計算機アーキテクチャ 139－15
（２０００．８．４）

EM-X と MD One を統合化した 粒子シミュレーション用並列計算機プロトタイプの構築

高田 亮[†,††] 清水 昭皓[††] 児玉祐悦[†††] 坂根 広史[†,†††] 佐谷野 健二[†††]
本多弘樹[†] 弓場敏嗣[†]

分子動力学や宇宙物理学のシミュレーション等に現れる粒子シミュレーションの大部分は単純な計算の繰り返しであり、ハードウェア化に向いている。ITL-MD One は、専用の演算 LSI を並列で用いることにより数十 GFlops という非常に高い実効性能を持つが、計算の柔軟性不足や、通信ボトルネックによる性能向上の頭打ち現象が問題点として挙げられている。一方 EM-X は、低オーバヘッドの細粒度通信が可能で、効率の良い汎用並列計算能力を持つとともにハードウェア拡張性に優れている。本研究では両者の利点を取り入れて、より優れた粒子シミュレーション用並列計算機を構築するために、両者のアーキテクチャを融合する方法を検討した。有効性を検証する為に、4 ノードの ITL-MD One を搭載したプロトタイプを作成し、80PE の EM-X に接続した。

A parallel particle simulation machine based on EM-X and MD One

Ryo Takata[†,††] Akihiro Shimizu[††] Yuetsu Kodama[†††] Hirofumi Sakane[†,†††]
Kenji Sayano[†††] Hiroki Honda[†] Toshitsugu Yuba[†]

Particle simulations such as molecular dynamics or gravitational N-body simulation consumes most of the machine time for simple iterations and therefore suitable for hardware acceleration. ITL-MD One has high peak performance of some tens of giga flops by utilizing multiple hardware pipelines to calculate the interactions between particles. But the following issues are known:(1)Host bus bottleneck occurs with increasing MD One boards.(2)O(N) processing on the host CPU becomes bottleneck as we speed up O(N^2) processing. EM-X is a parallel machine that has message based efficient remote memory access mechanism. We propose a hybrid parallel machine based on EM-X and MD One to overcome these problems A prototype machine with 4 ITL-MD One has been build and connected to the existing 80PE EM-X.

1．はじめに

粒子シミュレーションは、分子動力学、天体、粉体等の広範囲のシミュレーションに適用されており、計算モデルの性質上、並列化に適している事が知られている。粒子シミュレーションでは計算時間の大半が粒子間の相互作用の計算に費やされるため、GRAPE[1]等の専用計算機による高速化が試みられてきた。しかし、これまでの粒子シミュレーション用並列計算機は、並列化効率が必ずしも最適化されていなかったり、ある特定の計算用途に特化されすぎていたりした。一方、汎用の並列計算機では、粒子シミュレーションを含む広汎なアプリケーションが

†電気通信大学大学院情報システム学研究科情報ネットワーク学
University of Electro-Communication , Department of Information and Communication Engineering
††株式会社画像技研
Image Technology Laboratory Corp.
†††通産省工業技術院電子技術総合研究所
National Institute of Electrotechnical Laboratory

実行されているが、コストパフォーマンスの面で、満足の行くものではなかった。

本論文では、これらの問題点を克服し、粒子シミュレーションを効率的に処理するための並列計算機を提案する。この並列計算機は、汎用並列計算機EM-Xと、粒子シミュレーション用並列計算機ITL-MD Oneを統合化したアーキテクチャを持ち、高い性能と柔軟性を併せ持っている。

2．背景

2．1粒子シミュレーション用並列計算機ITL-MD One

ITL-MD Oneは、粒子間の相互作用をハードウェアで高速に計算するLSI（以下MD Chip）をボードあたり最大4個搭載し、このボードを複数枚PCIバスに実装して使用する。ボード上には粒子座標、係数（電荷等）を保持するためのメモリを持つ。ホスト計算機にはワークステーション又はパーソナルコンピュータが使用され、ボードのI/Oの制御と、その他の計算処理（三体力などの近接作用、粒子の移動、温度、エネルギーなど物理量の計算など）を行う[2]。ITL-MD Oneの基本的なアーキテクチャはGRAPEによっている。

2．2 EM-X

EM-Xは、電子技術総合研究所で開発された汎用並列計算機であり、細粒度通信機能をいかして柔軟な並列処理が可能であることが知られている[3]。1995年より、80PEのプロトタイプマシンが稼動しており、粒子シミュレーションを含む各種のプログラムで、並列性能が評価されている。

