パケット送受信の NIC の動作について

2013/10/15

増田陽介

1 はじめに

増田の研究テーマである「MSI のルーティング変更による NIC の移譲」を実現するために，必要な改修箇所箇所について調査している．＜資料 236-03 ＞でパケット送受信時における，NIC ドライバの処理流れについて調査した結果を述べた．この際パケット送受信時における NIC の動作について，十分に理解できていない箇所があった．そこで，パケット送受信時における NIC の動作について調査した．その結果，NIC の移譲時に問題になると考えられる処理が存在した．本資料では，パケット送受信時における NIC の動作について述べる．また，NIC 移譲時に発生すると考えられる問題について述べる．

2 調査環境

表 1 に調査環境を示す．

表 1　調査環境

項目名	環境
OS	Fedora 14
カーネル	Linux カーネル 3.0.8 64bit
CPU	Intel(R) Core(TM) Core i7-870 @ 2.93GHz
メモリ	2GB
NIC	RTL8168d/8111d
NIC デバイスドライバ	RTL8169

3 データ構造

パケット送受信時に，NIC および NIC ドライバが使用するデータ構造について説明する．

(1) 送信管理表

パケットバッファへの物理アドレスなど，パケットを送信するために必要な情報を持つ．NIC は送信管理票からバッファの情報を取得し，送信を行う．全部で 64 個存在し，NIC ドライバの初期化時に配列として確保される．この配列を送信管理表配列と呼ぶ．送信管理表は 256 バイト境界でアラインされている．送信するパケットは分割されている場合がある．送信管理表は，分

割したパケットの先頭，末尾であることを示す情報を持つ．送信管理表配列はリングバッファとなっており，リングバッファの終端となる送信管理表は，終端であることを示す情報を持つ．

(2) 受信管理表
パケットバッファへの物理アドレスなど，パケットを受診するために必要な情報を持つ．NIC は受信管理表からバッファの情報を取得し，受信を行う．全部で 256 個存在し，NIC ドライバの初期化時に配列として確保される．この配列を受信管理表配列と呼ぶ．受信管理表は 256 バイト境界でアラインされている．送信するパケットは分割されている場合がある．送信管理表は，分割したパケットの先頭，末尾であることを示す情報を持つ．送信管理表配列はリングバッファとなっており，リングバッファの終端となる送信管理表は，終端であることを示す情報を持つ．

4 レジスタ

パケット送受信時に使用される NIC のレジスタについて説明する．

(1) Transmit Priority Polling Register(TPPoll)
NIC の送信状態を表す．このレジスタが 1 の場合，NIC は送信処理を行う．送信処理が終了すると，NIC はこのレジスタを 0 にする．MMIO 領域にマッピングされている．

(2) Transmit Normal Priority Descripters(TNPDS)
送信管理表配列の先頭アドレスを持つ．MMIO 領域にマッピングされている．

(3) Receive Descriptors Start Address Register(RDSAR)
受信管理表配列の先頭アドレスを持つ．MMIO 領域にマッピングされている．

(4) 送信管理表アドレスレジスタ
パケット送信処理時に, NIC が取得する送受信管理表のアドレスを持つ．NIC 起動時に, TNPDS の内容で初期化される．MMIO 領域にマッピングされていない．

(5) 受信管理表アドレスレジスタ
パケット受信処理時に，NIC が取得する受信管理表のアドレスを持つ．NIC 起動時に，RDSAR の内容で初期化される．MMIO 領域にマッピングされていない．

5 NIC の動作について

5.1 パケット送信時における NIC の動作

NIC ドライバが MMIO を用いて，TPPoll を 1 にすることで，NIC は送信処理を開始する．パケット送信時における NIC の動作について図 2 に示し，以下で説明する．

