マルチリンクEthernet環境におけるSMPクラスタの性能評価

小林　智史† SHAN Axida† 吉永　努†
†電気通信大学　大学院情報システム学研究科

1. はじめに

近年、超並列計算機に代わってパーソナルコンピュータなどの一般的に普及しているコンピュータをネットワークで繋ぎ、1つの計算機として利用するPCクラスタシステムが注目されている。また、プロセッサが複数個搭載されているSMPコンピュータも一般的となってきている。

それらのネットワークとして、Myrinetや Infiniband, 10Gigabit Ethernet等、低レイテンシ、高バンド幅のものも登場している。しかし、これらのネットワークインターフェイスは高価である。一般的なPC環境においては1Gigabit(Gb) Ethernetが実装されていることが多く、安価に利用可能である。そのため、このようなEthernetを束ねマルチリンクにすることより高いバンド幅を実現する提案や実装が行われている。

本論文ではリンク数を増設することによる計算性能への効果を測定した。また、性能の違うノードで構成されたクラスタでの実験を行った。

2. マルチリンク

Ethernetにおいて、マルチリンクを実現する実装としてボンディングを行うLink Aggregation(IEEE802.3)やOpenMPI[1]、Scoreの通信ライブラリであるPM/Ethernetなどがあげられる。本実験では、容易に導入することが可能なOpenMPIを使用した。

3. ベンチマーク

3.1 NetPIPE[2]

NetPIPE (NETwork Protocol Independent Performanve Evaluator) は通信性能を計測するベンチマークである。MPIを使用した計測が行え、実際の並列計算に与える影響を反映したものである。

3.2 NPB[3]

NPB(NAS Parallel BenchMark)はNASA Ames Research Centerで開発された熱流体関連の科学技術計算ためのベンチマークである。様々なベンチマークプログラムが含まれており、より総合的な判断を行うことができる。計算結果として得られるMop/s(Mega Operation per second)値はEPとISを除き、ほぼMFlop/sと同値である。EPとISでの値は乱数発生数と整数の数である。

4. 実験

まず、スループット評価のため、OpenMPIの提供するマルチリンク通信で1～4リンクをトランキングした際の、メッセージサイズを変化させた場合の通信性能の変化をNetPIPEを用いて測定した。次に、NBPver3.3を用いて問題クラスA,B,Cについて OpenMPIの提供するマルチリンク通信を行った場合(1～4リンク/ノード)、及び4リンク/ノードの設定でノード上の4つのプロセッサが使用するリンクを静的に割り当てた場合(staticと呼ぶ)の実験を行った。

4.1 実験環境

表.1に実験に使用した2組のクラスタを示す。

表1. 測定環境

クラスタ1	16ノード
CPU	Intel Xeon 5160 3.00GHz
L2キャッシュ	4MB
Memory	4GB /ノード
OS	CentOS 4.6
MPI	OpenMPI 1.2.7
コンパイラ	Intel compiler10.1.018
NIC	Gb Ethernet(PCI-X x2, PCI-E x2)

クラスタ2	16ノード
CPU	Intel Xeon 7030 2.8GHz
L2キャッシュ	2MB
Memory	4GB /ノード
OS	Fedora Core 5
MPI	OpenMPI 1.2.4
コンパイラ	Intel compiler9.0
NIC	Gb Ethernet(PCI-E x4)

4.2 実験結果

図.1にクラスタ1で測定したNETPIPEの測定結果を示す。リンク数増加によりバンド幅が向上している。1リンクの最大スループット894MB/sに対して、2リンクで198%、3リンクで221%、4リンクで235%の性能を示した。これより、一対一通信での性能では、1リンクから2リンクにした時が最も通信性能が向上したこと

Performance evaluation of SMP Clusters with multilink Ethernet
Satoshi Kobayashi† Shan Axida† Tutomu Yoshinaga†
† Graduate School of Information System, The University of Electro-Communications

がわかる。

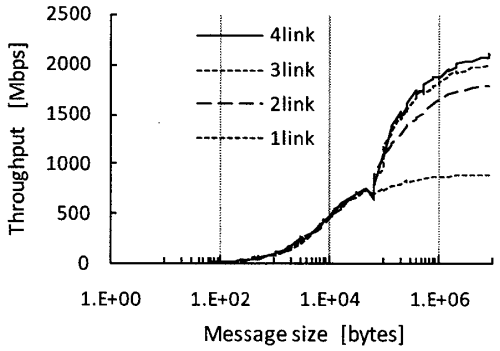

図1．リンク数によるスループットの測定(クラスタ1)

