7L−1

# 階層構造を持つ超並列計算機ＭＡＮＤＡＬＡの構成

加納 卓也　Andrew FLAVELL　広田 勝久　藤本 茂訓　高橋 義造
徳島大学工学部知能情報工学科

## I. はじめに

近年、並列処理の研究が盛んに行われ、要素プロセッサが数千台以上の規模の超並列計算機が注目されるようになったが、その現実的なアーキテクチャはいまだ開発されるにいたっていない。我々の提案するＭＡＮＤＡＬＡは、超並列化の際の最大の障壁である空間的な構成を考慮した階層構造を持つ相互結合網を用い、高速なプロセッサ間通信を実現するためにルータ、ＮＩＵなどの専用ハードウェアを使用した、ＭＩＭＤ分散メモリ型の超並列計算機である。本稿では、これらのアーキテクチャについての検討、評価を行い、超並列計算機の現実的なアーキテクチャであることを示す。

## II. 超並列計算機ＭＡＮＤＡＬＡ

図１にＭＡＮＤＡＬＡの構成図を示す。ＭＡＮＤＡＬＡは、超並列化の際に最大の障壁となるノード間の空間的な接続を可能とするため、物理的な階層構造を考慮した構成を持つことをその特徴としている。

超並列計算機は、コストパフォーマンス、フォールトトレランスなどの観点より、高密度集積化およびモジュール化が大前提となる。つまり、１つのＶＬＳＩチップの中に複数の要素プロセッサを組み込む必要が生じ、それらのチップはいくつかのまとまりで基板上に実装され、また、それらがマザーボードなどにより接続されることになる。ここに、物理的、階層的なクラスタリングが発生することになる。

ＭＡＮＤＡＬＡは、この階層構造に効率よくマッピングされうる構造を持っている。つまり、ＭＡＮＤＡＬＡの構成自身が、階層的にクラスタリングされており、実に自然にモジュール化することができるのである。

## III. 相互結合網

超並列計算機の各ノードは、なんらかのクラスタリングを行われることになるものと思われるが、そのクラスタから外部に出されるリンクの数によってシステム構成の容易さなどが評価できる。ここでは、このリンクの数をクラスタ結合次数[1]と呼ぶことにする。このクラスタ結合次数は、システム全体のノードをC台づつクラスタとしてまとめ、そのクラスタをさらにC組まとめて、‥‥と繰り返した場合の、レベルL、クラスタサイズC、全体のノード数Nなどの関数として表すことができる。例えば、バイナリハイパーキューブのレベルLのクラスタの結合次数は$(\log N-L\log C)C^{L}$であり、トーラスの場合には$4C^{L/2}$と表せる。これらは、クラスタのレベルLの増加にともない大幅に増加することになるため、超並列化は困難であるといえる。

ＭＡＮＤＡＬＡの相互結合網[2]は、再帰的に完全結合を構成する階層構造を持っており、クラスタ結合次数がクラスタサイズCに等しい小さな定数となるため、超並列計算機の階層構造に適したトポロジであるといえる。また、繰り返し構造を持つため、いくらでも大きなネットワークを簡単に構築できるとい

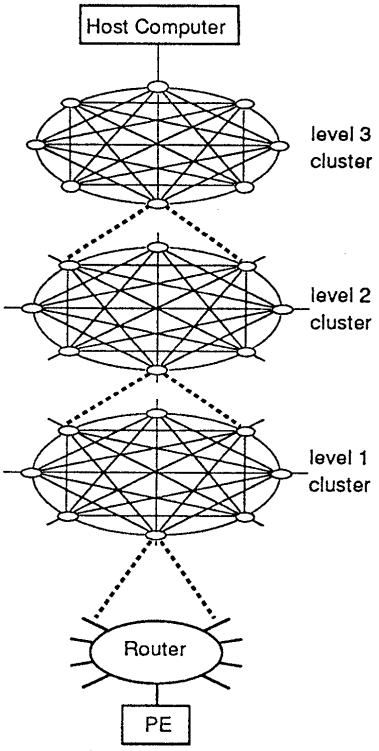


図１　ＭＡＮＤＡＬＡの構成
(C=8, L=3, N=512 の場合)

---

Organization of Hierarchically Structured Massively Parallel Computer MANDALA. Takuya KANOH, Andrew FLAVELL, Katsuhisa HIROTA, Shigenori FUJIMOTO and Yoshizo TAKAHASHI. Department of Information Science and Intelligent Systems, University of Tokushima.

