2H-05

オンボードCPUによるTCP通信高速化方式の実装に関する一検討

長谷川 輝之　　長谷川 亨　　加藤 聰彦
(株) KDD 研究所

1. はじめに

近年、ギガビットイーサネット (GbE) の普及に伴い、端末もギガビットの帯域が利用可能になっている。しかし、GbE における IP の最大パケット長 (MTU サイズ) は標準で 1500 バイトであるため、現状の端末では、パケット単位で行われるプロトコル処理や端末本体と通信ボード間のデータ転送処理がオーバヘッドとなる。

これに対して筆者等は、通信ボード上の CPU を利用して、TCP まで考慮したパケット分割・組立を行うことで端末に大きな MTU サイズを提供し、パケット単位の処理オーバヘッドを削減する方式を検討している[1][2]。本稿では、市販の GbE ボードを対象に、本方式の実装方法について検討した結果を示す。

2. 動作手順

本方式の動作手順は以下の通りである (図 1参照)。

(1) GbE ボードから IP モジュールへ大きな MTU サイズ (8800 バイトとする) を通知する。
(2) TCP では、通信開始時に MTU サイズに基づき最大セグメントサイズ (MSS) を交換する。GbE ボードは MSS を更新して、自身または相手端末の TCP モジュールに適切な MSS を通知する (図 1(a))。
(3) TCP モジュールは大きな MSS でプロトコル処理を行う。GbE ボードは、TCP モジュールから渡されたパケットを相手から通知された MSS に従って分割し、各々に適切な TCP/IP ヘッダを設定して送信する (図 1(b))。
(4) GbE ボードは、可能な限り大きな受信パケットに組み立ててから TCP モジュールに通知する (図 1(c))。
(5) 本方式では、TCP モジュールにおける 1 パケットがネットワーク上の複数パケットに対応する。従って本機能の無い既存端末との通信では、TCP モジュールが受信する ACK の数は多く、逆に、送信する ACK の数は少なくなる。TCP の輻輳制御や再送制御への影響を抑えるため、GbE ボードで ACK の数と内容を調整する (図 1(d))。

3. 実装方法

3.1 GbE ボード

本方式を、オンボード CPU を搭載した市販 GbE ボード (Alteon WebSystems ACEnic) 上に、そのファームウェアとデバイスドライバを改修して実現することとした。本ボードにおけるパケット送受信フローを図 2に示す。送受信フローは、ハードウェアによる処理ならびに CPU による処理から構成される。CPU での処理に必要なファームウェアはホストからダウンロードされる。

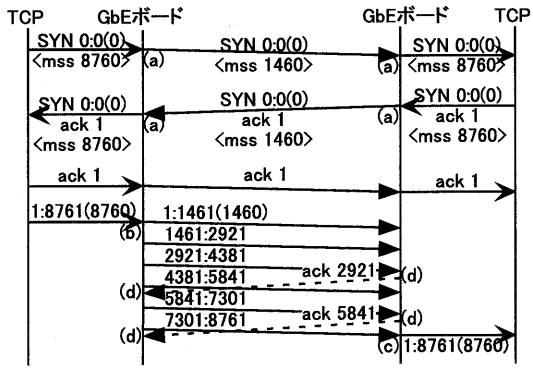

図 1: 通信シーケンス例

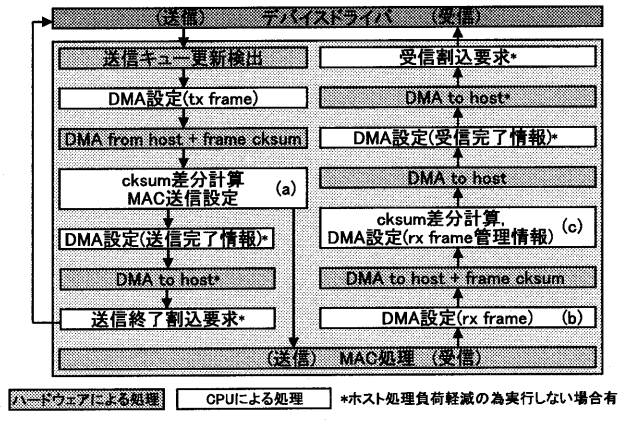

図 2: GbE ボードにおけるパケット送受信フロー

送受信処理の実装方法を検討するにあたり、本ボードの持つ以下の特徴を考慮した。

(1) CPU から MAC 処理部に渡す送信パケットは連続したオンボードメモリ領域に置かなければならない。
(2) TCP チェックサムの計算は、DMA 転送の際にハードウェアで送受信フレーム全体のチェックサムを計算し、その結果に対して、CPU で不要部分を差分計算することで実現している。従って、送信/受信パケットのチェックサム付加/検証は、DMA 転送完了後に行う必要がある。
(3) 図 2において CPU が行う各処理は、非同期かつ排他的に起動・実行される。

"A Study on Implementation of TCP Throughput Accelaration Mechanisms using On-board CPU"
Teruyuki HASEGAWA, Toru HASEGAWA and Toshihiko KATO
KDD R & D Laboratories, Inc.

