ネットワークトポロジを考慮したバンド幅推定の高速化手法

長沼 翔† 高橋 慧‡ 柴田 剛志†† 田浦 健次朗‡ 近山 隆‡‡

†東京大学工学部 ‡東京大学大学院情報理工学系研究科 ††東京大学大学院工学系研究科 ‡‡東京大学大学院新領域創成科学研究科

広域分散環境でのデータインテンシブなアプリケーションではデータ転送がボトルネックとなる。しかし広域分散環境ではバンド幅が場所によって大きく異なり、効率的かつ計画的にデータ転送を行う必要がある。この為には実ネットワークトポロジとバンド幅の情報が欠かせない。既存の手法では測定に時間がかかるうえ、実ネットワークトポロジのような詳細な情報は得られない。本研究では、ネットワークトポロジを考慮したバンド幅の推定を並列に測定することで高速に行う手法を提案する。

1 はじめに

1.1 背景

地理的に遠く離れた様々な環境のクラスタをネットワーク接続し、一つの大きな計算資源として一つの計算を進める、並列分散処理が可能となっている。並列分散計算を行うことによって安価に高性能な計算機資源を得ることができる。これによって、自然言語処理や遺伝子解析などの、これまで望めなかった大規模な処理が可能になっている。

このような分散環境において、各ノード間を結ぶリンクは、バンド幅の広いリンクと狭いリンクが混在している。ある並列分散プログラムを実装する際に、これらを考慮して通信を行わなければ、実行効率は思うようにあがらない。

この事態を避けるためには、アルゴリズム等の他にネットワークを考慮したデータ転送のスケジューリングが重要であり、そしてそれを組むためには、何らかの手法で全てのリンクのバンド幅を推定しておく必要がある。

1.2 関連研究

代表的なバンド幅推定手法に Iperf[1] がある。Iperf は各ノードのキュー待ち遅延やフォワーディング処理遅延等のネットワーク機器の遅延は一切考えず、データ転送時間がデータサイズとバンド幅の比で定まるとしている。よってバンド幅は、送ったデータサイズを転送時間で割ることで得られる。

Iperf ではデータサイズと転送時間を大きくすることで前提条件で無視していた遅延誤差を小さくし、正確でばらつきの小さい推定結果を得ることができる。

しかし、測定する 2 つの計算ノード間に複数のリンクが存在する場合、この手法ではそれらのボトルネックリンクの値しか得らず、ネットワークトポロジ上のどこのリンクのバンド幅を測定しているのかが分からない。

Pathchar[2] は、更に詳しくネットワークをモデル化し、2 ノード間でやりとりしたパケットの挙動からバンド幅を推定する手法である。この手法は、原理上大量のパケットのやりとりが必要なので測定に時間がかかるほか、パケットの挙動の微妙な変化や仮定ネットワークモデルの差異から、測定結果の精度は低く、ばらつきも大きい。また、推定方法の特性上、スイッチングハブ等のネットワークレイヤ 2 以下のノードが測定している 2 ノード間に存在すると正しい推定ができないという問題点もある。

Pathchar と似た考え方を取り入れ改良した、Packet-Pair 推定法がある。この手法は、二つの同サイズのパケットをあるノードから別のノードへ連続送信し、各パケットの到着時刻から 2 ノード間のバンド幅を推定する。Pathchar に比べ測定時間が短いほか、ネットワークレイヤ 2 以下のノードをまたいでも正確に測定できる。しかし、原理上 2 ノード間のボトルネックリンクしか測ることが出来ず、Pathchar 同様測定結果の精度は低く、ばらつきも大きい。Packet-pair 推定法を原理とした代表的なプログラムには Nettimer[3] がある。

1.3 本研究の貢献

本研究では、ネットワークトポロジの全リンクにバンド幅情報を関連付けた、バンド幅トポロジを短時間かつ高精度に構築する手法を提案する。得られたバンド幅トポロジは、データインテンシブな分散アプリケーションにおける効率の良いデータ転送スケジューリングの設計の基盤となる。

この手法では、トポロジ情報が与えられ、以下の 2 で述べる原理に従って推定を進めることで前述の Iperf や Pathchar 、Nettimer で見たような短所を避け、バンド幅トポロジを出力する。

2 原理

本手法は、Iperf のように実際にネットワークトラフィックを発生させる基本原理でバンド幅を推定していくが、トラフィックの流しかたを様々に工夫することで、以下に述べる二つの特徴的な推定手法が可能となっている。本手法は、これら二つを組み合わせて推定を行っていく。

2.1 基本ノードセットの測定

図 1 のように 1 スイッチに 3 ノードが接続された構成を基本ノードセットと呼ぶ。

ここに 3 本のリンクが存在するが、それぞれのバンド幅の値の組合せは、表 1 の左半分に示す 4 ケースだけを考えればよいことが分かる。これは、$a < b < c$ としてよい。

すなわち、各種測定の結果と表 1 右半分を照らし合わせながら有り得るケースを絞り込んでいけば、3 つの

Improving Efficiency of Network Bandwidth Estimation Using Topology Information
by Sho Naganuma†, Kei Takahashi‡, Takeshi Shibata††, Kenjiro Taura‡ and Takashi Chikayama‡‡ (University of Tokyo)

