社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－ARC－167 (3)
2006－HPC－105 (3)
2006／2／27

メモリ周りに制約を有するMPUにおけるプリフェッチ機能付メモリモジュールの意義

田邊 昇[†] 羅 徴哲[††] 並木 美太郎[††]
中條 拓伯[††] 天野 英晴[†††]

CPUやFPUをチップに大量に詰め込むことは技術的ハードルは低く、発熱問題を主因としてその流れは今後のトレンドとしてしばらく続くと思われる。一方、メモリ周りはその進歩に十分についていくのが困難になってくる。その典型的な例としてCell Broadband Engine(CBE)をHPC応用に適用するケースを取り上げて説明し、その解決策の一つとしてDIMMnetを位置づける。DIMMnet-3をCellリファレンスセット(CRS)に装着することにより予測される効果を考察した結果、使わない場合に対して最大約20万倍という性能差がつくケースが想定できることを示す。

Meaning of a Memory Module with Prefetching Functions for MPU with restrictions on Its Memory System

NOBORU TANABE,[†] ZHENGZHE LUO,[††] MITARO NAMIKI,[††] HIRONORI NAKAJO[††]
and HIDEHARU AMANO [†††]

A technical hurdle of integrating chip with a lot of CPUs and FPUs is low. It seems that such a tendency continues as a future trend for a while because of a heat problem. On the other hand, it becomes difficult that memory system follows the progress enough. As the typical example, a case to apply Cell Broadband Engine (CBE) to HPC applications is mentioned. We characterize DIMMnet as one of the solutions of the problem. As a result of having estimation for an effect by plugging DIMMnet-3 to Cell Reference Set (CRS), up to about 200,000 times acceleration ratio can be assumed.

1. はじめに

Graphic Processor Unit(GPU)は汎用PCに用いられるものでも、例えばnVidia Geforce7800GTXでは300GFLOPSを超えており、ゲーム機に用いられるGPUは1TFLOPSを超える。GPUを画像処理以外に用いる試み(GPGPU:General Purpose Computation on GPU)も各所で行なわれているものの、倍精度演算が全くできないということは、一般的なHigh Performance Computing(HPC)用途にとっては実用上大いに問題がある。

一方、Cell Broadband Engine(CBE)のように、最新のMPUでは複数のCPUコアを内蔵し、それらの32bit浮動小数演算性能は100GFLOPSを超えている。GPUと同様の理由でこれらはHPC用途にはほとんど寄与しない。

しかし、パイプライン化されていないものの倍精度浮動小数演算器をCBEは多数有しており、それらの合計はベクトル型スーパーコンピュータSX-8[2]やCray X1[1]の単体プロセッサ性能を超える素晴しいものであり、CBEをHPC用途に使うことは工夫次第では有望であると考える。

このようにCPUやFPUを1チップに大量に詰め込むことの技術的ハードルは低く、既に実用化されている。Intelの周波数向上主義からマルチコア化への方針転換でも決定的になったように、発熱問題を主因としてその流れは今後のトレンドとしてしばらく続くと考えられる。

しかし、メモリ周りは容量とバンド幅の両立を、低コストで実現していくことは難しい。MPUが多用される用途ではそのどちらかをないがしろにしても問題無い性能が出てしまうことが多々あるため、HPC用途を主目的にしないMPUではバンド幅やメモリ容量などがコストとの兼ね合いから制約を温存してしまう。

よって、チップ内の肥大化傾向にある演算能力をHPC用途でメモリシステムが引き出せるようにしていくことが課題になってくる。その典型的な例としてCBEをHPC応用に適用するケースをあげることができる。CBEを用いた東芝のCellリファレンスセット(CRS)[6][7]では25.6GB/sという高めの主記憶バンド幅、512MBという低めの主記憶容量を有するが、その容量を超えるアプリケーションを実行し始めた瞬間、仮想記憶によるページングのためのハードディスク(HDD)に対するランダムアクセスによって実用にならなくなるほど失速する、いわゆるスラッシング状態に陥ることは必至である。

CRS本来の設計意図からすれば倍精度のFPUを主目的に使うことや、512MB以上の主記憶を要求するアプリケーションの実行は想定外の使用方法であろう。しかしPCクラスタとは元々文房具に過ぎなかったパーソナルコンピュータ(PC)を、想定外のHPC用途に使えるように工夫する技術であった。PCクラスタの歴史がそうであったように、本論文ではCBEに代表される新型MPUの想定外の使用をする方法について検討する。

本論文では、まず今後のMPUのトレンドを考察し、現在から近未来にかけてのMPUをHPC用途に利用する際に直面すると考えられるメモリシステム上の制約を示す指標を列挙する。その指標の具体的なシステムにおける状況を確認する。その上で、CRSのHPC応用転用支援策としてiRAM[11]とDIMMnet[12][13][14][15]の利用を位置づける。次にその予測される効果を述べる。DIMMnetをCRSに適用するための改造方法について述べ、最後にまとめる。

2. 予測される今後のMPUのトレンド

これまでのMPUはPentium4に代表されるように、1個のCPUコアの性能を多くの電力とトランジスタをつぎこんで高速化するというアプローチをとってきた。しかしIntelの周波数向上主義からマルチコア化への方針転換でも決定的になったように、発熱問題を主因としてマルチコア化[16]は今後のトレンドとしてしばらく続くと考えられる。GPUには合計で1.8TFLOPSものFPUを搭載している例[8]もあるように、CPUやFPUを1チップに大量に詰め込むことの技術的ハードルは低く、既に実用化されている。マルチコア化が進展する

† (株)東芝、研究開発センター
Corporate Research and Developpment Center, Toshiba
†† 東京農工大学
Tokyo University of Agriculture and Technology
††† 慶應義塾大学
Keio University

