次世代マルチコアプロセッサ開発のための PS3 クラスタシステムの構築

篠原 啓志 † 大津 金光 † 横田 隆史 † 馬場 敬信 †
† 宇都宮大学工学部情報工学科

1 はじめに

近年，シングルプロセッサの性能向上は消費電力や発熱などの問題から頭打ちとなっており，プロセッサの更なる性能向上を目指すために 1 つのチップ上に複数のプロセッサコアを搭載するマルチコア化が進んでいる．次世代マルチコアプロセッサではコア数が増大したメニコアになり，各コアの内部構造も複雑化していくと予想される．次世代マルチコアプロセッサの開発に必要な評価データを集める時，すべてのコアのシミュレーションを内部構造まで詳細に行うと膨大な時間がかかるという問題がある．本研究では，次世代マルチコアプロセッサの評価にチップあたりの処理性能の高い PS3 を使用することを提案する．PS3 を複数台接続したクラスタ構成にすることでアーキテクチャの評価に要する時間の短縮を図る．PS3 クラスタを用いた計算機合成ホログラムの研究 [1] は存在するが，我々はこれを次世代マルチコアの開発に活用できないかと考えた．

2 PS3 クラスタ使用検討

マルチコアプロセッサとして 1 チップあたりの処理性能が跳び抜けて高い Cell プロセッサ (Cell Broadband Engine) に着目した．Cell プロセッサを搭載したソニーコンピュータエンタテイメント社の PS3 (PLAYSTATION 3) を用いる．価格対性能比において PS3 は市販 PC よりも優れており，高い演算処理能力を安価に入手することができる．

安価に高い演算処理能力を手に入れられる他の方法として GPGPU がある．GPGPU は多数のストリーミングプロセッサを持ち，数単位にグループ化したものに同じ命令を発行して処理を行うものだが，条件分岐が異なるケースで大幅に性能効率が落ちるのでプロセッサシミュレーションに向いていないと考えられる．

またプロセッサシミュレーションの別の方法として FPGA 上にハードウェアとして実現して動かす方法がある．しかし，FPGA は集積可能なハードウェア資源の制限からプロセッサコア数が多いものは実現できないという問題がある．また，ハードウェアの開発自体に多大なる時間的コストを要する．以上より次世代マルチコアプロセッサの評価環境として，PS3 クラスタは有用であると考えられる．

3 PS3 クラスタシステム

PS3 クラスタシステム構成を図 1 に示す．PS3 クラスタはフロントエンドの PC 1 台，バックエンドの PS3 16 台から構成され，それぞれ 1000Base-TX のスイッチングハブで接続されている．また，各ノード間の通信に MPI (Message Passing Interface) を用いる．

3.1 PC

PS3 クラスタシステムのフロントエンドに PC を用いる．PC はクラスタシステムでバックエンドである

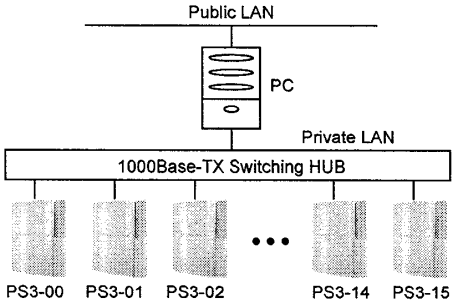

図 1: クラスタシステム構成

PS3 の各ノードとファイルの配布・収集，ジョブの投入・管理を行う．Cell クロスコンパイラ環境を構築して Cell アプリケーションの開発を PC 上で行うことができる．主な仕様は表 1 の PC の欄に示す通りである．

