広島工業大学紀要研究編
第47巻（2013）127–131
論　　文

大規模データ収集システムにおける通信の効率化に関する研究

長坂　康史*・沖　恭志**

（平成24年10月31日受付）

A study on an improvement of communication efficiency in a large-scale data acquisition system

Yasushi NAGASAKA and Kyoshi OKI

(Received Oct. 31, 2012)

Abstract

A data acquisition, DAQ, system with using the Internet technology is used in many fields these days. But the Internet is a general purpose technology and suitable to transfer data randomly. Moreover, as a lot of clients and servers are connected into the Internet, the network topology of it is N to M, i.e., the N source computers are connected into the M destination ones. On the other hand, the topology of DAQ systems is based on N to 1, because a measurement is performed with using several detectors and their data have to be sent to one destination computer to analyze them. The Internet technology is not efficient for transferring in N to 1, and such traffic occurs congestion in the network. We, therefore, focus on congestion control mechanisms in the TCP protocol to avoid the congestion and propose the efficient mechanism for a large-scale DAQ system.

Key Words: network traffic, congestion control, data acquisition system

1．はじめに

現在，TCP/IP プロコルはインターネットを中心にして多くのネットワークで使用されている。インターネットの使用目的は Web ページの閲覧や電子メール，ファイル共有，IP 電話など多岐にわたる。そのため HTTP や SMTP，POP3，IMAP，FTP など様々なアプリケーションプロトコルが混在する環境下で，平均的にみたときに，より高い性能を出せるよう進化してきた。

これらのプロトコルを実現するために，TCP/IP プロトコルのトランスポート層には TCP と UDP の二つのプロトコルがある。TCP プロトコルには信頼性や公平性が高いという特徴があり，相手との確実な通信を行う時に利用される。一方，UDP には信頼性は低いが通信速度が速いという特徴があり，リアルタイム性が必要とされるときなどに利用される。

TCP プロコトルには，その特徴である信頼性や公平性を実現し，維持するための機能の一つとして，輻輳制御の仕組みが組込まれている。これはネットワークが混雑し通信が滞るのを防ぐための機能である。この輻輳制御もあらゆる環境で使用されることが想定されており，汎用性の高い機能となっている。

一方，物理学実験などの検出器を利用した実験でも検出器から得られるデータの収集にはコンピュータが利用される。さらにコンピュータを複数利用する場合には，ネットワークを構築しデータ通信をする必要がある。特に，近年，質量の起源であると言われているヒッグス粒子の探索などで注目を集めている高エネルギー物理学実験では検出器から得られるデータは膨大な量となる。さらに，検出器の規模の拡大に伴い，得られるデータ量はますます膨大なものとなってきている。そのため，検出器で得られるデータを効率良く，また，高速に収集することが要求されている。

* 広島工業大学情報学部情報工学科
** 広島工業大学工学系研究科情報システム科学専攻（現　日本アイ・ビー・エム共同ソリューション・サービス株式会社）

このデータを収集するために数千ノードを閉じたネットワークで構成するデータ収集システムがある。これらの実験では，検出器が検出したデータをイベントと呼び，このイベントが発生するタイミングでそのイベントに関する情報を一斉に一箇所に集める必要がある。

このような特殊な環境では，膨大なデータを効率良く収集する上で，汎用的な TCP/IP の輻輳制御が問題となる場合がある。このことから，データ収集システム全体の処理能力の向上には，TCP/IP を適用システムに最適化する必要があると考える。そこで本研究では，データ収集システムに最適な輻輳制御の検討を行い，より効率よく，また，高速なデータ収集を可能にすることを目的とする。

2．対象システム

2.1　ATLAS 実験

本研究では ATLAS 実験で使用されるデータ収集システムである TDAQ システム（ATLAS Trigger and Data Acquisition System）を対象とする。