とこれまで演算器以外の部分に割り当てられていたリソースが演算器に割り当てられるようになるので、FLOPS に代表される演算能力指標は飛躍的に向上していく傾向をみせるものと考えられる。

3. メモリシステム上の制約を示す指標

マルチコア化が進展すると FLOPS に代表される演算能力指標は飛躍的に向上する。それに対してメモリシステムが、とりわけ HPC 用途においては十分についていけなくなってくる。本章では、現在から近未来にかけての MPU を HPC 用途に利用する際に直面すると考えられるメモリシステム上の制約を示す指標について述べる。

3.1 FLOPS あたりの主記憶容量

主記憶バンド幅を確保しようとするとメモリの周波数を上げることになるので、ここ数年メモリの周波数は毎年のように上がってきた。しかし、現状の主流である DDR や DDR2 はバス構造の配線上で一回しか分岐できない。これ以上の周波数を目指そうとすると、XDR DRAM や GDDR3 のように分岐構造が許されないポイント to ポイント接続にならざるをえなくなる。ピン数には限りがあるので自ずと主記憶容量に厳しい制約がかかってくる。

3.2 FLOPS あたりの主記憶バンド幅

ともかくバンド幅は上記の手法を使うなどして確保しておかないと、ストリーム処理のようなあまり大容量のメモリを必要としないケースにおいても性能が十分に出ないことになるので演算能力に応じたバンド幅が優先的に割り当てられることになる。よって FLOPS あたりの主記憶バンド幅が維持できていないシステムは GPU のように演算器がパイプライン的にチェインされて主記憶を経由しなくても演算が続行できる場合に限定される。ここでは HPC 用途を考えるので FLOPS は倍精度の浮動小数演算とする。

3.3 主記憶より下の階層の半導体記憶容量

十分に FLOPS あたりの主記憶容量が維持されていないシステムでは FLOPS 値に見合った実行時間を求めて大きな HPC 応用を実行させようとすると、データが主記憶より下の階層に溢れ出してくる。スーパーコンピュータでも半導体拡張記憶を I/O に接続してこの問題に対応するケースがあった。その際、その半導体記憶容量がさらに足りないと、HDD まで速度限定要因が落ちてくるので、主記憶より下の階層の半導体記憶容量が重要になる。

3.4 主記憶より下の階層のバンド幅

主記憶から溢れた HPC 応用では主記憶より下の階層のバンド幅で性能が律速される。このバンド幅は I/O インタフェースの最大転送速度で近似的に制約を代表させる。半導体記憶への連続的なアクセスとする場合に限り、この制約は実行性能を規定する。

3.5 主記憶より下の階層のレイテンシ

上記のバンド幅は通常の仮想記憶のページング動作における HDD においては得られず、実際に得られる実効バンド幅は主記憶より下の階層のレイテンシも併せて考える必要がある。

3.6 TLB がカバー可能な空間サイズ

大きな多次元配列を様々な方向にアクセスする場合など HPC 応用では TLB ミスやページフォルトが頻繁に発生するケースがあり、そのような状況を考慮せずに設計されたプロセッサにおいては HPC 応用で思ったほど性能が出ないという結果になる。その現象に関する指標としては TLB がカバー可能な空間サイズによって MPU の大容量メモリに対する設計ポリシーが浮き彫りになる。

4. 具体的システム上での指標状況確認

表 1 に具体的システム上での前章で述べた指標の状況を示す。

4.1 ベクトル型スーパーコンピュータ

ベクトル型スーパーコンピュータは従来の HPC 向け計算機の主役であったが、最近では PC クラスタや BlueGene/L 等の並列機に押されている状況にある。しかし NEC と Cray の 2 社によって現在も新製品が開発され市場に投入されている。その理由は十分に投資されたメモリシステムに支えられ、目先のピーク FLOPS よりも自分が抱えているアプリケーションの実効性能やプログラマビリティを優先する人々のニーズを掴み続けているからと考えられる。これらのマシンは過去からの FLOPS とバンド幅や主記憶容量のバランスを維持している。問題は狭くなっている市場と高騰する開発費から価格を安くすることができない点にある。よってベクトル型スーパーコンピュータの指標に近いものを安価に作ることが目標になる。

4.2 COTS PC 向け MPU

COTS PC 向け MPU、とりわけ Intel の Pentium4 Xeon がここ数年の HPC プラットフォームでは PC クラスタとして大きな比率を占めてきた。これまでは多くの台数を安定稼働させることが優先されていたので ECC 等に対応しているサーバー向けの Xeon が用いられることが多かった。今後は Dual コア化された Pentium D や AMD の Athlon64 といった MPU もその高いパフォーマンスおよび高いコストパフォーマンスにより、HPC 用途でも単体利用や少ない台数のクラスタに一層使われていくと考えられる。

その代表として Pentium D の指標状況を表 1 に載せてある。COTS は改良頻度が高いので部分的にはサーバー系よりも一見性能が高く見えることがある。しかし、DTLB がカバーできる空間の大きさやキャッシュのサイズなど COTS 向けの製品として節約されている部分があるため、ここに示されている数字が全てを表していないことに注意すべきである。

4.3 サーバ向け MPU

サーバ向け MPU は COTS PC 向け MPU と比較していくつかの点で HPC に向いている性質を持つ。ここでは TLB に焦点を当てて解説する。

4.3.1 Intel Itanium 2

Intel の Itanium2 は近年 Top500 の上位を占めるシステム上で徐々にそのシェアを上げてきている。その主たる理由はキャッシュサイズが大きいことによる高い実効性能や主記憶サイズが大きな製品が得やすい点と、TLB のページサイズが可変であり最低 4KB から最大 4GB という巨大なページサイズをサポートしていることから、大容量のメモリを必要とする HPC 用途に向くためと考えられる。

