base-m n-cube の実現方式に関するシミュレーション*

1H−7

川倉 康嗣・田邊 昇・鈴岡 節†

(株) 東芝　総合研究所‡

1　はじめに

共有メモリを持たない並列計算機において、汎用性のあるネットワークとしてbinary n-cube (hyper cube) が知られているが、大規模なシステムを構築するには、実装が困難になるという欠点がある。その欠点を補い、かつ、同等以上の性能を持つネットワークとして、binary n-cubeの2進をn進に拡張したネットワークbase-m n-cubeを提案した[1]。さらに、base-m n-cubeを採用した高並列計算機Prodigyを製作し、その有効性を確認した[2]。また、R256[3]、H2P[4]でも同様の考え方をしたネットワークを採用している。

本発表では、より実用的であるbase-m n-cubeマシンを構築するために、CPUおよびルータにおける様々な選択に対してシステム性能を定量的に評価したので、その結果について報告する。

2　CPUおよびルータにおける方式

1. 通信の準備にかかる時間

通信の準備作業とは、CPUがルータに受信パケットのメモリ上での置き場所を指示すること、受信パケットを格納するメモリ領域の管理、パケット送受信の完了をルータからCPUに通知することなどがある。このような準備作業を割り込みとソフトウェアにより行えば、柔軟な処理をできるが時間がかかる。これに対し、ハードウェアで行えば、固定された処理となるが高速に行える。

2. ネットワークに対するCPUの速度

高並列計算機システムの実用的な性能を向上させるためには、ネットワークの転送速度とCPUの処理速度とがつりあっている方が望ましい。ただし、このバランスは、高並列計算機システム上で実行するアプリケーションに大きく依存する。

3. CPU・ルータ間のメモリアクセス競合

CPUとルータとの間にあるメモリに対して、それぞれが独立にアクセスできる方式と排他的にしかアクセスできない方式とがある。特別のハードウェアを用いてメモリを独立にアクセスできるようにすれば、ハードウェアコストはかかるが、性能は高い。これに対して、ルータがメモリにアクセスしている時には、CPUにメモリをアクセスさせない方式では、性能は犠牲になるが、ハードウェアコストはかからない。

3　シミュレーション

1. 通信準備時間

通信準備に必要な時間は、CPUでの命令数としてとらえた（通信時間＝命令数×1命令の実行時間）。また、送信時と受信時のそれぞれに、この時間だけ通信準備のために必要である。送信時または受信時に必要な時間を1～100ステップの間で変化させた。

2. CPUの処理速度

シミュレーションでは、CPUで1命令の実行にかかるクロック数を変化させて実験する。このクロックとは、パケット1バイト（1ビット）をルータ間で1リンク分転送するためのクロックと同一である。なお、通信準備をソフトウェアで行っている場合には、そのための時間も短縮される。1命令の実行に必要なクロック数を10クロック、5クロック、1クロックと変化させて実験した。

3. CPUでの処理量

CPUでの処理量は、アプリケーションに依存するため一概には言えないが、非常に処理量が少ない場合として受信パケット1バイトあたり1命令、処理量が多い場合として1バイト当たり20命令、その中間として5命令、10命令の場合を調べた。

4. CPUとルータのメモリアクセス

同時アクセス可能な方式と排他的にしかアクセスできない方式とがある。シミュレーションでは、ルータがメモリへのパスを獲得した時から解放するまでの時間分だけCPUでの処理時間を延長することにより同時にアクセスできない場合と同じ状況と、両方が同時にアクセスできる状況とを作り出す。