ATLAS 実験とは欧州合同原子核研究機構（CERN）の大型ハドロン衝突型加速器（LHC: Large Hadron Collider）で行われている史上最高エネルギーの素粒子実験である。LHC はスイス・ジュネーブ郊外の地下にある円形の加速器である。LHC には ATLAS 実験の他にいくつかの実験グループがあるが，ATLAS 実験では質量の起源であると言われているヒッグス粒子や，宇宙の暗黒物質の解明につながる超対称性粒子の発見などを目的としている。近年ではヒッグス粒子の探索で大きな注目を集めている。

ATLAS 実験で利用される検出器は，縦 25 m，横 44 m と巨大なもので，内部はピクセル検出器，半導体飛跡検出器，遷移輻射飛跡検出器，ソレノイド電磁石，液体アルゴン電磁カロリメータ，タイルカロリメータ，ミューオンシステム（トロイド電磁石，精密ミューオン検出器，ミューオントリガー検出器）などで構成されており，検出器からのデータの読み出しチャンネル数は合計で 1 億 4 千万チャンネルにもなる。

2.2　ATLAS TDAQ

TDAQ のシステム概略図を図 1 に示す。このシステムは数千ノードで構成される。また，LHC 加速器のビーム衝突頻度は 40 MHz にもなり，測定したすべてのデータを収集・保存することができないため，TDAQ では必要なデータを収集するため，3 段階のトリガ，すなわち，フィルタを設けている。

最初のトリガは Level-1（LVL1）で，熱量計やミューオン検出器などから得られるデータから不要なデータをハードウェアレベルで削除することで，イベントレートを 100 kHz とする。この LVL1が条件を満たすと，検出器毎に用意されている Read-Out Driver（ROD）がデータを収集し，Read-Out System（ROS）に収集データを送る。ROS は約150台のコンピュータによって構成されている。また，イベントデータの収集と並行して，LVL1からのトリガ情報をもとに，二つ目のトリガである Level-2（LVL2）にデータを送る。LVL2では部分的に ROS からイベントデータを収集し，解析した結果，必要であれば，データをその後の DataFlow Manager（DFM）に送る。DFM は，SFI に必要なデータの情報を知らせ，ROS からイベントデータを取得させる。SFI は約100台のコンピュータで構成されている。このときのイベントレートは，LVL2トリガによって 3 kHz 程度となっている。その後，イベントデータは，最後のトリガである EF によって選別され，イベントレートとしては 200 Hz 程度となり後段の SFO へ送られ，最終的には，ストレージに記録される。

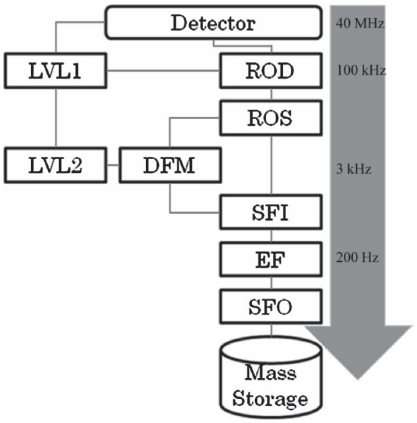

図 1．TDAQ システムの概略図

2.3　課　題

TDAQ を構成する数千ノードは全て閉じたネットワークで繋げられている。このシステムでは，TCP と UDP の両プロトコルが利用されている。

イベントが発生すると，そのイベントに関するすべての情報を同じタイミングで一箇所に集めなければならないことから，多くのコンピュータから一斉に一箇所へデータが送信されることになり，輻輳が発生し通信効率が落ちてしまう。 このように特殊な環境のため，効率良くデータ通信を行うためには，汎用的な TCP/IP を何も変更なく利用しているだけでは難しい。

