

メモリスロット装着型通信機構による大規模仮想共有メモリの研究開発 (0222029) Research and Development of Large Scale Virtual Shared-Memory with a Communication Mechanism Plugged into a Memory Slot

研究代表者

中條 拓伯 東京農工大学 大学院 共生科学技術研究院
Hironori Nakajo Institute of Symbiotic Science and Technology, Graduate School, Tokyo University of
Agriculture and Technology

研究分担者

天野 英晴[†] 田邊 昇^{††} 齋藤 彰一^{†††} 國枝 義敏^{††††}
Hideharu Amano[†] Noboru Tanabe^{††} Shoichi Saito^{†††} Yoshitoshi Kunieda^{††††}
[†]慶應義塾大学 理工学部 ^{††}株式会社 東芝 研究開発センター
^{†††}和歌山大学 システム工学部 ^{††††}立命館大学 情報理工学部
[†] Faculty of Science and Technology, Keio University
^{††} Corporate Research and Development Center, Toshiba Corporation
^{†††} Faculty of Systems Engineering, Wakayama University
^{††††} College of Information Science and Engineering, Ritsumeikan University

研究期間 平成 14 年度～平成 18 年度

本研究開発の概要

本研究では、パーソナルコンピュータ(PC)を高速なネットワークで接続するクラスタシステムの通信機構として、メモリスロットに装着するネットワークインタフェースの設計・開発を行い、大規模仮想共有メモリを構築する。そして、今後のブロードバンドインターネットに対応した、高速計算機群ならびに高速アクセス可能な大規模共有メモリを実現するための基礎技術を構築し、その有用性を評価することを目的とした。FPGA によるプロトタイプを基に、PC の DDR から DDR2 への進歩に適合しつつ、プロトタイプを作成し、Cell プロセッサシステムへの接続も可能とし、要素技術の有用性を実証した。

Abstract

In this research project, we have designed and developed network interface which is plugged into a memory slot as a communication mechanism for a cluster system in which personal computers are connected via a fast network. Using this system we have tried to implement a large scale virtual shared-memory system. The purpose is constructing the basic technology in order to implement fast and large capacity of shared-memory as well as evaluating its usefulness. The prototyped implemented systems can be applied to evolution of memory technologies from DDR to DDR2 as well as connected to a Cell based processing system with using FPGAs in order to validate usability of essential technologies in our systems.

1. まえがき

我々は、10 年以上も改定されない PCI バス等の標準入出力バスによるボトルネック回避のために、メモリスロットに搭載されるネットワークインタフェースカード(NIC)を設計し、その上での通信機構を検討してきた。本研究では、DIMMnet-2, 3 の開発および実装を行い、その上で大規模仮想共有メモリの実現することで、効率的なクラスタシステムを構築する。そして、今後のブロードバンドインターネットに対応した、高速計算機群ならびに高速アクセス可能な大規模共有メモリを実現するための基礎技術を構築し、その有用性を評価することを目的とする。

2. 研究内容及び成果

2.1 DIMMnet-2 & 3 プロトタイプの開発

我々は、前期で Xilinx 社の大規模 FPGA を搭載したプロトタイプボード DIMMnet-2 を開発した。Infiniband の NIC として動作するとともに、後述するような高性能メモリモジュールとしても動作する。100MHz の DDR メモリスロット上での動作ではあるが、実際に複数の PC 間でのデータ転送に成功した。DDR

メモリスロットにメモリ以外を接続した動作は世界初である。

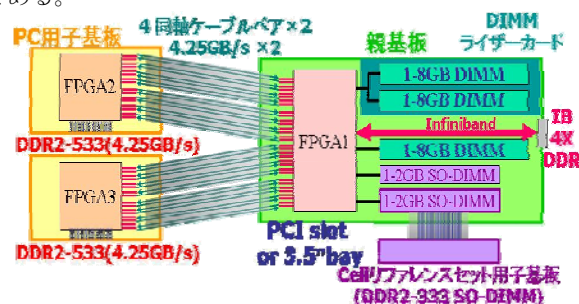


図 1 DIMMnet-3 の構成

後期に作成した DIMMnet-3 の PC 用子基板では、新しい世代の FPGA の初期ロットを用いて、200MHz の DDR2 への対処が可能という実験結果を得た。DDR2 メモリスロットにメモリ以外を接続した動作も世界初である。現状普及している FPGA ではそれ以上の周波数で動作が期待できる。

DIMMnet-3 親基板は単体で調整用に従来の DIMMnet-2 同様の一体型としての動作ができる。さら

に子基板と組み合わせることで PCI 型とハードディスクドライブ型の実装形態を取れる。その場合は DIMM ライザーカードを装着することで大幅な大容量化に対応している。子基板は通常の PC の DDR2 ソケット用の他に、東芝 Cell リファレンスセットの SO-DIMM ソケットに装着可能なもの作成し、適用範囲を広げた。

DIMMnet-3 によれば通常 4GB しかない安価な PC の主記憶容量を RAM ディスクドライブ経由で 58GB まで拡張させることができる。それらが高速な Infiniband スイッチで結合され、巨大な共有メモリを構築できる。この技術を応用すれば、下図のように性能低下せずに Google サーバのような Web 検索サーバの消費電力を一桁低下できると考えられる。

