

MAPOS のソフトウェアドライバの構成

4 G-4

吉田敏明

ベルクマイクロシステムズ開発部

1 はじめに

MAPOS¹⁾ は専用線の物理層プロトコルである SONET / SDH の上に HDLC のフレームを載せる低オーバーヘッドのリンクレイヤプロトコルである。また MAPOS はコネクションレスであり、ブロードキャスト/マルチキャストが可能である。本稿では、PCI Bus の MAPOS インタフェース・カード²⁾ を FreeBSD³⁾ で利用するために独自に開発したデバイスドライバの構成とその性能評価を示す。

2 デバイスドライバの構成

デバイスドライバでは NSP(Node Switch Protocol)⁴⁾、ARP(Address Resolution Protocol)⁵⁾ の処理を行う。

インタフェースの MAPOS 物理アドレスは NSP によって割り当てられる。回線の切断・接続の度に MAPOS 物理アドレスの割り当てが実行され、アドレスが割り当てられるまでは、他のデバイスドライバの処理は行われない。

IP アドレスから MAPOS 物理アドレスへのアドレス解決のために ARP が利用される。Ethernet の ARP^{6), 7)} とほとんど同じ処理を行うが、古い ARP エントリを明示的に消去させるために UNARP という機能が付加される。

IP は通常、上位層のプロトコル処理⁷⁾ と共に使われる。

送信の場合、ソケット層ではユーザ空間からカーネル空間の mbuf 構造体にデータをコピーする。mbuf のクラスタのサイズは 2K バイトで、データサイズが大きければ、TCP 出力ルーチンはこの単位で呼び出される。UDP 出力ルーチンはユーザが指定したサイズをコピーした後に呼び出されるが、ソケットバッファの大きさを

越える事はできない。ソケットバッファの最大値は通常 256K バイトである。データのコピーは通信速度を支配する要因の一つと考えられる。

UDP 出力ルーチンではヘッダを付加し、チェックサムを計算して IP 出力ルーチンを呼び出す。TCP でもチェックサムの計算が行なわれるが、チェックサムの計算は速度を支配するもう一つの要因と考えられる。

IP 出力ルーチンは IP ヘッダを形成してデバイスドライバを呼び出すが、データサイズが大きい場合には分割を行なう。

受信プロトコル処理はデバイスドライバの割り込み処理からソフトウェア割り込みという形で呼び出される。

2.1 デバイス制御

ハードウェアは、CPU メモリの任意のアドレス上のデータ(送信パケット)を送信、または受信パケットを CPU メモリ上の領域に DMA(Direct Memory Access) 転送することができる。

DMA コントローラはアドレス、長さを記述した送信バッファ・ディスクリプタを見て送信 DMA 転送を行うが、一つのパケットが複数のディスクリプタに分割されていても構わない。

デバイスドライバは上位プロトコルからデータを mbuf 構造体によって渡される。mbuf 構造体のチェーンで表現されるデータは、通常連続領域には存在していない。連続領域にコピーして一つの送信バッファ・ディスクリプタで記述することも考えられるが、データサイズが大きい場合にはコピーに時間がかかることが予想される。実際にはデバイスドライバは mbuf 構造体のチェーンを送信バッファ・ディスクリプタのチェーンとして記述する。ただし、小さいデータが複数の送信バッファ・ディスクリプタに分割された場合、DMA コントローラのオーバーヘッドは大きくなる。

デバイスドライバは受信バッファとして最大パケット長(MTU (Maximum Transmission Unit) サイズ)の連続領域を予め準備して、受信バッファ・ディスクリプタに記述しておく。

MAPOS Device Driver

Toshiaki YOSHIDA

Werk Mikro Systems Ltd.

250-1, Mikajiri, Kumagaya-shi, Saitama-ken, 360 Japan.

