バッファ監視とクライアントの優先度に基づく動画像情報多重アクセス制御方式

藤井 寛[†] 石川 篤[†] 櫻井紀彦[†]

ビデオ・オン・デマンド等の動画提供サービスのためのサーバには多数のクライアントからの同時アクセスを可能にする機構が必要である．それには動画データを格納するディスク装置の高性能化に加えて，ディスク装置上のデータ・ストリームへの同時アクセスを制御する多重アクセス・スケジューリングが重要である．本論文では高いデータ転送効率とクライアントへの短い応答時間を同時に実現する多重アクセス・スケジューリング方式について述べる．本方式は，連続転送しているデータ・ストリームのセグメントの転送のように応答時間が重要でない要求をSCANに基づいてスケジューリングすることでアクセス・オーバーヘッドを減少させる．逆に，新しいデータ・ストリーム転送開始のようにクライアントへの短い応答時間を必要とする要求には優先権を与えて，この要求のサービス順序を繰り上げることで応答時間を短縮する．また，クライアントへの動画データの連続的な供給を中断せずに多重アクセスできる条件はクライアントとディスクの間に置かれたバッファに存在するデータ量に依存するが，本論文の方式ではバッファのデータ量を最適化するようにデータ転送量を決定することで多重アクセス性能を向上する．

Multiple-Access Scheduling for Moving Picture Information Based on the State of Buffers and Priority of Clients

HIROSHI FUJII,[†] ATSUSHI ISHIKAWA[†] and NORIHIKO SAKURAI[†]

Video servers for VOD service must have a mechanism to allow multiple access from many users. In this mechanism, advances in disk technology have improved performance considerably, however, to meet the requirements of VOD services the multiple-access scheduling method for data streams of moving pictures on disks must also be improved. This paper proposes a multiple-access scheduling method that simultaneously achieves efficient data transfer and quick response to clients. This method improves disk access efficiency using SCAN to schedule requests for which the response time is not important, such as requests for a data segment in the middle of a data stream. By prioritizing requests, the method also shortens the response time for requests that require a quick response, such as those for new data streams. The condition in which moving pictures data can be transferred without interruption depends on the amount of data in the buffers between the disks and clients. Our method improves the multiple-access efficiency by determining the amount of data transferred from the disk so that the amount of data in the buffers may be optimized.

1. はじめに

磁気ディスク装置の大容量化と高速化，ディジタル画像圧縮符号化方式の進歩，および高速ネットワークの出現にともない，ビデオ・オン・デマンド等の対話型動画提供サービスが技術的に可能になっている[1)~3)]．これらのサービスにおいて，複数クライアントが独立に，それぞれ異なった動画や音声といった連続メディアのデータ・ストリームを任意のタイミングで要求する．よって，動画提供サービスのためのサーバには蓄積された大量の動画データに対する多重アクセスが可能な機構が必要であり，そのためには高速データ転送が可能な蓄積装置と実時間性を満足するデータ転送制御が重要である．つまり，動画を再生中のすべてのクライアントへデータ・ストリームを一定速度で供給し，再生を継続させなければならない．

また，ワークステーションやパーソナル・コンピュータ上で動画を取り扱うための，マルチメディア・オペレーティングシステム[4)]が出現し，これに対応した，ディスク上の連続メディアへの効率良い多重アクセス制御が必要となってきている．そこでは連続メディア

† NTT 情報通信研究所
NTT Information and Communication Systems Laboratories

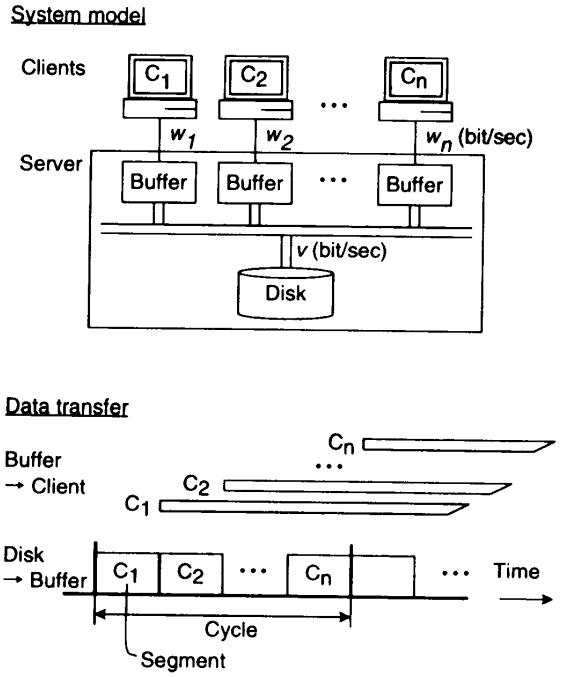

図1 動画データへの多重アクセス
Fig. 1 Multiple-access to moving picture data.

とともに文字，数値データの入出力も同時に扱うことができる機構が必要である．

図1が多重アクセスの模式図である．ディスク上に動画データが格納され，表示系であるクライアントへデータを供給する．ディスクの転送速度が動画のビットレートより大きいとし，この差を利用して多重アクセスを行う．速度の差はクライアントとディスクの間に置かれたバッファで吸収する．ディスクから周期的に大量のデータをバッファに転送し，クライアントの必要とする速度で連続的にバッファからクライアントへ転送する．クライアントへの一定速度のデータ転送を保証するには，バッファのアンダーフローおよび溢れを防止しなければならない．

動画サーバのデータ蓄積装置としては複数のディスクに並列アクセスしてデータ転送を高速化するディスクアレー[5]が主流である[6)~8)]．一定速度のデータ供給を行う，ディスクの多重アクセス・スケジューリング方式としては，オーバーヘッドを減らして同時アクセス可能なクライアント数を増加させるためにSCAN[9]に基づく方式が多く用いられている[10)~12)]．また，データ転送のために固定長タイムスロットを周期的に割り当てることで複数のデータ・ストリームの転送を同期させ，ディスクアレー中の各ディスクからつねに別々のデータ・ストリームを取り出すことで応答時間を短縮する方式も考案されている[13)]．

