科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　　６　日現在

機関番号：　１２６０８
研究種目：　基盤研究(C)
研究期間：　2010～2013
課題番号：　２２５０００５４

研究課題名（和文）システム性能向上を目的とした超高速ネットワークの自律制御手法の研究

研究課題名（英文）Study of the autonomous path control method for the purpose of the effectiveness performance enhancement of the data transmission

研究代表者
友石　正彦（Tomoishi, Masahiko）

東京工業大学・大学マネジメントセンター・教授

研究者番号：６０２６２２８４

交付決定額（研究期間全体）：（直接経費）　　2,500,000円、（間接経費）　　750,000円

研究成果の概要（和文）：実生活におけるの通信においては、実際には通信路が混んでいないのに、通信が遅くなることがある。その原因として、たくさんの通信が一つのサーバに集まることによる相互作用が考えられる。
本研究では、そういったことが本当に起こることを実験環境で再現することで確認し、いろいろな場合で再現させることで起こり方を特定し、通信路にある機械の設定のみを変更することで、全体の通信性能が上昇する方法を提案することを目的とする。

研究成果の概要（英文）：In the thing communication in everyday Internet life, communication may become slow though the communication channel is not congested in fact.
In this study, we confirm it by reproducing that such a matter is really generated in our making experiment environment, find how to happen according to making it reappear by various cases, and are intended proposing how the whole communication performance rises by changing only a setup of the machines in a communication channel.

研究分野：　総合領域

科研費の分科・細目：　情報学 計算機システム・ネットワーク

キーワード：　ネットワーク運用技術

１．研究開始当初の背景

インターネットにおけるネットワーク伝送技術の進歩は著しく、特にその伝送速度は飛躍的に伸びた。現在 Ethernet 規格では 10Gbps のネットワークインタフェースを搭載したネットワーク機器が市販されており、IEEE 802.3 ワーキンググループにより 40Gbps および 100Gbps の伝送速度を実現する規格の標準化作業が 2009 年に行われた。

一方で、そのような 超高速ネットワークを利用したシステムやアプリケーションにおいて、通信性能上の問題が発生するようになった。特に、高度な科学計算を行う高信頼計算システムでは、数値演算や冗長化を含めた記憶装置へのアクセス処理が頻繁に発生するし、通信回線ではなくシステムを構成するノード、I/O 性能が過負荷状態に陥った結果、通信性能が出ないという状況が、実システムにおいては多く発生していた。一般的にこのような過負荷状態のノードは最適な通信性能を発揮できず、結果としてディスクなどの計算機資源も効率的に利用できていなかった。ネットワークの高速化の傾向から、ディスク I/O や CPU の処理能力のようなネットワーク以外のノード性能の方が低いという不均衡は今後も続くと考えられる。

計算機システムがネットワークに接続されることが当然となった今、これは早急に解決されるべき問題であった。

２．研究の目的

ネットワーク観測に基づくシステム性能向上を実現するネットワーク自律制御手法の確立:

- 超高速ネットワークの観測情報からネットワークを利用するシステムの稼働状況を推測する手法
- 推測した稼働状況をもとに、システム性能の向上を目的としてネットワークを制御する手法

の 2 点について研究し、全体としてネットワーク観測とそれに基づくネットワーク制御のみで完結する、ネットワーク型システム性能最適化手法を確立することを目指す。すなわち、システムを構成するエンドノードである計算機側には何ら修正を加えないことが制約条件とし、そのことによりエンドノードですでに研究が行われている手法との独立性を保つとともに将来の協調を狙う。
ここでのシステムとは、高速に大容量のデータを授受するようなものを指し、ネットワークの伝送速度は 10Gbps 以上のものを想定する。
本研究の独特の着想は、ネットワーク観測のみで得られる情報がシステムの性能と相関を持っていることを仮定している点である。前述のようにシステムから能動的に制御を行う解決策の提案もあるが、その場合では各システムに対して個別に制御処理を定義・実装しなければならず、複数ユーザ、多様なアプリケーションが並列でバックボーンを利用している場合の実展開においては対応できない障壁がある。
我々の考える手法では、推測されるシステム稼働状況の情報の精度では劣る可能性があるが、観測及び制御の対象がネットワーク内で閉じているために、管理ネットワーク内で多様なシステムに一元的に対応でき、実展開の容易性の点で優れている。
この優位性を考慮して、本研究ではこのような方法を選択した。
また、手法が確率されれば、既存のノードにおける制御手法とも協調動作が可能であると考えている。