また，今後イベントレートがさらに高くなることが予想されており，それらの要求に対応するためにより効率良く通信を行うシステムが必要となる。

そこで，本研究では，TCP の輻輳制御方式に注目し，システムに最適な輻輳制御方式を選択することで通信効率の向上に繋げる。

3．輻輳制御

ネットワークが混雑している輻輳状態では，その状態が長く続くと通信は滞ることになる。これを回避するための機能がTCPプロトコルに組込まれている輻輳制御機能である。TCPはエンドツーエンドで通信を行うので途中のノードの情報を持たない。そこでTCPは一度に送るデータ量を，始めは少なくし，次第に増やしていく方式を取っている。パケットロスの検出などにより，輻輳を検知すると再び一度に送るデータ量を減らす。このような輻輳制御方式を採用しているが，これまでに，特定の環境に特化した改良型の方式が提案されている。

今回はScientific Linux CERN（SLC）5.7 kernel 2.6.18でモジュールとして利用可能な以下の10の輻輳制御方式について検討する。なお，SLC5.7標準の輻輳制御方式はBICである。

- Binary Increase Congestion（BIC）control[1]
- CUBIC TCP
- High Speed TCP
- H-TCP
- TCP-Hybla congestion control algorithm
- TCP Low Priority
- Scalable TCP
- TCP Vegas[2]
- TCP Veno
- TCP Westwood+

これらの中でVegasのRTTを利用して輻輳制御を行う方式がデータ収集システムに向いていると考えた。

4．性能評価実験

4.1　性能評価実験概要

データ収集システムではイベント発生時には複数のコンピュータから1台のコンピュータに一斉にデータが送られる。この時輻輳が発生し，輻輳制御方式による差が表れると考えられる。そこでデータ収集システムに類似した環境を構築し，輻輳制御方式の違いによる性能差の測定を行う。

性能評価はネットワークのベンチマークソフトであるnetperfを利用して実際に測定するとともに，ネットワークシミュレータであるNS2を利用して，シミュレーションでの検証も実施する。netperfを利用した測定では，複数のコンピュータから1台のコンピュータに対して一斉にデータを送信しその時のスループットを各輻輳制御方式で測定する。NS2を利用したシミュレーションでも同様に，複数のノードから1つのノードに向けて一斉にデータを送り，複数の条件でスループットを比較する。

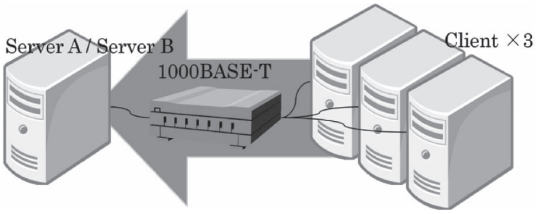

図2．実験の概略図

4.2　各輻輳制御方式でのスループット測定

4.2.1　実験方法

netperfを使用し3台のコンピュータから1台のコンピュータに向けて一斉にデータを送り，スループットの計測を行う。計測時間は60秒で，スイッチとServer間のスループットを求める。送信データサイズは1 Byteから128 KByteとして計測した。実験の概略図を図2に示す。Serverには比較的高性能なものと低性能なものの2台を用意した。これはServerの性能差により輻輳制御方式の差がより顕著に表れるのではないかと考えたからである。

なお，使用機器は以下の通りである。

(1) Server A
CPU: AMD Athlon 64 X2 DualCore Processor 6000+
メモリ：4 GB
NIC: Realtek Semiconductor Co., Ltd. RTL8111/8168B PCI Express Gigabit Ethernet controller

(2) Server B
CPU: Intel Celeron CPU 1200 MHz
メモリ：192 MB
NIC: Realtek Semiconductor Co., Ltd. RTL-8169 Gigabit Ethernet

(3) Client×3台
CPU：Intel Xenon 2.40 GHz 2コア
メモリ：1024 MB
NIC: Intel Corporation 82540EM Gigabit Ethernet Controller