SCAN方式は単体ディスクおよびディスクアレーに適用可能で，転送効率の点で優れているが[14),15)]，固定順序でタイムスロットを割り当てる場合に比べてクライアントへの応答時間が長くなり，必要なバッファ容量も増加する．クライアントの増加にともないこれは無視できなくなるが，応答時間およびバッファ容量を小さくするには転送効率を下げねばならない．よって，多数のユーザに効率良くサービスするには，短い応答時間を保ちつつ，限られた容量のバッファを用いて高いアクセス効率を実現する手法が必要である．

本論文では我々の開発した，バッファに存在するデータ量とユーザの要求する応答時間に基づいて連続メディアのデータ・ストリームの転送制御を行う多重アクセススケジューリング方式，Variable Time-Slot Scheduling（VTSS）について述べる．本方式はSCAN方式の欠点を克服するために，短い応答時間を必要とするクライアントの要求に対しては，他のクライアントのバッファの状態を検査して，一定速度のデータ転送が中断しないことが保証されるとき，高優先度を与えて応答時間を短縮する．また，バッファの空き容量に基づいて転送量を決定し，バッファを有効利用する．

2. VTSSに基づく多重アクセス

ディスクの転送速度を v (bit/sec)，クライアント C_i の要求するデータ・ストリームのビットレートを w_i (bit/sec) $(v > w_i, i = 1, 2, \cdots)$ とする．ここで各データ・ストリームをセグメントに分割し，v と w_i の差を利用して周期的にセグメントを転送することでデータ・ストリームの多重アクセスが可能になる．つまり，VTSSではクライアントのデータ・ストリーム転送要求はセグメントを転送する入出力要求の系列に変換される．この入出力要求系列自体はクライアントによって発生されるが，系列中の各入出力要求，つまりセグメントの転送順序および大きさはVTSSが自律的に決定し，クライアントによる直接の制御は行われない．

動画を再生するクライアント等，ディスクからの一定速度のデータ供給を必要とするものを実時間クライアントと呼ぶこととする．このうちデータ・ストリーム転送を要求したが，まだ実際に転送は行われていないクライアントは遷移中であり，遷移クライアントと呼ぶ．たとえば，動画の再生開始や再生場面の変更を要求中のクライアントは遷移中である．クライアントへの応答時間はデータ・ストリーム転送要求の発行から，そのクライアントがデータを使用可能になるまでの時間と定義する．遷移中でない実時間クライアントは定常クライアントと呼ぶ．遷移クライアントが存在しな

いとき多重アクセスは定常状態にあるといい，それ以外のときは遷移状態にあるという．また，非実時間クライアントはそれ自身がバッファのアンダーフローや溢れを制御可能で，リアルタイムのデータ転送を必要としないクライアントである．バッチ処理でディスク上の動画データを転送する場合がこれに相当する．

多重アクセスで最も重要なのは応答時間および同時アクセスしているデータ・ストリームの数を表す多重度である．VTSS は以下の手続きに基づいて，多重度と応答時間に対してそれぞれ相反する影響を与えるセグメントの順序および大きさを動的に決定することで，データ・ストリームの多重アクセスにおける高多重度と短い応答時間を可能にする．

[VTSS の手続き]

1 [スケジューリング点] 1-1, 1-2, 1-3 を実行する．

1-1 クライアントの状態およびバッファに存在するデータ量を調べる．

1-2 遷移クライアントが存在すれば，スケジューラビリティを 3 章の多重度判定法によって判定する．

1-3 各データ・ストリームのサービス順序とセグメント長を 4 章の順序付け法と 5 章のセグメント化手続きに従って決定する．

2 1-3 の条件に従い，各データ・ストリームに対して 2-1, 2-2 を実行する．

2-1 バッファ-ディスク間のセグメントの転送を行う．

2-2 転送したデータ量をバッファに通知する．対応するクライアントが遷移中ならばバッファのデータが即利用可能かどうかを通知する．

3 すべてのクライアントにサイクルの終了を通知し，次のスケジューリング点へ行く． □

上記手続において，スケジューリング点から次のスケジューリング点までをサイクルと呼ぶ．

定常クライアントはデータ・ストリームのビットレートに応じた速度で連続的にバッファからデータを取り除く．バッファが空になるとエラーが発生する．遷移クライアントはバッファのデータが使用可能になってからデータを使用開始し，定常クライアントになる．VTSS の手続きのステップ 2-2 でデータ使用可能の通知があればその時点からデータを使用開始する．通知がなければサイクル終了時点からデータを使用開始する．

3. サイクル長と多重度

1 サイクルで転送可能なデータ量は，サイクル長 T，ディスクの転送速度 v，およびオーバーヘッドに依存する．多重度 n のときのオーバーヘッドは，シークおよび回転待ち時間とその他のソフトウェア等のオーバーヘッドからなり，確率変数 X_n で表される．

i トラックのシーク時間を $a\sqrt{i}+b$，最大回転待ち時間を c，1 入出力あたりのその他のオーバーヘッドを h とし，すべてのシリンダが等確率でアクセスされると仮定すると，アクセス順序が SCAN で決定されるときの X_n の平均 $\mu(X_n)$ は次のようになる（式 (13) を参照）．

$$\mu(X_n)=\left(\frac{c}{2}+b+h\right)n + \frac{4n\sqrt{n}(n+1)}{(2n+1)(2n+3)}a\sqrt{N} \qquad (1)$$

N はディスクのシリンダ数である．n の増加にともない X_n のばらつきは相対的に小さくなる．そこで 1 サイクルの転送量を求める際にオーバーヘッドの値として $\mu(X_n)$ を用いる．ディスク上のアクセス位置の分布に偏りがなければ，多重度 n のとき長さ T のサイクルで転送可能なデータ量の合計 $D_n(T)$ は

$$D_n(T)=\{T-\mu(X_n)\}v \qquad (2)$$

である．また，平均転送速度は

$$\{1-\mu(X_n)/T\}v \qquad (3)$$

である．ここで，

$$T\sum_{k=1}^{n}w_k \le D_n(T) \qquad (4)$$

ならば 1 サイクルですべてのクライアントにそのサイクルでの消費量以上のデータを転送できる．アクセス位置の分布に偏りがあるとき X_n の平均は $\mu(X_n)$ より小さくなるから，上記の条件でアクセス可能である．サイクル途中でのバッファ・アンダーフローを防ぐにはサイクル開始点で各クライアント C_i のバッファに Tw_i 以上のデータが必要である．したがって新たなデータ・ストリームが要求されたとき，次のようにサイクル開始点で可能な多重度を判定する．