特に TLB は二階層になっており L1DTLB が 32 エントリ、L2DTLB が 128 エントリと COTS PC 向けの MPU と比べ多くのエントリを有する。4GB のページサイズを用いれば 512GB もの空間を L2DTLB のミスなしでアクセスできる。しかし現実にはページフォルト時のダメージやメモリ利用効率の観点からそのような使い方ができるケースはまれと考えられる。

4.3.2 富士通 SPARC64 V

SPARC64 V[9] の TLB は、mTLB(main TLB) と μTLB(micro TLB) の二階層から構成される。ページサイズは 8K バイト、64K バイト、512K バイト、4M バイトの 4 種類をサポートする。mTLB は、RAM 部と CAM 部から構成される。RAM 部は、1024 エントリ、2 ウェイ・セットアソシアティブ構成である。RAM 部は 8K バイトページ専用である。CAM 部は、32 エントリのフルアソシアティブ構成である。8K バイト以外のページサイズのエントリは CAM 部に登録する。μ TLB は L1 キャッシュをアクセスするパイプラインから高速に索引するための mTLB の部分コピーで、32 エントリのフルアソシアティブ構成である。

よって、8KB のページが 1024 エントリで 8MB,32 エントリの μ TLB や mTLB が 4M バイトのページサイズでカバーする空間サイズは 128MB で、合計 136MB であり、Itanium2 ほど極端ではないが COTS 用 MPU よりは格段に大きい。

4.4 Cell Broadband Engine

CBE は IBM,Sony,Sony Computer Entertainment, 東芝の 4 社が共同で開発したヘテロ型マルチコアプロセッサである。8 個の独立した浮動小数点演算コア (SPU) と Power ベースのコア (PPU) を持ち、4GHz を超えるクロックスピード（動作周波数）と表 1 に示されるようなスーパーコンピュータ並みの浮動小数点演算性能を実現している。そのメモリシステムは XDR DRAM をポイント to ポイントで直結するという極めて GPU に近い作りとなっている。

CBE の本論文において取り上げている指標上の特徴は、FLOPS あたりの主記憶容量において顕著である。他のシステムが概ね 1 近辺にあるのに対し、CBE は倍精度の場合でもその 1/50 に過ぎない。ここに CBE が多くの HPC 応用においてそのままでは利用不可能な問題があり、本論文が解決していくべき問題の原点がある。

リアルタイムに映像程度のバンド幅のデータが流れていく応用では主記憶はバッファとして機能すれば良くなるので、この

表 1 具体的システム上での指標状況

	Cray X1E[1]	SX8i[2]	Itanium 2	Pentium D	Cell reference set[6][7]	nVidia RSX[8]
COTS?	Custom	Custom	Server	COTS	COTS	GPU
周波数	1.13GHz	2GHz	1.6GHz	3.2GHz	4GHz[3][4]	550MHz
ピーク FLOPS(64bit)	18GFLOPS	16GFLOPS	7.2GFLOPS	12.8GFLOPS	>26GFLOPS[3][4]	0FLOPS
ピーク FLOPS(32bit)			7.2GFLOPS	25.6GFLOPS	>256GFLOPS[3][4]	1.8TFLOPS
主記憶バンド幅	34GB/s	64GB/s	10.7GB/s	10.7GB/s(i955)	25.6GB/s	22.4GB/s
主記憶容量	16GB	16GB	8GB	8GB	512MB	256MB
主記憶容量/FLOPS	0.84	1	1.11	0.62	0.02	n.a.
主記憶バンド幅/FLOPS	1.89	4	1.49	0.84	1	n.a.
下層の半導体記憶容量	半導体拡張記憶?	半導体拡張記憶?	0	0	512MB(SO-DIMM)	0
下層のバンド幅	4.8GB/s	3.2GB/s	150MB/s	300MB/s	2.6GB/s(DDR2) 133MB/s(IDE)	?
下層の遅延	?	?	10ms	10ms	?(DDR2) 10ms(IDE)	10ms
DTLB がカバー可能な空間	?	?	512GB	256KB	?	n.a.

問題は多くの場合顕在化しない。しかし HPC 応用のように 3 次元空間の全状態を保持し時間発展するようなケースではこの問題が顕在化する。

4.4.1 Mercury CBE ベースブレード

Mercury CBE ベースブレード[10] は 1 枚のブレードに 2 個の CBE を搭載し、それらが相互に FLEXIO によって 20GB/s 全二重接続されている。ブレード間は PCI Express x4 経由でブレードあたり 2 枚の Infiniband 4X HCA により接続される。拡張用のメモリソケットは無い。そのため DIMMnet を装着する手段は無い。

4.4.2 東芝 Cell リファレンスセット

東芝 Cell リファレンスセット (CRS) は 2006 年 4 月以降に販売予定である Cell を搭載したデジタルメディア機器などのソフトウェア開発用のキットである。I/O ブリッジとして各種音声/画像インターフェイスを搭載した Super Companion chip を使用している。HPC 用途にはそれらのインタフェースはほとんど意味をなさないが DDR2 の SO-DIMM ソケットを 1 個搭載しているので DIMMnet を装着できる可能性がある。HDD 用に IDE インタフェース (ATA133) が搭載されているので主記憶や SO-DIMM からはみでるような大きなデータを扱う HPC 用途では IDE 経由の HDD がボトルネックとなる。

4.5 GPU

GPU の特徴は単精度の浮動小数演算能力が桁違いに高く、それに対して主記憶容量はかなり少ない。主記憶 (ビデオメモリ) は nVidia GeForce7800GTX のように CBE 同様の 512MB を持つものもある。ここに掲げた指標で見る限り前記の CBE は GPU に近い性質を有する。

5. CBE の HPC 応用転用支援策

本章では近未来の MPU のトレンドを色濃く反映していると考えられる MPU として CBE をとりあげ、これを HPC 用途に転用することを目指して、改良を加える具体的方策を考察する。

