ギガビットレートに対応したパケットヘッダ収集方式

6L−03

長谷川 輝之　　大岸 智彦　　長谷川 亨

(株) KDDI 研究所

1. はじめに

近年、ギガビットイーサネット (GbE) の普及により、その運用・監視を目的とした、低コストかつ拡張性の高いモニタシステムの実現が求められている。このためには、PC 等の市販ハードウェア上で動作するソフトウェアベースのモニタシステムをギガビットレートに対応させる必要がある。しかし、広く利用されている TCPdump[1] 等の libpcap[2] ベースのソフトウェアでは、高性能なハードウェアを用いた場合も、GbE ラインレートでパケットヘッダを収集すること自体が困難である。これに対して筆者らは、市販 NIC の DMA 機能を利用したヘッダ抽出方式とネットワークを介したヘッダ情報転送方式を特長とする、ギガビットレートに対応したパケットヘッダ収集方式を提案している [3]。本稿では、提案方式の概要とこれに基づき実装したヘッダ収集システムの性能評価結果について述べる。

2. ヘッダ収集方式

図 1に提案するヘッダ収集方式の全体構成を示す。ヘッダ収集は、DMA 機能を有する 64bit/66MHz PCI 対応 GbE NIC と、そのデバイスドライバに以下の改修を加えた専用ドライバを組み合わせて実現する。

(1) パケット全体を受信する代わりに、各受信パケットのヘッダ部分を連続するメモリ領域上に抽出する。

(2) 抽出したヘッダ情報を、UDP パケットとして自身またはネットワーク経由で他ホストへ転送する。

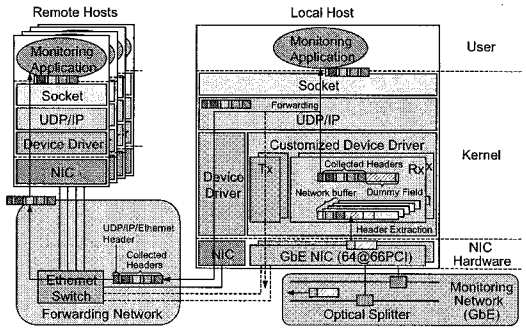

図 1: 全体構成

"A Method of Packet Header Collection with Gigabit-Rate Capability"
Teruyuki HASEGAWA, Tomohiko OGISHI, Toru HASEGAWA
KDDI R & D Laboratories Inc.

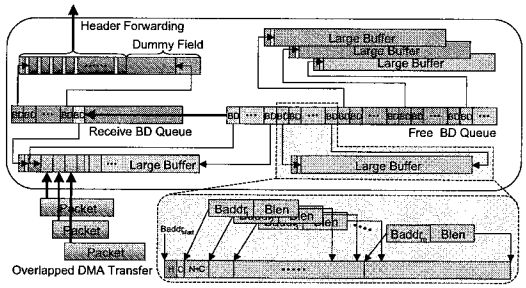

図 2: ヘッダ抽出時の処理の流れ

図 2にヘッダ抽出時の処理の流れを示す。処理内容は以下の通りである。

(1) 通常、NIC とそのドライバでは、受信パケット毎に管理情報 (BD: Buffer Descriptor) を用意し、BD 毎に DMA 転送先のバッファを割り当てる。これに対して本方式では、複数の BD を大きな共通バッファに対応させる。

(2) NIC は、BD に登録された転送開始アドレスに従って DMA 転送を開始する。そこで、各 BD の転送開始アドレスをヘッダ (図中 N) とその制御情報 (図中 C) 分だけずらして設定する。これにより、ヘッダ以外の部分が重複して DMA 転送されるため、共通バッファ上にはヘッダのみが自動的に展開される。

(3) ドライバでは、各ヘッダの制御情報、ならびに共通バッファの先頭に確保した UDP/IP/Ethernet ヘッダ領域 (図中 H) を設定した上で、IP 経由で自身または他ホストに転送する。

3. 性能評価

提案方式の有効性を確認するため、Linux 2.4.x 上で動作するヘッダ収集システムを実装し、その性能評価を行った。本システムの構成を以下に示す。

- **ヘッダ収集ドライバ**: 市販 NIC (Syskonnect SK-98xx) 用ドライバを基に実装され、提案方式に従ったヘッダ抽出・転送を行う。
- **ファイル記録アプリケーション**: ヘッダ収集ドライバから転送されるヘッダ情報を UDP ソケットから受信し、ファイルシステムに記録する。

性能評価に使用したハードウェアの仕様を表 1に、試験環境を図 3に示す。試験では、表 2に示す 6 種類のシステム構成について、トラヒックテスタ (ANTARAnet FLAME THROWER) から GbE ラインレートで固定

表 2: システム構成

	構成 1	構成 2	構成 3	構成 4	構成 5	構成 6
OS	Linux 2.4.2				FreeBSD 4.2	
デバイスドライバ	ヘッダ収集ドライバ		標準ドライバ			
ヘッダ抽出用 PC	PC1					
アプリケーション (AP)	ファイル記録 AP		TCPdump with libpcap			
ファイルシステム (FS)	RAID			/dev/null	RAID	/dev/null
AP と FS が動作する PC	PC2	PC1				