[多重度判定法] T をどの定常クライアントのバッファも空にならない最大時間とする．このとき式 (4) が成立しなければクライアントの新たな要求は拒否される．(ここで最大サイクル長 $T_{\max}$ を設定して T の最大値を $T_{\max}$ に制限することもできる.) □

X_n は確率変数であり，実際は 1 サイクル $T\sum_{k=1}^{n}w_k$ のデータの転送に要する時間は T 以上になりうる．しかし，ここですべてのバッファに余分なデータがあれば X_n の分散を吸収できるため，アンダーフローを起こさずに多重度 n でアクセス可能である[14]．

4. 順序付け法

4.1 データ・ストリームの優先度と順序

すべてのデータ・ストリームが連続的に流れ続ける定常状態では，オーバーヘッドを減らして転送効率を上げ，多重度を向上させることが最も重要である．これにはディスクのシーク時間を短縮するSCANによってアクセス順序を決定することが有効である[14]．一方，遷移状態では遷移クライアントへの応答時間も重要である．遷移クライアントのデータ要求と同時にそのバッファの古いデータは無効になるため，新たなデータのバッファへの転送が完了するまで遷移クライアントはデータを使用できない．そこで，応答時間を短縮するために遷移クライアントに優先度を与えてセグメントの転送順序を繰り上げる．

データ・ストリームの順序はスケジューリング点で優先度に基づいて次のように決定される．

[順序付け法]

1 同一優先度を持つクライアントが要求するデータ・ストリームごとにグループ化する．

2 まだ順序付けられていない最も高い優先度のグループをSCANによって直前のグループの次に順序づける．

3 2に戻る． □

優先度はスケジューリング点で以下の基準で与える．

- 定常クライアントには標準優先度を与える．
- 遷移クライアントにはその重要度に応じて優先度を与える．
  - 重要度 = 高い：つねに高い優先度を与える．応答時間はつねに短縮されるが，定常クライアントはバッファ・アンダーフローを起こす可能性がある．
  - 重要度 = 標準：4.2節で述べるように，すべての定常クライアントのバッファに，アンダーフローを起こさないだけの十分なデータが存在するときのみ高い優先度を与え，応答時間は短縮される．
  - 重要度 = 低い：標準優先度を保つ．
  - 重要度 = 非常に低い：つねに低い優先度を与える．直後の連続した数サイクルでデータ転送を行わない．
- 非実時間クライアントにはつねに低い優先度を与える．

クライアントの重要度は動画に支払われる対価に依存，または再生開始よりも一時停止後の再開を重要視するといったように任意の基準で定義できる．

4.2 優先度上昇の判定

式(2)，(4)から次の式が得られる．

$$T \geq \frac{v}{v - \sum_{k=1}^{n} w_k}\mu(X_n) = \frac{v}{v-w}\mu(X_n) \quad \text{ただし } w = \sum_{k=1}^{n} w_k \tag{5}$$

これは多重度 n でデータ・ストリームのビットレートの合計が w のときに，ディスクからのデータ供給量がクライアントのデータ消費量の総和を上回るためのサイクル長 T の条件である．各バッファにはアンダーフローを防ぐためにそのサイクルでの消費量以上のデータが必要である．よってサイクル開始時の C_i のバッファ中のデータ量の下限は次のようになる．

$$f_{LB}(w_i, w) = \frac{w_i v}{v-w}\mu(X_n) \tag{6}$$

さらに，高優先度のクライアント C_i は応答時間を短縮するためにサイクルの途中でバッファのデータを使用開始するから，次のサイクル開始時に式(6)がつねに成立するためには転送するセグメント長は $2f_{LB}(w_i, w)$ 以上でなければならない．

遷移状態のスケジューリング点ではすべての標準重要度の遷移クライアントが高い優先度を持つと仮定して，5章のセグメント化手続きに従ってセグメント長を計算する．ただし，標準未満の優先度のクライアントのセグメント長は0に固定する．ここでこれらの仮に高優先度を与えられた遷移クライアント C_i それぞれに $2f_{LB}(w_i, w)$ 以上の大きさのセグメントを割当て可能ならば実際に C_i に高優先度を与える．さもなければ仮の高優先度はすべて取り消し，各 C_i に標準優先度を与えて再びセグメント長を計算する．この判定においてオーバーヘッドの分散を吸収するために余裕 f_M を持たせうる．このとき，$f_M + 2f_{LB}(w_i, w)$ のセグメントが割当て不可能ならば高優先度は与えられない．

図2では t_0 で C_2 が新たなデータ・ストリームを要求している．上の図で C_2 の優先度は標準で，$C_1 \sim C_5$ のアクセス順序はSCANに基づいている．C_2 はデータを次のサイクル開始まで使用できないため応答時間は $t_1' - t_0$ である．下の図では C_2 に高優先度を与えており，セグメントは第2サイクルの最初に転送されて転送終了直後にデータを使用開始している．よって応答時間は $t_1 - t_0$ で，上図のものより短い．