5.1 CBE の HPC 転用の意義

CBE は単精度浮動小数演算性能が 256GFLOPS と著しく高いことは有名で、この影になってパイプライン化されていない倍精度浮動小数演算は目立たなかった。むしろ単精度と倍精度の性能差に目を奪われ、倍精度が 10 倍遅いから CBE は HPC 用途には向かないとする意見も良く聞かれる。

しかし、周波数の高さと演算器の個数の多さから Cray X1 や NEC SX-8 の要素プロセッサよりも CBE の倍精度浮動小数演算性能は高く、CBE を HPC 用途に使うことは工夫次第では有望であると考える。

5.2 CBE の HPC 転用時の課題

科学技術計算においては主記憶バンド幅が極めて重要で、性能のボトルネックになる。主記憶容量が足りないと通常は仮想記憶を通して HDD へのアクセスとなるために実効的な主記憶バンド幅が劇的に低下する。

1FLOPS あたり 1B/s というベクトルマシン上で維持されてきたバランス指標からすれば CBE の主記憶は 25.6GB/s なので倍精度浮動小数演算能力とはバランスしているかに見える。しかし東芝 CRS では 512MB という GPU 並みの CBE のメモリシステムに対して、1FLOPS あたり 1 バイトというもう一つのバランス指標と照らし合わせれば、20GFLOPS 程度でも演算能力が多すぎると言える。つまり、CBE の HPC 応用においては演算能力ネックが生じることは少なく、メモリシステムの改善がそのまま実効性能に現れるはずであることを意味する。

CRS の主記憶は 512MB しかないので、それ以上の大きさの配列をアクセスするような HPC 用途ではスラッシングのため殆ど実行不能な状態に陥ることは必至であり、その改善が CBE の HPC 転用においては最大の課題である。

そこで、その課題に対し比較的簡単に採用できる解決策として、以下の二つのハードウェアを利用することを考える。

5.3 i-RAM の活用

5.3.1 i-RAM の概要

i-RAM[11] とは GIGABYTE 社が出荷している SATA インタフェースによって 4 本の DDR DIMM をアクセス可能にしたバッテリーバックアップ付のシリコンディスクである。最大搭載可能メモリ容量は 4GB である。PCI バスを支えおよびスタンバイ電源供給用に用いる。

5.3.2 スワップファイルの高速化

i-RAM は SATA なので SATA〜IDE 変換器を用いれば CRS にも装着が可能と考えられる。i-RAM は装着するホスト (例えば CRS) から見ればハード的には IDE の HDD として見えるので特にドライバは不要である。HDD にスワップファイルを作る要領で i-RAM によりスワップファイルの高速化ができると考えられる。特に遅延時間が HDD と比べ劇的に短くなるので、ページング時の性能低下は劇的に改善するはずである。ただし、バンド幅は SATA の 1.5Gbps(187.5MB/s) が上限なので、これを主記憶のつもりでアクセスさせれば HPC 応用では性能低下は不可避と考えられる。

5.4 DIMMnet の活用

5.4.1 DIMMnet の概要

DIMMnet は DIMM スロットに装着される PC クラスタ構築用ネットワークインタフェースである。DDR に対応した DIMMnet-2[12][13][14] からはベクトル型のプリフェッチ機能を有するメモリモジュールを兼ねるようになった。現在開発中の DIMMnet-3[15] は DDR2 スロットに対応する。よって、CRS にも利用可能することをめざしている。搭載メモリ容量は 4GB の DIMM を 4 枚まで装着して 16GB まで実装可能としている。不連続アクセスを連続化してキャッシュベースの MPU では非常に効率が悪かった NAS CG[13] 等の HPC 応用や主記憶データベース検索[14] において実効的なバンド幅を大幅に改善できる機能を有する。

5.4.2 RAM ディスクドライバを介したスワップ利用

CBE のメモリシステム改良における第一の DIMMnet の利用方法は、DIMMnet を i-RAM のようにスワップファイルの置き場所として用いる方法である。DIMMnet はホストからハード的には DIMM(主記憶) に見えているため、そのままではスワップファイルを置くことができない。

そこで DIMMnet では DIMMnet 上のメモリ領域の全部または一部をハードディスクとして見せかける RAM ディスクドライバを提供し、その仮想化されたディスク上にスワップ領域

を確保する。

このソフトウェア的改良は既に Pentium4 ベース PC 上の DIMMnet-2 においては完了しており、実機上での動作が確認されている。その詳細は別途報告[17]されることとなっている。

CRS において何らかの制約があるかどうかの調査と実装は今後の課題であるが、Linux ベースで CRS が動作している現状から、移植の実現可能性は高いと考えている。

5.4.3 不連続アクセス連続化機能利用

CBE のメモリシステム改良における第二の DIMMnet の利用方法は、DIMMnet-2 提唱以来の本来の使用方法である、不連続アクセス連続化機能の利用である。

本機能は、DIMMnet 上の大容量半導体記憶 (DIMMnet-2 上では 2 枚の DDR 型 SO-DIMM) 上に確保された配列に対し、等間隔または間接ベクトルロードストアコマンドを用いて DIMMnet 上のベクトルレジスタに使用するデータのみを収集 (gather) したり、ベクトルレジスタから大容量半導体記憶上の不連続な位置に分配 (scater) 格納できる。

よって、DDR、DDR2 とバースト長が延びる傾向にある DRAM の進化とともに拡大する主記憶直前の階層のキャッシュのラインサイズの増加にともなう悪影響を不連続アクセスが多用される応用において軽減する。たとえば本来 128 バイト単位でアクセスしなければならないキャッシュ可能領域に対して 4 バイト分しか使うデータが無いような不連続アクセスをした場合、実効的なバンド幅は 32 倍に向上する。