表 1: ハードウェア仕様

	PC1	PC2
Vendor	Supermicro	Dell
Chipset	ServerWorks HE-SL	
CPU	Intel Pentium III 1GHz×2	Intel Pentium III 866MHz×1
Memory	512MB	128MB
RAID system	AMI Series 493 + IBM DDYS-T36950×4	ArenaII + IBM DTLA-307075×4
NIC (抽出)	SysKonnect SK-9843	−
NIC (転送)	Intel Pro/1000F	

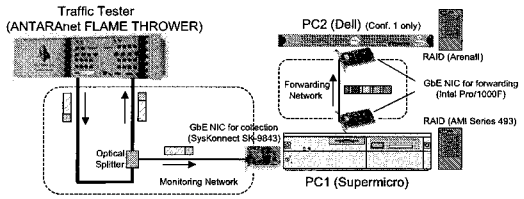

図 3: 試験環境

長の Ethernet フレームを 10,000,000 個送信し、40 バイトの TCP/IP ヘッダの収集とファイルへの記録を行った。なお、他ホストへのヘッダ情報転送の有無による違いを評価するため、提案方式として 2 種類の構成 (構成 1、2) を試験した。一方、既存方式については、Linux (構成 3、4) に加えて FreeBSD (構成 5、6) を用いた構成についても試験を行った。この内構成 4、6 では、PC1 におけるディスクアクセスが無い構成 1 との比較を行うため、仮想的なファイル (null デバイス) を使用しディスクアクセスを避けることとした。

図 4に各構成におけるヘッダ収集率を、図 5に PC1、PC2 における CPU 使用率を示す。これらの試験結果から以下の知見が得られた。

(1) 図 4より、全てのフレームサイズにおいて、提案方式が既存方式を上回るヘッダ収集率を達成している。提案方式の内、構成 2 では、384 バイト以上のフレームサイズにおいて GbE ラインレートでのヘッダ収集が可能である。

(2) 図 5より、全てのフレームサイズにおいて、提案方式は既存方式よりも格段に低い CPU 使用率でヘッダ収集が可能である。これは、提案方式が、収集ヘッダの解析処理に対してより多くの CPU リソースを割り当て可能であることを意味する。

(3) 図 5において、構成 1 における PC2 と構成 2 における PC1 を比較した所、PC1 に比べハードウェアの性能が劣っている PC2 の方が、より低い CPU

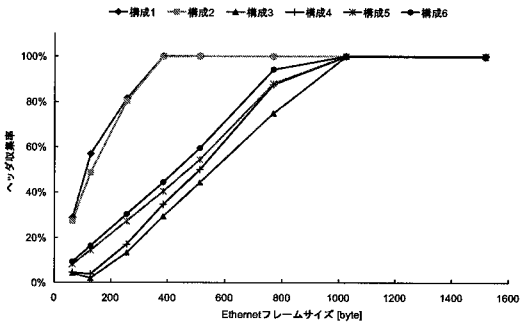

図 4: ヘッダ収集率

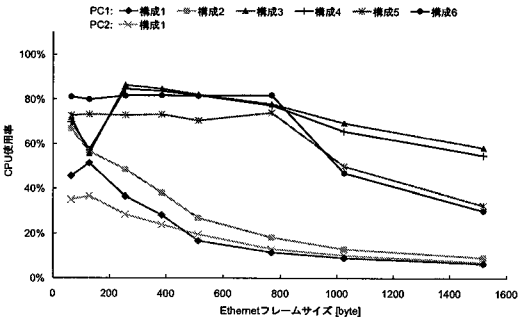

図 5: CPU 使用率

使用率で動作していることが判明した。これは、他ホストにヘッダ情報転送を行うことで、ヘッダ収集のオーバヘッドを解析処理から切り離し可能であることを表している。

4. おわりに

本稿では、ギガビットレートに対応したパケットヘッダ収集方式について述べると共に、提案方式に基づき実装したヘッダ収集システムの性能評価を通じて、その有効性を明らかにした。最後に日頃御指導頂く KDDI 研究所浅見所長に感謝します。

参考文献

[1] V. Jacobson, C. Leres and S. McCanne. TCPdump 3.4, Lawrence Berkeley National Laboratory, Berkeley, CA, June 1998. http://www-nrg.ee.lbl.gov/.

[2] S. McCanne, C. Leres, and V. Jacobson. LIBPCAP, Network Research Group, Lawrence Berkeley National Laboratory. ftp://ftp.ee.lbl.gov/libpcap.tar.Z.

[3] T. Hasegawa, T. Ogishi, T. Hasegawa, "A Framework for Gigabit-Rate Packet Header Collection to Realize Cost-Effective Internet Monitoring System," 情処論, Vol.42 No.12, Dec. 2001.