遷移クライアントの優先度を上げると応答時間は短縮される．しかしSCANによる最適な順序が乱れてサイクル長が増加し，かつ遷移クライアントへの転送

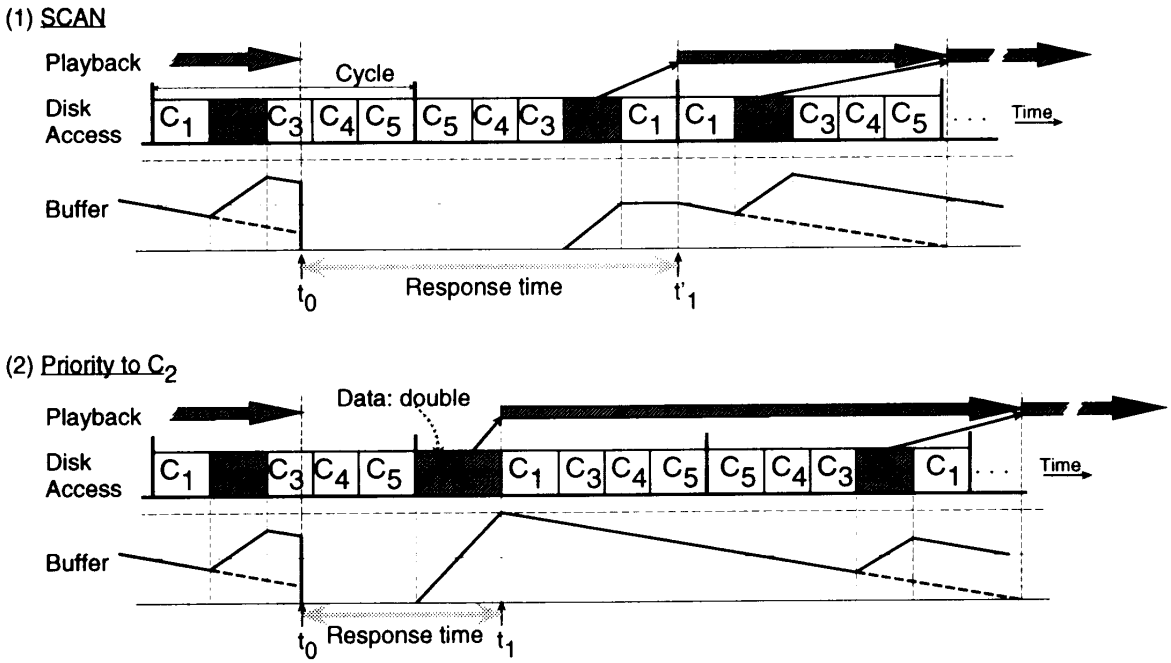

図2 クライアント C_2 への優先度
Fig. 2 Priority given to client C_2.

量が増加するためサイクルが長くなる．よって応答時間を短縮するためにはバッファにデータが十分ある必要がある．

4.3 低優先度クライアント

低優先度の遷移クライアントは応答時間の重要度が低く，データ要求直後の一定期間転送を行わない．

サーバが可能な最大多重度近くで多重アクセスされているとき，アクセス時間の分布によって定常クライアントのバッファのデータが欠乏することがある．このとき，多重度を下げてバッファのデータ量を増加させるために低優先度のクライアントのサービスを遅延する．遅延の量は，あらかじめ設定した最大遅延量と，定常クライアントのバッファのデータ量がしきい値を超えるまでの時間のうち短い方として決まる．

5. セグメント化手続き

式(3)が示すとおり，セグメントが大きくサイクルが長いほどオーバーヘッドが減少し，データ転送効率は向上する．ただしバッファ・アンダーフローを防ぐために，最大サイクル長は全定常クライアントのバッファのデータ量のうち最小のものでおさえられる．T をアンダーフローが発生しない最大サイクル時間，S を T で転送可能なデータ量とする．サイクル開始点で各セグメントのアドレスが分かるため S の値は計算できる☆．ここで s_{dis} $(\leq S)$ のデータをセグメントに分割する．分割量は次のサイクルでの T が最大になるように決める．また，転送効率を保つためにセグメント長に下限 s_{lb} を設定する．さらに，サイクル開始時にバッファに上限（$f_{i,UB}$）を越えるデータを持つクライアントへはセグメントを割り当てない．$f_{i,UB}$ はバッファ容量から s_{lb} を引いたもの等に設定される．

データ・ストリームのビットレートは画像サイズや解像度といった素材の性質や再生速度，QoS 等の OS の制御によって変化するから，バッファの状態はデータ量の絶対値ではなくビットレートとの比で評価する．

VTSS はディスクからのデータ・ストリームの読み出しだけでなく，録画等のディスクへの書き込みもスケジューリングできる．録画クライアントに対してはバッファの溢れを防止するためにデータ量の代わりに空き領域について考慮する．

セグメント化手続きを以下に示す．

[セグメント化手続き]

1 $C_1, C_2, \cdots, C_n$ を実時間クライアント，b_i，f_i，w_i をそれぞれ C_i のバッファ容量，バッファに存在するデータ量，データ・ストリームのビットレートとする．ここで，

☆ 実際には回転待ち時間の正確な予測は困難で，S はある分布をなすが，この分散は吸収可能であると仮定する．

$$F_i = \begin{cases} f_i, \\ \quad C_i \text{ がディスクから読み出すとき} \\ b_i - f_i, \\ \quad C_i \text{ がディスクへ書き込むとき} \end{cases} \tag{7}$$

$$T_i = F_i/w_i \quad (i = 1, 2, \cdots, n) \tag{8}$$

と F_i, T_i を定義し，T を T_i の最小値とする．また S を 1 サイクル T で転送可能なデータ量とする．

2 各クライアント C_i に対し，セグメント長 s_i を

$$(F_i + s_i)/w_i \quad (i = 1, 2, \cdots, n) \tag{9}$$

のうち最小のものが，以下の条件の下で最大になるように決める．

$$\begin{cases} s_i = 0 \text{ または } s_i \geq s_{lb}, \\ \quad \text{ただし } F_i \geq f_{i,UB} \text{ のとき } s_i = 0 \\ s_{dis} = \sum_{k=1}^{n} s_k \text{ として, } s_{dis} \leq S \\ f_{LB}(w_i, w) + w_i(\mu(X_n) + s_{dis}/v) \\ \quad \leq F_i + s_i \\ \quad \leq b_i \end{cases} \tag{10}$$

□

ここで，s_{dis} が小さいほど応答時間は短く，s_{dis} が大きいほど転送効率は良い．$s_{dis} = S$ のとき式 (10) の第 3 式の左辺は $2f_{LB}(w_i, w)$ となる．アクセスの効率化のため，s_i の大きさは実際にはブロック単位で決められる．