CBE の汎用プロセッサである PPU(Power Processor Unit) はキャッシュベースであるため Pentium4 同様に上記の理論がそのまま適用できる。一方、CBE の演算処理の中核を担う SPU(Synagetic Processor Unit) はキャッシュベースではない。しかし、SPU のローカルメモリと主記憶間の DMA 転送を 128 バイト単位で実行するのも 1 バイト単位で実行するのもほとんど同じ時間がかかるため、上記の理論がほぼそのまま SPU にも適用できると考えられる。

一方、Pentium4 ベースの PC に DIMMnet-2 を装着した場合は、NAS CG ベンチマークや Wisconsin ベンチマークにおいてキャッシュラインをフラッシュする命令を実行して一貫性を保持した場合は 1/3 から 1/2 に加速率が低下する現象が報告されている。CBE の SPU においてはキャッシュベースではないためこのような一貫性をとる必要がなく、上記の性能低下は生じない。ただし、それ以外の隠れた性能低下要因があるかどうかの確認は今後の課題である。

6. 期待される効果

本章では DIMMnet-3 を CRS に装着しない場合に対する装着した場合の前述の評価指標への効果について考察する。

6.1 FLOPS あたりの主記憶容量

厳密に言うと主記憶容量そのものは変化しないが、仮想記憶を介して HDD をアクセスするのと比較して主記憶アクセスに近い性質（バンド幅およびレイテンシ）を有する半導体記憶容量が DIMMnet-3 を装着した場合、ノードあたり最大 16GB 増加する。この時点で素の CELL(512MB) に比べて、本指標は 32 倍の改善がなされる。

さらに、DIMMnet-3 は DDR 版 Infiniband4X(2GB/s) で互いに接続され、ローカルの SO-DIMM 領域とリモートの SO-DIMM 領域をアクセスするのにそれほど大きな差は無い。

既に DIMMnet-2 上ではリモートノードの SO-DIMM をソフト的にはローカルの主記憶に見せるソフトが完成している。そのような意味からは DIMMnet-3 を用いる場合は 16GB × ノード数が主記憶容量に相当すると見ることも可能である。

6.2 FLOPS あたりの主記憶バンド幅

厳密に言うと主記憶容量そのものは変化しないが、NAS CG ベンチマークや主記憶データベースに対する Wisconsin ベンチマークのように不連続アクセスを多用するアプリケーションから見た場合の SO-DIMM へのバンド幅は DIMMnet-3 の不連続アクセス連続化機能を用いると大幅に向上する。

具体的には、CBE の場合は 128 バイトのアクセスを行うのも 1 バイトのアクセスを行うのも同じバンド幅を消費するので、最も効果が顕著な場合 (4 バイトの属性値のフルスキャンを主記憶データベースに対して行なう場合に見かけ上のバンド幅は 32 倍増加する。技術計算においては殆ど倍精度浮動小数 (8 バイト) 単位でのアクセスを行うが、その場合でも見かけ上のバンド幅は 16 倍増加し、83.2GB/s 相当として機能する。

一方、CBE ベースのシステムで 512MB 以上の配列全体をアクセスするようなケースでは、実効バンド幅は HDD のランダムアクセス時のバンド幅まで主記憶バンド幅が低下した状態 (スラッシング状態) になる。HDD の平均シーク時間を 5ms とすると 10ms で 4KB のページのスワップアウト・スワップインをするので、実効バンド幅は 400KB/s に過ぎない。これは CRS の SO-DIMM のバンド幅 (2.6GB/s) より 6500 倍低下することを意味する。

上記は CPU が使用するデータが 4KB のページにぎっしり詰まっている場合の値であるが、NAS CG ベンチマーク class C では 128 バイトのキャッシュライン上に 1 個程度しか有効なデータはなく、その状態と比べると DIMMnet-3 を装着すると 8 バイトデータのアクセス時で 10 万倍以上の実効バンド幅で動作すると考えられる。

例えば NAS CG ベンチマーク Class C のように、主記憶に対するランダムアクセスバンド幅が支配的で 2GB 程度の主記憶を必要とするアプリケーションを実行させる場合を考える。それを、通常の PC で 2GB の主記憶を持たせたもので 20 分くらいで終るものが、素の CRS では実質的には実行不可能である。しかし、DIMMnet-3 は PC を上回る計算速度まで CRS の実効性能を引き上げることができると考えられる。

6.3 主記憶より下の階層の半導体記憶容量

Marcury の CBE ベースブレードでは主記憶直下の階層は隣接 CBE の主記憶となり、その容量は 512MB である。ただし、FLOPS あたりの記憶容量の向上には全く寄与しない。

素の CRS の場合は SO-DIMM の容量が 512MB で、主記憶と同容量しかない。一方、DIMMnet-3 を装着する場合は最大 16GB で、32 倍の容量を有する。よって多くのプログラムが殆ど HDD アクセスなしに実行できるようになると考えられる。

6.4 主記憶より下の階層のバンド幅

Marcury の CBE ベースブレードでは主記憶直下の階層は隣接 CBE の主記憶となり、そのバンド幅は 25.6GB/s である。隣接 CBE 間の接続は 20GB/s 全二重なので、読書きいずれかに偏る利用をした場合は 20GB/s となる。その下は IDE の HDD となるため、100MB/s 程度と考えられる。

CRS の主記憶直下の階層は SO-DIMM があり、そのバンド幅は 2.6GB/s と主記憶の 1/10 に過ぎない。その下は IDE の HDD となるため、100MB/s 程度と考えられる。しかし DIMMnet-3 を装着した場合は CRS の SO-DIMM 内で納まる範囲のデータ量であれば Wisconsin ベンチマークのような応用では 32 倍、8 バイトの倍精度浮動小数不連続アクセスを主体とする NAS CG ベンチマーク ClassA のような応用では 16 倍に実効的な主記憶バンド幅が向上する。さらに SO-DIMM 内で納まらないデータ量であると、この指標は 4 バイトアクセス時に 800 倍、8 バイトアクセス時に 400 倍の差となる。