(1) 送信管理表を取得
ポインタレジスタに格納されているアドレスに DMA を用いてアクセスし，送信管理表を取得する．

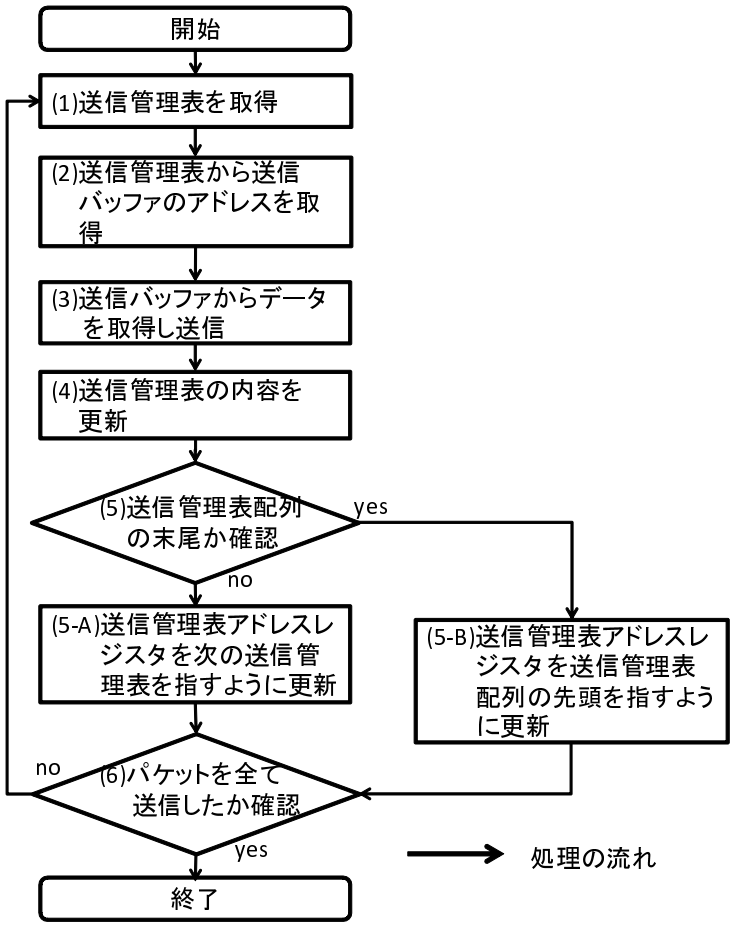

図 1　パケット送信時における NIC の動作

(2) 送信管理表から送信バッファのアドレスを取得

送信管理表を参照し，送信データが格納されているバッファのアドレスを取得する．送信バッファは DMA 領域にマッピングされており，送信データが格納されている．

(3) 送信バッファから送信データを取得し送信

DMA を用いて送信バッファから送信データを取得し，送信する．

(4) 送信管理表の内容を更新

送信に使用した送信管理表の内容を更新する．

(5) 送信管理表配列の末尾か確認

取得した送信管理表が，送信管理表配列の末尾にあるものか否かを確認する．

（A）送信管理表アドレスレジスタを次の送信管理表を指すように更新

送信管理表配列は，メモリ上の連続した領域に確保されている．送信管理表は 256 バイト境界でアラインされている．よって，送信管理表アドレスレジスタに 256 バイト加算することで，送信管理表アドレスレジスタは次の送信管理表を指すこととなる．送信管理表アドレスレジスタを更新する様子を図 1 に示す．

