計算機アーキテクチャ 85－1
（1990．11．21）

資源共有型並列計算機“砂丘”

荒川 修・橋本 正巳・井上 倫夫・小林 康浩
鳥取大学工学部

本研究では，資源共有型マルチマイクロプロセッサシステム“砂丘”を開発している．本システムは，６４台の汎用マイクロプロセッサを接続可能で，６４Ｍバイトの容量の主メモリを持つ．並列処理プログラムの同期を容易に実現するために，システムコントローラ（割り込み制御回路），排他制御回路などのハードウェアを特別に設けている．また，演算結果を分かりやすく表示するためのグラフィックディスプレイ，マンマシンインターフェースとシステムの制御を行うためのフロントエンドプロセッサを設けている．本報告では，フロントエンドプロセッサ，システムエリア及び同期通信方式について報告する．

The Multiprocessor System "SAKYU" of the Shared-Memory Type

Osamu ARAKAWA・Masami HASHIMOTO・Michio INOUE・Yasuhiro KOBAYASHI
Faculty of Engineering,Tottori University

Our multiprocessor system, called "SAKYU", consists of 64 microprocessrs and has main memories (shared type) of 64MB a total. To achieve parallel processing regularly, various ideas are taken into this system, and are realized as hardware; e.g., System-Controller, Semaphore, Arbiter of memory contention. As the I/O equipment several graphic displays are furnished. Front-End-Processors are provided for practising good man-machine interface and for controlling the whole system easily.

Among them, the function of the Front-End-Processors, the role of System-Area devices and the procedure of the syncronized communication is described in this paper.

1．はじめに

筆者らは，実験室レベルで特定ユーザーが利用する数値シミュレーションマシンとして，並列計算機“砂丘”を開発している[8]～[10]．“砂丘”は，複数のマイクロプロセッサ（μP）を密に結合したマルチμPシステムである．一般に，この種のシステムでは，メモリを共有することによる各μPから共有メモリへのアクセスから競合する機会が増し，実質的に携わるμPの数が増えないことになる．μPの接続台数が，数十～百台程度であれば，アクセスパスを分散することによりアクセス競合を緩和ないし回避することが可能である[10]．共有メモリ方式は，データの授受が高速かつ容易に行えることから，さまざまな処理アルゴリズムに対応できるような利便性のよいシステムの構築が可能である．

本報告では，まず現在開発中の並列計算機“砂丘”のシステム構成を述べ，次に同期通信方式について述べる．またシステムの稼働率とテストプログラムによる実測値について述べる．

2．システム構成

図１は，筆者らが製作を進めている並列計算機“砂丘”のシステム構成図である．本システムは，最大６４台のプロセッサユニット（ＰＵ）が接続可能な密結合型マルチμＰシステムである．その

