキャタとギャザ）

$$T = (Ts * p + Tt * N) \qquad \text{(1)}$$

T　：分散データの転送時間
Ts：データ転送のためのセットアップ時間
Tt：1ワード転送にかかる時間
N　：総転送ワード数
p　：セル数

(1)式から，転送するデータ量を一定とした場合，セル数が増加するほど，転送のセットアップによるオーバーヘッドが増加する．CAP-Ⅱでは，各セルにデータの分配収集を高速化するスキャタ・ギャザ機能を付加することにより，ホスト側のデータ転送セットアップ時間をセル台数に関係なく一定にしている．

グローバル通信は，Broadcast network(Bネット）で行う[3]．共通バスは，物理的に近距離にある少数のセルを接続するには適している．しかし，数百以上のセルを1つの共通バスに接続することは困難である．リング状のネットワークは，セル間の物理的距離を延ばすことができる．しかし，セル数が増加すると通信のレイテンシが大きくなる．Bネットは，物理的に局所的に存在できる部分（バックパネル内）のセルについては，共通バス接続を行い，距離の離れるところ（バックパネル間）にはリング構造を取り入れた複合ネットワークで実現する．

2.3.　ローカル通信

セル間のローカルな通信は，任意のセル間で同時に並行して行われる．データ転送能力が大きいこと，レイテンシが小さいこと，同時に通信が発生してもデッドロックの発生やスループット低下がないことが重要である．

通信路自身のデータ転送速度を高めるためには，ネットワークのデータ通信路のビット幅を広くし，通信レートを上げればよい．レイテンシを小さくするためには，データ転送のセットアップやルーティングのオーバーヘッドを極力小さくする必要がある．ローカル通信ネットワークでは，デッドロックの起きないルーティングアルゴリズムをハードウェア化することが必要である．

ネットワークのトポロジによって，1セルから出る通信ポートの数は異なる．1セルから入出力できる信号線の数と速度を一定とすると，個々の通信ポートのデータ転送速度Fは，通信ポートの数に反比例する．通信ポートのデータ転送速度は，1セルから出るポート数が少ない程高くすることができる．

$$F = npin * f / (nport * W) \qquad \text{(2)}$$

nport：1ノード当たりのポート数
npin：1セルから入出力できる信号数．
W：ワード幅　　f：信号の速度
F：ポート当たりのデータ転送速度

レイテンシは，ワームホールルーティング[4]では，セル間の距離とデータ長の和に比例するので，セル間の距離が短いほど小さくなる．1セルでのルーティングにかかる時間を1ワード転送の時間と同じとすると，レイテンシLは次のようになる．

$$L = (D + N) / F \qquad \text{(3)}$$

L：レイテンシ　　D：ノード間の距離
N：転送するデータ数

(2)，(3)式より，レイテンシLは，nport *(D＋N）に比例する値となる．二次元トーラス，三次元トーラス，ハイパーキューブの場合についてこの値を比較すると，表1のようになる．

データ転送速度は，ポート数の少ない低次元メッシュの方が常に高速である．レイテンシは，セル数が数百程度までは，2次元がすぐれており，千台以上になると3次元のほうが小さくなる．転送データ数が大きくなるとこの差は小さくなる．ハイパキューブはセル間距離が小さいが，データ転送幅も小さいのでレイテンシは小さくならない．

表1　通信のレイテンシ

トポロジ		二次元		三次元		ﾊｲﾊﾟｰｷｭｰﾌﾞ	
ｾﾙ間最大距離		$\sqrt{p}$		$1.5\sqrt[3]{p}$		log p	
ポート数		4		6		log p	
転送ﾃﾞｰﾀ 数N		8	16	8	16	8	16
レイテンシ	p＝ 64	64	96	84	132	84	132
	p＝ 128	77	109	93	141	105	161
	p＝ 256	96	128	105	153	128	192
	p＝ 512	123	155	120	168	153	225
	p＝1024	160	192	139	187	180	260

p：セル数

CAP-Ⅱでは，二次元トーラスのトポロジを持つTorus Network （Tネット）によりローカル通信を行う．Tネット

では，ワームホールルーティングと構造化バッファプールのアルゴリズムを組み合わせた方式でデータのルーティングを行う．[5]