（B）送信管理表アドレスレジスタを送信管理表配列の先頭アドレスを指すように更新

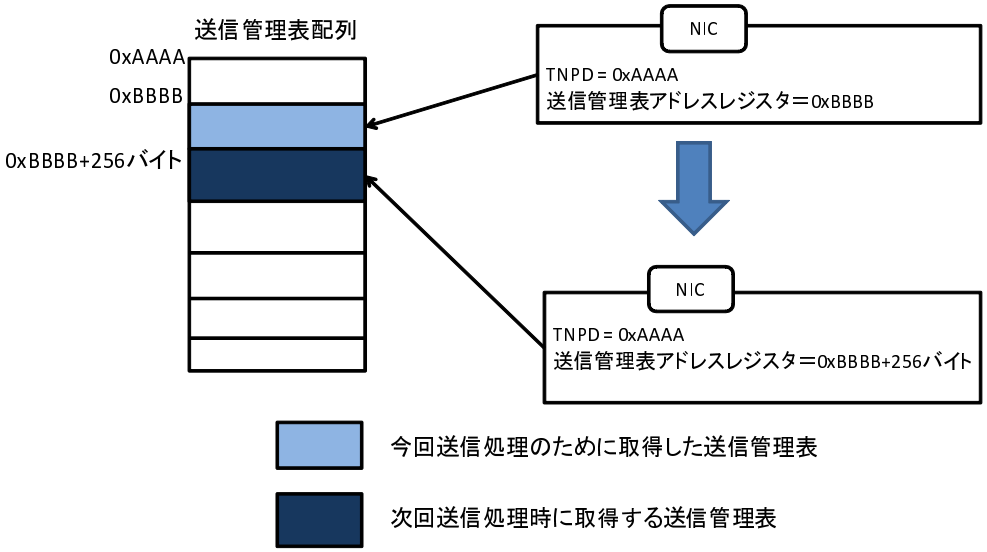

図 2　送信管理表配列の末尾でない場合における送信管理表アドレスレジスタの更新

送信管理表はリングバッファである．このため，取得した送信管理表が送信管理表配列の終端であった場合，次回は送信管理表配列の先頭にある送信管理表を取得する．送信管理表の先頭アドレスを格納している TNPD の内容を，送信管理表アドレスレジスタに書き込む．送信管理表アドレスレジスタに TNPD の内容を書き込む様子を図 3 に示す．

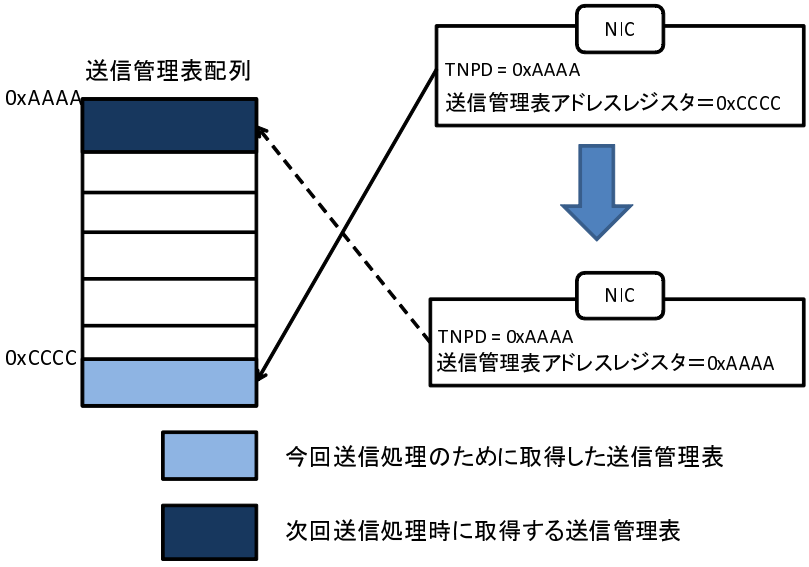

図 3　送信管理表配列の末尾である場合における送信管理表アドレスレジスタの更新

(6) パケットを全て送信したか確認

パケットサイズが大きい場合，分割してバッファに配置される．送信管理表には送信データが分割したパケットの最後尾であるか否かという情報が格納されている．送信したデータが最後尾だった場合，TPPoll を 0 にし，送信処理を終了する．この際，割り込みを発生させる．

