InfiniBand を用いた遠隔メモリアクセスの性能

岩井田　匡俊　　鈴木　悠一郎　　緑川　博子

The Performance of Remote Memory Access over InfiniBand

MASATOSHI IWAIDA, YUICHIRO SUZUKI and HIROKO MIDORIKAWA

1. はじめに

科学技術計算分野では，問題サイズの大規模化などにより，近年大容量メモリの必要性が高まっている．64bitOS により大規模なアドレス空間が利用可能になったが，メモリ消費電力の観点からも 1 ノードあたりのメモリ搭載量には限りがある．筆者らは，遠隔メモリを利用し，逐次プログラムに仮想的な大容量メモリを提供する分散大容量メモリシステム DLM(Distributed Large Memory)を構築している．本報告では，マルチスレッドプログラムに対応した DLM[1]を用い，OpenMP 並列プログラム，マルチスレッド実装ライブラリ利用の逐次プログラムなどの応用処理を，CPU，ネットワークなどの世代の異なる 3 種のクラスタシステムで実行し，その性能を評価した．さらに，各クラスタにおける実際のネットワーク性能について調査した．

2. 実行環境

ここでは，T2K 東大 HA8000 クラスタ(T2K)，FDA クラスタ(FDA)，crest クラスタ(CR)の 3 つのクラスタを用いた．T2K はノード（4CPU, 16 コア, AMD 4Core Opteron 8356@2.3GHz），ネットワーク（Myrinet-10Gx2(20Gbps)，MPICH-MX）である．FDA は，ノード（2CPU，12 コア，Intel Xeon CX5680@3.33GHz），ネットワーク(4xQDR(32Gbps)，IPoIB，MPICH2)である．CR は，ノード（2CPU，16 コア，Intel Xeon CPU E5-2687W@3.1GHz），ネットワーク(4xFDR (54.4Gbps), MVAPICH2)である．

3. 性能評価

3.1 評価プログラム

評価プログラムには，OpenMP プログラムとして(1)正方行列積(4096×4096，ブロック化 16×16)と(2)ステンシル処理(8192×8192，マスクサイズ 15×15)，マルチスレッド版ライブラリ関数使用例として FFTW[2]を用いた(3)3D-FFT(2048×2048×1024)プログラムを用いた．ここでは，ローカルメモリのみで実行した場合（ローカルメモリ率 100%）とユーザプログラムが使用する全体のデータ量の 20%をローカルメモリにおき 80%を遠隔メモリにおいた場合（ローカルメモリ率 20%，ローカルメモリサイズの 5 倍の仮想メモリ利用に対応する）との性能評価を示す．性能向上比には，ローカルメモリのみ使用した 1 スレッドの実行時間を基準としている．

3.2 スレッド並列による性能向上比

図 1-図 3 は，正方行列積計算，ステンシル処理，FFT 計算におけるローカルメモリ率 100%(実線)と 20%(点線)の場合のスレッド並列性能向上比を示す．各応用でローカルメモリのみを利用し 1 スレッド時の実行時間を示したのが図 4 である．

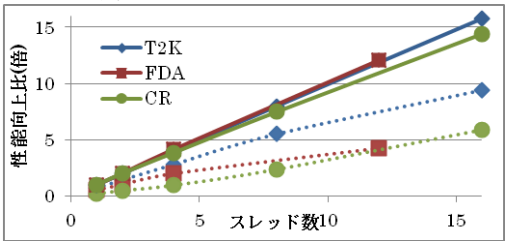

図 1.正方行列積計算，性能向上比(倍)

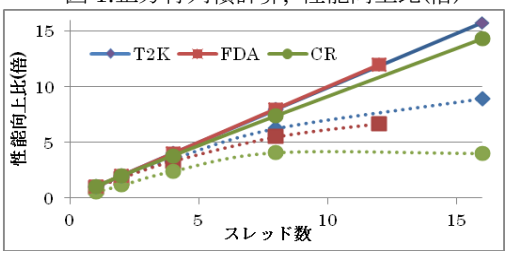

図 2.ステンシル処理，性能向上比(倍)

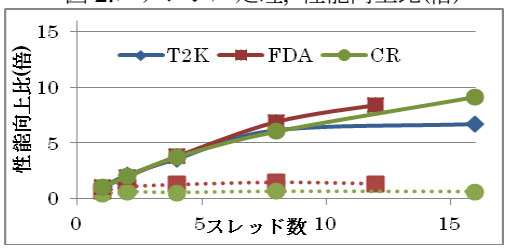

図 3.FFT 計算，性能向上比(倍)

成蹊大学理工学研究科理工学専攻
Graduate school of Science and Technology, Seikei University
独立行政法人科学技術振興機構, CREST
Japan Science and Technology Agency, CREST

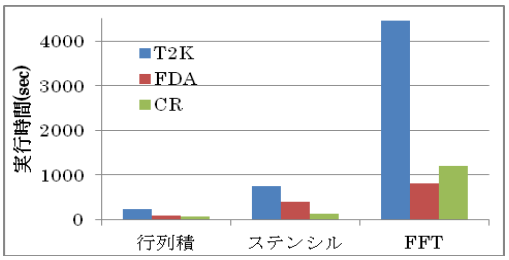

図 4.ローカルメモリ率 100%の実行時間(sec)

