5F－6

連続メディア処理向きマイクロカーネルの開発(3)
－メモリ管理の開発－

中原雅彦，岩嵜正明，竹内　理，中野隆裕，芹沢　一
(株)日立製作所　システム開発研究所

1．はじめに

連続メディア処理では、入出力スループットの向上やQOS（Quality of Service）の保証が必要であるが、この実現には、OSのメモリ管理にもこの要求に対応する機能が必要になる。しかし、既存OSのメモリ管理が提供するAPI（Application Program Interface）では十分な機能を提供できない。

本稿では、上記の課題に配慮した連続メディア向きマイクロカーネルHiTactixのメモリ管理について述べる。

2．メモリ管理の課題

連続メディア処理に要求される機能を満たすためには、メモリ管理が以下の課題を解決する必要がある。

（1）仮想記憶における時間予測不能の問題

連続メディア処理においては予測不能な資源待ちを発生させないようにする必要があるが[1]、例えばオン・デマンド・ページング機能は、予測不能のページングを発生させる。これを避けるためにページ・ワイヤリング等の措置が必要になる。また、コピー・オン・ライト機能では、データ書き込みの際に物理メモリコピーが発生し、これが予測不能な待ち時間発生の要因となる。

（2）空間構成と性能の問題

一般に多重アドレス空間機能を導入すると、プロセス当りの消費メモリ量は増加し、ユーザ・アプリケーションの実行性能は低下する。メモリ消費量増加の要因は、ページテーブル等のメモリ管理情報の増加である。また、実行性能低下の要因は、空間間を渡る制御やデータアクセスに伴う空間切り替えオーバヘッド、TLBミス率やキャッシュミス率の増加である。

（3）物理メモリのページ分割問題

一般の仮想記憶方式では、物理メモリはページ単位に管理され、また、ユーザが物理メモリを直接操作するAPIは提供されていない。このため、物理アドレスの連続性はページ単位にしか保証されない。２ページ以上のデータをDMA転送する場合、１ページ毎にDMA転送を行うコマンドチェインを作成するか、１ページ毎に入出力処理を行う必要がある。この結果、入出力処理に伴うオーバヘッドが大きくなる。

（4）メモリコピー・オーバヘッドの問題

既存OSの入出力APIは、ユーザ・バッファの開始アドレスとサイズを指定させる。このインタフェースでは、開始アドレスやサイズがページ境界やワード境界に整合している保証がない等、DMAバッファとして使用できない場合がある。このような場合にカーネル・バッファとユーザ・バッファ間でメモリコピーが必要となり、これが性能上のボトルネックとなる。

3．メモリ管理の概要

HiTactixでは、ユーザ・アプリケーション間のメモリ保護、及び、アドレス空間内のフラグメント化を防止するために、多重アドレス空間構成を導入している。また、ページング機能及びコピー・オン・ライト機能を省くことにより、予測不能なオーバヘッドの発生をなくしている。

また、HiTactixの仮想アドレス空間内は「領域」を単位として管理している。領域とは、使用目的毎に仮想アドレス空間上の範囲を切り出した論理的単位である。

更に、物理ページの割り当ては「物理ページ集合」を単位として行う。物理ページ集合は、１ページ以上の物理ページを一つの資源として、領域内の連続した仮想アドレスにマップするための管理単位である。

以下では、第２節の（2）～（4）に対する解決について述べる。

（1）領域共有と物理ページ集合共有

HiTactixでは、多重アドレス空間構成による性能低下を回避するため、領域及び物理ページ集合を仮想アドレス空間間で共有できる機能を設けた。空間間で領域及び物理ページ集合を共有することにより、キャッシュミスの発生を低減している。また、共有した領域及び物理ページ集合はメモリ管理内の管理情報も共有する。管理情報の共有により、管理情報に必要な物理メモリ量を削減している。

更に、一体型カーネルと同様に、各プロセスの仮想アドレス空間を、ユーザ領域とシステム領域に分割し、システム領域を全プロセスで共有する構成とした（図１）。この

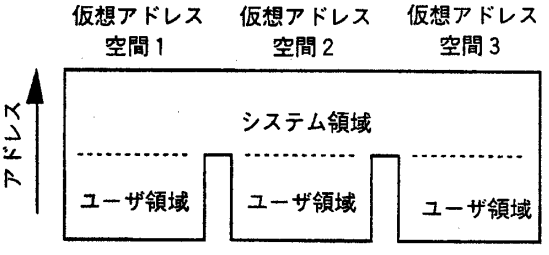

図１　空間構成

A Micro-kernel for Continuous Media Processing
－ Memory Manager －
Masahiko NAKAHARA, Masaaki IWASAKI,
Tadashi TAKEUCHI, Takahiro NAKANO
and Kazuyoshi SERIZAWA
Systems Development Laboratory, Hitachi Ltd.