5.2 パケット受信時における NIC の動作

パケット受信時における NIC の動作について図 4 に示し，以下で説明する．

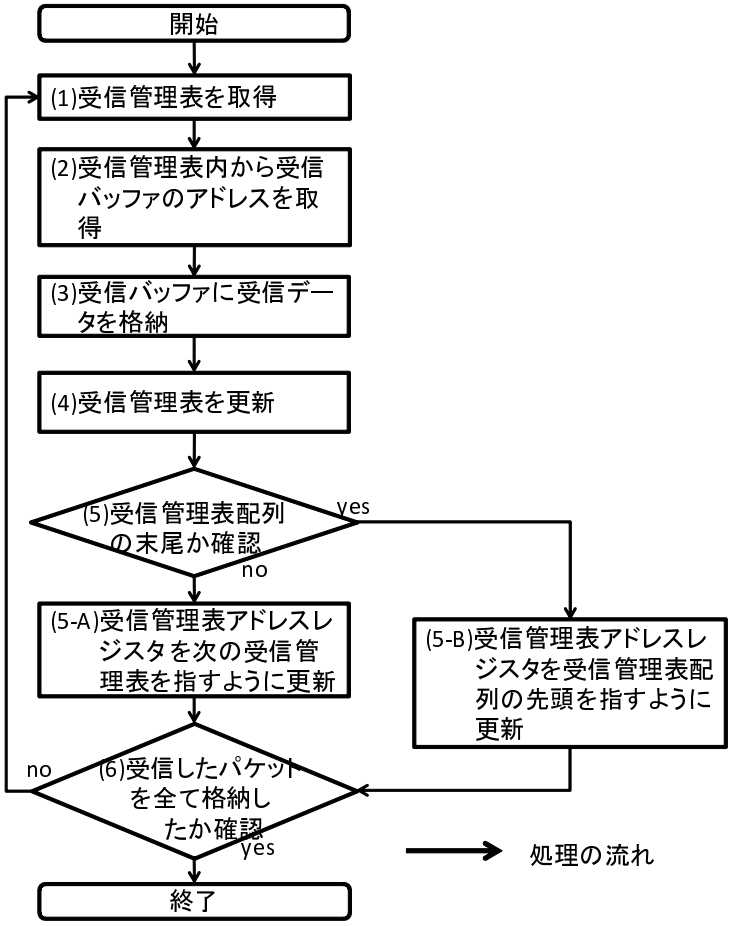

図 4　パケット受信時における NIC の動作

(1) 受信管理表を取得
受信管理表アドレスレジスタにに格納されているアドレスに DMA を用いてアクセスし，受信管理表を取得する．

(2) 受信管理表から受信バッファのアドレスを取得
受信管理表を参照し，受信データを格納する受信バッファのアドレスを取得する．

(3) 受信バッファに受信データを格納
DMA を用いて，受信バッファに受信データを格納する．

(4) 受信管理表の内容を更新
受信管理表の内容を更新する．

(5) 受信管理表配列の末尾か確認
取得した受信管理表が，受信管理表配列の末尾のものか否かを確認する．

（A）受信管理表アドレスレジスタを次の受信管理表を指すよう更新

受信管理表配列は，メモリ上の連続したメモリ領域に確保されている．受信管理表は 256 バイト境界でアラインされている．よって，受信管理表アドレスレジスタに 256 バイトを加算することで，受信管理表アドレスレジスタは次の受信管理表を指すようになる．受信管理表アドレスレジスタを更新する様子を図 5 に示す．

