様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月　 ６ 日現在

機関番号：１５３０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2022
課題番号：１９Ｋ１１９２６

研究課題名（和文）データセンターネットワークにおけるスループット急落の回避に関する研究

研究課題名（英文）Studies on avoidance of throughput degradation in data center networks

研究代表者
横平　徳美（Yokohira, Tokumi）

岡山大学・ヘルスシステム統合科学学域・教授

研究者番号：５０２２０５６２

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：データセンターネットワークで使用されている分散ファイルシステムにおいては、トランスポート層プロトコルとしてTCP (Transmission Control Protocol)を用いている場合、多数のサーバとクライアント間の通信において、スループットが急落するという現象（インキャスト）が生起するという問題点があった。本研究では、インキャストを回避するための方法を考案し、それらの有効性をシミュレーションにより確認した。

研究成果の学術的意義や社会的意義
AIを駆使して大量のデータを分析することにより、意志決定や戦略を実行するデータ駆動型社会に急速に移行しようとしている。大量のデータを分析するために、1つのファイル全体を1つのサーバに格納する形式の集中ファイルシステムを使用した場合、ディスク入出力がボトルネックとなり、速度を向上させるのは難しい。そこで、最近のデータセンターでは、1つのファイルを多数のユニットに分割し、それらを別々のサーバに格納する形式の分散ファイルシステムを採用するようになって来ている。本研究の成果は、分散ファイルシステムの性能を高く維持するために利用できるものであって、その学術的意義や社会的意義は大きいと考えられる

研究成果の概要（英文）：n distributed file systems used in data center networks, when TCP (Transmission Control Protocol) is used as the transport layer protocol, there is a problem that the throughput drops sharply when communicating between many servers and clients. The problem is called Incast. In this study, we have proposed methods to avoid Incast and investigated their effectiveness through simulations.

研究分野：情報ネットワーク

キーワード：インキャスト　データセンターネットワーク　分散ファイルシステム

１．研究開始当初の背景

データセンターネットワークでは、分散ファイルシステムが用いられているが、分散ファイルシステムでは、ファイル入出力を要求するクライアントとそのファイルを構成する各ユニットを保持するサーバとの通信は、通常、インターネット標準のトランスポートプロトコルであるTCP (Transmission Control Protocol)が用いられる。しかし、このような分散ファイルシステムと TCP の組合せの環境では、以下に述べるような深刻な問題が発生する可能性がある。

クライアントがあるファイルを読み込もうとする場合、クライアントの TCP は、そのファイルを構成する各ユニットを保有しているサーバの TCP とコネクションを設定し、そのコネクションを使用して、各サーバが対応するユニットを送信することで、クライアントは所望のファイルを得ることができる。この際、すべてのサーバがほぼ同時にユニットをクライアントに送ろうとするので、ネットワーク内部の一部のリンクには多くのパケットが集中的に到着することになり(図 1)、それらのリンクの接続されたポートのバッファ容量(バッファサイズ)が十分でない場合、多くのパケットがバッファーオーバーフローにより失われてしまう可能性がある。特に、クライアントの接続されるリンクでは、このような可能性が高くなる。このとき、いくつかのサーバの TCP では、タイムアウトが起こるまで待ってから失われたパケットを再送するが、この待ち時間(タイムアウト時間)の最小値は TCP の標準実装では 200ms であって、典型的なデータセンターネットワークにおけるラウンドトリップタイム(以下、RTT と略記する)である約 200μs に比較して極めて大きい。このため、1 つのファイルを構成する全ユニットをクライアントが受け取るまでの時間が大きくなり、スループットで見れば 0 に近い極端に小さい値になってしまう。このような現象は TCP-Incast (以下、インキャストと呼ぶ)と呼ばれ、データセンターネットワークにおける解決すべき重要な課題の一つとなっている。

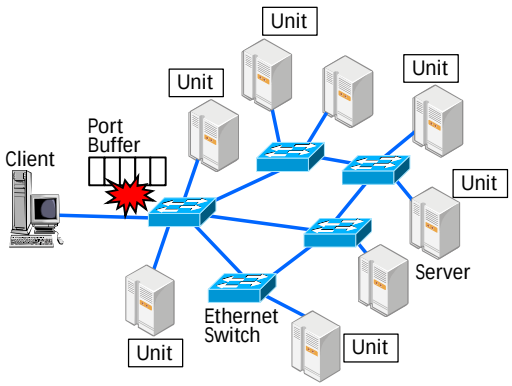

図１　クライアントリンクへのトラヒック集中

２．研究の目的

本研究の目的は、TCP 通信に基づく分散ファイルシステムを用いるデータセンターネットワークのインキャスト回避法を確立することである。

インキャストの起こる根本的な原因は、多くのサーバからのパケット送信が一部のリンクに集中することである。そこで、本研究では、同時に通信するサーバ数を制限することで（すなわち、同時に設定するコネクションの本数を制限することで)、インキャストを回避しようとする。例えば、あるファイルの読み込みのために 50 個のサーバと通信する必要があるとき、50 本のコネクションを同時に設定するのではなく、5 本ずつ 10 回に分けてコネクションを設定することにする。以下、この方法を直列化法と呼ぶ。このように直列化した場合、一般のネットワークでは、50 本のコネクションの遅延の面での公平性の問題、すなわち、あるコネクションが別のコネクションより極端に終了時間が遅いといったことが問題となる。しかし、分散ファイルシステムの場合、50 本全部のコネクションの送信がいつ終了するのかが重要であって、個々のコネクションの終了時間は気にする必要はないので公平性の問題は起こらない。