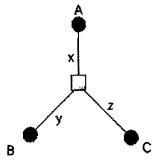

図 1: 基本ノードセット

表 1: 有り得る組み合わせと観測されるべき値

	実値			観測値					
	x	y	z	A-B	A-C	B-C	A-BC	B-AC	C-AB
case 1	a	a	a	a	a	a	a	a	a
case 2	a	a	b	a	a	a	a	a	min(2a, b)
case 3	a	b	b	a	a	b	a	b	b
case 4	a	b	c	a	a	b	a	b	min(a+b, c)

バンド幅の組がどのケースに属し、それぞれ値はどれ程かを推定することができる。ここで言う各種測定とは表 1 右上部に記述してある 6 種の測定であり、例えば A-B とはノード A、B 間の測定値、A-BC とはノード A から B、C にトラフィックを枝分かれさせて流したときに A が観測した値である。ネットワークトポロジを基本ノードセットの連続として捉え、与えられたトポロジ情報を基本ノードセットに分解して測定を開始すれば、各基本ノードセットを並列に測定することができる。

2.2 ネットワークトラフィックを束ねた測定

例えば、図 2 のようなバンド幅をそれぞれ持ったネットワークを考える。中央のリンク (太線、10) のバンド幅を測定しようとした場合、いずれの左右のノードの組合せで測定しても、結果はボトルネックリンクの値、5 となる。そこでノード A からノード C へ、ノード B からノード D へそれぞれ一斉にトラフィックを束ねるように発生させれば、両測定とも値 5 が観測され、すなわち中央のリンクはそれらを足した 10、と推定することができる。

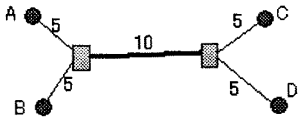

図 2: ネットワーク構成の例

3 実験と結果の評価

3.1 実験環境

実験環境として InTrigger プラットフォーム*の 5 つのクラスタの 283 計算ノードを使用した。今回、これらの計算ノードを論理トポロジ上でネットワーク接続する為のルータやスイッチ類は 19 個、測定すべきリンクは 301 本であった。

*https://www.logos.ic.i.u-tokyo.ac.jp/intrigger/

3.2 実験結果と評価

まず、測定で得られた値について、今回は測定結果の精度が良く、ばらつきも少ない Iperf の測定結果を正解として評価した。本手法の測定結果は Iperf の測定結果とほぼ一致していることが確かめられた。本手法の測定の基本原理は、Iperf と同様の方法をとっているので、Iperf 同様ばらつきの少ない測定結果が得られた。また、トポロジ上でより WAN 側のリンクには、より広いバンド幅を持つリンクが設置されていることが多いが、Iperf では測定している 2 計算ノード間のボトルネックリンクの値しか得られないという問題があった。しかし本手法では、2.1 や 2.2 で述べた原理により、このような状況にあるバンド幅の値を正しく掴むことができた。

次に、測定にかかった時間について、今回、301 本のリンクのバンド幅を全て決定するのに要した時間は、およそ 1 分 10 秒〜1 分 20 秒であった。これは、283 計算ノードの全対全の組み合わせにたいして Iperf を実行することを考えれば、所要時間も作業の手間も大幅に削減できたことが分かる。また、本手法の測定に要する時間は、ネットワークトポロジをツリー構造と見たときのツリーの深さにのみ比例するので、ノードの増加にたいしてスケーラブルな手法といえる。

4 おわりに

本稿では、トポロジを考慮してネットワークトラフィックの流し方を様々に工夫してバンド幅測定を進めていくことによって、ネットワークトポロジの全リンクにバンド幅情報を関連付けた、バンド幅トポロジを短時間かつ高精度に構築する手法を提案した。また実際に実装を行い、複数クラスタにまたがる環境下において、本手法が既存手法と比べより正確なバンド幅トポロジを構築し、しかも短時間かつ少ない手間で結果を得ることができることを確かめた。

現段階では、各所測定の基本部分に要する時間を、1 秒や 2 秒などと経験的な数字で決めているが、この時間をどこまで切り詰められるかということが、より短時間で終了させるかの糸口になっている。ただし、測定に要する時間と結果の精度はトレードオフになっているということに気をつけなければならない。また、与えられたトポロジを、深さが最小なツリーとなるように再構築することでも、測定時間を短くすることができる。今後はこれらの部分を改良し、精度を保ったままより短時間に測定が終了できることを目指していく。

謝辞 本研究の一部は文部科学省科学研究費補助金特定領域研究「情報爆発に対応する新 IT 基盤研究プラットフォームの構築」の助成を得て行われた.

参考文献

[1] Iperf (version 1.7.0), http://www.dast.nlanr.net/projects/Iperf/ .

[2] V. Jacobson, "pathchar - a tool to infer characteristics of Internet paths," ftp://ftp.ee.lbl.gov/pathchar/, April 21, 1997.

[3] K. Lai and M. Baker, "Nettimer: A tool for measuring bottleneck link bandwidth," Proceedings of the 3rd conference on USENIX Symposium on Internet ech-nologies and Systems - Volume 3, p1-5, January 29, 2001.