2.4. 同期とステータス

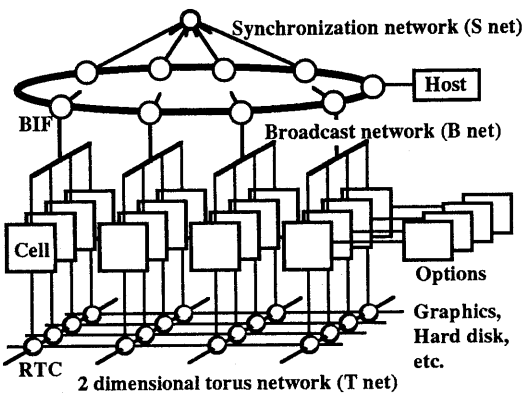
並列処理においては，処理の流れのある時点でセルを同期させることが必要になる．同期は，最もレベルの低い同期プリミティブである．より高度な同期プリミテブ（しばしば遅くなりがち）を提供するよりも，単純で高速なものを提供した方が有用であると考えた．

同期機能の他に，全セルの状態のAND 条件を与えるステータス機能も必要である．例えば，反復法の計算では，各セルが独立に反復計算を行うが，計算の終了時に全てのセルで結果が収束しているかどうかの判断が必要である．同期とステータスを組み合わせて，同期が確立した時の状態を全セルに自動的に通知することにより，この判断を高速化することができる．

CAP-Ⅱの同期・ステータス系は，ツリー状のSynchronization network （Sネット）で構成される．全セルからのデータのAND 条件を全セルに分配することにより，一定時間で同期やステータスの検出ができる．

2.5. ＣＡＰ－Ⅱの構成

図2にCAP-Ⅱのシステム構成を示す．セル同士は，Tネットによって4つの隣接するセルと接続している．Bネットによって，全セルとホスト計算機が接続している．Bネットは，論理的には1つの共通バスであるが，実際には，リング&階層バスで実現されている．また同期・ステータス用にSネットがある．



Cell: Cell processor　BIF: Broadcast network interface
RTC: Routing controller

図2　　ＣＡＰ－Ⅱのハードウェア構成

3．ホストインターフェースの構成

CAP-Ⅱのホスト計算機には，良好なソフトウェア開発環境をもつEWS を使用する．アプリケーションのトータルの処理時間を短縮するために，ホストの性能も十分高いことが必要である．CAP-Ⅱのホストインターフェースは，図3のようにEWS 側のVME-IFとCAP-Ⅱ側のＢネットインターフェース（BNIF）から構成される．BNIFには，32MBのローカルメモリとDMACを搭載しＢネットとのデータ転送を高速化する．