どの応用においても，FDA や CR は T2K より性能向上比が低い．これは，基準となるローカルメモリ 100%時の性能が，FDA や CR では T2K に比べ高速化している一方で，次節で述べるように，ネットワーク実質性能が低いことにより，遠隔メモリとローカルメモリバンド幅の性能比がさらに拡大していることが原因である．

3.3 遠隔メモリバンド幅

表 1 は Stream ベンチマーク[3](Triad)の性能で，遠隔メモリの括弧内は各クラスタのローカルメモリバンド幅に対する性能比である．これをみると，18%から 2%へと性能比は次第に大きくなっている．さらに，利用ネットワークの公称バンド幅が増加しているにも関わらず，遠隔メモリの実測バンド幅が低下していることがわかる．

表 1. Stream の Triad 時のメモリバンド幅(MB/s)

	ローカルメモリ	遠隔メモリ	ネットワーク
T2K	2700	493(18%)	Myri-10G x2 2.5GB/s (理論値)
FDA	9196	399(4%)	IB 4x QDR 4GB/s (理論値)
CR	15409	277(2%)	IB 4x FDR 6.8GB/s (理論値)

4. 2side-MPI 通信によるネットワーク性能の実際

IB FDR を持つ CR において，1MB データを pingpong 通信 100 回(以下 PP)の遅延時間（往復時間/2）の計測を行った（MVAPICH2-1.9a2 利用）．同じデータを繰り返し送る（キャッシュ効果あり）場合，平均バンド幅は 4644MB/s で，偏差も小さい．一方，DLM に近い動作環境で毎回違うデータを送受信する PP（キャッシュ効果なし）では，平均バンド幅は約 1471MB/s と約 1/3 にまで低下し，毎回の遅延時間は 4 倍以上のばらつきが生じた．（表 2 参照）．図 5 は遅延時間の実測系列を示す．MPICH-3.0.2 (IPoIB)で同じ実験を行った結果が図 6 である．図 5 ほどのスパイクはないが，この場合にも同じ傾向がみられる．一方，T2K では，バンド幅平均は約 978MB/s と低いが，図 7 のように遅延時間は常に安定している．

表 2. CR pingpong 通信による遅延時間(msec)

1MB PP 通信片道時間	Max	Min	Ave	偏差
キャッシュ効果あり	0.43	0.21	0.23	0.04
キャッシュ効果なし図 5	1.87	0.43	0.71	0.28

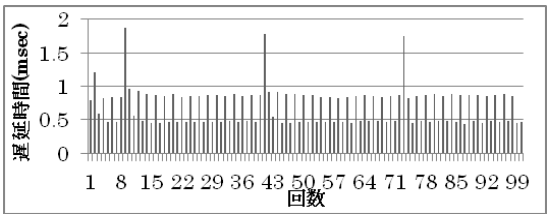

図 5. CR, MVAPICH2-1.9a2 での遅延時間

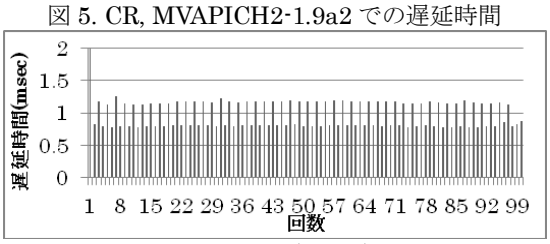

図 6. CR, MPICH-3.0.2 (IPoIB)での遅延時間

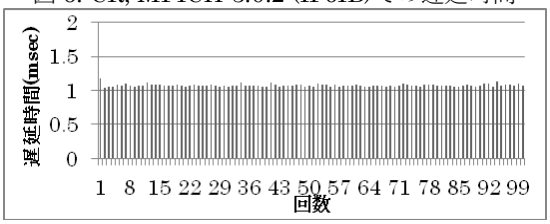

図 7. T2K, MPICH-MX での遅延時間

5. まとめ

クラスタノード内の CPU，メモリバンド幅の高速化は著しいが，ネットワークの高速化も一定レベルで進歩していると考えられてきた．しかし性能指標とされる MPI ベンチマークの多くは，キャッシュ効果の高い現実離れしたデータ送受信環境での最良値を示しており，現実の応用での性能を反映していない．特に InfiniBand は CPU コアやキャッシュ環境などの環境にも影響されやすく性能の不安定さがあることが，今回の使用環境では特に顕著に表れた．

参 考 文 献

[1] 鈴木他，"マルチスレッドプログラムのための遠隔メモリ利用による仮想大容量メモリシステムの設計と初期評価", 情処 HPC 研究会 Vol.2011-HPC-132, No.13, pp.1-6, November, 2011

[2] Fastest Fourier Transform in the West [Online] http://www.fftw.org/, 2011

[3] Stream benchmark http://www.streambench.org/