L-028

大規模データ収集システムにおける無損失転送技術を用いた通信効率の向上に関する研究
A Study on improvement of a communication efficiency using the lossless transmission

山木戸 啓亮† 長坂 康史† 堀 裕貴 † 西村 俊彦†
Keisuke Yamakido Yasushi Nagasaka Hiroki Hori Toshihiko Nishimura

1． まえがき

ATLAS 実験[1]-[3]などの高エネルギー物理学実験では，巨大な測定器で測定されるデータを数千規模のコンピュータによって構成された大規模データ収集システムで処理している．このシステムでは，多数のスイッチを介した並列処理を行っており，測定データを一つの解析専用コンピュータへと集めるために多対一の TCP/IP 通信を行っている．

しかし，各通信経路の I/O ボトルネックなどが招くデータの輻輳によってパケットロスが散発的に発生すると，TCP コネクションでは再送処理や輻輳制御が行われ，通信効率が低下してしまうという問題がある [4]．

そこで，本研究では，大規模データ収集システムに関する輻輳問題を解決する方法として Data Center Bridging (DCB) 技術の導入を提案し，通信中に発生するパケットロスの抑制や，各フローの通信速度公平性を確保することで，通信効率を高めることを目的とする．

2． Data Center Bridging (DCB)

一般的な LAN 環境でのパケットロス対策では，TCP のベストエフォート型データ通信を基準とする一方で，SAN（Storage Area Network）環境では，ストレージのバックアップ時にパケットロスが与える影響が大きいため，FibreChannel などのロスレスプロトコルが用いられてきた．

また，多くのデータセンタでは，LAN や SAN の混在化によって，通信環境が複雑化し運用や保守が困難になっている．このため，複雑化する通信環境に対して，ネットワーク機器の統合を図るために，IEEE802.1 WG 内の DCB Task Group[5] において DCB の開発や標準化が行われている．

DCB とは，高い信頼性が要求される FibreChannel のデータを，信頼性が低いイーサネット上で通信するための技術である.この技術は，FibreChannel over Ethernet (FCoE) のために開発されたロスレスイーサネット技術の一つである．DCB は，従来の イーサネットの機能を拡張した一連の規格で構成され，主要な規格として，優先度に基づいてトラフィックを区別する Priority-based Flow Control (PFC)や，輻輳の検知と制御を行う Congestion Notification (CN)，優先度に基づいた通信帯域の保障を行う Enhanced Transmission Selection (ETS)と呼ばれる規格がある．

3. 大規模データ収集システムの通信特性

3.1. データ収集トラフィック

高エネルギー物理学実験の大規模データ収集システムは，測定データの収集に多数のネットワークスイッチとそれらに接続された多数のコンピュータで構成される．実験で得られた測定データは，これらの多数のコンピュータ上に散在しており，最終的にはそれを一箇所にまとめ，処理する．

また，同時刻に発生したデータをまとめて処理するために，ネットワークに接続された複数台のコンピュータ上に散在する測定データを，同じタイミングで一斉に一台の解析コンピュータに集める必要がある．

このため，大規模データ収集システムの通信特性は，多対一の同期トラフィックとなる．また，システムの通信効率を高めるためには，通信遅延や同期性の崩壊を引き起こす輻輳やパケットロスの発生を抑制しなければならない．

3.2. 既存システムの課題

ATLAS 実験で構築されている大規模データ収集システムの Testbed において，一般的な TCP の QoS 制御を用いた場合の通信速度の測定を行った．測定環境は，一台のスイッチを介して四台のコンピュータが接続されており，三台のコンピュータが同時に一台のコンピュータへデータを 60 秒間送信している．この三対一の通信環境における測定結果を図 1 に示す．

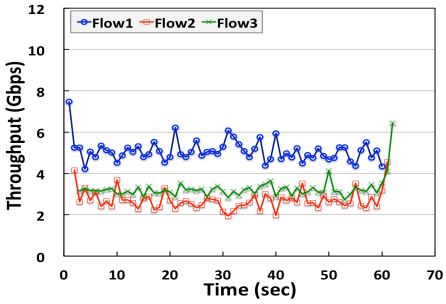

図 1 既存システムでの通信速度(多対一環境)

測定の結果，各フローの平均通信速度は 3.903±0.689 Gbps，パケットロスが各フロー合計で 5,442 Frame 発生した．また，各フローでは輻輳によるパケットロスが散発的に発生し，各フローそれぞれの輻輳ウィンドウが非同期的に減少してしまうことから，各フローの通信速度に公平性がなく，輻輳がシステムに悪影響を及ぼしている．

4. 提案手法

大規模データ収集システムの現状より，輻輳によるパケットロスが通信速度と公平性を低下させてしまうことから，可能な限りパケットロスを抑制しなければならない．

このため，本研究では，大規模データ収集システムの通信効率を高めるために，システム内のスイッチングネットワークに対して，DCB によるロスレス化を提案する．

具体的には，システム内のネットワークに DCB をサポートしたスイッチや NIC を実装することでロスレス・ネットワークを構築する．このネットワークは，イーサネットがベースとなっており，従来の環境で使用していた TCP/IP 通信を行う機器の可用性を維持することが出来る．また，ホスト間の各コネクションでは，IEEE802.1Q の VLAN タグ内にある 3bit 長の PCP(Priority Code Point)を有効化した．この PCP によって，各フローは最大 7 つの優先度を持つことができ，前述の PFC や ETS の機能を組わせることによって，優先度別に最低通信帯域の保証や Pause フレームの柔軟な制御が可能になり，既存システムの QoS を拡張することが出来る．