3.2 PS3

クラスタのバックエンドとして価格対性能比に優れている PS3 を用いる．PS3 は高い演算処理性能を持つ Cell プロセッサを搭載している．Cell プロセッサはメインプロセッサコアの役割を持つ PPE (PowerPC Processor Element) とサブプロセッサの役割を持つ SPE (Synergistic Processor Element) 複数個が高速なバスである EIB (Element Interconnect Bus) で接続されている．PPE は 64 ビット PowerPC アーキテクチャの汎用プロセッサであり，OS の駆動やアプリケーションの実行，及び SPE の制御を行う．SPE はデータ演算処理に特化したシンプルなプロセッサコアであり，強力な SIMD 演算能力を持つ．SPE はそれぞれ 256KB のローカルストアと呼ばれる専用メモリを持ち，メインメモリへのアクセスはローカルストアを介してのみ行うことができる．PS3 の Cell プロセッサは PPE，SPE 共に 3.2 GHz で動作し，Fedora 9 から 1 個の PPE と 6 個の SPE を扱うことができる．メインメモリとして 256MB の XDR DRAM を搭載しており，ネットワークはオンボード NIC である．開発環境として，CellSDK3.1 と MPICH2 をインストールして用いる．

表 1: PS3 クラスタの環境

ノード	PC	PS3
CPU	Pentium 4	Cell Broadband Engine
動作周波数	3.2 GHz	3.2 GHz
メモリ	DDR2 SDRAM 4.0 GB	XDR DRAM 256 MB
OS	Fedora 10 / x86	Fedora 9 / ppc
ネットワーク	オンボード NIC Broadcom BCM5751	オンボード NIC
開発環境	CellSDK 3.1	CellSDK 3.1
通信ライブラリ	MPICH2 1.0.8	MPICH2 1.0.8
ハブ	FXG-24IRM (24 ポート 1000Base-TX HUB)	

3.3 システムの特徴

本クラスタは PS3 のユーザアプリケーションが使用可能なメモリ領域を多く確保するためにクラスタによく用いられる NFS などのネットワークファイルシステ

Construction of a PS3 Cluster System for Development of Next Generation Multicore Processors
†Keishi Shinohara, Kanemitsu Ootsu, Takashi Yokota and Takanobu Baba
Department of Information Science, Faculty of Engineering, Utsunomiya University (†)

ムを使用しない．PC 上で開発されたアプリケーションやデータファイルは各 PS3 ノードにリモートコピーすることで一括して配布する．

4 PS3 クラスタでの並列処理

次世代マルチコアプロセッサはコア数が増大してメニコアとなるが，それらすべてのプロセッサコアを高速かつ詳細にシミュレーションするためには，例えばプロセッサコア 1 個のシミュレーションを 1 個の SPE に割り当てるなどの並列化を行う必要がある．本節では PS3 クラスタシステムにおける並列処理方法を説明する．

4.1 全体の流れ

まず，フロントエンドの PC からバックエンドの各 PS3 にジョブを投入する．投入されたジョブは MPI によって各 PPE で並列に実行される．各ノードの PPE は主に SPE の管理，各ノードとの通信を行い，与えられたジョブを分割して各 SPE プログラムを実行する．SPE で加工されたデータはホストのメインメモリに書き戻すことで PPE が扱えるようになる．各 PPE に SPE によって書き戻されたデータを 1 台の PS3 に集めることで分散した処理結果をまとめ，最終的な結果を得ることができる．

4.2 MPI を用いた並列処理

MPI は分散メモリ型の並列計算機で複数のプロセス間でのデータをやりとりするために用いるメッセージパッシングライブラリである．MPI を用いたノード間での並列処理を行う．まず，各ノードの PPE は MPI によって 0 から始まる整数の番号が割り当てられる．この番号によってジョブの分割，処理の分岐，メッセージの送信先や受信先を指定することができる．本稿の PS3 クラスタシステムでは MPI による並列処理で最大 16 ノードを使用可能である．

4.3 SPE での並列処理

ノード内での SPE コア間による並列処理を行う．各ノードの SPE は MPI による並列化ができないため，PPE からコントロールしなければ扱うことができない．PPE から SPE を扱う時に必要になるのが SPE 開発用ライブラリである libspe2 である．libspe2 は CellSDK に含まれている．SPE での並列処理で 6 コアまで使用することができ，MPI による並列化と合わせると最大 96 コア使用可能である．