図 3 に上記手続きに基づく 1 サイクルのデータ転送の様子とバッファの状態を示す．5 つの同一ビットレートのデータ・ストリームがディスクから転送されている．バッファの 1 ブロックは 1 単位時間で消費され，多重度 5 のとき 1 単位時間に 1 サイクル合計少なくとも 5 ブロックのデータを転送できる．サイクル開始時のバッファのデータ量は C_4 が 2 ブロックで最も少ないから $T = 2$，よって $S = 10$．そこで $s_{dis} = 10$ としてこれを 5 つのセグメントに分割する．サイクル終了時のバッファのデータ量の最小値を最大にするには転送後のデータ量を平均化すればよい．つまり C_1, C_3, C_5 に 2 ブロックずつ，C_2 に 1 ブロック，C_4 に 3 ブロックを分配する．1 サイクル合計 10 ブロックの転送時間は 2 単位時間より短く，その間のデータ消費量は 2 ブロック以下である．よって転送後の各バッファには 3 ブロック以上のデータが存在するから，次のサイクルで $T \geq 3$ となり T は増加する．

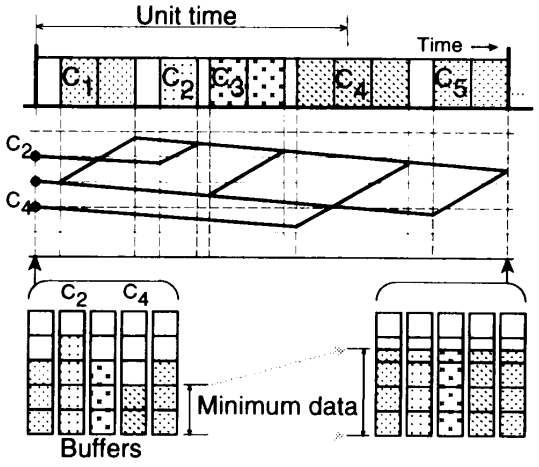

図 3 データ転送とバッファの状態
Fig. 3 Data transfer and status of buffers.

転送効率に加え，前章で述べたようにセグメントの順序決定の自由度もバッファのデータ量に依存するが，セグメント長を最適化することで効率良くバッファ中のデータ量を増加させることができる．

圧縮符号化された動画像等のデータ・ストリームはビットレート一定とは限らない．これに本手法を適用するには平均ビットレートを用いる．サイクル長と同程度の時間において平均ビットレートが一定であればビットレート一定の場合と同様に扱える．サイクル長より長い時間においてビットレートが変動する場合，平均ビットレートを用いてスケジューリングすることにより低ビットレートのところでバッファのデータ量が増加し，高ビットレートのところでこれが消費される．一定のサイクルでビットレートが変動する場合，バッファが最大ビットレート時においてアンダーフローを起こさない十分な容量を持てば，データ・ストリームへアクセスする初期の，バッファ中のデータ量が十分でないときにアンダーフローが起こる可能性はあるが，それ以降のサイクルにおけるアンダーフローの発生は抑えられる．また初期におけるバッファ・アンダーフローの抑制には，遷移クライアントへのセグメントサイズの余裕 f_M を増加させることが有効である．

6. 副スケジューリング点

セグメント長の増加にともなって転送効率が向上する反面，応答時間が長くなる．そこで応答時間を短く保つためにサイクル中のセグメントの間に副スケジューリング点を挿入する．定常状態では副スケジューリング点は機能しない．あるクライアントが新たなデータ・ストリームを要求したとき，スケジューリング点より近くに副スケジューリング点が存在すれば，そのうち

最も近いものが活性化され，遷移クライアントの要求を含めてその点において再スケジューリングが行われる．このときセグメント長および順序はスケジューリング点と同様に決定される．ただし，すでにそのサイクルで副スケジューリング点以前に転送されているデータ・ストリームのセグメント長は0に固定される．

副スケジューリング点以降における動作は遷移クライアント C_i に割当て可能なセグメント長 s_i によって次の3つに分かれる（図4）．

- $s_i < f_{LB}(w_i, w)$ のとき，そのサイクルでは C_i にセグメントを割り当てない．新たにセグメントを割り当てられる遷移クライアントが存在しないとき，再スケジューリングはすべて無効になる（図4(A)）．
- $f_{LB}(w_i, w) \leq s_i < 2f_{LB}(w_i, w)$ のとき，大きさ s_i のセグメントが転送されるが，データは次のサイクルの開始まで使用できない（図4(B)）．
- $s_i \geq 2f_{LB}(w_i, w)$ の場合，大きさ s_i のセグメントが転送され，データ転送完了直後からバッファのデータは使用可能になる（図4(C)）．

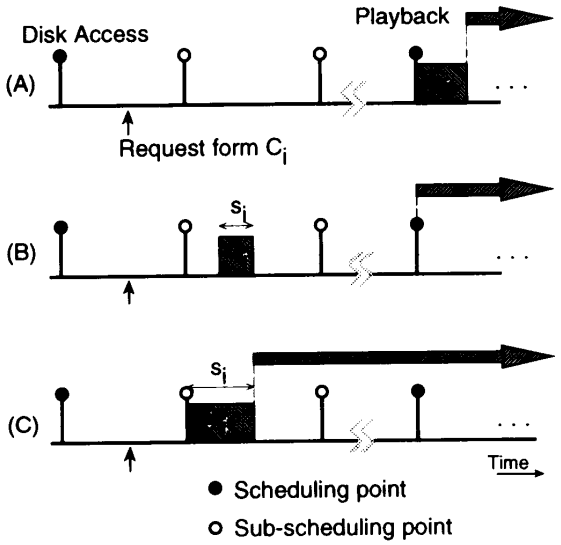

図4 副スケジューリング点
Fig. 4 Sub-scheduling point.