2．3 ITL-MD Oneの性能解析とその問題点

ITL-MD Oneのアーキテクチャ上の問題点については、性能解析モデルと実測値を元に文献[4]で議論されている。以下にその要点をまとめる。

（1）粒子座標のボードへの転送オーバヘッド

現状のPCIバスを応用したシステムでは、ホスト計算機と専用ボード間のブロードキャストができず、全てのボードに全粒子を転送しているため、ボードの枚数を増やしていくとオーバヘッドが大きくなる。

（2）MD ChipのI/O処理のボトルネック

MD Chipのレジスタに相互作用を受ける粒子の座標を書き込んだり計算された相互作用を読み出したりするI/O処理が、一つのホストCPUでシーケンシャルに行われる事によるボトルネックが存在する。

（3）ホスト計算機でのO(N)処理のボトルネック

近接粒子相互作用の計算や、温度制御、圧力制御、粒子座標の更新等が一つのホストCPUでシーケンシャルに行われる事によるボトルネックが存在する。

（4）パイプライン構造による応用範囲の限定

ITL-MD Oneで使用している専用計算ハードウェア(MD Chip)は、パイプラインの構造が基本的に固定であり、解ける問題の形も限定されている。

3．EM-XとITL-MD Oneの統合アーキテクチャ

3．1 設計思想

EM-XとITL−MD Oneの統合アーキテクチャでは、前章で述べた問題点の（1）（2）（3）を解決する事を目標とする。（4）の問題は、専用計算LSIのアーキテクチャの改良により、別途解決する必要がある。

EM-Xは、強力な細粒度通信機構を有しており、高い並列性能を有している。このネットワークアーキテクチャの上に、計算エンジンとして、ITL-MD Oneのような相互作用を計算するパイプラインを搭載する事により、高い並列性能と、柔軟性を併せ持った粒子シミュレーション用並列計算機を構築する事が、本計算機のねらいである。

3．2 計算機構成

3．2．1 ネットワーク構成

並列計算機のネットワーク構成は、EM-X に準じ、サーキュラオメガネットワークとする。ノードの最大数は、EM-X をベースとしたネットワークでの PE の番号付けのビット長の関係で、1024 となるが、このフィールドを拡張することも可能である。図1に PE の構成を示す。

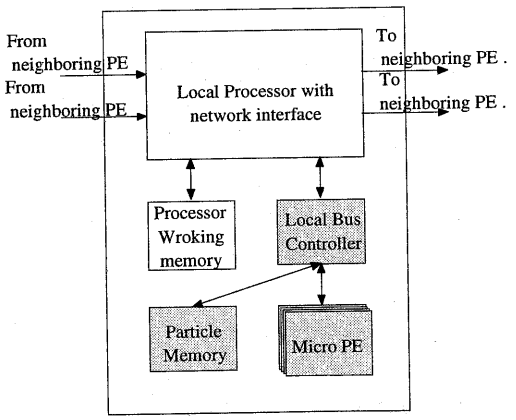

図1 PE 構成図

3．2．2 PE 構成

図1における PE の構成で、影付の部分が、新たに追加された部分である。ローカル CPU は EM-X の要素プロセッサ(EMC-Y)のように、命令レベルで細粒度通信機能をサポートする。ローカル CPU は、ネットワークインターフェースの他に2つのローカルバスを持っている。一方は、ローカル CPU のワークメモリに接続されており、CPU より任意のタイミングでアクセス可能である。もう一方は、バスコントローラを介して、粒子座標メモリと、専用計算パイプライン（以後 Micro PE と称する）に接続されている。2 バス構成をとることにより、計算を続行しながら、粒子座標や演算結果の相互作用を送受信する事が可能である。

3．3 特長

以下に本計算機の特長について、前章において指摘されていた問題点の解決方法に重点を置きながら以下に述べる。

3．3．1 ブロードキャスト機能

本計算機のサーキュラオメガネットワークは、ソフトウェアによるデータのコピーを行うことにより、ブロードキャストをエミュレートすることが可能である。このネットワークは、EAM 法における近接粒子座標の交換など、粒子シミュレーションに現れる多様な通信処理を効率的に処理することが可能である。