(4) スイッチ
Planex FXG-16IRM
パケットバッファ：340 KB
スイッチングファブリック：32 Gbps

4.2.2　結　果

netperfを利用した性能測定の送信データサイズに対するスループットを図3，および図4に示す。

どちらのServerでもBICとVegas以外の輻輳制御方式はBICとほぼ同一の結果となったためグラフでは省略している。Vegasは高性能なServer Aを利用した場合は低いスループットとなり，逆に低性能なServer Bでは高いスループットとなった。

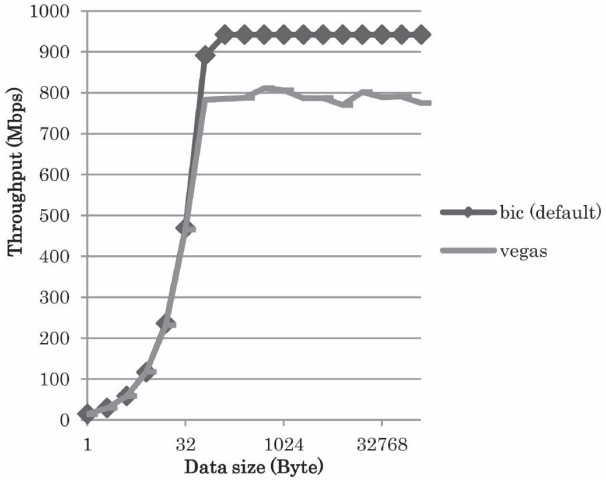

図 3．Server A の結果

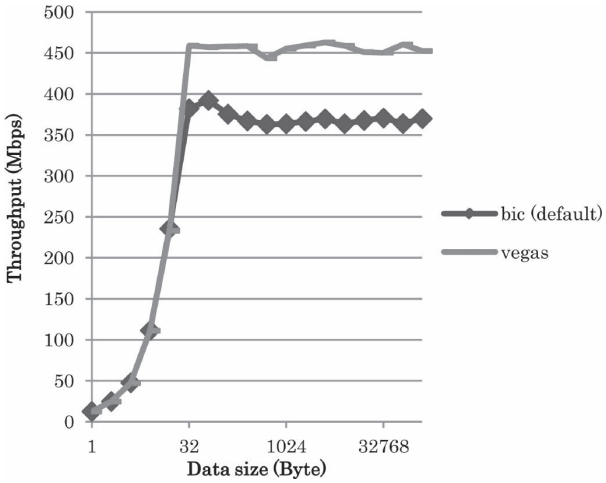

図 4．Server B の結果

4.3　シミュレーションによるスループットの比較

4.3.1　実験方法

複数の条件下でシミュレーションを行う。そしてシミュレーション結果からスイッチと受信ノード間のスループットを求め比較する。

シミュレーションを行うのは複数台（1，2，3，5，10台）対 1 台の通信である。ネットワークの帯域は 1000 Mbps，遅延は 0.1～50 ms（0.1, 0.5, 1, 5, 10, 50 ms）とする。輻輳制御方式は，netperf を利用した性能評価で他の輻輳制御方式と大きく異なった結果を出した Vegas と，SLC 5.7で標準の輻輳制御方式である BIC を対象とする。またデータ収集システムと同様に同じタイミングで一斉にデータが送信されるように一定間隔ごとにデータ送信を行う。この間隔を 0.1～50 ms（0.1, 0.5, 1, 5, 10, 50 ms）とする。

4.3.2　結　果

シミュレーションの送信ノードを 3 台としたときの遅延時間と送信間隔ごとのスループットを図 5 と図 6 に示す。また，表 1 に示した各条件でスループットを比較し，BIC のスループットが高い場合は正の値，Vegas のスループットが高い場合は負の値で，Vegas のスループットと BIC のスループットの差を表している。