3.2 送信処理

3.1(1),(2) を考慮し、送信時は、分割する各パケット毎に DMA 転送を行うことで、各パケットのヘッダ領域確保とチェックサム生成を実現した。また、ヘッダ設定等の処理を、チェックサムが確定する図 2(a) 内に導入した。送信時の処理の流れを以下に示す。

(1) デバイスドライバでは、各パケットの DMA 転送要求を設定し送信キューに登録する。この際、図 3(a) に示すように、先頭パケット以外は、パケットに連続したヘッダ領域確保のため、転送開始アドレスをヘッダ長だけ先頭方向に移動させる。

(2) 送信キューに登録された内容に従って各パケットが DMA 転送され、同時にフレームチェックサムも計算される。

(3) DMA 転送完了後、図 2(a) が起動される。CPU は、各パケットのヘッダ領域を、フレームチェックサムと先頭パケットのヘッダ情報を基に更新した上で MAC 処理部に渡す (図 3(b))。また、送信シーケンス番号等のプロトコル情報も更新する。

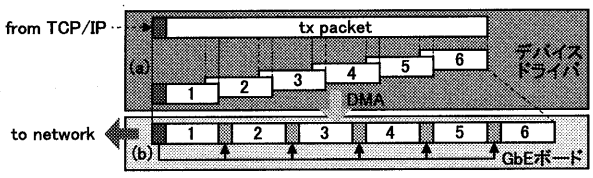

図 3: 分割・送信処理

3.3 受信処理

受信側では、受信パケットの組立処理と、ACK 調整 (2.(5) 参照) 等の TCP を考慮したプロトコル処理を行う。これらの処理は、TCP チェックサムを検証した上で開始すべきであるが、そのためには 3.1(2) より受信パケットのホストへの DMA 転送が必要となる。そこでチェックサム検証前に、図 2(b) 内で DMA 転送を利用してホストメモリ上で組立処理を実現する。また、図 2(c) 内でチェックサム検証を含むプロトコル処理を実現する。受信時の処理の流れを以下に示す。

(1) デバイスドライバは、IP に通知した MTU サイズに応じた大きさを持つ受信バッファを複数用意する。

(2) パケット受信により MAC 処理部は図 2(b) を起動する。CPU は受信パケットの属するコネクションを識別する。図 4(a) に示すように、同一コネクションに属する連続したパケットに対しては、(先頭パケットを除き) ヘッダを削除した上で、同じ受信バッファ内で連結されるように DMA 転送要求を設定する。ここで設定情報内に、受信パケットの検証に必要な削除部分のチェックサムや、プロトコル処理に必要なシーケンス番号、ACK 番号、ウインドウサイズを記録しておく。

(3) DMA 転送完了後、図 2(c) が起動される。CPU ではチェックサムを検証し、正しい場合は、連結部分の長さとチェックサムをオンボードメモリ内に設けた組立用ヘッダに反映し、TCP を考慮したプロトコル処理を開始する (図 4(b))。

(4) デバイスドライバへの受信パケットの通知は、一定時間経過後、もしくは、受信バッファ内でパケット連結が不可能となる直前で行う。ここで同時に、組立用ヘッダからヘッダ更新のための情報も通知し、ドライバにヘッダを更新させる (図 4(c))。

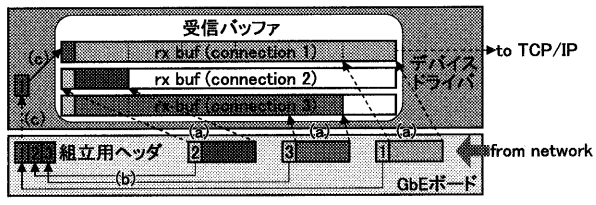

図 4: 受信・組立処理

一方、上記 (3) において誤りを検出した場合の処理手順を図 5を例に説明する。

(a) 組立処理により、連続したシーケンス番号を持つパケットは連結される (図 5(a))。パケット 201:301 の DMA 転送要求後、プロトコル処理が起動される。

(b) プロトコル処理では、パケット 101:201 の誤りを検出するとこれを破棄し、組立処理側に連結中止を要求する (図 5(b))。また 3.1(3) の理由から、パケット 201:301 のように誤りパケットへ連結される場合もある。これも同様に破棄し、次の連結パケットがないことを確認してから、未通知の連続受信パケット 1:101 を通知する (図 5(c))。

(c) 組立処理では以降のパケットを連結しない (図 5(d))。各パケットはプロトコル処理を経由して単独でドライバへ通知される。

(d) プロトコル処理では、誤り回復を確認すると、組み立て処理側に連結再開を要求する (図 5(e))。

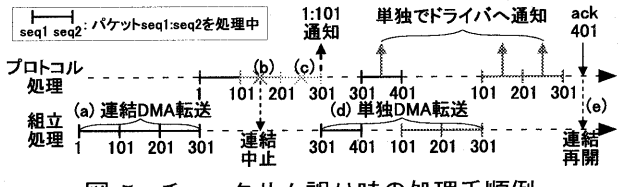

図 5: チェックサム誤り時の処理手順例

4. おわりに

本稿では、オンボード CPU を用いた TCP 通信高速化手法について、市販 GbE ボードを対象に、その特徴を考慮した実装方法を検討した結果を示した。最後に日頃御指導頂く KDD 研究所 秋葉所長に感謝します。

参考文献

[1] 長谷川, 長谷川, 加藤, “オンボード CPU によるパケット分割・組立機能を用いた TCP 通信高速化の検討,” 信学総合大会, D-6-17, Mar. 2000.

[2] 長谷川, 長谷川, 加藤, “オンボード CPU を用いた TCP 通信高速化方式の検討,” DICOMO2000 シンポジウム論文集, pp. 505-510, June 2000.