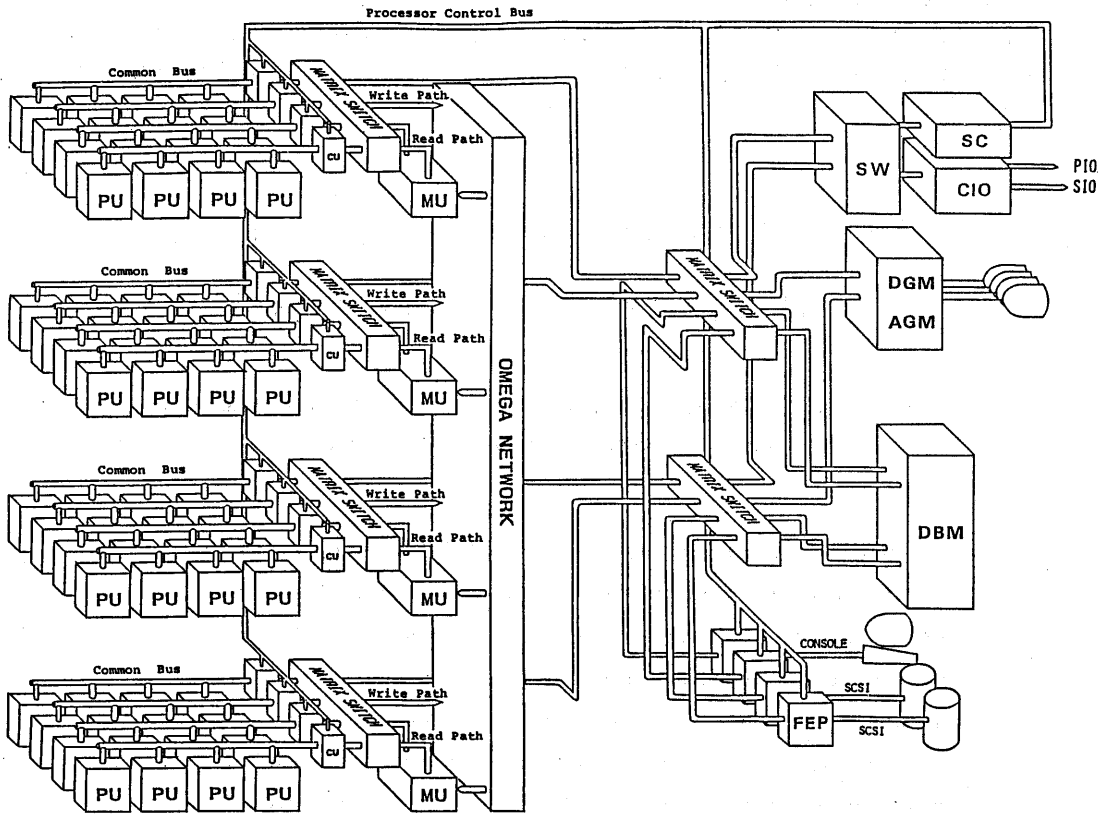

図１　並列計算機“砂丘”のシステム構成

構成はバスコントロールユニット（ＢＣＵ）と最大４台のＰＵを１グループとし（単一バスで接続），それらの４グループをまとめて１ブロックとする．システム全体で４ブロックをマトリックススイッチ（ＭＴＸ）を介して主メモリ，オメガネットワーク，システムエリアに接続する．ＭＴＸは４×４のバススイッチである．４入力は，すべてＢＣＵに接続する．４出力は，１出力がシステムエリアのアクセスに，他の３出力が主メモリのアクセスに使用される．主メモリのアクセス（３出力）は，２出力が２インターリーブ構成のリードアクセスパスとして使用され，別の１出力がライトアクセスパスとしてオメガネットワークに接続される．

２．１　主メモリ

主メモリは，合計６４Ｍバイトの容量を持ち，演算データを格納する．通常のプログラムでは，ＰＵの共有メモリアクセスは，リードアクセスの方がライトアクセスより多く行われる．このため主メモリを４つのユニットに分割し，各ＰＵブロック毎に置く．各メモリユニットのリードアクセスパスは２インターリーブでＭＴＸの２出力と接続した．このため同じＰＵブロック内でのみリードアクセス可能である．またライトアクセスパスはオメガネットワークを通して全てのメモリユニットへ接続される．ライトアクセスパスにオメガネットワークを使用したことにより，主メモリは次の様な利用形態が考えられる．

（１）マルチリード・ワンライト方式

オメガネットワークの一斉放送モードを用いることにより４つの主メモリユニットの内容を同一に保持することができる．全ＰＵが同じデータで一斉に処理をするときに，見かけ上のメモリ容量は４分の１になるが，同一ＰＵブロックからのリードアクセスの競合のみを考えれば良く，アクセス競合を緩和できる．

（２）パイプライン接続

あるＰＵブロックで所定の処理した結果を，隣のＰＵブロックのメモリユニットにライトアクセスする．そしてライトされたデータに新たな処理をし，結果をさらに隣のメモリユニットにライトアクセスするということを反復する．

（３）個別接続

それぞれのＰＵブロックのライトアクセスを，リードアクセスで使用するのと同じメモリユニットに対して実行する．つまり各ブロックが４つのメモリユニットをそれぞれ独立して使用する．

２．２　フロントエンドプロセッサ（ＦＥＰ）

ＰＵが演算専用なのに対して，ＦＥＰはマンマシンインターフェースやシステムの制御の部分を受け持つ．ＦＥＰにも複数のプロセッサを用意してそれぞれの機能を分散させる．マンマシンインターフェースを受け持つＦＥＰは，ユーザーからの要求の受け付けやプロセスの制御を行う．入出力装置を制御するＦＥＰは，プリンタなどのキャラクタ型デバイスと，ハードディスクなどのブロック型デバイス向けに個別に用意する．