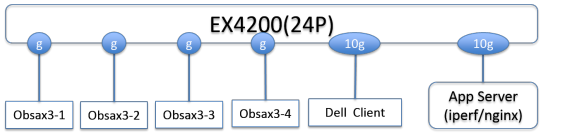
３．研究の方法

実験を担当する共同研究者が途中退職したため、実験の規模を縮小し、「高速ネットワークにおけるシステム性能の不均衡により発生するネットワーク性能の劣化を事例として示す」こと迄を目標に以下の実験を行った。

具体的には、ターゲットなるスイッチ Juniper 社製 EX4200 対して、
- 10Gb で接続されたサーバ 1 台～2 台
- 10Gb で接続された高速クライアント 1 台
- Gb で接続されたクライアント 1～4 台

を接続した環境を準備した。基本的ネットワーク性能について、ほぼ理論値通りに性能が出ることを確認した接続環境に、 http サーバを立て、実アプリケーションでの通信においては、帯域に余裕があっても性能低下があることを示す。さらに、ネットワークにおける性能向上制御の第一歩として、一部クライアントにある制限を加えた方が全体性能の出ることがあることを示す。実験環境における機器の性能、および、接続は以下の通り:

機器	CPU	RAM	Ports	OS
Obsax3-1	1.33GHz	3G	1G x 4	Linux 3.2.40
Obsax3-2	1.33GHz	1G	1G x 2	Linux 3.2.40
Obsax3-3	1.33GHZ	3G	1G x 4	Linux 3.2.40
Obsax3-4	1.33GHz	3G	1G x 4	Linux 3.2.40
Dell Client	2.93GHz	8G	1G x 2 + 10G x 1	Linux 3.2.0
App Server	2.93GHz	8G	1G x 2 + 10G x 1	Linux 3.2.0



(1) iperf による基本性能測定

iperf によって実験環境の基本的な性能測定を行った:

- サーバ： iperf 2.0.5 Server
- クライアント: iperf 2.0.5 Client

サーバ側の機械は Dell サーバで固定し、クライアント側を以下のように変化させた上で以下のそれぞれの通信パターンで 10 回ず

つ測定を行った:
- Obsax3 単体による単一セッション
- Obsax3 単体による複数セッション
- Dell Client 単体による単一セッション
- Obsax3 複数台による複数セッション
- Obsax3 複数台+Dell Client による複数セッション

性能評価は帯域性能による。

(2) アプリケーションによる測定 1
http サーバに対して、専用の測定ソフトウェアを使用して性能測定を行った:
- サーバ: nginx
  - Worker_connections: 2560
  - Ulimit -n 100000
  - Index.html: 151bytes
- クライアント: ApacheBench(ab)
  ab -n 500000 -c XX http://appserver/index.html

サーバ側の機械は Dell サーバ 1 台、もしくは、Dell 2 台をサーバとし、クライアント側のパターンを変化させ、5 分間隔(セッションの残りがないことは確認し)で、10 回測定を行った:
- サーバ 1 台とクライアント Obsax3 1 台
- サーバ 2 台とクライアント 2 台、別 VLAN での並行実行
- サーバ 2 台とクライアント 4 台、1:2 を別 VLAN で並行実行
- 前実験の開始時間を少しだけずらす

性能比較は秒当たりのリクエスト処理数による。

(3) アプリケーションによる測定 2
ファイルサイズ(index.html)を 1430 バイトとして、より大きな帯域時の性能を 10G クライアントの通信性能を制限しながら測定。
- Gb clients: ab -n 500000 -c 100
- 10G client: ab -n 2500000 -c 1000
- 10G client 接続ポートで帯域制限

性能比較は秒当たりのリクエスト処理数と通信帯域による。

４．研究成果

(1)iperf による基本性能測定
- Gb クライアント 1:1 単体毎
  Gb 間平均: 941Mbps
  10Gb 間: 9410Mbps(タグなし),
  9410Mbps(tag native), 6942Mbps(タグ)