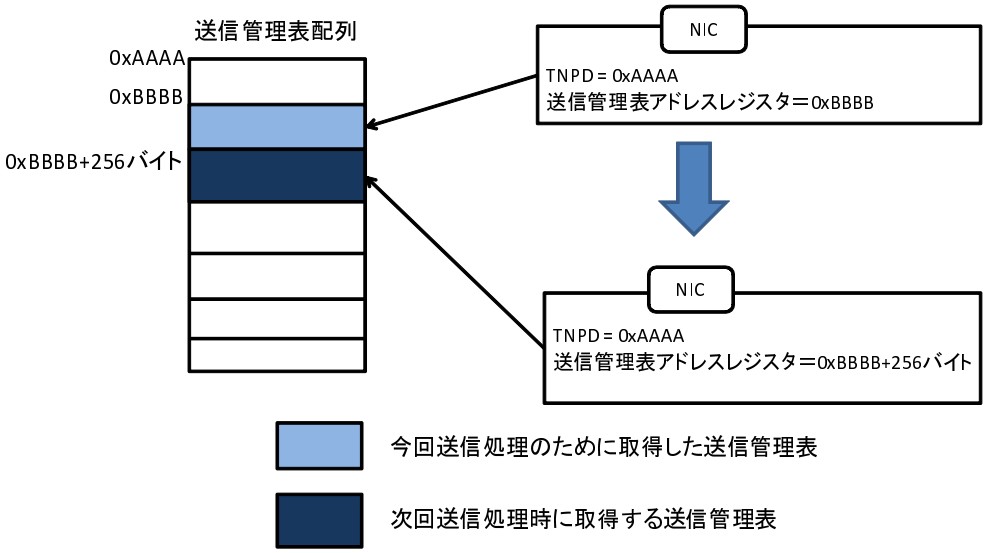

図 5　受信管理表配列の末尾でない場合における受信管理表アドレスレジスタの更新

（B）受信管理表アドレスレジスタを受信管理表配列の先頭を指すよう更新

受信管理表はリングバッファである．このため，取得した受信管理表がリングバッファの終端であった場合，次回は受信管理表配列の先頭にある受信管理表を取得する．受信管理表配列の先頭アドレスを格納している RDSAR の内容を，受信管理表アドレスレジスタに書き込む．受信管理表アドレスレジスタに RDSAR の内容を書き込む様子を図 6 に示す．

(6) 受信したパケットを全て受信バッファに格納したか確認

受信したパケットを全て受信バッファに格納した場合，NIC は受信処理を完了し，割り込みを発生させる．

6　NIC 移譲時に発生すると考えられる問題

4 章で，パケット送受信時の NIC の動作について述べた．パケット送受信時の NIC の動作で，MSI のルーティング変更による NIC 移譲の際に発生する問題の要因を以下に述べる．