２．３　システムエリア

２段目ＭＴＸの出力に接続されているメモリ領域をシステムエリアと呼ぶ．このエリアは大きく分けてグラフィックメモリ，システムワークエリアから成る．２段目ＭＴＸの入力には，ＰＵ側からは１段目ＭＴＸの出力と接続し，ＦＥＰもここに接続する．このエリアは全てのプロセッサからアクセス可能である．

（１）グラフィックメモリ

グラフィックディスプレイが膨大な計算結果を分かりやすく図示するために用意される．デジタルグラフィックは主に計算結果をグラフなど２次元的に図示するために用い，アナロググラフィックは主に色の違いや濃淡で３次元的なものを表すのに用いる．

（３）システムワークエリア

このエリアには，並列処理プログラムを効率よく実行するためのハードウェアとして，システムコントローラ，メイルボックス，排他制御回路などを置く．また，他の機器との入出力を行うためのコミュニケーションＩ／Ｏが設けてある．

3．同期・通信方式

マルチプロセッサシステムでは，複数のプロセス相互の協調をとりながら，まとまった仕事を遂行させなければならない．そのためには，プロセスの実行を調整して正しい結果となるよう実行順序を保障しなければならない．それには割当てられたプロセスの同期と，互いの進行状況や中間結果を他のプロセスに通報することである．

本システムではこれら同期・通信を効率よく行うための専用回路として，システムコントローラ，排他制御回路，プロセス番号割当回路を設け同期・通信操作に伴うオーバーヘッドを小さくなるようにした．

（１）システムコントローラ

システムコントローラとして，割り込み制御回路を設けており，全てのμＰ群に接続している．各μＰは，この制御回路及びシステムワークエリア内のメイルボックスを用いることにより，プロセス起動，μＰ間の通信，同期等を行うことが可能になる．また，上記の操作を行うμＰは，制御回路内に設けた同期フラグを管理することにより，処理の終了を確認することができる．

（２）排他制御回路

排他制御を実現するために，共有変数に対して行う不可分的なリード・モディファイ・ライト動作をハードウェアで実現した．共有エリアの一部を排他制御用メモリとして確保し，このメモリに対してはリードとライト動作を不可分的に実行する．具体的には，μＰのリード時に出力データのＭＳＢの値に応じて，回路内の制御部で自動的にライトサイクルが実行され，排他制御が行われる．

（３）プロセス番号割当回路

多数のプロセスを限られた数のμＰで効率よく並列処理するには，各μＰの処理するプロセスの番号管理が必要である．本システムではこの機能をハードウェア化している．共有変数を順序づけし，その番号をカウンターで管理し，μＰからのリード動作が終わると自動的に指示値を増すようにした．これによりμＰは共有変数に対して不可分的なリード・モディファイ・ライト動作を行わなくてよく，リード動作だけで番号管理ができる．

4．検討

資源共有型のマルチμＰシステムでは，原理的に共有バス上でアクセス競合が発生する．したがって，各μＰの共有資源アクセス要求を能率よく制御し，競合による待ち時間の累積を如何に抑制するかがシステムを効率よく運用する上で重要な課題である．

本システムで扱う問題は，特に限定せずに科学技術計算全般を予定している．例えば，ある事象をモデル化し，偏微分方程式による数値シミュレーションを行ったり，一次元・二次元のＦＦＴを行うような場合には，大きな配列データを主メモリ中に置くことになる．各μＰはこの配列からデータを読み込み，演算を行った後に，配列のデータを更新することを繰り返す．このような計算を行う場合，各μＰが分担する作業は，ほぼ均等に割りつけられて（アルゴリズム上では最適）いる．したがって，アクセス競合が頻繁に起こり易く，起これば全μＰにアクセス待ちが生じ，処理速度の向上は望めない．アクセス競合の機会を減らすため，主メモリにはマルチリード・ワンライトメモリ方式，２インターリーブ方式，２ポート方式を採用している．

μＰが共有メモリをアクセスするとき，共有メモリ本来のアクセス時間，共有バスでの信号遅延時間及びアクセス競合調停のための待ち時間が必要となる．μＰの持つ能力をどれくらい利用できたかを表すために，ここでは，稼働率を純粋に計算に専念できた時間の割合として定義し，検討を述べる．

