L-017

データセンタネットワークにおけるトラフィック優先度を考慮した動的ポーズ時間設定方式

A Study on Dynamic Pause Time Control Considering Traffic Priority in Data Center Networks

早坂 光雄† 大島 訓†
Mitsuo Hayasaka Satoshi Oshima

1. まえがき

データセンターにおいてパケットロスのない高品質なネットワークを構築するために、データリンク層におけるストップ・リスタート型のフロー制御 Priority-based Flow Control (PFC) が提案されている。この方式は受信側から停止時間を示した制御パケットを送信側に転送することにより、指定した停止時間だけ送信を停止させる方式であるが，各トラフィックの優先度毎に制御パケットが送信されるため，制御パケット数が過剰に多くなる課題があった。そこで，本稿では、各トラフィックの輻輳状態に応じて停止時間を自動計算することにより，制御パケット数を減少させパケットロスを回避する方式を提案し，その有効性を示す。

2. 従来方式

PFC は一つの物理ポートに論理的な複数の送受信キューを持たせて，論理キュー毎にポーズ処理を行うものである。各論理キューはプライオリティと呼ばれ，ある特定のフローで輻輳が発生した場合，受信側のポートは輻輳が発生したプライオリティに制御パケットである PAUSE フレームを送信する。このとき，PAUSE フレームが送信されたプライオリティ以外のトラフィックはそのまま送信可能である[1]。

図 1 に、各論理キューで稼働する従来方式の Pause On Off Control (POOC)のバッファ管理方法を示す[2]。ルータ／スイッチのバッファに高・低の 2 つの閾値を設定する。キュー長が高閾値を超えたら、ポーズ時間に最大値 65535 を指定したポーズフレームを送り、送信停止させバッファオーバフローを回避する。キュー長が低閾値まで減少したら、ポーズ時間に最低値 0 を指定したポーズフレームを送信側に送り、送信停止をキャンセルさせバッファアンダフローを回避する。

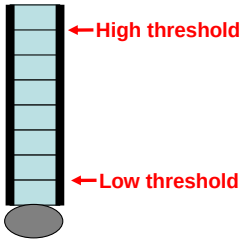

図 1　ポーズ on/off 制御方式

POOC では、キュー長が高閾値を超えた時および低閾値を下回った時の両方で，制御パケットであるポーズフレームを送出する。従って、制御パケットのオーバヘッドがネットワーク輻輳を増長させてしまう問題があり、RFC1889 でも制御パケット数を減少させる必要性が指摘されている[3]。

3. 提案方式

輻輳状況に従って、ポーズ時間を自動計算し、送信側の送信停止を行う。ポーズ停止時間が経過すれば送信が再開されるため、バッファオーバフローとアンダフローを避けるようにポーズ時間を決定する必要がある。

図 2 に提案方式である Counter-based Dynamic Pause Time Control (C-DPTC)のバッファ管理方法を示す。バッファに 1 つの閾値を設定し、キュー長が閾値を超えたらポーズフレームを送出する。ターゲットキュー長は、ポーズ機能の送信停止により閾値から減少させるキュー長であり、Pkt_{diff} はその差分である。これらの値は、閾値が決定すれば求まる固定値である。

ポーズ時間は，各論理キューで管理している送受信パケットカウンタに着目し，(1), (2)を用いて計算する。

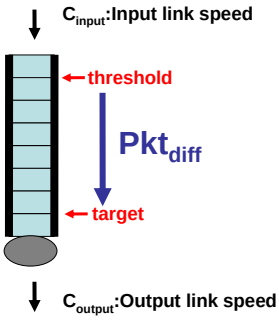

図 2　提案方式

$$P_{time} = R * \frac{4Pkt_{diff}}{(PktCnt_{cur} - PktCnt_{sent})} * Pkt_{diff} \quad (1)$$

$$T_{pause} = \frac{P_{time} * PktSize_{ave}}{512} = \frac{R * Pkt_{diff}^{2} * PktSize_{ave}}{64(PktCnt_{cur} - PktCnt_{sent})} \quad (2)$$

(1)はキュー長からターゲットキュー長まで減少させるために必要となるポーズ時間を示す。$PktCnt_{cur}$ はキュー長が閾値に達した時のパケットカウンタ値であり，$PktCnt_{sent}$ は最後にポーズフレームを送出した時のパケットカウンタ値である。パケットカウンタは単調増加する値であり、この差分が小さい場合には、大量のトラフィックが到着したと判断し大きいポーズ時間となるように計

†（株）日立製作所, 横浜研究所.
Hitachi, Ltd. Yokohama Research Laboratory.