Gb クライアントについては、回線性能が出ることを、10G クライアントについてはタグ使用時に 7G 程度の性能が出ることを確認した。

- Gb クライアント単体 複数セッション

機器・セッション	セッション1	セッション2	セッション3	合計
Obsax3-1	720.8Mbps	683.5Mbps	693.6Mbps	2097.9Mbps
Obsax3-3	686.9Mbps	685.3Mbps	738.5Mbps	2110.7Mbps

今回使用するクライアントの性能では、複数ポート使用時には回線性能合計値までは出ないことを確認した。

- Gb4 台同時

機器・セッション	セッション1	セッション2	セッション3	合計
Obsax3-1	939.7Mbps	X	X	
Obsax3-2	940Mbps	X	X	
Obsax3-3	X	940.9Mbps	X	
Obsax3-4	X	X	940Mbps	
合計				3760.6Mbps

クライアントを独立にすれば、それぞれ独立のときと同様の性能が出ることを確認した。

- Gb4 台、10G クライアント 1 台

機器・セッション	セッション1	セッション2	セッション3	合計
Obsax3-1	938.7Mbps	X	X	
Obsax3-2	939.4Mbps	X	X	
Obsax3-3	X	940.6Mbps	X	
Obsax3-4	X	939.7Mbps	X	
Dell Client (10G)	X	X	5631Mbps	
合計				9389.4Mbps

Gb は単体とほぼ同じ、10G クライアントは単体のときの 80%程度の性能となりサーバ側のポート性能 10G 近くの通信性能が出た。

- Gb クライアントのセッションを増やす

機器・セッション	セッション1	セッション2	セッション3	合計
Obsax3-1	674Mbps	706.3Mbps	684Mbps	
Obsax3-2	940.6Mbps	X	X	
Obsax3-3	700.7Mbps	762.7Mbps	614Mbps	
Obsax3-4	X	939.8Mbps	X	
Dell Client (1G)	X	X	941Mbps	
合計				6963.1Mbps

機器・セッション	セッション1	セッション2	セッション3	合計
Obsax3-1	802.8Mbps	833.3Mbps	X	
Obsax3-2	940Mbps	X	X	
Obsax3-3	X	938.1Mbps	936.5Mbps	
Obsax3-4	X	X	940.3Mbps	
合計				5391Mbps

単体測定時と同じくクライアント単体の性能はポート理論値合計を下まわるが他のクライアントの性能には影響しなかった。

(2) アプリケーションによる測定 1
- 単体毎

機器・セッション数	1000	500	300	200	100	50
Obsax3-1 (RPS)	6819	6747	7142	X	X	X
Obsax3-3 (RPS)	7126	7400	7667	7831	7962	7995
Obsax3-4 (RPS)	6735	7068	7251	7388	7516	7549

単体性能について、セッション数 50 迄は数が減るほど出ることを確認した。

- クライアント Gb サーバ 1:1 を 2 組

機器・セッション数	1000	500	200	100	50
Obsax3-1 (RPS)	6375	6600	6599	6836	6909
Obsax3-3 (RPS)	6797	6587	7092	7329	7272

並行で通信がある場合には、ない場合より少し性能が低下し、かつ、セッション数の減少に対して単調に性能が増加しなくなった。

- クライアントサーバ 2:1 を 2 組 x 2 回

機器・セッション数	1000	644	500	200	100	80	60	50	30
Obsax3-1 (RPS)	6125	6101	6243	6495	6584	6731	6488	6406	6502
Obsax3-4 (RPS)	6417	6248	6337	6721	6699	6853	6804	6658	6640
Obsax3-2 (RPS)	6455	6287	6317	6539	6915	6893	6920	6847	6841
Obsax3-3 (RPS)	6461	6302	6284	6580	6958	7077	6847	6954	6838

機器・セッション数	100	90	85	80	75	70	65	60	55
Obsax3-1 (RPS)	5893	5761	5791	6297	6292	6014	6195	6020	5808
Obsax3-4 (RPS)	5527	5428	5649	5824	5929	5592	5876	5875	5975
Obsax3-2 (RPS)	5676	5573	5700	5780	5872	5637	5867	5743	5790
Obsax3-3 (RPS)	5873	5789	5902	6169	6039	6074	6279	6168	5954

