HiTactix/Symbioseの開発（4）

2Z－4　－ストライプト・メディア・ファイルシステム－

児玉昇司，岩嵜正明，中原雅彦，田口しほ子，竹内理，中野隆裕，川田容子
（株）日立製作所システム開発研究所

1. はじめに

近年、大容量ディスク・ドライブやFibre Channelなどの高速データ転送インタフェースが出現してきた。これらのハードウェアを用いて、MPEG形式などに圧縮されたビデオ・データや音楽データなど（本論文ではコンテンツと呼ぶ）を同時に多数配信可能なストリーム・サーバが実現可能となりつつある。

しかしながら，その実現には，以下の要件を満たすファイルシステムを開発する必要がある．

（要件1）ファイル・ストライピングによる入出力スループットの向上：各コンテンツを複数のディスクにストライプ状に分割して格納し，単体ディスク・ドライブの限界を超えるスループット性能を実現する．

（要件2）QoSを保証した多重ストリーム同時入出力：高多重ストリーム入出力時に，各ストリーム毎に遅延時間を保証する．

（要件3）ストリーム入出力と非周期的入出力との混在：ストリーム・サーバ稼働中のコンテンツ入れ替え作業のような，突発的な非周期的入出力要求が発生しても，ストリーム入出力のQoSを保証する．

本論文では，上記課題を解決するファイルシステム「ストライプト・メディア・ファイルシステム（SMFS）」の設計と評価について述べる．

2. 従来研究

従来，QoSを保証した入出力スケジューリング方式として，SCAN-EDFアルゴリズム[2]や位相シフト時分割ビデオ多重アクセス方式[3]が提案されている．SCAN-EDFアルゴリズムは，入出力の完了時刻（デッドライン）を各入出力要求毎に指定し，最も早いデッドラインの入出力要求を優先的に実行するアルゴリズムであり，上記の要件2を満足する．また[3]の方式は，複数ディスクにコンテンツをストライプ状に格納することで要件1を満足し，さらにストリーム・サーバ内で周期的なタイムスロット列を発生させ，各タイム・スロット内で1本のストリーム入出力を実行する方式であり，要件2を満足する．しかしどちらの方式も要件3を満足していない．

本論文では，[3]の方式を拡張し，未使用のタイムスロットを非周期的入出力要求に割り当て可能なスケジューリング・アルゴリズムを提案する．

A Development of HiTactix/Symbiose(4) - Striped Media File System -
Syouji Kodama, Masaaki Iwasaki, Masahiko Nakahara, Shihoko Taguchi, Tadashi Takeuchi, Takahiro Nakano, Yoko Kawata
Systems Development Laboratory, Hitachi, Ltd.

3. ストライプト・メディア・ファイルシステム

3.1. システム構成

この節では，2章で述べたファイルシステムを実現するSMFSの構成について述べる．SMFSのシステム構成を図1に示す．SMFSは周期的入出力発行機構，入出力予約時間テーブル管理機構，入出力予約時間テーブル，及び入出力要求待ち行列から成る．各コンテンツは複数のディスクにストライプ状に分割して格納する．

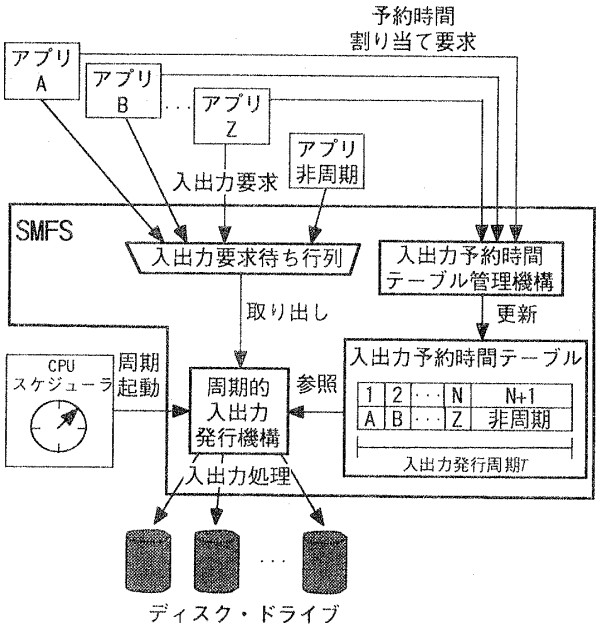

図1　SMFSの構成図

(1)周期的入出力発行機構

この機構は，HiTactixが提供する高精度スケジューラ[1]を使用して，各ストリーム毎の入出力要求を，入出力発行周期T内で必ず1回実行することを保証する．各ストリーム毎に割り当てる1周期内の入出力処理時間を，そのストリームの予約時間と呼ぶ．