4．１　並列動作時の各μＰの稼働率

筆者らは，各μＰの待ち時間の累積が最大となる（均等負荷で動作している）ときのシステムの性能を陽的に解析できることを示した[6]．そこでシステムのハードウェア特性（共有メモリのアクセス時間，サイクル時間等）及びソフトウェア特性（タスクの処理時間，共有メモリのアクセス回数）とを用いて，並列動作時の各μＰの平均稼働率を次のように表すことができる．

$$P(n,Th) = \frac{100}{1 + \frac{tac + m_n \cdot t_s}{Th}} \quad [\%] \quad (1)$$

ここで,

$$m_n = \begin{cases} 0 & (n \leq i_B \cdot n_0) \\ \frac{n}{i_B} - n_0 & (n > i_B \cdot n_0) \end{cases} \quad (2)$$

ただし,

n ：同時に動作しているμPの台数

Th ：タスクの平均シンクタイム［S］
（μPが共有メモリをアクセスする平均時間間隔）

t_s ：共有資源のサイクル時間［S］

t_{ac} ：共有資源をアクセスするとき各μPが要する平均アクセス時間

$$t_{ac} = t_B + t_A + \frac{1}{2} t_s \quad (3)$$

t_A ：共有資源のアクセス時間［S］

t_B ：BCU，MTX等での遅延時間［S］

i_B ：共有バスのインターリーブ数

n_0 ：同一の共有バスでアクセス競合による待ち時間の累積なしに動作できるμPの台数

$$n_0 = 1 + \frac{Th}{t_s} \quad (4)$$

4．2　マルチリード・ワンライトメモリ方式による稼働率

マルチリード・ワンライトメモリ方式では，並列稼働が可能なμPの台数及び稼働率は，主メモリへのリードとライトのアクセス比に依存する．主メモリへのアクセス全体を1としたときのリードアクセスの割合をrとすれば，（4）式から，リードアクセスにおいて競合による待ち時間の累積なしに並列稼働が可能なμPの台数は，主メモリを4ユニット設け，それぞれのリードバスを2インターリーブとしたので，