実際にはページングが頻繁に発生した場合はランダムアクセスになるため HDD は 100MB/s ものバンド幅を到底維持できない。よって上記の指標上の比率はかなり控えめな倍率を示しており、後述の指標と組合せて考える必要がある。

6.5 主記憶より下の階層のレイテンシ

CRS の主記憶直下の階層は SO-DIMM であり、そのアクセスレイテンシは実際に実機上で測定してみないと現状では判らない。ただし、20GFLOPS 程度の能力を持つマシン上で PC を差し置いて実行させる技術計算のジョブが 1GB に収まるものではなく、実質的には HDD へのページング時のバンド幅が実効バンド幅となってしまう。

その実効バンド幅と前述の「主記憶より下の階層のバンド幅」の間には大きな隔たりがあることが容易に想像つくが、その原因を構成する要因が「主記憶より下の階層のレイテンシ」である。つまりページング時にはランダムアクセスとなるため、「主記憶より下の階層のレイテンシ」= HDD の平均シーク時間×2 程度となる。よってこの指標と前述の「主記憶より下の階層のバンド幅」を組み合わせた時に、実際に得られる倍率に近いものとなる。

ランダムアクセス時の HDD の平均シーク時間は最大シーク時間 (約 10ms) の半分で、主記憶や SO-DIMM バッファへのアクセス遅延とは 25000 倍程度の差があると考えられる。

6.6 TLB がカバー可能な空間サイズ

TLB がカバー可能な空間サイズは限度があり、その点については今後 CBE の仕様を調査しないと現状がつかめない。例えば Pentium4 に近い一階層の数十エントリレベルの TLB であるならば 3 次元配列における転置をさせた場合などのように毎回 TLB ミスを発生し、悲劇的な性能低下を起こすであろ

うし、Itanium2 や SPARC64V のように大きなページサイズでかつ容量も大きいのであれば、そこまで TLB ミスの問題が性能に大きな影響を及ぼさないケースも考えられる。

ただし、DIMMnet を用いる場合にはたとえ CBE の TLB が小さい場合だったとしても、ユーザ空間用には Window メモリや制御レジスタをマップするのに十分な 1 エントリ程度しか TLB エントリを消費しないので、全く心配はいらない。

7. DIMMnet の CBE 適用向け改造法

本章では DIMMnet を CRS に適用可能とするためのハード面およびソフト面からの改造方法に関して考察する。

7.1 ハード面での改造

7.1.1 DDR2 スロットとの接続手法案

(1) アクティブケーブル引き出し法

図 1 に示すように DIMM 型ホストインタフェース部のロジックを分離し、複数チップ構成で DIMMnet-3 を構成する方法を SWoPP'05 において提案した。基本的にはこの方法をそのままで、サイズのみ SO-DIMM とした子基板を作成して CRS に用いる案である。

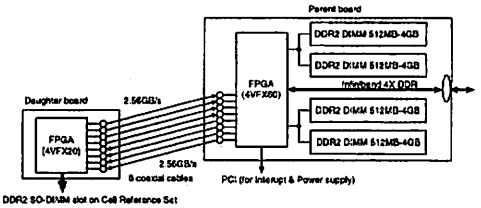

図 1 アクティブケーブル引き出し法

問題として考えられるのは、CRS は図 2 に示すように SO-DIMM タイプのマザーボードに平行に装着するタイプのソケットを採用しているので、基板面積が少なく、必要な回路を全て装着した上で、同軸ケーブルを引き出せるかという点である。

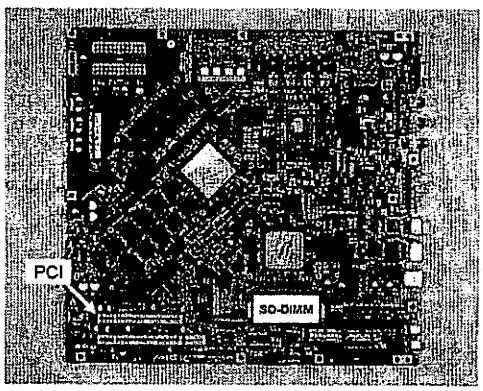

図 2 東芝 Cell リファレンスセット (CRS) の基板上の SO-DIMM と PCI の配置

(2) パッシブエクステンダー基板法

図 3 に示すように、SO-DIMM 側に装着される基板には FPGA などの能動的回路は一切載せず、SO-DIMM 形状の装着部を有する基板上に基板対基板タイプのコネクタのみ搭載した受動的なプリント基板により、PCI バス上に装着する親基板との間を接続する方法が考えられる。

この方法の利点は、以下の 2 つがある。

・SO-DIMM 側の基板の面積の制約を受けにくい

・FPGA 分割に伴う設計コストや、グルー回路や、ケーブルが排除される。

一方、この方法の欠点は以に示す 3 つがある。

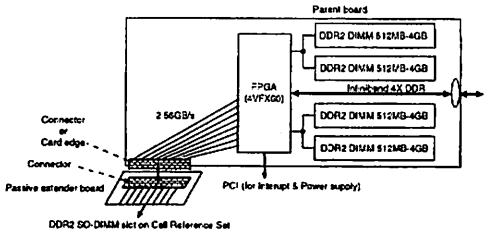

図 3 パッシブエクステンダー基板法

・配線長が延びるのでタイミング的に間に合わなくなる可能性がある。

・基板形状がおそらく現行の CRS 固有の物となるので、他のシステムにはほとんど流用できない。

・エクステンダー基板がマザーボード上の部品と機械的に干渉する可能性がある。

(3) パッシブケーブル引き出し法

図 4 に示すように、SO-DIMM 側に装着される基板には FPGA などの能動的回路は一切載せずに、SO-DIMM 形状から基板対ケーブルタイプのコネクタのみ搭載した受動的なプリント基板とし、PCI バス上に装着する親基板との間をケーブルで接続する方法が考えられる。