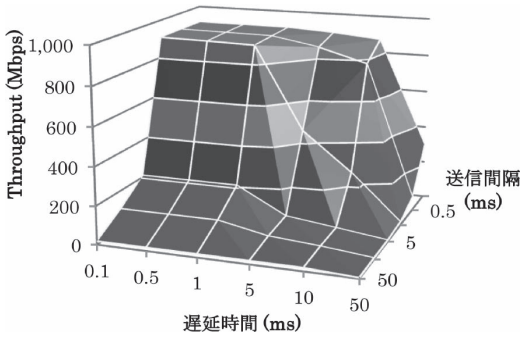

図 5．BIC の結果

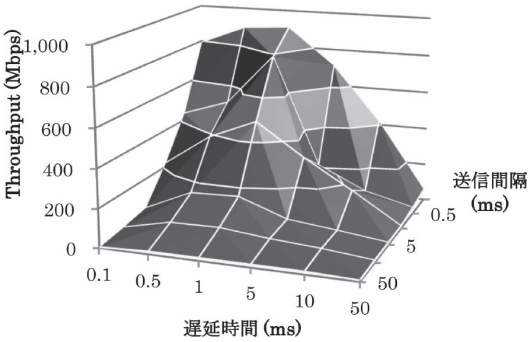

図 6．Vegas の結果

表 1．BIC と Vegas の比較

送信間隔（ms） ＼ 遅延時間（ms）	0.1	0.5	1	5	10	50
0.1	−115.27	−16.29	+0.16	−180.68	−450.56	−240.74
0.5	−471.62	−386.29	−112.80	−643.64	−689.92	−43.50
1	−688.97	−513.32	−403.83	−135.26	−84.24	−5.45
5	−166.05	−54.04	−46.79	−1.98	+2.45	+4.90
10	−82.92	−27.16	−23.62	+1.83	+3.55	+4.56
50	−14.14	−3.38	−2.93	−2.69	−2.70	+2.79

4.4　考　察

標準の BIC は高速だが，距離の長いネットワーク，すなわち，遅延時間が大きいネットワークに最適化されたものであり，閉じたネットワークである TDAQ には適さないと考えていた。しかし，4.1の実験では Server A と Server B では反対の結果となった。性能の高い Server A では Vegas だけが低く，性能の低い Server B では Vegas だけが高いスループットを記録した。また，シミュレーションの結果から，多くの条件で BIC が性能の高い結果となったが一部では RTT から輻輳を予測し，制御を行う Vegas のほうが高性

能であった。特に遅延時間，送信間隔共に大きい場合は，パケットロスではなく遅延時間から輻輳を検知する Vegas のほうが高い性能を記録する傾向にあることがわかった。

5．まとめ

本研究では，CERN の LHC で行われている ATLAS 実験で使用されているデータ収集システム TDAQ を対象に，データ収集システムに最適な TCP の輻輳制御方式の検討を行った。

性能評価実験をネットワークのベンチマークソフトである netperf とネットワークシミュレータである NS2を利用して行った。netperf は複数台のコンピュータから同時に 1 台へ向けてデータを送信し，各輻輳制御でスループットを計測した。結果は条件により異なるものだったが Vegas が他の輻輳制御方式に対して大きく異なる結果を得た。次にシミュレーションを利用した性能評価実験は Vegas と BIC に着目して行った。複数の送信ノードから指定した送信間隔で受信ノードにデータを送ることを想定した。実験の結果，全体的に BIC が高性能であることがわかったが，送信ノード数に関わらずネットワークの遅延，および，送信間隔が共に長いときは Vegas の性能が高くなる傾向にあることがわかった。

今後はより実環境に近い設定でシミュレーションを行うことや，実際に高校率な通信が行えるかの確認を行う必要がある。また，TDAQ 向けに改良を加えていくことも必要である。

文　献

［1］Lisong Xu, Khaled Harfoush, and Injong Rhee, “Binary Increase Congestion Control for Fast, Long Distance Networks”, Infocom, IEEE,（2004）.

［2］Brakmo, L. S., Peterson, L. L., “TCP Vegas: End to End Congestion Avoidance on a Global Internet”, IEEE Journal on Selected Areas in Communication, Vol. 13, No. 8, pp. 1465–1480,（1995）.