4.4 SIMD 演算による並列処理

SIMD 命令を用いた演算での並列処理を行う．SIMD 演算は複数のデータに対する処理を 1 命令で行う演算であり，同じ処理を少ない命令数で実行することで処理時間の短縮を図ることができる．SPE は 128bit の SIMD 演算ユニットであるので，通常 32bit データを扱う場合 4 並列で演算を行うことができる．

5 評価

プロセッサシミュレーションでは，整数演算処理能力が極めて重要になる．そこで本稿では，PS3 クラスタシステムの整数演算処理能力を評価するために N-Queens 問題を用いる．N-Queens 問題は，N × N のチェス板に N 個の Queen をお互いに攻撃できないように配置する仕方がいくつあるか総数を求めるプログラムである．また，ノード間通信性能として 1 対 1 通信でのレイテンシの評価を行う．一般に，メッセージパッシングによるノード間通信はノード内での PPE や SPE といったプロセッサコア間の通信に比べると圧倒的に低速であるため，本クラスタの通信性能を評価する．

5.1 N-Queens による整数演算性能

N-Queens 問題を MPI を用いてノード間で並列化，ノード内のコア間で並列化を行い，本クラスタシステムの整数演算性能を評価する．PC プログラムと PS3 プログラムは共に最適化オプション-O3 を適用した．本クラスタの具体的な実行手順は次の通りである．

1. MPI からの番号で各 PPE に処理データを分割
2. メインメモリの処理データを SPE 6 個分に分割
3. 各 SPE コンテキストを起動
4. メインメモリからローカルストアへの DMA 転送
5. SPE でデータを処理し，結果を PPE に転送
6. PPE で各 SPE での結果を受け取る
7. 1 個の PPE へ結果を集める

18 Queens の結果を表 2 に示す．SPE の苦手とする条件分岐の多いアプリケーションであったが，PS3 ノード数 1 においては 4.6 倍，ノード数 16 においておよそ 48 倍 PC に対して高速に実行できることを確認した．また PS3 ノード数 1 に対してノード数 16 では 10 倍以上高速化した．アプリケーションの並列化効率が高いことから，さらにノード数を増やしても速度向上を達成できると考えられる．

表 2: 18 Queens の処理時間

	PS3		PC
計算ノード数	1	16	
処理時間 (秒)	112.22	10.89	519.78

5.2 ノード間通信性能

PS3 ノード間においてノード間通信の基本となる 1 対 1 通信でのレイテンシを計測する．計測方法はピンポン通信プログラムを用いて行い，2 ノード間で 4 バイトのデータを 10000 回送信・受信を繰り返すのに要した時間を 100 回計測して，その平均値を求める．この値を 2 で割ることで 1 回あたりの通信に要する時間を求める．

計測した時間の平均は 4018.91 ミリ秒であった．この値から 1 回あたりの通信時間は 0.20 ミリ秒と算出される．この値は 2 ノード間のみの通信時間であるため，多ノードが同時に通信を行い，ネットワークが混雑した場合は通信レイテンシが大きくなると考えられる．

6 おわりに

本稿では，次世代マルチコアプロセッサ開発のための PS3 クラスタシステムを構築し，その整数演算性能と基本的なノード間通信性能の評価を行った．PS3 クラスタシステムは並列化によってその性能を引き出すことができ，メニコアにおけるプロセッサシミュレーションでも並列化を行い，整数演算処理能力を発揮することで高速なシミュレーションが行える可能性を示した．

今後の課題として，マルチコアプロセッサシミュレータを本クラスタシステム上で並列に実行するための並列化の検討と性能評価を行うことが挙げられる．

謝辞

本研究は，一部日本学術振興会科学研究費補助金（基盤研究 (B)18300014，同 (C)19500037，同 (C)20500047）および宇都宮大学重点推進研究プロジェクトの援助による．

参考文献

[1] 荒井大輔，“Cell を用いたクラスタシステムによる計算機合成ホログラムの高速化”，Cell スピードチャレンジ 2008.