３．研究の方法

本研究の開始以前において、筆者らは、既に、クライアントとすべてのサーバが１つのイーサネットスイッチに接続されており、かつ、すべてのリンクの伝送速度が互いに等しい環境を対象に、上述した直列化法を提案し、シミュレーション実験により、その有効性を確認している。しかし、このような環境を対象とした方法は、小規模な(サーバ数の少ない)分散ファイルシステムに対しては適用可能であるが、大規模な分散ファイルシステムの場合、通常、クライアントとサーバは互いに別々のイーサネットスイッチに接続され、かつ、リンクの伝送速度も異なる場合が多いので、これまでのコネクション直列化法をそのまま適用することはできない。そこで、本研究では、各リンクの伝送速度が与えられた時、インキャトを回避し、かつ、スループットを最大化できるように、従来の直列化法を拡張することを検討した。拡張した方法を拡張直列化法と呼ぶ。この拡張直列化法では、全リンクの伝送速度の最大値を直列化法における伝送速度とみなし、直列化法を適用した。

また、拡張直列化法を使用しても、インキャストを完全に回避することが難しくなる場合も想定して、別のインキャスト回避法についても検討した。従来、高精度のカーネルタイマーを使用

して、最小タイムアウト値を従来の 200 ミリ秒から数 100 マイクロ秒にすることにより、再送待ち時間を短縮して、インキャストを回避しようとする方法(以下、FGTCP と呼ぶ)が提案されていたが、この方法でも、完全にインキャストを回避することはできなかった。そこで本研究では、FGTCP をベースに、4 つの方法、すなわち LNRTCP、HYBTCP、AHTCP、NOBTCP を提案した。通常の TCP 再送待ち時間は指数的に増加するが、LNRTCP は再送待ち時間を線形的に増加させるものであり、HYBTCP は指数的な増加と線形的な増加を併用するものである。AHTCP と NOBTCP は、クライアントにおいて全コネクションのスループットを計測することを前提として、スループットが下がり始めたら、サーバに対してパケットの再送を要求するものである。AHTCP と NOBTCP は、この再送要求の方法が異なっており、AHTCP は特別な確認応答(ACK)パケットを使用し、NOBTCP は、新たな TCP オプションを使用する。

本研究では、上述した拡張直列化法、LNRTCP、HYBTCP、AHTCP、NOBTCP を、国立研究開発法人情報通信研究機構（NICT）が提供している、PC サーバを相互接続スイッチで接続したテストベッド設備「StarBED」に実装して、その性能を評価することも試みようとした。

４．研究成果

拡張直列化法については、従来の直列化法と同程度の性能が得られると予想できたので、LNRTCP、HYBTCP、AHTCP、NOBTCP について、ネットワークシミュレータ NS2 を使用して性能を比較した。比較の一例として、図 2 のネットワーク形態についての結果を図 3 に示す。LNRTCP と AHTCP が良好な Goodput（アプリケーションレベルのスループットに相当する）が得られている。

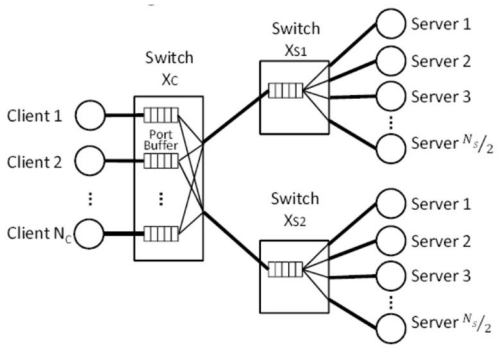

図 2　ネットワーク構成

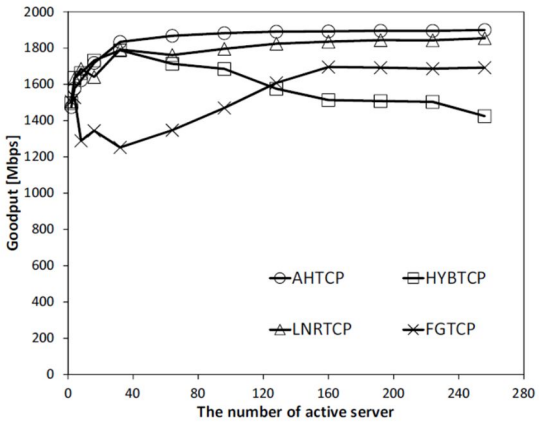

図 3　比較結果

また、StarBED における実験では、通常の TCP 通信を用いて、インキャストが生起するかどうかの実験を行ったが、インキャストは生起しなかった。その原因としては、StarBED は大規模イーサネットスイッチを使用しており、そのポートバッファも大容量であり、そのため、ポートバッファでのパケットロスが生起しないからであろうと予想した。このため、本研究で考案したインキャスト回避を性能を実験的に確認することはできなかった。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 濱田泰誠，樽谷優弥，福島行信，横平徳美
２．発表標題 TCPインキャスト回避法の性能比較
３．学会等名 電子情報通信学会2022年総合大会
４．発表年 2022年

１．発表者名 Shiden Kishimoto, Shigeyuki Osada, Yuya Tarutani, Yukinobu Fukushima and Tokumi Yokohira
２．発表標題 A TCP Incast Avoidance Method Based on Retransmission Requests from a Client
３．学会等名 International Conference on ICT Convergence 2019 (ICTC 2019)（国際学会）
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	福島　行信 (Fukushima Yukinobu) (00432625)	岡山大学・自然科学学域・准教授 (15301)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関