$$n_R = 8 \left(1 + \frac{Th/r}{t_s} \right) / r \quad (5)$$

と表せる．ライトアクセスにおける台数は，

$$n_W = \left(1 + \frac{Th/(1-r)}{t_s} \right) / (1-r) \quad (6)$$

と表せる．（5）（6）式の関係を図2に示す．

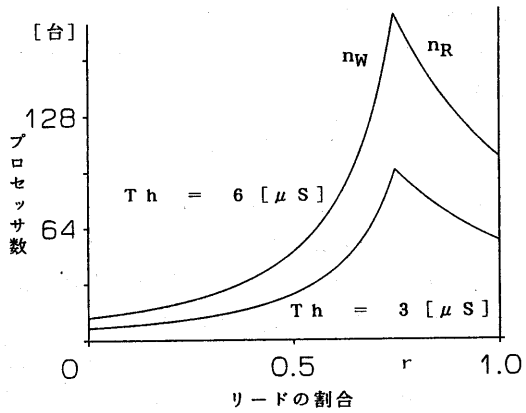

図2　メインメモリ利用時のリード・ライト比と競合を起こさずに稼働できるμP数

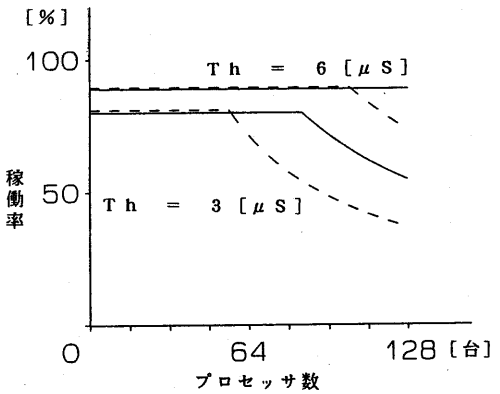

図3　メインメモリ使用時のμP数と稼働率

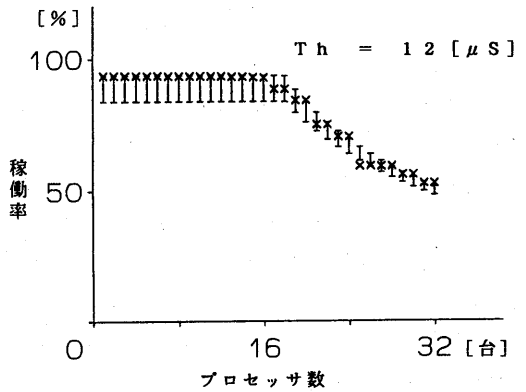

図4　並列計算機“砂丘”のシステムメモリにおける稼働率の実測値と理論値

この図からｒ＝０．７５付近であるとき，最もアクセス競合が起こりにくいことがわかる．また稼働率を表す式は，

$$P(n) = \frac{100}{1 + \frac{t_{ac} + m \cdot t_s}{t_h / r} + \frac{t_{ac}' + m' \cdot t_s}{t_h / (1-r)}} \quad (7)$$

と表せる．（７）式において，ｒ＝０．８のときのμＰの台数と稼働率の関係を図３に示す．破線は共有メモリを８インターリーブ方式で構成した場合の稼働率である．この図より接続するμＰの台数が同じ場合，マルチリード・ワンライトメモリ方式は８インターリーブ方式に比べてシンクタイムをより短くできることがわかる．

４．３　稼働率の実測値

３２台のμＰを１６台のＢＣＵに２台ずつ接続し，システムメモリを使用してシンクタイム１２［μＳ］のときの本システムの稼働率を測定した．その結果を示したものが図４で×印が実測値を表している．Ｉ印は理論値である．Ｉ印の上端は各μＰがシステムメモリをアクセスするときのアクセスタイムの最短の場合を示しており，下端は最長の場合を示している．この図より稼働率の実測値が理論値とよく一致していることがわかる．先の（１）式はμＰの能力の利用率を表している．この式をもとにシステムのハードウェア特性の評価を行うことができる．具体的にはバスの多重度，高速メモリの評価等をシステムの基本設計時に行うことができる．

５．おわりに

以上，資源共有型並列計算機“砂丘”の共有メモリの制御方式，機能及びμＰ間の同期・通信方式を紹介した．数値シミュレーション等を行う場合には主メモリに対するアクセスが頻発することから，競合を抑制し主メモリを能率よく共同利用するために，マルチリード・ワンライトメモリ方式の採用が効果的であることを示した．また稼働率の実測値が理論値とよく一致することも示した．

今後の課題は本システムを使用して各種の具体的な問題に対する評価をすることである．

参考文献

［１］　白川　他：“並列計算機ＰＡＸ－１２８”通信学論，Vol. J67-D, No.8, pp.853-860, Aug. (1984)

［２］　出口　他：“コンピュータグラフィックシステムＬＩＮＫＳ－１における画像生成の高速化手法”　情報処理論文誌，Vol.25, No.6, pp.944-952, Nov. (1984)

［３］　Rodrigue. G. : “Parallel Computation” Academic Press (1982)

［４］　Paker. Y: “Multi-microprocessor” Academic Press (1983)

［５］　井上・小林：“マルチマイクロプロセッサシステムα－16のアーキテクチャ”情報処理論文誌，Vol.25, No.4, pp.632-639, July (1984)

［６］　井上・小林：“α－16マルチマイクロプロセッサシステムの性能評価”情報処理論文誌，Vol.25, No.4, pp.640-646, July (1984)

［７］　井上・小林：“資源共有型マルチマイクロプロセッサシステムにおけるアクセス競合の調停について”　電子情報通信，回路とシステム研究会資料，CAS84-206, pp.9-16 (1984)

［８］　山根　他：“並列計算機“砂丘”のハードウェアアーキテクチャ”　電子情報通信，コンピュータシステム研究会資料，CPSY87-3, pp.15-22, June (1987)

［９］　加納　他：“マルチマイクロプロセッサシステムの大容量共有メモリの一構成法”情報処理学会　計算機アーキテクチャ研究会資料，CA-66-2, pp.1-8, July (1987)

［10］　井上　他：“マルチプロセッサシステム“砂丘”の共有メモリアーキテクチャについて”　情報処理学会　計算機アーキテクチャ研究会資料，CA-66-2, pp.9-16, Nov. (1989)