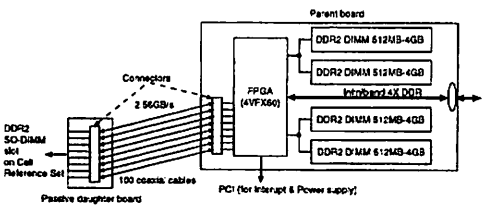

図 4 パッシブケーブル引き出し法

この方法の利点は、以下の 2 つがある。

・SO-DIMM 側の基板の面積の制約を受けにくい。

・FPGA 分割に伴う設計コストやグルー回路が排除される。

・SO-DIMM スロットを有する複数種類のマザーボード上で利用可能になる。

一方、この方法の欠点は以に示す 2 つがある。

・配線長が延びるのでタイミング的に間に合わなくなる可能性がある。

・ケーブルコストが多少かかる。

この信号線を通過する周波数は CRS の場合 2.6GB/s という公開されている SO-DIMM スロットのバンド幅から逆算すると、166MHz に過ぎないので、この程度の信号を通す比較的安価なケーブルは携帯電話やノート PC 内のヒンジ部などで用いられているような汎用品として存在すると考えられる。

また、図 2 に示されるように現行の CRS 上では SO-DIMM スロットと PCI バススロットは極めて近接しており、配線長の増加は 10cm 以内 (時間換算で 0.5ns 以内) に抑えられるのではないかと考えている。

7.1.2 Window の個数の増加

CBE の中には 8 個の SPU が利用可能であり、そこに存在する 8 個の倍精度浮動小数演算器の利用を考えるならば、適切にジョブを分割し、並列実行をさせなければならない。

DIMMnet-2 には Read および Write の各方向に 4 セットの Window を実装していたが、それでは 8 個の SPU から独立にアクセスしてくる要求には耐えられない。よって、CBE 対応とするためには DIMMnet-3 の Window の個数を増加させる必要がある。いかなるケースで何個あれば良いのかという点については定量的な評価を行なう必要があり、それは今後の課題である。

なお、現在の DIMMnet-2 の実装では FPGA のブロック RAM の使い方に 32 倍の無駄があり、RAM 上でプロセスをマルチプレクスすることで、使用ブロック RAM 数を増やさずに 32 プロセス分実装することが可能と考えられるので、Window にアクセスする相手が 8 倍に増えても 1 タイミングでは 1 つ

の相手からのアクセスになれば問題は無いものと考えられる。

7.2 ソフト面での改造

7.2.1 プリフェッチ命令を DMA コマンドに変更

現在、Pentium4 ベース PC 上で稼動中の DIMMnet-2 は、DIMMnet-2 とホスト CPU の間のデータ転送におもにプリフェッチ命令を利用して、CPU が実際にデータを使用する前に L2 キャッシュまで DIMMnet-2 の Read Window 上のデータを先取りするようにしている。

しかし、CBE の SPU にはキャッシュのかわりに通常のアドレス可能な RAM であるローカルメモリが装備されており、DMA コントローラによって主記憶や SO-DIMM との間のデータ転送を行なう。このため前述の Pentium4 上での操作と同等なデータ転送を DMA コマンドで置き換える改造が必要である。

この際、Pentium4 と CBE では主記憶レイテンシやバンド幅等のハードウェアパラメータが異なるため、Pentium4 用にソフトウェアパイプラインが生み出すプリフェッチタイミングが手動で調整されているケースでは、DMA コマンドを実行するタイミングを新たに調整しなおす必要がある。

7.2.2 キャッシュラインフラッシュ命令の削除

Pentium4 や CBE の PPU が DIMMnet-3 のホストとして Read Window をアクセスする場合は、利用が終った Read Window を再利用する前に対応するキャッシュラインのフラッシュを行なう命令を実行する必要がある。

一方、CBE の SPU にはキャッシュのかわりに通常のアドレス可能な RAM であるローカルメモリが装備されている。このため Pentium4 の命令セットがサポートしていた CLFLUSH 命令に相当するキャッシュラインフラッシュを行なう命令が SPU には存在しないし、それを使う必要が無い。

よって、SPU が Read Window をアクセスするケースに相当するプログラムから Pentium4 向けに挿入していた CLFLUSH 命令をコメントアウトするなどして削除する必要がある。

この変更は、Pentium4 ベースの PC 上では CLFLUSH の実行によってアプリケーションの実効性能が 1/3 から 1/2 に低下していたことから、CBE が Pentium4 よりも DIMMnet-3 との相性が良く、高い加速率が得られることが予想される。

実際には、Pentium4 で CLFLUSH 命令が思わぬ性能低下をもたらしたのと同様に CBE においても思わぬ性能低下要因が潜んでいる可能性があり、その確認は今後の課題である。

7.2.3 複数の SPU へのジョブの分割

現在 DIMMnet-2 を装着して稼動している Pentium4 はシングルコアの CPU であり、プログラムはマルチコアには全く対応していない。

一方、CBE の中の 8 個の SPU に存在する 8 個の倍精度浮動小数演算器の利用を考えるならば、適切にジョブを分割し、並列実行させなければならない。SPU に勝手に SO-DIMM をアクセスさせに行くと転送経路の効率が低下する可能性があるので、CRS 上のスケジューリングや帯域予約といったミドルウェア [18] の活用を検討する必要があり、今後の課題である。

8. おわりに

本論文では、まず今後の MPU のトレンドを考察し、現在から近未来にかけての MPU を HPC 用途に利用する際に直面すると考えられるメモリシステム上の制約を示す指標を列挙し、その指標の具体的なシステムにおける状況を確認した。