定常クライアントへはスケジューリング点と同様に決定したセグメント長を割り当てて転送する．

図5において，$C_1 \sim C_4$ がアクセスしている．t_1，t_2，t_3 がスケジューリング点である．各クライアントへ2ブロックのデータが転送されており，サイクル長は1ブロックずつ転送する場合の約2倍である．図では2クライアントごとに t'_1，t'_2 で示される副スケジューリング点を挿入している．

ここで $t_1 \sim t_2$ は定常状態である．よって，t'_1 の副スケジューリング点は機能しない．C_4 は t_r で動画の新たな場面を要求している．t_r に最も近い副スケジューリング点 t'_2 で C_4 は C_1，C_2 とともに再スケジューリングされる．なぜなら，t'_2 までに C_3 への転送のみが正当に完了しているからである．また，サイクル $t_2 \sim t_3$ では t'_2 で C_4 が再びアクセスするから，定常クライアントのバッファが空になる前にサイクルを終了するために C_1，C_2 への転送量は減少する．これは次サイクルの可能な最大サイクル長の減少につながる．

7. 解　　析

遷移クライアントの優先度を上げられるかどうかはサイクル長と定常クライアントのバッファ内のデータ量に依存する．本章ではこの点について解析する．

モデルとして表1に示す仕様の磁気ディスクに格納されたビットレート1.5 Mbit/secのデータ・ストリームへの多重アクセスを考える．各トラックは等確率でアクセスされ，標準的サイクル長を1秒とする．式(4)から，このときの最大多重度は35である．クライアン

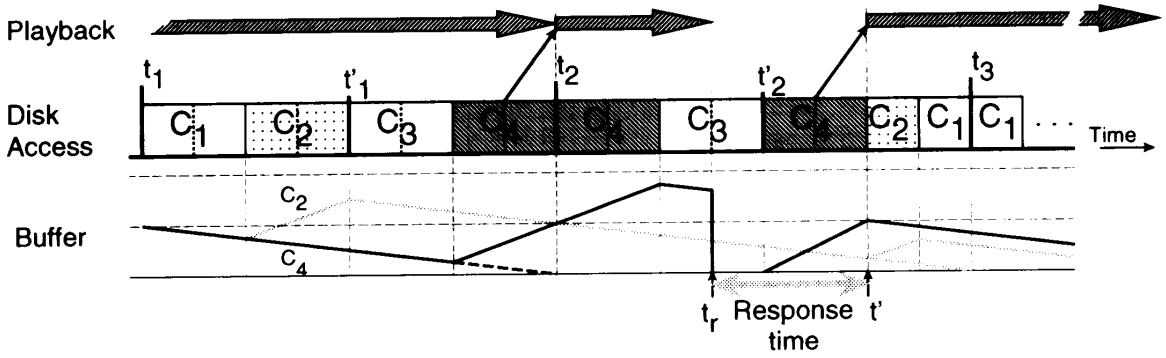

図5 副スケジューリング点の動作例
Fig. 5 Action at sub-scheduling point.

表1 磁気ディスクの仕様
Table 1 Specifications of the magnetic disk.

パラメータ		値	
容量		約 1.4	Gbyte
シリンダ容量		660	Kbyte
シリンダ数	$N=$	2235	
転送速度	v(byte/s) $=$	$20\times1024^2\times8$	byte/s
オーバーヘッド	h(ms) $=$	5	ms
最大回転待ち時間	c(ms) $=$	16.7	ms
シーク時間		0	ms ($n=0$)
(n トラック)	$a\sqrt{n}+b$(ms) $=$	$0.45\sqrt{n}+1.95$	ms ($n>0$)

トの総数がその1.5倍程度の50，各クライアントは平均100秒の指数分布に従う間隔で遷移状態になるとしたとき，長さ1秒のサイクルで同時に4以上のクライアントが遷移状態になる確率は約0.16%，長さ2秒のサイクルで約1.7%である．よって以下の議論では1サイクルに存在する遷移クライアント数は最大多重度に比べて十分小さいと仮定する．また，シーク距離 i に対して $r=i/N$ が連続と見なせると仮定して，シーク距離を r，シーク時間を $t=a'\sqrt{r}+b=a\sqrt{N}\sqrt{r}+b$ と表現する．

遷移クライアントに高優先度を与えるとアクセス時間とデータ転送時間が増加する．多重度一定の場合，シーク時間の定数部分と回転待ち時間の分布は不変である．よってアクセス時間の増加は遷移クライアントのアクセスにおけるシーク時間の $a'\sqrt{r}$ の項から，定常クライアント間のシーク時間の $a'\sqrt{r}$ の項の減少量を引いた値になる．以下においてシーク時間の第1項のみ考慮し $t'=a'\sqrt{r}$ とする．

m 遷移クライアントによるシークはSCANによる m 回と，遷移クライアントと定常クライアント間の1回からなる．SCANでスケジューリングされた n 回のランダムなアクセスのシークの合計距離 R の密度関数は，

$$n(n+1)(1-R)R^{n-1},\quad 0\le R\le 1 \tag{11}$$

である（式の導出は付録に示す）．ここで n 回のシーク距離はすべて等しいと仮定する．この仮定に基づく単一のシーク時間 t' の密度関数は，$t'=a'\sqrt{R/n}$ から，

$$2(n+1)\left(\frac{n}{a'^2}\right)^{n+1}(a'^2-nt'^2)t'^{2n-1},\quad 0\le t'\le a'/\sqrt{n} \tag{12}$$

となる．なお，シーク距離の和が一定とすると，距離を等分したときにシーク時間の和は最大になる．n シークの平均時間は単一シーク時間の平均の n 倍で，

$$t_{\mathrm{seek}}(n)=\frac{4n\sqrt{n}(n+1)}{(2n+1)(2n+3)}a' \tag{13}$$

である．また，$n=1,2,3$ のときの単一シーク時間の分散はそれぞれ

$$\frac{11}{225}a'^2\approx 22.1,\quad \frac{73}{4900}a'^2\approx 6.74,\quad \frac{43}{6615}a'^2\approx 2.94 \tag{14}$$