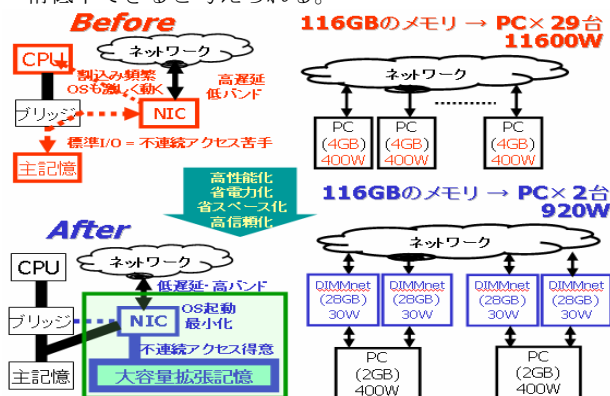


図2 DIMMnet-3による検索サーバの改良例

2.2 ベクトルアーキテクチャを導入したメモリモジュール

我々は、「ベクトルアーキテクチャを導入したメモリモジュール」というコンセプトを提案した。メモリモジュールの中にベクトルロードストアコマンドやインタリードメモリといったベクトル型スーパーコンピュータに類似した機構を取り込み、メモリモジュール以外の全ての部品を COTS しながら、汎用 CPU の不連続アクセスとキャッシュの不整合の問題を解決できる。

我々はリストアクセスが最大の性能ネックとなっている NAS CG ベンチマークや、主記憶上にデータベースが配置された時に等間隔アクセスが性能ネックとなる Wisconsin ベンチマークを用いて評価を行った。

その結果、キャッシュの弊害が顕著に現れる CG Class C において、DIMMnet-2 を装着しない Pentium4 ベース PC に対して約 3 倍の高速化を確認した。

さらに Wisconsin ベンチマークにおいては約 5 倍の高速化を確認した。Pentium4 ベース PC 上で DIMMnet を利用する際は使用済みキャッシュラインの破棄操作が性能向上のボトルネックであった。一方、Cell に内蔵される SPU のようなキャッシュアーキテクチャではないプロセッサに適用する場合を想定した条件下では約 25 倍の高速化が得られることが判った。

このベクトルコマンドを C 言語のプログラムから生成できるコンパイラを開発し、上記の機構をより簡単に利用できるようになっている。

2.3 ハードウェアによるメッセージ交換の加速

メッセージ交換による並列アプリケーションプログラムインタフェース(API)のデファクトである MPI を様々な方向からハードウェアにより加速する機構を提案し、DIMMnet-2 上に実装し、評価した。

IPUSH[2]は MPI のような送信側と受信側の連携で行われる 2sided 通信を、RDMA のような送信側のみの都合で行う通信と同様の高い性能をメッセージ長が短い場合でも実現できることが判った。この機構は国際会

議の席上で MPI の創始者である Gropp 博士より大変興味深いとのコメントを得た。

さらに、VDDC[3]は派生データ型通信という不連続なデータ通信をハードウェアで高速化する方式である。キャッシュミスが多発して低速だった通信が 2.2 の研究同様に高速化する。我々のアプローチを GA/ARMCI(米国のメジャーな API)の創始者である Nieplocha 博士より大変興味深いとのコメントを得た。

3. むすび

本研究は、ネットワークインフラ系サーバのみならず、PC 単体の高性能化と大容量化や、HPC 系・データベース系の各種サーバの高性能化・大容量化・省電力化を実現する社会的影響力が大きい成果を挙げたと考える。

ASIC 開発費の高騰により LSI 化は見送られたため、大規模システム上での実証や商品レベルの完成度は達成できなかったものの、実機を併用した高い有用性を示す研究結果は、国内の受賞や、海外の一流の研究者からも期待の声が寄せられており、学術的な評価も高い。

【誌上発表リスト】

- [1] 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 (2005.8)
- [2] 北村、宮部、中條、田邊、天野“メッセージパッシングモデルを支援するパケット受信機構の DIMMnet-2 への実装と評価”、情報処理学会論文誌コンピューティングシステム、Vol.47, No.SIG12(ACS15), pp.59-73 (2006.9)
- [3] 宮部、宮代、北村、田邊、中條、天野、“MPI 派生データ型通信支援機構の DIMMnet-2 への実装と評価”、情報処理学会論文誌コンピューティングシステム、ACS 論文誌第 19 号 および SACSIS2007 (2007 採録決定)

【申請特許リスト】

- [1] 田邊 昇、情報処理装置及びデータ転送方法、日本、2003 年 03 月 07 日
- [2] 田邊 昇、通信システム及び通信方法、日本、2003 年 3 月 7 日
- [3] 田邊 昇、情報処理装置、情報処理システム、DMA コントローラおよび情報処理方法、日本、2005 年 08 月 02 日

【登録特許リスト】

- [1] 田邊 昇、情報処理装置及び再送制御方法、日本、出願日 2003 年 03 月 07 日、登録日 2005 年 10 月 07 日、登録番号 P3727928
- [2] 田邊 昇、メッセージ通信方法並びに受信側通信装置及び送信側通信装置”、日本、出願日 2004 年 03 月 23 日、登録日 2005 年 10 月 28 日、登録番号 P3735633

【受賞リスト】

- [1] 北村、伊豆、伊沢、宮代、宮部、渡邊、大塚、濱田、田邊、中條、天野：第 12 回 FPGA/PLD Design Conference ユーザ・プレゼンテーション 優秀論文賞、“FPGA を用いたメモリスロット装着型ネットワークインタフェースの設計”、受賞年月日 2005 年 1 月 28 日
- [2] 田邊：情報処理学会山下記念研究賞、“プリフェッチ機能付きメモリモジュールによる不連続アクセスの連続化”、受賞年月日 2005 年 7 月 27 日

【本研究開発課題を掲載したホームページ】

<http://www.tuat.ac.jp/~nakajo>