また、CBE に内蔵される倍精度浮動小数演算器の性能が最新のベクトル型スーパーコンピュータの単体プロセッサに匹敵することを指摘し、CBE の HPC 応用の有望さを主張した。

その上で、CBE の HPC 応用は悲観的にならざるを得ない前記の指標が見受けられたが、その解決策として iRAM と DIMMnet の利用を提案した。

特に DIMMnet-3 を CRS に装着することにより予測される効果を考察した。その結果、使わない場合に対して最大約 20 万倍という性能差がつくケースが想定できることを示した。

また、DIMMnet-3 においてを CRS に適用するためのハード面とソフト面の両面からの改造方法について考察した。種々の検討課題が残っているが、全体として有望であると考える。

謝辞 本研究は総務省戦略的情報通信研究開発推進制度 (SCOPE) の一環として行われたものである。医用画像処理における大容量メモリの必要性についてご教授いただきました大阪大学萩原教授、Itanium2 上で実行時間の 83%がメモリアクセスに消費されるアプリケーションのソースコードをご提供いただきました東北大学情報シナジーセンターの小林教授,SPARC64V の TLB についてご教授いただきました富士通の井上氏に感謝いたします。DIMMnet-2 の開発に関する議論にご参加いただいている慶應義塾大学の西謙師、渡辺氏、大塚氏、北村氏、宮代氏、宮部氏、伊沢氏、東京農工大学の浜田氏、荒木氏、木立氏、森氏、金井氏、立命館大学の国枝教授、和歌山大学の齋藤講師、日立 IT 社の上嶋氏、今城氏、岩田氏に感謝いたします。DIMMnet-3 開発用 FPGA の入手に際してご協力をいただいた日本 XILINX 社の吉沢社長、東芝セミコンダクター社の各務氏、斉藤氏、安藤氏に感謝いたします。

参考文献

1) Cray Inc. "CRAY X1E DATASHEET", http://www.cray.com/downloads/X1E_datasheet.pdf (Jan. 2005)
2) NEC : "Personal Supercomputer SX-8i", catalog (2005)
3) Cell Broadband Engine Architecture, http://www.ibm.com/developerworks/power/cell, (Aug. 2005)
4) B. Flachs, S. Asano, S. H. Dhong, H. P. Hofstee. G. Gervais, R. Kim, T. Le, P. Liu, J. Leenstra, J. Liberty, B. Michael, H-J. Oh, S. M. Mueller, O. Takahashi, A. Hatakeyama, Y. Watanabe, N. Yano. "A Streaming Processor Unit for a CELL Processor", IEEE International Solid-State Circuits Symposium, pp.134-135. (Feb. 2005)
5) J. A. Kahle, M. N. Day, H. P. Hofstee, C. R. Johns, T. R. Maeurer, D. Shippy "Introduction to the Cell multiprocessor", IBM J. Research & Development Vol. 49 No. 4/5, pp.589-604 (Jul. 2005)
6) 東芝 : "次世代プロセッサ Cell のチップセットおよびリファレンスセットの開発・販売について", 東芝プレスリリース (2005.9.20) http://www.toshiba.co.jp/about/press/2005_09/pr_j2002.htm
7) 増渕 : "Cell プロジェクトは "第 2 幕"へ, 開発環境から応用まで全体像を明らかに", 日経エレクトロニクス 2005 年 9 月 26 日号 http://www.semicon.toshiba.co.jp/solution/keytech/keytech_5.html
8) SCE : "次世代コンピュータエンタテインメントシステム PLAYSTATION 3 を 2006 年春発売予定", Sony Computer Entertainment Inc. プレスリリース (2005.5.17) http://www.scei.co.jp/corporate/release/pdf/050517.pdf
9) 富士通 "UNIX サーバ用プロセッサ「SPARC64(TM) V」", http://primeserver.fujitsu.com/primepower/catalog/data/pdf/sparc64_v_j.pdf (Aug. 2004)
10) Mercury Computer Systems Inc. : "Dual Cell-Based Blade DATASHEET", http://www.mc.com
11) GIGABYTE TECHNOLOGY : "i-RAM", http://www.gigabyte.co.jp/nippon/i-ram/iram-m.html
12) 田邊, 濱田, 中條, 天野 : "メモリスロット装着型ネットワークインタフェース DIMMnet-2 の構想", 情報処理学会計算機アーキテクチャ研究会, 2003-ARC-152, pp.61-66 (Mar. 2003)
13) 田邊, 安藤, 箱崎, 土肥, 中條, 天野 : "プリフェッチ機能を有するメモリモジュールによる PC 上での間接参照の高速化", 情報処理学会論文誌コンピューティングシステム,Vol.46, No.SIG12 (ACS11), pp.1-12 (Aug. 2005)
14) 田邊, 羅, 中條, 箱崎, 安藤, 土肥, 北村, 天野 : "プリフェッチ機能を有するメモリモジュールによる等間隔アクセスの高速化", ハイパフォーマンスコンピューティングと計算科学シンポジウム (HPCS2006), pp.55-62 (Jan. 2006)
15) 田邊, 羅, 濱田, 中條, 北村, 宮代, 宮部, 天野 : "DIMM スロット装着型デバイス DIMMnet-2 の改良方針", 情報処理学会計算機アーキテクチャ研究会, 2005-ARC-164, pp.127-132 (Aug. 2005)
16) 笠原, 木村 : "マルチコア化するマイクロプロセッサ", 情報処理, Vol.47, No.1 pp.10-16 (Jan. 2006)
17) 金井, 森, 荒大, 田邊, 中條, 並木 : "コモディティ OS と単一仮想記憶管理によるクラスタシステムの構築法", 情報処理学会論文誌システム評価研究会 (発表予定 Mar. 2005)
18) 前田, 雨宮 : "ヘテロマルチコアプロセッサ Cell 上でのスレッド実行環境", 情報処理, Vol.47, No.1 pp.34-40 (Jan. 2006)