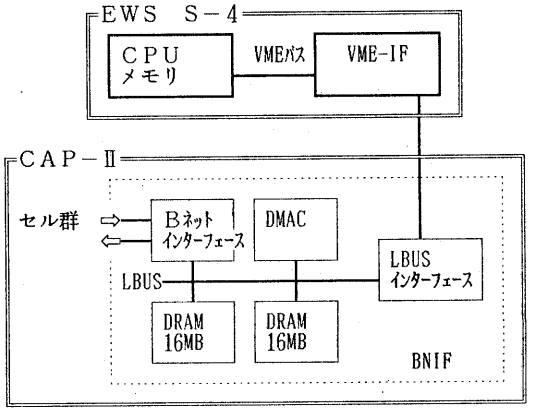


図3　ホストＩＦの構成

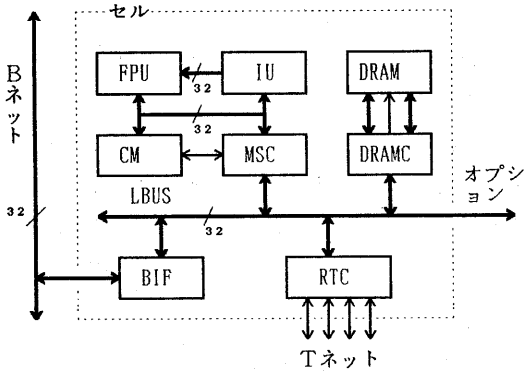
4．セルの構成

4.1. セルの構成

システムをコンパクトに実現するために，セルの機能を6チップのLSI と汎用メモリで実現する．図4にセルのハードウェア構成を示す．プロセッサは，整数演算・論理演算・制御を行うIUと浮動小数点演算器FPU から成り，キャッシュメモリCMと接続される．DRAMコントローラは，16MBのメモリの制御とエラーの検出・訂正を行う．

通信インターフェースとして，メッセージコントローラMSC [6]，ルーティングコントローラRTC，Ｂネットインターフェース BIF があり，DRAMC とともにLBUSと呼ぶ内部バスにより接続する．

LBUSは，32ビット幅のアドレス・データ多重の同期式バスである．LBUSは，各セルごとにコネクタを介して外部に取り出されており，画像入出力デバイスや高速Ｉ／Ｏインターフェース，ディスクインターフェース，拡張メモリ，ベクタープロセッサなど種々のオプションハードウェアの付加が可能である．



IU　：整数演算ユニット
FPU　：浮動小数点演算ユニット
DRAMC　：メモリコントローラ
MSC　：メッセージコントローラ
BIF　：Bネットインターフェース
RTC　：ルーティングコントローラ
CM　：キャッシュメモリ
DRAM　：メモリ

図4　セルの構成

4.2.　セルプロセッサ

セルプロセッサのCPU には，EWS で使用され，良質なコンパイラなどのソフトウェア開発環境が整っているSPARC を採用する．SPARC のようなRISC (Reduced Instruction Set Computer) のCPU では，ほとんどの命令が1クロックで実行される．この高性能を生かすためには，毎クロックごとに新しい命令を供給し続けることが必要である．このために，高いデータ転送能力を持ったデータバスや主記憶が要求される．

SPARC-IUからのメモリシステムへの最大の要求は，1クロック動作のメモリを実現することである．高速でかつ大容量のメモリを妥当なコストで実現するためには，キャッシュメモリの使用が必要となる．キャッシュのヒット率が高ければ，プロセッサは，ほとんどのメモリアクセスをキャッシュに対して行うので等価的に1クロックのメモリシステムが実現できる．ヒット率が低い場合には，主記憶やデータバスの転送速度の影響が出てくる．

4.3.　セルの通信インターフェース

通信を高速化するためには，送信や受信のセットアップ時間の短縮，データ転送と演算の並列実行が必要である．MSC は，7チャネルのDMA コントローラとキャッシュコントローラから構成される．転送セットアップのオーバーヘッドを減少しデータ転送処理をIUからオフロードする．MSC 内のDMACは，飛び飛びに存在するデータを転送したり，リスト構造や間接データを直接転送することができる．また，キャッシュに乗っている可能性の高い小さなデータを直接RTC やBIF に送り出すことにより，小さなメッセージの高速通信を可能にしている．

BIF は，Bネットとのデータ転送を行う．システム起動時のブートプログラムの自動ロード機能やスキャターやギャザのための機能，同期・ステータス系の機能を備えている．RTC は，トーラスネットワーク上でのデータの中継を行う．

5．セルのメモリシステム

5.1キャッシュの構成

キャッシュメモリ設計の際のトレードオフについては，いままでに種々の報告がなされている[7]．CAP-Ⅱのキャッシュシステムを設計する上で，実装面積が小さいことと性能が高いことを目標とした．