† 広島工業大学, Hiroshima Institute of Technology

5. 性能評価

DCB の有効性を調査するために評価実験を行った．DCB を Linux 上に実装し，イーサネットで構成された通信環境での通信効率や信頼性の評価を行った．

評価を行うにあたり，スイッチを介しパケットを中継する環境を構築した．また，これらの各通信環境において，DCB の有効性を検証するために，DCB 機能を無効に設定した従来のイーサネット環境と，DCB 機能を有効に設定した環境で性能比較を行った．この性能比較では，多対一の環境を構築し，10 Gbps 環境の通信速度とパケットロス数の測定を行った．なお，通信速度を測定するネットワークベンチマークツールには iperf3[6]を用いた．

5.1. 性能評価概要

実験では，DCB の規格に対応した Brocade 社の NIC BR1860 を搭載した四台のホストと同じく Brocade 社製の DCB 規格に対応したスイッチ VDX6740T-1G を利用した．実験環境を図 2 に，使用する PC の性能を表 1 に示す．

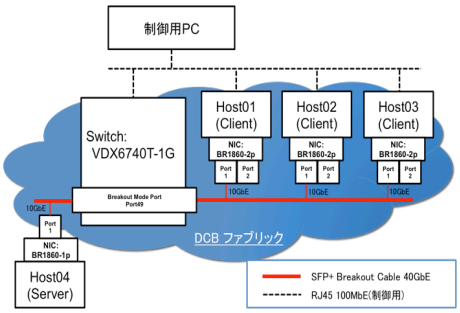

図 2　実験環境

表 1　PC および通信機器(Host01-04)

CPU / RAM	Intel(R) Xeon(R) CPU E5-1410 v2 / 8192 MB
NIC / Cable	Brocade BR1860 / Brocade SFP+
OS	ScientificLinux6.6 X64_86
SWITCH	VDX6740T-1G

5.2. 測定結果

図 2 に示す実験環境において，三対一の測定を行った．この測定では，Client の Host01-03 から Server の Host04 に向けて 600 秒間の測定を行った．また，本環境での測定結果を表 2 に，DCB を無効にした場合の通信速度を図 3 に，DCB を有効にした場合の通信速度を図 4 に示す．

表 2　三対一環境の測定結果

DCB	平均通信速度	パケットロス数
有効	3.139±0.015 Gbps	0 Packets
無効	3.139±0.698 Gbps	8629 Packets

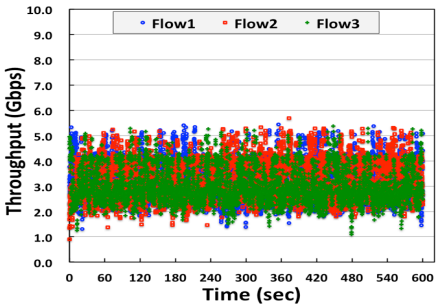

図 3 三対一環境のフロー別通信速度 (DCB:無効)

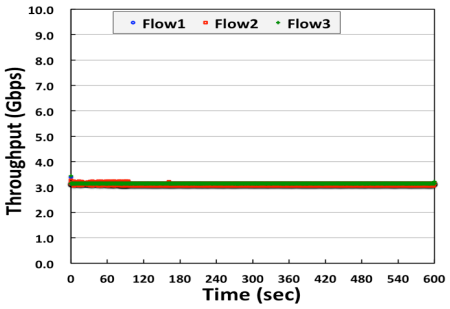

図 4 三対一環境の通信速度 (DCB:有効)

5.3. 考察

三対一の通信環境では， DCB の有効・無効で平均的な通信速度に差異は認められなかった．また，全フロー合計のパケットロス数に関して，表 2 より，DCB 無効では，パケットロスが多発した結果となったが，有効時には 0 回となっている．

各フローの通信速度公平性に関しては，図 3 の DCB 無効時のばらつきが 0.698 Gbps と非常に大きいのに対して，DCB 有効では 0.015 Gbps と対照的に小さく，図 4 の様に各フローの公平性が得られている結果となった．これらの結果より，パケットロスの発生を限りなく抑えることが出来たため，公平性に関しても高い結果が得られたと考える．

6. まとめ

本研究では，大規模データ収集システムが抱えるパケットロスや通信速度の非公平性などの問題に対して，ロスレスイーサネット技術である DCB を用いて解決を図ることを提案した．性能評価では，実験環境において多対一通信の場合の通信速度や各フローの公平性について確認を行った．測定結果より，高い輻輳制御によるパケットロスの抑制や，通信速度の公平性を確保した．このことから，大規模データ収集システムに DCB を実装することで，通信効率を高めることが出来ると考える．

今後はシステムとしての有効性を追求するために，ホスト数及びスイッチの台数を増やし，より多くのフローが存在する環境での評価を行い，数千台規模の機器が通信を行う大規模データ収集システムに対するより深い考察が必要であると考える．

参考文献

[1] Jinlong Zhang, et al., "ATLAS Data Acquisition", Real Time Conference, pp.240-243,May 2009.

[2] I. Riu, et al., "Integration of the Trigger and Data Acquisition Systems in ATLAS", Nuclear Science, IEEE Transactions on, Vol.55, pp.106-112, Feb 2008.

[3] J. Vermeulen, et al., "ATLAS DataFlow: The Read-Out Subsystem, Results From Trigger and Data-Acquisition System Testbed Studies and From Modeing", Nuclear Science, IEEE Transactions on, Vol.53, pp.912-917, June 2006.

[4] 沖恭志，長坂康史,"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"，広島工業大学紀要 研究編，Vol.47，pp.127-131，2013.

[5] Data Center Bridging Task Group, http://www.ieee802.org/1/pages/dcbridges.html/ (Accessed 2015-01-21)

[6] iperf3, http://software.es.net/iperf/ (Accessed 2014-08-19)