3．3．2 並列 I/O

MicroPE を各 PE に搭載したことにより MicroPE に対する I/O 処理が並列に実行可能となり、並列度を増した場合でも、このことに起因する性能低下がない。しかし、MicroPE にリードライトする情報（粒子座標、相互作用など）を局所的に処理出来るかどうかは、実行する粒子シミュレーションのアルゴリズムに依存しているため、実際のシミュレーションを走行させて、検証することが必要である。

3．2．3 O(N)処理の並列化

従来、ホストの CPU でシーケンシャルに処理していた、近接相互作用の計算、温度制御、圧力制御、粒子座標の更新などの処理を、マルチ CPU で並列実行できる。また、要素プロセッサのマルチスレッド機能を利用すれば、粒子座標などの通信や、専用計算パイプラインの I/O 処理と、これらの計算をオーバーラップさせることが可能である。

3．3．4 スケーラビリティ

これまでの粒子シミュレーション用並列計算機では、専用計算ハードウェアの部分を増設して、性能を上げる事が議論されてきたが、実際のアプリケーションでは、ホスト計算機での計算時間が無視できないケースも多く、性能の頭打ちがおきるか、粒子数が非常に多い場合でないと性能が出ないという傾向があった。本計算機では、専用計算パイプラインを搭載した PE と通常の PE を、ネットワーク上で自由に混在させることが可能で、その比率を設計時に決定することにより、ターゲット問題にあわせた装置構成が可能である。一方 Micro PE を搭載した

ノードの構成においては、ローカルバスの転送効率を考慮する必要があるものの、Micro PE の増設が容易で、従来の粒子シミュレーション用専用計算機のように、単純に専用計算ハードウェアを増設して性能向上することが可能である。

3．4 性能解析モデルにおける評価

本計算機における性能向上を文献４の性能解析モデルを使用して検証した。トータルの計算時間として、専用計算機で、粒子の相互作用を計算する部分を取り出し、比較した。性能解析で使用するバスの転送速度等のパラメータは、アーキテクチャの比較を行うという意味で原則同じとしたが、MD Chip レジスタ読み出しに関しては、ローカルに行える事から、ランダムライトの転送速度と同じとした。改善点は以下の通りである。

（1） 各 PE からブロードキャストで複製したデータを並列で与えることにより、粒子座標の転送時間からボード枚数のファクターを消去できる。

（2） 各 PE で I/O を並列に行うことにより、専用計算パイプラインのレジスタの書き込み時間からボード枚数のファクターを消去できる。

（3） 初期化は各 PE の CPU が担当するため、並列に実行でき、やはりボード枚数のファクターが消去できる。

上記の性能モデルに基づき、性能予測を行った結果を図２に示す。従来の MD One では頭打ちとなっていた 10PE 以上のシステムにおいても、PE 数に応じたパフォーマンスの向上が観察される。

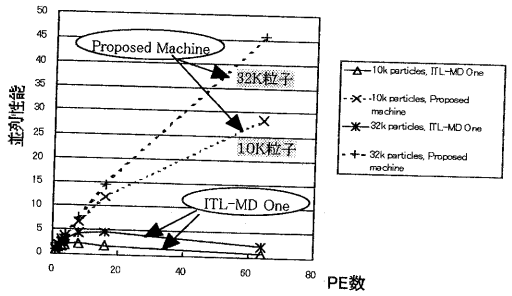

図２ 並列性能の比較

ITL-MD One に比べて改善されているとはいえ、並列性能の台数効果が、頭打ちになる傾向が見られるのは、個々の MicroPE に全粒子座標を転送しているためである。パイプラインによる計算と通信を並列で実行できる本計算機の場合は、計算する粒子を分担し、部分和を求めてゆく方式が、より効率的に実行できる可能性があり検討が必要である。

一方周期境界条件下でクーロン力を求める場合、Ewald 法などが使用でき、担当する粒子または波数ベクトルの局在化により、このオーバヘッドを低減できる。この場合、粒子座標の交換や、途中の計算結果の通信と加算などを、EM-X の細粒度通信機能を利用して、効率良く処理できることが期待されるが、性能解析モデル上での確認が必要である。また、EAM 法などの近接粒子のみ計算するケースについての性能解析も必要と考えられる。