(問題の要因)　NIC 移譲時に，送信管理表アドレスレジスタ，受信管理表アドレスレジスタの内容を書き換えることができない．

この要因により，発生する問題について図 7 で示し以下で説明する．

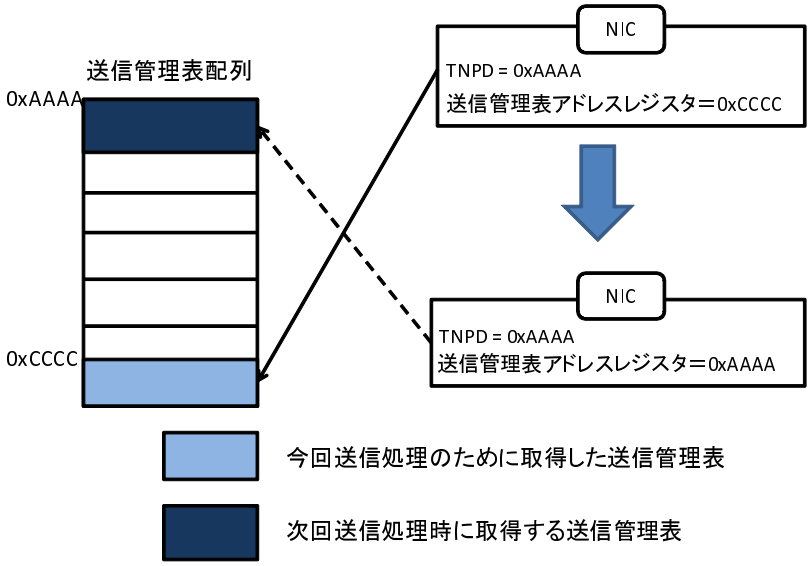

図 6　受信管理表配列の末尾である場合における受信管理表アドレスレジスタの更新

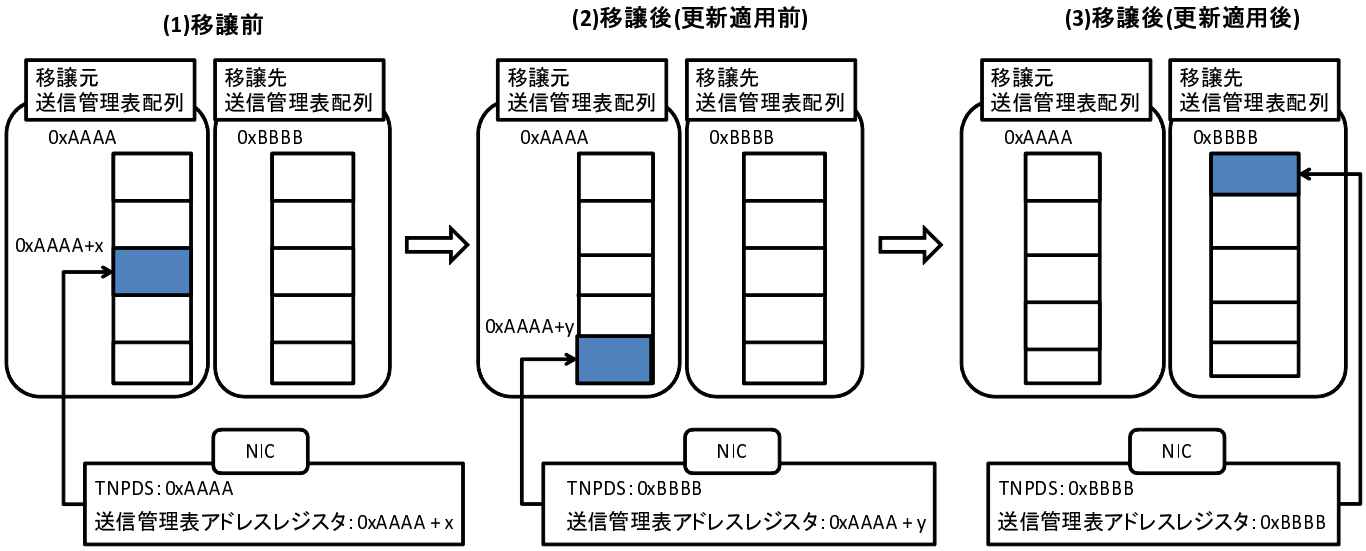

図 7　NIC 移譲時に発生する問題

(1) 移譲前

NIC の TNPDS は移譲元送信管理表配列の先頭アドレス 0xAAAA を指している．送信管理表アドレスレジスタは 0xAAAA から x バイト進んだアドレスを指している．

(2) 移譲後 (更新適用前)

NIC を移譲する．これにより，TNPDS は移譲先送信管理表配列の先頭アドレス 0xBBBB を指すようになった．しかし，送信管理表アドレスレジスタは送信管理表配列の末尾に到達するまで，TNPDS の内容を書きこまれない．よって，送信管理表アドレスレジスタは移譲元送信管理表配列を参照し続けている．

(3) 移譲後 (更新適用後)

NIC が移譲元送信管理表配列の末尾を参照したことにより，送信管理表アドレスレジスタに TNPDS の内容が書き込まれる．これにより，NIC は移譲先送信管理表配列の使用を開始する．NIC が送信管理表配列の末尾を参照するまで，TNDPS の変更が反映されないという問題がある．

7 おわりに

本資料では，パケット送受信時における NIC の動作と，NIC 移譲時に発生すると考えられる問題について述べた．NIC 移譲時に送信管理表アドレスレジスタ，受信管理表アドレスレジスタが変更できないことにより，TNPDS，RDSAR の変更が即座に反映されないという問題があると考えられる．今後は，問題への対処を検討する．