性能とピークセッション数に揺れがあるがほぼ 80 周辺で最高値が出ることを確認した。

(3) アプリケーションによる測定 2
- 10G クライアント単体

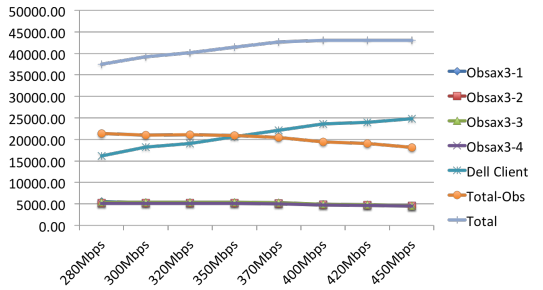
機器・セッション数	8000	7500	7000	6500	6000	5500
Dell Client (RPS)	27241	26193	27755	26919	27108	25044

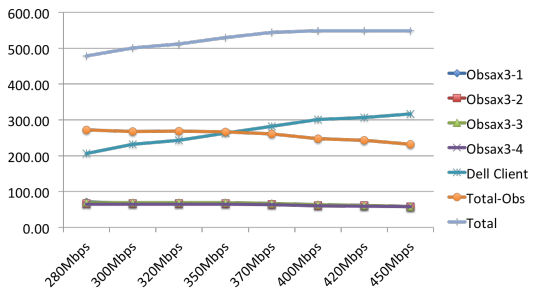
取得ファイルを大きくしたことでピーク性能の出るセッションは 7000 周辺になった。

- 全クライアント

横軸は 10G クライアントへの帯域制限:

機器・帯域	280Mbps	300Mbps	320Mbps	350Mbps	370Mbps	400Mbps	420Mbps	450Mbps
Obsax3-1	5595	5238	5321	5207	5096	4886	4757	4521
Obsax3-2	5235	5235	5232	5218	5114	4831	4771	4566
Obsax3-3	5416	5424	5408	5386	5286	4973	4842	4566
Obsax3-4	5089	5092	5095	5077	4988	4731	4655	4505
Dell Client	16143	18213	19055	20588	22138	23595	24011	24825
Obs合計	21335	20989	21056	20888	20484	19421	19025	18158
総合計	37478	39202	40111	41476	42622	43016	43036	42983





帯域、RPS ともに、帯域制限がない場合には 10G クライアント 1 台で他 Gb クライアント 4 台を上まわる。また、10G クライアントへの制限を掛ける場合には、より帯域を絞る程、Gb クライアントの合計性能は上昇する。しかし、システム全体、つまり、10G+Gb の合計性能については、10G ポートを 420Mbps に制限したときに(Gb 合計、10G のどちらでもピークでもないにもかかわらず)、ピーク性能となり、制限がその上でも下でもサーバとの通信性能の合計は低下する。

(4) 考察と今後

実アプリケーション利用時のネットワーク性能は、サーバ、クライアント、および、アプリケーションソフトウェアに依存し、基本的性能が出ることが確認されている環境においても、十分に出ないことを確認した。
また、並行通信があるとその影響も少なからず出ることが確認できた。

さらに、そのような状況下で、性能の違うクライアントを同時に利用している場合に、高性能クライアントの帯域を通信環境において制限すれば、それ以外のクライアントの通信性能は、当然上がるが、システム全体の通信性能は、下がるのではなく、帯域前よりも向上する点があることが実験で確認できた。

つまり、クライアントサーバの環境の通信設定を変更しなくとも、アプリケーション、通信環境に応じて、通信路の制限によって性能向上はありうる。

今回はその確認までとなったため、十分な知見が得られたとは言えなかったが、以下の条件を変化させたときに同様なシステム全体の通信性能を上げる通信路における制限を定式化したいと考えている:

- 転送ファイルの大きさ(ランダムも含む)
- クライアント数
- 並行通信数
- 別アプリケーション
- 高負荷時(10G を使い切るような通信を並行で行った場合)

5. 主な発表論文等

(研究代表者、研究分担者及び連携研究者には下線)

〔雑誌論文〕(計 0 件)

〔学会発表〕(計 0 件)

〔その他〕
ホームページ等

6. 研究組織

(1)研究代表者
友石 正彦 (TOMOISHI, Masahiko)
東京工業大学・大学マネジメントセンター・教授
研究者番号: 60262284

(2)研究分担者 平成 25 年 3 月 22 日辞退
益井 賢次 (MASUI, Kenji)
東京工業大学・学術国際情報センター・特任助教
研究者番号: 60531340