である．遷移クライアントと定常クライアント間のシークは $n=1$ の場合に相当する．よって，式(13)からアクセス時間は平均

$$t_{\mathrm{seek}}(m)+t_{\mathrm{seek}}(1) \tag{15}$$

増加する．多重度 $n_{\max}$ のとき定常クライアント間のシーク時間は式(13)から，

$$t_{\mathrm{seek}}(n_{\max})-t_{\mathrm{seek}}(n_{\max}-m-1) \tag{16}$$

減少し，転送時間は遷移クライアントの要求するデータ・ストリームのビットレートを $w_1,w_2,\cdots,w_m$ (bit/sec) としたとき，

$$\frac{\sum_{k=1}^{m}w_k\times 1000}{v} \tag{17}$$

増加する．

式(15)，(16)，(17)から，m 個の遷移クライアントが存在するとき，サイクル長は

$$\begin{aligned}&t_{\mathrm{increase}}(n_{\max},m)\\&=t_{\mathrm{seek}}(m)+t_{\mathrm{seek}}(1)\\&-\{t_{\mathrm{seek}}(n_{\max})-t_{\mathrm{seek}}(n_{\max}-m-1)\}\\&+\frac{\sum_{k=1}^{m}w_k\times 1000}{v}\end{aligned} \tag{18}$$

増加する．

$t_{\mathrm{seek}}(n)$ は実際の平均 n シーク時間よりもわずかに大きく，また，クライアント数の増加にともないSCANによるシーク距離の分散は小さくなる．よって，実際に増加する平均シーク時間は式(18)に示されるよりもわずかに小さい．各シーク時間の分布を正規分布で近似できると仮定すると，式(14)および回転待ち時間の分散 $c^2/12\approx 23.2$ から，遷移クライアントに高優先度を与えたとき，サイクル長の分散 σ^2 はたかだか $(c^2/12)\times 2$ 程度増えるだけである．これは定常状態におけるサイクル長の分散 $(c^2/12)n_{\max}$ [14]

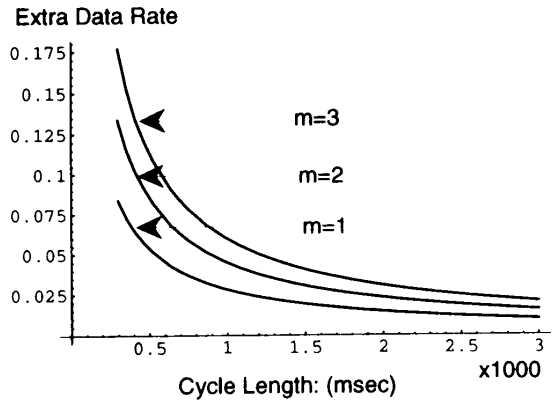

図6 遷移クライアントに高い優先度をあたえるためのデータ量
Fig. 6 Amount of data to prioritize transitional clients.

に比べて，$n_{\max} = 10$ 程度以上のとき十分小さいと考えられる．

以上の議論から，遷移クライアントによってサイクル長の分布は最大 $t_{\text{increase}}(n_{\max}, m)$ 右へずれる．つまり，定常状態におけるサイクル長を T としたとき，定常クライアントのバッファのデータ量が定常状態における下限の

$$1 + \frac{t_{\text{increase}}(n_{\max}, m)}{T} \tag{19}$$

倍のとき m 遷移クライアントに優先度を与えることができる．また，このとき定常クライアントがアンダーフローを起こす確率は定常状態においてバッファのデータ量が下限の場合よりも小さい．

本章のモデルにおける，遷移クライアントに高優先度を与えられる条件を**図6**に示す．横軸は遷移クライアントに高優先度を与えないときのサイクル長，縦軸はそのときの最大多重度において m 遷移クライアント（$m = 1,2,3$）に高優先度を与えるために必要な，定常クライアントのバッファの余分なデータ量を定常状態における下限との比で表したもの（$t_{\text{increase}}(n_{\max}, m)/T$）である．この図から $m \leq 3$ のとき遷移クライアントのアクセスを優先するのに必要となる余分なデータ量はきわめて少ないことが分かる．

8. おわりに

動画の再生開始等，素早い応答を必要とするデータ転送要求に優先度を与えることで，クライアントへの応答時間を短縮する動画データへの多重アクセス・スケジューリング方式VTSSについて述べた．本方式はSCAN方式を基本として，バッファに存在するデータ量を監視しつつ特定クライアントへのデータ転送を優先することで，バッファがアンダーフローを起こさないようにスケジューリングする．同時に，データ転送量をすべてのバッファのデータ量が平均化するように決定する．1クライアントに連続転送可能なデータ量はバッファに存在するデータ量に依存し，連続転送量が多いほどアクセス・オーバーヘッドが相対的に減少して効率が良いので，これは多重アクセス効率の向上に効果的である．

VTSSは，クライアントごとに要求される動画の品質やサイズが異なる場合や文字，数値データを同時にアクセスする場合にも対応できる．さらに，クライアントが動画の再生だけでなく録画も行う場合もVTSSは多重アクセスを制御可能である．

動画提供サービスではQoS（quality of service）の制御もまた重要な要素である．QoSを効果的に制御するには動画データのディスク上の配置，バッファのデータ量といった多重アクセス・スケジューラの持つ情報をオペレーティング・システムに通知する機能が追加される必要がある．これは早送り，巻き戻しといった特殊再生に対応するためにも重要である．

参考文献

1) Sincoskie, W.D.: Video on Demand: Is It Feasile?, *IEEE GLOBECOM '90*, pp.201–205, New York (1990).
2) Kotani, N., Kishigami, J., Sakurai, N. and Ishikawa, A.: MAMI: The New Direction of Interactive TV, *Proc. 15th PTC 1993*, pp.328–333, Honolulu (1993).
3) Miller, G., Baber, G. and Gilliland, M.: News On-Demand for Multimedia Networks, *ACM Multimedia 93*, pp.383–392, Anaheim (1993).
4) Steinmets, R.: Analyzing the Multimedia Operating System, *IEEE Multimedia*, Vol.2, No.1, pp.68–84 (1995).
5) Patterson, D.A., Gibson, G. and Kats, R.H.: A Case for Redundant Arrays of Inexpensive Disks (RAID), *ACM SIGMOD Conference*, Chicago (1988).
6) Ishikawa, A., Kishigami, J., Sakurai, N. and Kotani, N.: Multiple-Access Moving Picture Information System (MAMI), *IEEE GLOBECOM '92*, Vol.2, pp.759–763, Orlando (1992).
7) 高倉 健，櫻井紀彦，石川 篤：映像情報サーバにおける連続データ転送方式の評価，第49回情報処理学会全国大会，pp.4-121–4-122 (1994).
8) Tobagi, F.A., Pang, J., Baird, R. and Gang, M.: Streaming RAID - A Disk Array Management System for Video Files, *ACM Multimedia 93*, pp.393–400 (1993).
9) Teorey, T.J. and Pinkerton, T.B.: A Comparative Analysis of Disk Scheduling Policies,