4．プロトタイプの構築

4．1 プロトタイプの構成

本アーキテクチャの有効性を検証するため、MD One/E ボードを搭載した４つの専用計算ノード（以下 MD-EXP と称する）を作成し、80PE の EM-X に接続した。EM-X のネットワークと PCI バスを接続するためのインターフェーボード（EM-X ネットワークインターフェースボード、以後 NIB と称する）を開発した。本装置は 80PE の EM-X と、4 ノードの MD-EXP で構成されている。図３はシステムラック (NEXCOM 社 PCK550/400R) の写真で、CPU、NIB、MD One/E、10Base-T Ethernet ボードが３セット収納されている。MD One/E ボードは、NIB 経由で PCI バスに接続できるため、各 NIB に複数枚実装する事が可能であるが、プロトタイプでは、ノードあたり１枚の構成となっている。

プロトタイプでは、専用計算ハードウェ（MD One/E）が、EM-X のネットワークの I/O ノードとして接続されている点が、提案する計算機の構成と大きく異なる。その結果 MD Chip

の I/O 処理は、リモート PE が、ネットワークを経由して行うことになる。また、要素プロセッサ及び専用計算 LSI は既存のものを使用している。

図４に MD-EXP の構成図を示す。

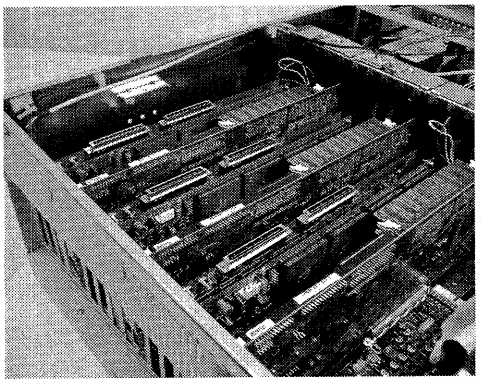

図３プロトタイプ外観図

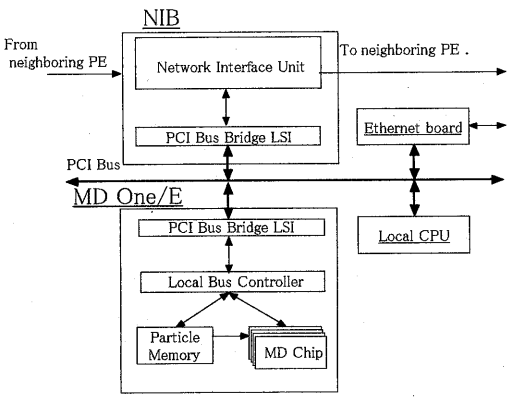

図４MD-EXP 構成図

（1）CPU ボード

CPU ボードは、PCI バスのシステムコントローラとしての機能と、初期化などの機能を担当する。CPU は MMX Pentium 233MHz で、64Mbyte の SDRAM と 4.3Gbyte の HDD を有する。OS は Vine Linux 1.1 (Kernel Version 2.0.36) を使用した。

（2）MD One/E ボード

MD One/ Eは、ITL-MD One/PCI のメモリを増設し、Linked Cell List のアドレッシング機能を搭載したものであり、4Gflops の Peak 性能と、200 万粒子の粒子メモリを持つ。MD One/ Eボードの詳細については、別の論文（報告）で詳細に述べる予定である。

（3） NIB

EM-X の Network に MD-EXP を接続するために開発したボードである。詳細は以下に説明する。

（4） Ethernet インターフェースボード

10Base-T の Ethernet Interface ボードである。メンテナンスバスの代用として使用する。

４．２ネットワーク接続

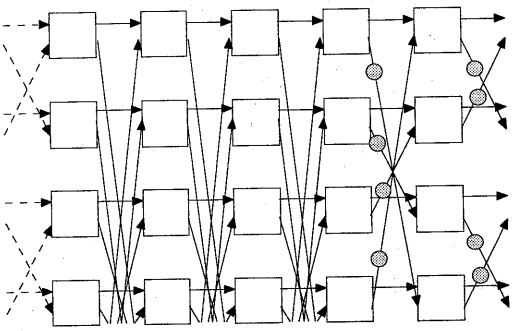

図５プロトタイプネットワーク接続

MD-EXP は EM-X のサーキュラオメガネットワーク上で、ケーブルで接続された任意のポイントに挿入することが可能である(図 5)。図では、EM-X の 80PE(16 グループ x 5 メンバ)のうち、0-3 グループのみが図示されている。図中の正方形は PE を表し、水平方向に並んだ 5 つの PE が一つのグループを形成して 1 枚の基板に実装されている。グループ間は、バックプレーン又はケーブルで接続されており、MD-EXP が挿入可能なケーブル接続の個所は円形で示されている。MD-EXP の NIB には空き番となっている適当なプロセッサ ID が選択され、設定される。メンテナンスネットワークは、接続されていない。