う優れたスケーラビリティを持っている。

さらに、各クラスタ間の結合が完全結合と密であるため、高速な通信が期待できる。これまでに発表された階層構造を持つ並列計算機は、階層の上位レベルの結合が比較的疎であるものが多いが、より上位のレベルでは、物理的な伝送速度が遅いと予想されるため、密な結合を用いて、できるだけ経由距離を短くする方がよい。また、実際の問題における通信にはなんらかの局所性が存在しているため、ローカルな接続も密である方がよい。小数のプロセッサしか必要としないような小さなプロセスを複数動かすような場合にも同じことが言える。最も密な接続法は、完全結合であるが、全てのノードを完全結合することはハード量の点から不可能である。しかし、これを階層的に用いることは現実的に可能であり、良好な性能を得るものと期待できる。

ＭＡＮＤＡＬＡのクラスタサイズは、できるだけ大きくとった方が、平均通信距離の点で有利となり、また、大規模な網を構成する際に無駄にレベルを増やさないですむ。クラスタサイズは、ノードの次数に等しいため、クラスタサイズを大きくするには各ノードの次数を大きくする必要がある。このため、各ノードに使用される接続装置は多くの接続を可能とするものが要求される。

ＭＡＮＤＡＬＡにおける要素プロセッサ間の通信は、パケット交換によるメッセージ交換によって行われる。このため、各プロセッシングノードは、パケット交換を実現する機構を持ち、中継、ルーティングを行う必要がある。ＭＡＮＤＡＬＡでは、これを専用のハードウェアであるルータにより行う。このルータのプロトタイプとして、別稿にて発表のＭＥＧＡルータ[3]を現在開発中である。

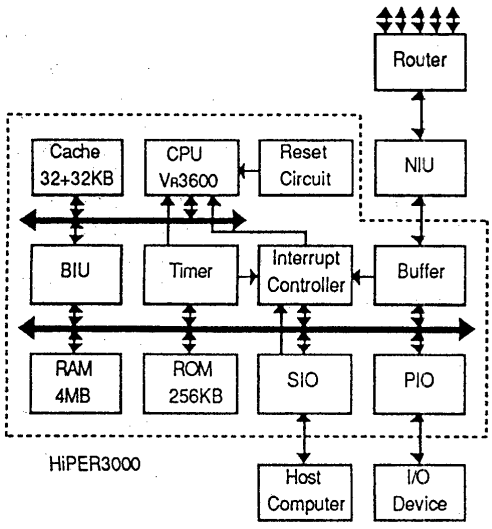


図２　ＨｉＰＥＲ３０００の構成

## IV. 要素プロセッサ

超並列計算機では、全体の性能を向上させるために、個々の要素プロセッサにも高性能なものを用いる必要がある。また、多数台必要となるため、コストと、サイズにも注意しなければならない。そこで、ＭＡＮＤＡＬＡでは、コストパフォーマンスに優れたＷＳクラスの性能を持つ要素プロセッサを使用する。

我々は、この要素プロセッサのプロトタイプとして、ＨｉＰＥＲ３０００（High-performance Processing Element utilizing R3000）を開発している。ＨｉＰＥＲ３０００は、図２のように構成され、ＭＰＵとしてＮＥＣの３２ビットＲＩＳＣプロセッサＶR３６００（ＦＰＵ内臓）を使用し、約２０MIPS／７Mflopsの性能を持つ。また、メインメモリ４Ｍバイト、キャッシュ６４Ｋバイトのほか、タイマ、パラレルＩ／Ｏなども持つ。

このＨｉＰＥＲ３０００を使用した場合の全体性能は、100%の台数効果を仮定すれば、1024台で、２０GIPS／７Gflops、65536台で、１.３TIPS／４５０Gflopsとなる。

また、要素プロセッサは、ＮＩＵによってルータと接続される。ＮＩＵ（Network Interface Unit）は、パケットの生成、組み立て、保留、管理などを、要素プロセッサとは独立に行う専用のハードウェアである。ＮＩＵは、現在ＬＣＡ（Logic Cell Array）にて開発中である。

## V. おわりに

ＭＡＮＤＡＬＡは、物理的な階層構造にマッチした構成をしており、超並列計算機のアーキテクチャとして、現実的かつ優れているといえる。また、ルータ、ＮＩＵなどの通信専用ハードウェアの採用は、超並列計算機の性能を決めるノード間通信時間を短縮する有効な方法であるといえる。さらに、高性能な要素プロセッサの使用は全体性能の向上に大きく貢献するものである。今後は、プロトタイプシステムの完成を目指し、各装置の開発をさらに進めていく予定である。

## 参考文献

[1] 加納卓也, 広田勝久, 藤本茂訓, Andrew Flavell, 高橋義造: 階層構造を持つ分散メモリ型超並列計算機ＭＡＮＤＡＬＡの設計, 情報処理学会第87回計算機アーキテクチャ研究会 No.11 (1992)

[2] Andrew Flavell, Takuya Kanoh, Yoshizo Takahashi : Mandala: An Interconnection Network For A Scalable Massively Parallel Computer, 情報処理学会第43回全国大会, 4Q-13 (1991)

[3] Andrew Flavell, Yoshizo Takahashi : The MEGA Router: A Hardware Message Passing Gate Array Router, 情報処理学会第45回全国大会, 7L-04 (1992)