Comm. ACM, Vol.15, No.3, pp.177–184 (1972).
10) Narasimha Reddy, A.L. and Wyllie, J.: Disk Scheduling in a Multimedia I/O System, *ACM Multimedia 93*, pp.225–233, Anaheim (1993).
11) Chen, M.-S., Kandlur, D.D. and Yu, P.S.: Optimization of the Grouped Sweeping Scheduling (GSS) with Heterogeneous Multimedia Streams, *ACM Multimedia 93*, pp.235–242, Anaheim (1993).
12) Fujii, H., Ishikawa, A., Kotani, N. and Sakurai, N.: Multimedia Server for On-Demand Services, *IS&T/SPIE symposium on electronic imaging science and technology*, San Jose (1994).
13) 阪本秀樹，西村一敏，中野博隆：ビデオ情報の大規模多重アクセス方式，電子情報通信学会論文誌，Vol.J78-D-II, No.1, pp.76–85 (1995).
14) 藤井 寛，石川 篤，櫻井紀彦：動画像情報への多重アクセススケジューリング方式とその評価，電子情報通信学会論文誌，Vol.J77-D-I, No.10, pp.729–736 (1994).
15) 梶谷浩一：動画サーバのためのディスクアレー管理法についての考察，電子情報通信学会論文誌，Vol.J77-D-I, No.1, pp.66–76 (1994).

付録 式(11)の導出

$\{L_1, L_1+1, \cdots, L_2\}$ (L_1, L_2 は整数) の中から重複を許して n 個とり，それらのうちの最大数と最小数の差が l となる順列の数を

$$ {}^{n}_{L_1,L_2}H_l \tag{20}$$

と書くとすると，n 個の値がすべて同じ値となるのは L 通りであるから，

$$ {}^{n}_{1,L}H_0 = L \tag{21}$$

また，

$$ {}^{n}_{1,L}H_l = {}^{n}_{1,l+1}H_l + {}^{n}_{2,l+2}H_l + \cdots + {}^{n}_{L-l,L}H_l. \tag{22}$$

すべての $k = 1, 2, \cdots, L-l$ に対して，

$$ {}^{n}_{k,k+l}H_l = {}^{n}_{1,l+1}H_l \tag{23}$$

であるから，

$$ {}^{n}_{1,L}H_l = (L-l){}^{n}_{1,l+1}H_l. \tag{24}$$

$\{1, 2, \cdots, L\}$ の中から重複を許して n 個とる順列は

$$ \sum_{k=0}^{L-l} {}^{n}_{1,L}H_k = L^n \tag{25}$$

通りである．よって，

$$ {}^{n}_{1,L}H_{L-1} = L^n - \sum_{k=0}^{L-2} {}^{n}_{1,L}H_k $$
$$ = L^n - \sum_{k=0}^{L-2} (L-K){}^{n}_{1,k+1}H_k \tag{26}$$

また，

$$ {}^{n}_{1,L-1}H_{L-2} = $$
$$ (L-1)^n - \sum_{k=0}^{L-3} (L-K-1){}^{n}_{1,k+1}H_k \tag{27}$$

式(26)，(27)より，

$$ \sum_{k=1}^{L} {}^{n}_{1,k}H_{k-1} $$
$$ = {}^{n}_{1,L}H_{L-1} + {}^{n}_{1,L-1}H_{L-2} + \sum_{k=0}^{L-3} {}^{n}_{1,k+1}H_k $$
$$ = L^n - (L-1)^n \tag{28}$$

が得られるから，

$$ {}^{n}_{1,L}H_{L-1} = \sum_{k=1}^{L} {}^{n}_{1,k}H_{k-1} - \sum_{k=1}^{L-1} {}^{n}_{1,k}H_{k-1} $$
$$ = L^n - 2(L-1)^n + (L-2)^n \tag{29}$$

式(24)より，

$$ {}^{n}_{1,L}H_l $$
$$ = (L-l)\{(l+1)^n - 2l^n + (l-1)^n)\}, \quad l \geq 1 \tag{30}$$

$\{1, 2, \cdots, L\}$ の中から重複を許して n 個とり，それらのうちの最大数と最小数の差が l 以下となる順列の数は

$$ \sum_{k=0}^{l} {}^{n}_{1,L}H_k = l^n + (L-l)\{(l+1)^n - l^n\} \tag{31}$$

$l = RL$ とおき，R を連続と見なすと，R の分布関数は

$$ F(R) = \lim_{L\to\infty} \frac{\sum_{k=0}^{l} {}^{n}_{1,L}H_k}{L^n} $$
$$ = -(n-1)R^n + nR^{n-1} \tag{32}$$

密度関数は

$$ f(R) = \frac{\partial F(R)}{\partial R} $$
$$ = n(n-1)(1-R)R^{n-2} \tag{33}$$

ここで，n を $n+1$ と置き換えると式(11)が得られる．

(平成7年8月25日受付)
(平成8年1月10日採録)

藤井 寛（正会員）

昭和41年生．平成元年京大・工・情報卒業．平成3年同大大学院修士課程修了．同年日本電信電話（株）入社．現在，動画像情報への多重アクセス方式の研究，映像情報流通方式の研究に従事．電子情報通信学会会員．

石川 篤（正会員）

昭和32年生．昭和55年日大・理工・電気工学卒業．同年日本電信電話公社（現NTT）入社．平成元年～2年（財）新世代コンピュータ技術開発機構へ出向．主に専用プロセッサ構成法の研究に従事．現在，映像情報サーバ構成法の研究に従事．テレビジョン学会会員．

櫻井 紀彦（正会員）

昭和31年生．昭和54年早大・理工・電気卒業．同年日本電信電話公社（現NTT）入社．以降，大型計算機の記憶階層構成の研究および記憶装置の開発に従事．電子情報通信学会会員．