算を行う。逆に差分が十分大きい場合には、小さいポーズ時間となるように計算する。R は，ネットワーク管理者がネットワークの構成などを考慮して設定する固定値である。(5)は、(4)で求めた値を実際のポーズフレームに設定するポーズ時間に変換する計算式である。

本方式において，変化するのはパケットカウンタ値のみであり，シンプルに計算が可能である。

4. 特性評価

従来方式 POOC と提案方式 C-DPTC の特性評価を行う。ここでは， 図 3 に示す入力ポート数N，出力ポート数Nのルータ/スイッチを想定する。表 1 に評価パラメータを示す。各ソースからは 3 種類のトラフィックが各論理キューに入力し，各キューからは Round-Robin によって，パケットが出力される。また，スイッチ側では各優先度に応じたキューが用意され，各キューで POOC または C-DPTC が稼働する。スイッチにおける論理キューの長さは，それぞれ 333 パケットとし，POOC における高閾値と低閾値はキュー長の 9/10 と 1/10 とした。C-DPTC のターゲットキュー長は POOC の低閾値と同じ設定とした。

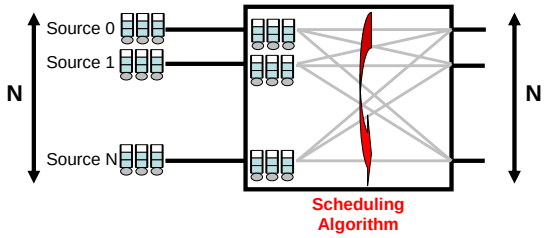

図 3　ネットワークトポロジー

表 1　評価パラメータ

項目	説明
スイッチタイプ	N×N スイッチ
バッファタイプ	FIFO
Port 毎の論理キュー数	3
論理キュー長	333 パケット
論理キュー間の Scheduling Algorithm	Round Robin
回線速度	1Gbps
パケットサイズ	1518 バイト
ポート間の Scheduling Algorithm	待ち時間の長いパケットを優先

特性評価より，パケットロスについては，従来方式および提案方式の両方で観察されなかった。ポーズ機能が有効に稼働したためである。図 4 と図 5 はネットワーク負荷 0.75、スイッチサイズ 32 の時の制御パケット数の変化とポーズフレームに設定されたポーズ時間の最大値・最小値を示したものである。それぞれポート毎に集計した値を用いており，制御パケット数はポート毎の平均を示している。図 4 より，提案方式 C-DPTC は，R の値が大きくなるにつれて制御パケット数を減少させ、POOC と比較して約 85%減少させている。図 5 より，R が大きくなると，C-DPTC は，最小値が最大のポーズ時間に近い値に設定されるようになる。つまり，トラフィックの状況に関係なく最大値に近い値が設定されることを示し，バッファアンダフローの可能性が大きくなることを示す。そのため、R はなるべく小さい値にすることが望まれる。

以上から，C-DPTC において，R に関するトレードオフの関係が存在する。両結果の評価から，R を 7 に設定することが、制御パケットのオーバヘッドを抑制し、安定的にバッファオーバフロー／アンダフローを回避させる設定値であるといえる。

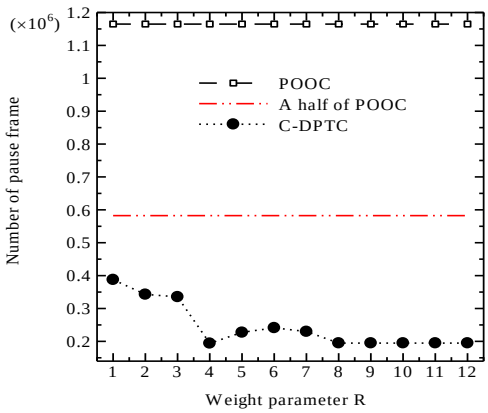

図 4　制御パケット数

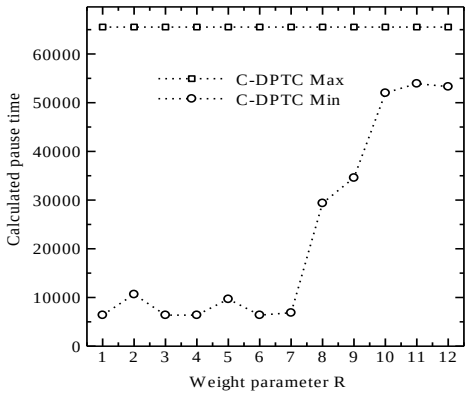

図 5　ポーズ設定時間の最大値・最小値

5. 結論

本稿では、データセンターにおける高品質なネットワーク構築のために，輻輳状況に応じてポーズ時間をシンプルに自動計算する方式を検討した。本方式は，特にトラフィック優先度を考慮した PFC が稼働している際に，パケットロスを回避しながら制御パケット数を大幅に減少させる方式であり，効率の良いフロー制御を実現する。

参考文献

[1] Annex 31 B, “Carrier sense multiple access with collision detection (CSMA/CD) access method and physical layer specification”, IEEE standard 802.3, 1998 Edition.

[2] Rich Seifert, “The Switch Book: The Complete Guide to LAN Switching Technology” Wiley, 2000.

[3] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, “RTP: A Transport Protocol for Real-Time Applications,” RFC 1889, January 1996.