3 性能評価

デバイスドライバの実行にかかる時間を測定した。評価には Pentium II 300MHz の PC に FreeBSD-2.2.5 を載せて使用した。MAPOS インタフェース・カードは OC-3c(156Mbps) のものである。カーネル、ドライバ中に CPU のタイム・スタンプ・カウンタの値を読み出してメモリ中に記録するコードを埋め込み、ドライバに特別な ioctl を追加して、その記録を読み出せるようにした。一か所に対して 256 回の値を記録して、読み出した後平均を取っている。転送は間隔を空けて行ない、DMA 転送による影響などはないようにしている。

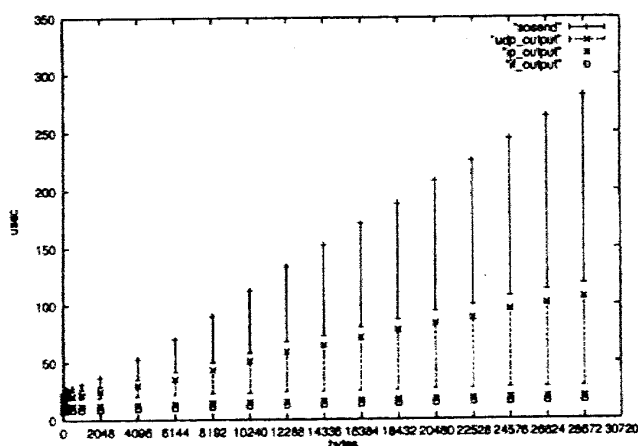


図 1: UDP 送信

UDP の送信の実行時間を図 1 に示す。ここで sosend、udp_output、ip_output、if_output は、それぞれソケット層、UDP、IP、デバイスドライバの出力ルーチンを示す。sosend の実線部分はデータのコピーにかかる時間を示している。また udp_output の点線部分はチェックサムの計算時間を示している。データサイズに比例してこれらの時間が大きくなる。データサイズが 28K バイトの時、sosend に対する if_output の割合は僅か 5.7% である。

この値から sosend の転送性能を計算すると、28K バイトの時の処理時間は 281μsec で 816Mbps、6K バイトの時は 70.3μsec で 699Mbps となり、OC-12c の回線速度に十分対応可能である。

4 おわりに

MAPOS プロトコルはシンプルであり、DMA コントローラの機能を活用することによって、高速なデバイスドライバを実現することができた。

今回、測定は OC-3c のインタフェース・カードで行なったが、処理性能的には OC-12c のカードを使用しても問題ないように見える。ただし、トラフィックが高い

状態では CPU と DMA コントローラで CPU メモリの競合が発生して性能が低下すること考えられ、今後検討していく予定である。

謝辞

いつもお世話になっております NTT 光ネットワークシステム研究所 分散ネットワークシステム研究部 高橋直久グループリーダー、村上健一郎グループリーダー、丸山 充氏ならびに並列分散アーキテクチャ研究グループの皆様、中央システム技研の小林正之氏に感謝致します。

参考文献

- 1) K. Murakami, M. Maruyama, "MAPOS - Multiple Access Protocol over SONET/SDH Version 1", IETF RFC-2171, June 1997.
- 2) 丸山, 川野, 八木, 村上, "Frame Switching 方式による通信インタフェースの実現と評価 -SONET-LAN-", 情報第 54 回大会, 6N-2, 1997.
- 3) FreeBSD source code, Release 2.2.5, <http://www.freebsd.org>.
- 4) K. Murakami, M. Maruyama, "A MAPOS version 1 Extension - Node Switch Protocol", IETF RFC-2173, June 1997.
- 5) K. Murakami, M. Maruyama, "IPv4 over MAPOS Version 1", IETF RFC-2176, June 1997.
- 6) W. Richard Stevens: "TCP /IP Illustrated, Volume1: The Protocols", ISBN 0-201-63346-9, Addison-Wesley, 1994.
- 7) Gary R. Wright and W. Richard Stevens: "TCP /IP Illustrated, Volume2: The Implementation", ISBN 0-201-63354-X, Addison-Wesley, 1995.