キャッシュ用のメモリは，外付けの汎用SRAMを使用する．実装面積の点で，データ幅を増やしてセットアソシアティブ構成にすることができない．反面，大容量の高速SRAMを使用できるので，メモリ容量でヒット率を稼ぐことが可能である．CAP-Ⅱでは，32K ×8 構成の汎用の高速スタティックRAM を4チップ使用して，128KB のダイレクトマップキャッシュを実現する．

5.2.　キャッシュ制御方式

キャッシュへの書き込みがあったとき，主記憶へそれを反映させるための方式として，ライトスルー方式とコピーバック方式がある．ライトスルー方式では，メモリへの書き込み操作がすべてミス扱いになるので，メモリアクセス全体に占める書き込みアクセスの割合以上にヒット率は上がらない．コピーバック方式には，このようなことはないので，高いヒット率を得ることができる．従って，メモリアクセスの頻度も，コピーバック方式の方が少ない．

並列処理では，プロセッサ間の通信が頻繁に起きる可能性があり，メモリアクセスの頻度は増えがちである．キャッシュとメモリとのトラッフィックは少ない程望ましいのでCAP-Ⅱでは，コピーバック方式を採用する．

5.3.　ＤＲＡＭＣと主記憶

プロセッサの性能は，メモリの平均アクセスタイムに比例する．キャッシュメモリのヒット率の上昇が頭打ちになると，平均アクセスタイムの短縮のためには，ミス処理の

時間を短縮する必要がでてくる．ミス処理の時間を短縮するためには，主記憶のアクセスタイムを短くすることが必要である．主記憶のアクセスタイムの短縮には，複数ワードのブロック転送機能や，メモリバンクのインターリーブが有効である．

CAP-Ⅱのセルの主記憶は，DRAMC と4MbitDRAM から構成される．DRAMC は，セルのローカルバスに直接接続され16 MBのメモリを制御するLSI である．ECC 機能を持ち，1bit の誤りを検出・訂正し，2bit誤りを検出する．1M×(32+8) bit のメモリバンクを４つ持ち，４重インターリーブ制御により連続したアドレスのアクセスを高速化している．小型化のためDRAMC は，CMOSのゲートアレイ上に18,000ゲート相当の回路として実現した．

表２　セルプロセッサ諸元

プロセッサ	SPARC　IU+FPU (25MHz)
ｷｬｯｼｭﾒﾓﾘ制御方式	ｺﾋﾟｰﾊﾞｯｸ　ﾀﾞｲﾚｸﾄﾏｯﾌﾟ　I/D 共有
キャッシュ容量	128KB ﾗｲﾝｻｲｽﾞ 16B
主記憶容量	16MB　4Mbit DRAM使用
主記憶制御方式	４ｳｪｲ ｲﾝﾀｰﾘｰﾌﾞ ECC 機能
アクセスタイム	リード　ライト ﾜｰﾄﾞｱｸｾｽ　400ns　160ns ﾌﾞﾛｯｸｱｸｾｽ　640ns　400ns
ＤＲＡＭＣ 付加機能	20ﾋﾞｯﾄﾀｲﾏｰ ×２　割り込み制御 ｽﾀｰﾄｱｯﾌﾟ 時の自動ｲﾆｼｬﾗｲｽﾞ

6．あとがき

本文では，並列計算機で高い通信性能を実現するためには，高速な同期機能，データの分散・収集を考慮したグローバル通信，デッドロックフリー，低レイテンシという特性のローカル通信が要求されることを示した．次に，高並列計算機CAP-Ⅱのアーキテクチャ，セルの構成，セルのメモリシステムの構成について述べた．

CAP-Ⅱでは，データ分散による並列化や少量データのランダムな通信といったアプリケーションの特性を考慮した通信性能の向上をねらっている．このために，Tネット上でのルーティングを行うRTC，Bネットと接続してデータの分散収集を行うBIF，データ転送を効率化するMSC，メモリを制御するDRAM C の４種類のLSI を開発した．CAP-Ⅱのセルは，SPARC-IU, FPU, MSC, RTC, BIF, DRAMCと汎用メモリからなり，コンパクトに実装している（図５）．