(2)入出力予約時間テーブル管理機構

コンテンツのファイル・オープン時に，そのコンテンツをストリームとして入出力処理するための予約時間を割り当てる機構である．

(3)入出力予約時間テーブル

各ストリーム毎に割り当てた予約時間を管理するテーブルである．非周期的な入出力要求に対しては，ストリーム毎に区別することなく，まとめて1つの予約時間を割り当てる．

(4)入出力要求待ち行列

アプリケーションがSMFSに発行した入出力要求を，SMFSが実行するまで待機させる待ち行列である．

3.2. 入出力スケジューリング・アルゴリズム

この節では，2章で述べたファイルシステムを実現するために必要な入出力予約時間テーブルの作成アルゴリズムについて説明する．

ストリームRを周期的に入出力するとき，入出力発行周期T内で入出力するデータ・サイズをSとする．ディスク・ドライブDがデータ・サイズSを入出力するのに必要な処理時間を$T_D(S)$†とする．これをストリームRに割り当てる予約時間の長さとする．すでに割り当て済みの予約時間の長さの総和をWとする．次の不等式が成立するとき，ストリームRに対して予約時間を割り当てる．

$$W + T_D(S) \leqq T \qquad (1)$$

次に未使用の予約時間の長さUを次式により求める．

$$U = T - (W + T_D(S)) \qquad (2)$$

長さUの予約時間を非周期的入出力要求に割り当てる．周期的入出力発行機構は，各周期内において，この予約時間U内に実行可能な非周期的入出力要求を入出力要求待ち行列からFIFO順序で取り出して実行する．これによって各ストリーム毎にディスク入出力のQoSを保証し，かつ非周期的な入出力要求を同時に処理可能となる．

4. 評価実験

この章では，3章で説明したSMFSをPC上に実装し，SMFSの有効性を実証する．

4.1. 実験環境

リアルタイム・カーネルHiTactix上にSMFSを実装し，これをPentium IIプロセッサ‡(266MHz)及び128Mバイトのメモリを搭載したPC-AT互換機上で動作させた．今回の実験では，SMFS上の各コンテンツをSCSI接続したディスク・ドライブ3台にストライピングする．

4.2. 実験内容

SMFSを用いて同時に12本のストリーム（各ストリームの再生レートは495Kバイト/秒）をリードするストリーム・サーバを製作し，その性能を測定した．ただしストリームの入出力と非周期的な入出力とが混在した場合の性能を測定するために，8本のストリームにはそれぞれ予約時間を割り当てるが，残りの4本は非周期的な入出力要求として処理する．12本のストリームをリードしている途中でSMFSに対して突発的なリードを行い，各ストリーム毎に1秒当たりのリード量を測定する．

4.3. 実験結果

図2及び図3に実験結果を示す．両グラフとも，横軸は経過時間（単位は秒），縦軸は1秒当たりのリード量（単位はKバイト）である．図2は非周期的な入出力要求として処理した4本のストリームのうち1本を代表している．図3は個々に予約時間を割り当てた8本のストリームのうち1本を代表している．突発的なリードは時刻37に開始し，時刻75で終了する．

図2は，ストリーム入出力を非周期的な入出力要求として処理すると，突発的なリードが発生している間は，一定レートでデータをリードできないこと示している．この理由は，ファイルシステムが突発的なリードを処理している間，非周期的な入出力要求として扱っているストリーム入出力の処理が遅延するからである．一方，図3は，各ストリーム入出力毎に予約時間を割り当てることにより，突発的なリードが発生しても，一定レートでデータをリード可能であることを示している．この結果からSMFSの有効性を示すことができた．

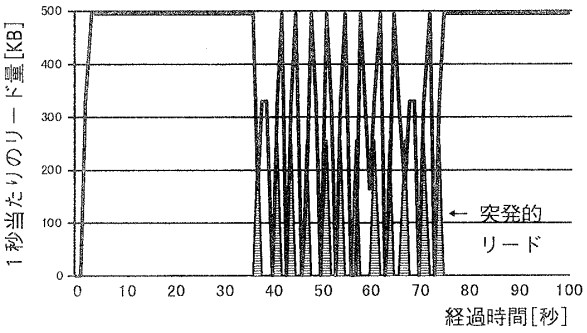

図2 非周期的入出力のリード量時間変化

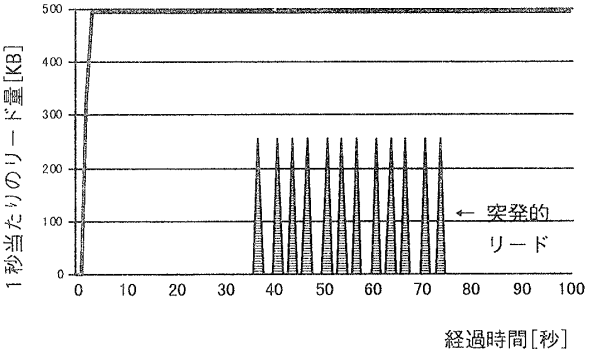

図3 ストリーム入出力のリード量時間変化

5. おわりに

本論文では，多数のコンテンツを同時に配信可能なストリーム・サーバ実現のために，以下の要件を満足するファイルシステムを設計し評価した．本ファイルシステムは，第1に，各コンテンツをストライピングし入出力スループットを向上させた．第2に，各ストリーム毎にディスク入出力のQoSを保証可能とした．第3に，各ストリーム毎のQoSを保証し，かつ突発的な非周期的入出力要求を同時に処理可能とした．

参考文献

[1]竹内　理他，「アイソクロナス・スケジューラの設計と性能評価」，情報処理学会研究報告，97-OS-74, Feb. 1997.

[2]A. L. Narasimha Reddy, et al., "I/O Issues in a Multimedia System", IEEE Computers, pp. 69-74, March 1994.

[3]阪本 秀樹他，「ビデオ情報の大規模多重アクセス方式」，信学論 D-II, Vol. J78-D-II, No. 1, pp. 76-85, 1995.

† $T_D(S)$は，入出力開始時でのディスク・ドライブのヘッダ位置に依存する．この論文では，$T_D(S)$として，実測データの平均値にマージンを加えた値を用いている．

‡ Pentium IIは米国Intel Corpの登録商標です．