４．３ NIB

図６に NIB の外観を示す。NIB は、Network Interface Unit (NIU) と、PCI Bus Bridge で構成されており、EM-X のネットワークと、PCI バスのブリッジ機能を提供する。

NIU は、EMC-Y のネットワークインターフ

ェースを構成する３つの Unit (SU:Switching Unit、IBU:Input Buffer Unit、OBU:Output Buffer Unit)のサブセット及び、Local Bus Interface Unit(LBU)からなっている。LBU は、EMC-Y での Memory Control Unit(MCU)に相当する。

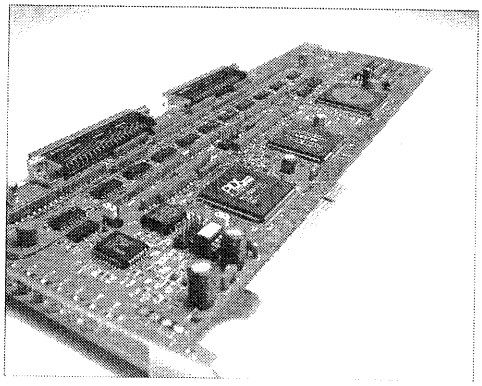

図 6 NIB 外観図

NIU は、FPGA(Altera 社製 Flex20K200) 上に実現されており、VerilogHDL 及び VHDL で記述されている。NIB の動作クロックは Network 側が 16MHz、Local Bu s側が 32MHz である。

４．４ 動作概要

EM-X のネットワークから見たバスアクセスは、SYSWR, SYSRD 及び DMA Read を使用して行われる。DMA Read は、NIB から任意の PE に対する SYSWR として実装した。DMA 起動は SYSRD により行われ、完了は Continuation として通知される。

PCI バスを経由した MD One/ Eボードへのアクセスのスループットはアクセス対象のメモリや、レジスタにより、若干変化するが、基本的には、EM-X のネットワークのスループットを阻害しないように設計されている。

４．５ソフトウェア

EM-X 上に MD-EXP 制御用のドライバ及びライブラリを移植した。MD-EXP 上のリソースにアクセスする部分では、インストラクション実行と、リモートメモリアクセスの性能差が少ない EM-X の特性を考え、ページングの判定をループの外側に出すなどの改良を行った。

５．おわりに

本論文では、既存の粒子シミュレーション用並列計算機の問題点を解決するために、汎用並列計算機 EM-X と粒子シミュレーション用並列計算機 ITL-MD One を統合した粒子シミュレーション用並列計算機のアーキテクチャについて提案し、その有効性について検討した。また、有効性を検証するため、80PE、4 ノードからなるプロトタイプを構築した。

今後は、プロトタイプ上で、各種のベンチマークを行ない、基礎的なパフォーマンスの評価を行うとともに、実際の粒子シミュレーションプログラムを移植して、トータル性能を評価する予定である。また、本アーキテクチャの検討を進め、より高速で高性能な、粒子シミュレーション用並列計算機の構築を目指す。

参考文献

[1] Junichiro Makino, Makoto Taiji, Toshikazu Ebisuzaki, Daiichiro Sugimoto, "GRAPE-4:A massively parallel special-purpose computer for collisional N-body simulations," The Astrophysical journal, 480: pp.432-446, 1997.

[2] （株）画像技研, "ITL-MD One 概説書," 1994.

[3] Kodama,Y., Sakane,H., Sato,M., Yamana,H., Sakai,S., and Yamaguchi,Y., "The EM-X Parallel Computer: Architecture and Basic Performance," Proc. 20th Int. Symp. on Computer Architecture, pp.14-23, 1995.

[4] 高田亮、本多弘樹、弓場敏嗣 "階層並列構造と演算チェインニング機構を持つ粒子シミュレーション用並列計算機の提案, " 情報処理学会研究報告 ARC-134-22(1999)、pp.127-132.