今後，種々のアプリケーションをCAP-Ⅱにインプリメントして，アーキテクチャの有効性を実証していきたい．より広い分野の問題に適用するためには，並列化コンパイラや並列プログラムデバッグツールといったソフトウェアの充実が課題となる．最後に，CAP-Ⅱを開発するにあたりご指導いただいた，当研究所ｼｽﾃﾑ研究部の石井部長，白石部長代理，㈱富士通）ｽｰﾊﾟｰｺﾝﾋﾟｭｰﾀ開発部）内田部長に深謝します．

表３　ＣＡＰ－Ⅱシステム諸元

システム	分散メモリ　MIMD並列計算機 64～1024セル構成
ﾌﾞﾛｰﾄﾞｷｬｽﾄ ﾈｯﾄﾜｰｸ	リング＋階層バス　50MB/s ｲﾆｼｬﾙﾌﾟﾛｸﾞﾗﾑﾛｰﾄﾞ，ﾌﾞﾛｰﾄﾞｷｬｽﾄ ｽｷｬﾀｰ，ｷﾞｬｻﾞ ｵﾍﾟﾚｰｼｮﾝ
ﾄｰﾗｽ ﾈｯﾄﾜｰｸ	二次元トーラス　25MB/s ﾜｰﾑﾎｰﾙﾙｰﾃｨﾝｸﾞ＋構造化ﾊﾞｯﾌｧﾌﾟｰﾙ 行または列へのﾌﾞﾛｰﾄﾞｷｬｽﾄ
同期ﾈｯﾄﾜｰｸ	同期＋ステータス　40ビット

参考文献

(1)佐藤他，”高並列計算機CAP-Ⅱによる三次元ｸﾞﾗﾌｨｸｽ”，本研究会投稿予定
(2)佐藤他，”並列計算機CAP による分子動力学計算”，電子情報通信学会研究会資料，CPSY89-24, pp. 57-62, 1989
(3)加藤他，”高並列計算機CAP-Ⅱのﾌﾞﾛｰﾄﾞｷｬｽﾄ・ﾈｯﾄﾜｰｸ ”，本研究会投稿予定
(4)W. J. Dally, "A VLSI Architecture for Concurrent Data Structures", Kluwer, Hingham, MA, 1987.
(5)堀江他，”高並列計算機CAP-Ⅱのﾙｰﾃｨﾝｸﾞ・ｺﾝﾄﾛｰﾗ”，本研究会投稿予定
(6)清水他，”高並列計算機CAP-Ⅱのﾒｯｾｰｼﾞｺﾝﾄﾛｰﾗ ”，本研究会投稿予定
(7)A. J. Smith, "Cache Memories", ACM Computing Surveys, Vol. 14, No. 3, Sept., 1982, pp. 473-530

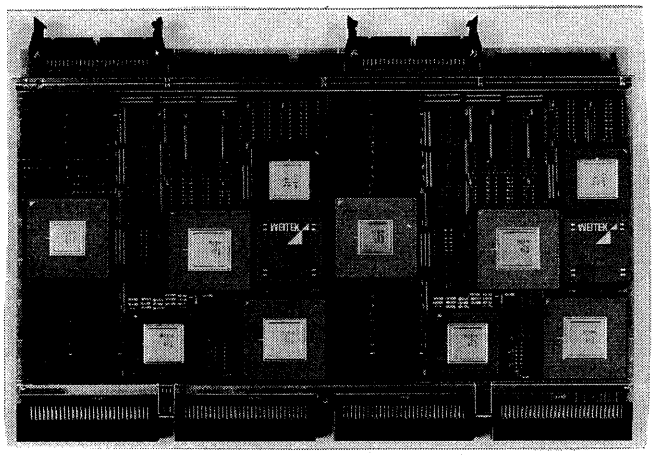


図５　CAP-Ⅱセルプリント板（２セル実装）