構成により、システムコールが発行されても空間の切り替えが発生せず、空間を渡るデータアクセスも発生しないため、TLBミスやキャッシュミスの発生を低減している。

（2）連続物理ページ

HiTactixでは、ページサイズ以上のアドレスが連続している物理メモリの確保を可能としている。ページサイズ以上の連続アドレスを持つ物理メモリを連続物理ページと呼ぶ。連続物理ページは物理ページ集合の一機能として提供している。この機能により、高速入出力デバイスが必要とする大容量DMA転送バッファを提供可能である。

（3）ダイレクト・バッファ・マッピング

HiTactixでは、ユーザ空間内のバッファにDMA転送を可能とするダイレクト・バッファ・マッピング機能を提供する（図2）。この機能により、カーネル・バッファとユーザ・バッファ間のメモリコピーが不要となる。2ページ以上の大きさを必要とする場合にも、先に述べた連続物理ページを使用することにより、ユーザ領域にDMAバッファを作ることが可能である。

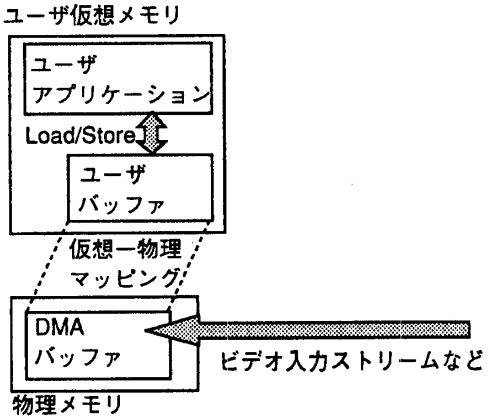

図2　ダイレクト・バッファ・マッピング

一般に、コピーオーバヘッドを削減する手法として、リマップによる仮想コピーやメモリ・マップト・ファイルがある。しかしながら、仮想コピーは、大容量コピーを行う場合には有効であるが、数ページ程度のコピーではリマップのコストが無視できなくなり、物理コピーよりもコストが大きくなる場合がある。また、メモリ・マップト・ファイルは仮想アドレス空間の大きさを越えるビデオデータ等を扱うことができない。

4．評価

本節では、入出力処理性能を物理コピー、仮想コピー、ダイレクト・バッファ・マッピングの各方式について比較する。具体的には、OS内のディスク用バッファからユーザ・バッファを介してネットワーク・デバイスのバッファへデータを渡す処理（図3）のコピーオーバヘッドについて考察する。ここでは、コピーオーバヘッド以外の性能に影響を与える要因は無視する。なお、測定はすべてHiTactixを実装したPC-AT互換機（Pentium：90MHz）上で行った。

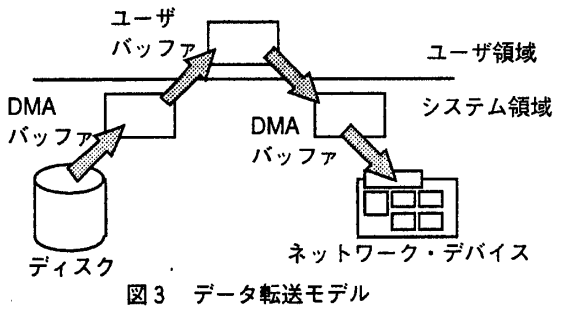

図3　データ転送モデル

（1）物理コピーの場合

物理コピーに要するコストは、4KB当たりベストケース（キャッシュヒット時）で約11,000サイクル、ワーストケース（キャッシュミス時）で約17,600サイクルであった。この値から換算すると、データ転送スループットはベストケースでも約17MB/s、ワーストケースでは約10MB/sが限界となる。

（2）仮想コピーの場合

表1はHiTactixのメモリ管理APIを使用してリマップ（アンマップとマップ）を行った場合のコストと、その値から換算したデータ転送スループットの上限値である。

表1から、小さいバッファしか使用できない場合、仮想コピーを行っても物理コピーと同程度～数倍の性能しか得られないことが判る。

表1　リマップコストとデータ転送スループット

サイズ	リマップコスト		スループット	
	ベストケース	ワーストケース	ベストケース	ワーストケース
4 KB	4,500ｻｲｸﾙ	17,000ｻｲｸﾙ	40MB/s	11MB/s
64KB	7,800ｻｲｸﾙ	20,400ｻｲｸﾙ	370MB/s	140MB/s
1MB	57,400ｻｲｸﾙ	71,600ｻｲｸﾙ	800MB/s	640MB/s

（3）ダイレクト・バッファ・マッピングの場合

ダイレクト・バッファ・マッピングを使用すれば、バッファ間コピーは1回も発生しない（図3に示す3個のバッファは同一となる）。したがって、本方式のコピーオーバヘッドは0であり、これがデータ転送スループットの限界要因にはならない。

5．おわりに

HiTactixのメモリ管理では、領域及び物理ページ集合を共有することで、キャッシュヒット率、TLBヒット率を向上させる空間構成を可能とした。また、ダイレクト・バッファ・マッピングの採用により、既存OSでは回避不能であったOSーユーザ・アプリケーション間のメモリコピー・オーバヘッドを解消した。

参考文献

[1]竹内他、連続メディア向きマイクロカーネルの開発(2)ーサイクリック・スケジューラの設計と実装ー、情報処理学会第53回全国大会予稿集、1996