---

*Simulation for various implementation of base-m n-cube network

†Yasushi Kawakura, Noboru Tanabe, Takashi Suzuoka

‡TOSHIBA Research and Development Center

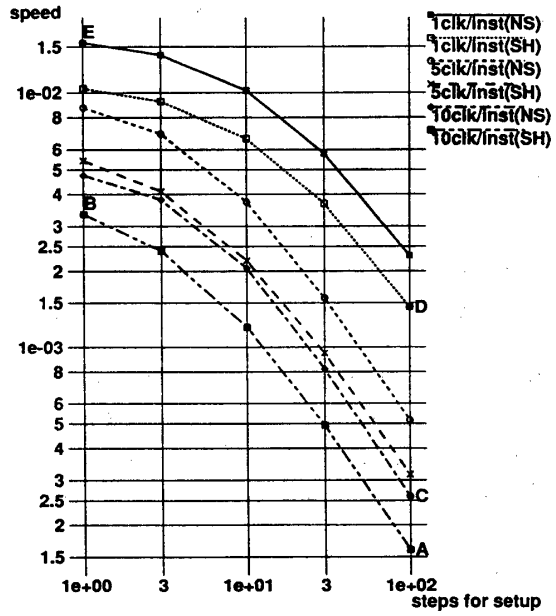

図1 パケット長:10バイト 処理量:1命令/バイト

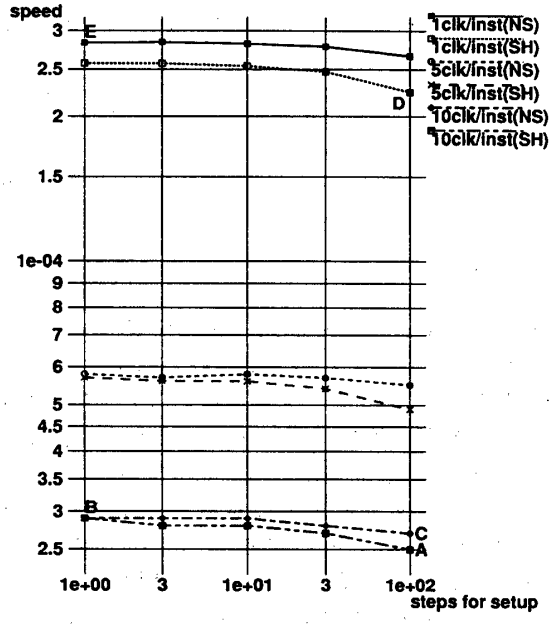

図2 パケット長:100バイト 処理量:20命令/バイト

4 結果

PE台数を512台、パケット数を512個に固定して、横軸に通信準備時間、縦軸にPE1台の単位時間当たりの処理パケット数をとり、結果をグラフに表した。図1は、パケット長が10バイト、CPUでの処理量が1バイトあたり1命令の場合で、図2は、パケット長が100バイト、CPUでの処理量が1バイトあたり20命令の場合である。

1枚のグラフは、CPUの処理速度と通信準備時間を変化させた状況を表している。また、グラフ中でNSはメモリに同時アクセス可能、SHは同時アクセス不可能の場合を表す。

図中で、A→Bは、通信準備時間を削減した場合、A→Cは、CPU・ルータ間のメモリを同時アクセス可能にした場合、A→Dは、10倍高速なCPUを用いた場合、A→Eは、これらをすべて同時に適用した場合である。

- 処理粒度が小さく、ネットワークが混んでいる場合には、通信準備時間を減少させると効果があり、逆に、処理粒度が大きく、ネットワークが空いている場合には、CPUの性能を向上させると効果がある。
- パケット長が短い場合や、CPUでの処理量が少ない場合には、通信準備時間を減少させると効果が大きい。
- CPU・ルータ間のメモリを両方から同時にアクセス可能としても、その効果はそれほど高くないが、CPUが高速になるほど効果が出てくる。

base-m n-cubeを採用した高並列計算機を構築する際に、1) 高性能CPUを採用する、2) 通信準備に必要な時間を短縮する、3)CPUとルータが通信用バッファに同時アクセス可能にする、という方式を採用すれば、それぞれ2～21倍の性能向上が見込めることがわかった。

5 おわりに

base-m n-cubeを用いて高並列計算機システムを構築する際の様々な実現方式の評価を行い、それぞれの性能向上に対する効果がわかった。

今回の実験結果を、実用的な高並列計算機システムを構築する際に、各種の実現方式のトレードオフの判断材料として活用する。

参考文献

[1] 鈴岡, 藤田, 中村, 小柳. 超並列AIマシンの構想. 第*35*回情処全大 *3C-5*, 1987.

[2] 田邊, 中村, 小柳. 高並列計算機Prodigyを用いた各種結合網の通信性能評価. 信学技報 *CPSY90-13*, pp. 7–12, 1990.

[3] T. Fukazawa, T. Kimura, M. Tomizawa, K.Takeda, and Y. Itoh. R256:a research parallel processor for scientific computation. In *ACM 0884-7495*, pp. 344–351, 1989.

[4] 中越, 田中, 濱中, 面田. 並列計算機H2Pの要素プロセッサ間非同期データ転送方式. 第*38*回情処全大 *6T-7*, 1989.