図.2 にネットワーク性能に強く依存する NPB ベンチマークコードのクラスタ 1 での測定結果を示す。

CG ではクラス B においては、2 リンクに増やすことにより性能が 57%向上した。一方で 3 リンク、4 リンクでは低下している。これは、コネクション確立やパケットの順序の入れ替わりやによるコストが増加したためと考えられる。また、クラス C においては、3 リンク、4 リンクに増やした際も向上している。計算量が増加し、パケット処理の影響が減少したためと考えられる。

FT では複素数型の大きな配列データを多く通信する。クラス B においては、リンク数を上げたことによるスループットの向上が効果的に働いたと考えらえる。クラス C においては、スループットの向上よりも通信量が増えたことによるパケットの処理によるコストがかかっていることが考えられる。

MG ではメッセージ長が非一様な通信が散発的に発生する。FT 同様にリンク数の増加によってスループットの向上が効果的に働いている一方で、リンク数が増えるとコネクション確立やパケット処理のコストが増加していることが考えられる。

IS と FT では、全ての問題サイズにおいて static が 4link よりも高い Mop/s を示している。どちらも MPI_Alltoall 通信をしており、プロセッサとネットワークが静的に接続されていることによってコストを下げることができたと考えられる。

SP と BT では、クラスタ 1 とクラスタ 2 において異なった傾向がみられた。クラスタ 2 ではリンク数が増えると性能向上が見られたが、クラスタ 1 ではみられなかった。これは、L2 キャッシュの大きさの違いによるものと考えられる。

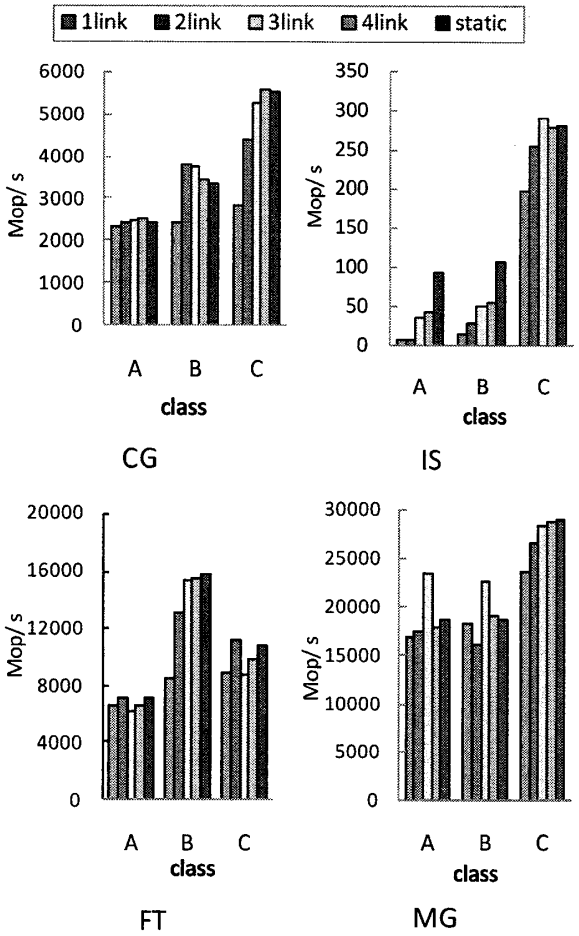

図.2 リンク数による NPB の測定(クラスタ 1)

5. おわりに

本論文では、SMP クラスタにおいてリンク数を上げることによる効果を NETPIPE、NPB を用いて測定した。通信量・通信頻度の高い計算においてはリンク数を上げることにより性能を向上させることができた。一方で、コネクション確立やパケットの順序入れ替えなどのコストがリンク数を上げると共にかかり性能に影響を及ぼすことがわかった。

参考文献

1)Richard L. Graham, Timothy S.Woodall, Jeffrey M.Squyers:"OpenMPI:A Flexible High Performance MPI", the 6th International Conference on Parallel Processing and Applied Mathematics,September 2005 pp228-239

2)NetPIPE:http://www.scl.ameslab.gov/netpipe/

3)NPB:http://www.nas.nasa.gov/Resources/Software/npb.html