マルチメディア通信と分散処理 75－11
（1996. 3. 1）

分散RAID型V.O.D.におけるデータ配置問題について

清水 洋 *1　中村俊一郎 *2　峯村治実 *2　山口智久 *2
渡辺尚 *1　水野忠則 *1

*1 静岡大学　　*2 三菱電機

ビデオストリームサーバの配信性能の向上を目指したRAID方式ビデオサーバの提案を行なう。RAID0, 4, 5型の各方式について、正常時、サーバ縮退時、ディスク縮退時のそれぞれについてシミュレーションを行なった。RAID0, 4型の2つの方式についてはほぼ期待通りの性能が観測できた。しかし、RAID5型については、ディスク縮退時の性能が期待に反したものであった。原因としてパリティセットに関する問題があることが分かった為、この問題についての1つの解法を示した。

Data Allocation Problem of Distributed RAID Style V.O.D

Hiroshi Shimizu*1
Shunichiro Nakamura*2　Harumi Minemura*2　Tomohisa Yamaguchi*2
Takashi Watanabe*1　Tadanori Mizuno*1

*1 Shizuoka University　　*2 Mitsubishi Electric Corp.

In this paper, we present a distributed RAID style video server that address the problem of increasing video stream supplying capability in VOD systems. We made simulation for a RAID0, 4 and 5 style video server at modes of normal, server degradation and disk degradation. A precise performance evaluation was made in RAID0 and 4 style. But that wasn't in a RAID5 style . The cause of this concerned a parity data allocation, and we present a solution about this problem.

1　はじめに

近年におけるマルチメディアブームの流れの中でV.O.D (Video on Demand) 等の実現を目的としたビデオサーバの研究開発が盛んになってきている。この背景としてはMPEG1、MPEG2といった映像／音声の圧縮技術の進歩、それらを処理するプロセッサが十分高性能になってきたこと、及びそれらを伝送するネットワークの高性能化が見えてきたこと等が挙げられる。

映像や音声を扱う場合には、当然実時間性が要求されるため、この点を重視して研究されているものが多い。それに対し、従来のファイルデータの転送と同様な形でデータ転送する方式のビデオサーバも提案されている。

本研究では、後者の方式のV.O.D.、つまり通常のサーバからイーサネットを介してクライアントにビデオストリームデータを供給するようなモデルを考え、コンピュータ上でシミュレーションを行ない、考察をする。

RAIDと呼ばれるディスク装置が、高速なデータ転送と耐故障性の向上を実現するものとして知られている。RAIDのコンセプトは、小型で安価なディスクを数台並べ、データを分散して記録し、並列にアクセスを行なうことである。並列化により、ディスク台数分のデータ転送速度の向上が見込まれることとなる。また、RAID3以上では、通常のデータの他にエラー訂正用コードを追加することにより耐故障性の向上も見込まれる。

本研究では、このRAIDと同様のアプローチをビデオサーバに対して行なう。つまり、RAIDでは複数のディスク装置を並べ並列にアクセスするのに対し、本研究のコンセプトは、安価なビデオ

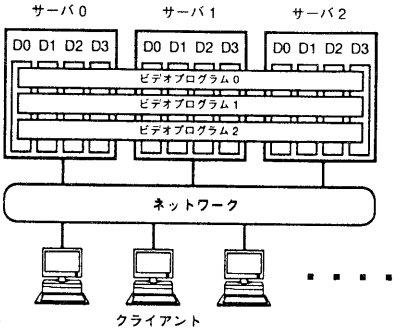

図 1: システムモデル

ストリームサーバを複数並べ、ビデオストリームデータを分散して記録し、これに並列にアクセスを行なうことである。これを「分散 RAID 方式」と呼ぶ [2]。

「分散 RAID 方式」は、並列アクセスによるビデオストリームデータ配信性能の向上と、また、RAID と同様のエラー訂正コードの追加により耐故障性の向上を見込める。

2 システムモデル

本研究で扱うシステムモデルを（図 1）に示す。複数の V.O.D. サーバが、複数のクライアントにイーサネットを介して接続される。サーバ、クライアント共に通常のパソコンを用いるものとする。それぞれのサーバは複数のディスク装置から構成される。ビデオストリームデータは分散 RAID 方式により、あらかじめ各サーバに格納されているものとする。

このモデルでは、各クライアント自身がデータの格納情報を持っており、必要とするデータブロックを格納しているサーバ、ディスクを求め、アクセスする形態をとる。ここでデータブロックとは、ビデオストリームデータを分割する単位のことである。

データの配置とシステムの故障に関する情報はあらかじめクライアントに知らされているものとする。

本研究では、システムが正常な場合の性能に加え、故障時の性能についても取り扱う。システムの故障には 2 つのモードを想定する。複数あるサーバのうちの 1 つが故障した状態を「サーバ縮退」、各ディスクのうちの 1 つが故障した状態を「ディスク縮退」と呼ぶ。

3 データ配置方法

本研究は、主に分散 RAID 方式 V.O.D. システムにおけるデータ配置方式についての問題を扱う。

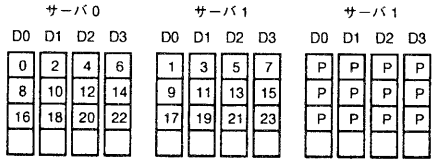

図 2: RAID4 型のデータ配置

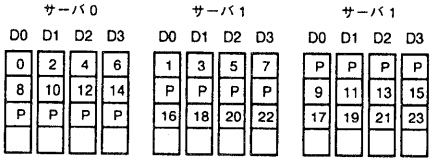

図 3: RAID5 型のデータ配置

データ配置方式は、システムの性能に大きな影響を与える。

本研究では以下の 3 つのデータ配置方式を考えている。便宜上 RAID0、4 、5 型と呼ぶが、厳密には通常の RAID とは異なるものである。ストライピングはすべてデータブロック単位である。

図の数字はデータブロックの番号、"P" はパリティを指す。すべてサーバ 3 台、サーバ当たりのディスク 4 台の場合の例である。

RAID0 型は、3 つの方式の中では最も単純な、ストライピングのみを行なう配置法である。データを分散することでの配信性能の向上は見込めるが、エラー回復コード（パリティ）をもたないため、耐故障性はない。

RAID4 型は、各サーバのうちの 1 つにパリティデータを持つ配置法である（図 2）。パリティを持つことで縮退時にもデータを回復することが可能であるが、パリティデータは故障が発生しない限りアクセスがない。つまり通常時は、RAID0 に比べ稼働サーバ数が減ることになり、配信性能は RAID0 に劣る。しかし同時に、正常時も縮退時も稼働クライアント数がかわらないため、常に一定の性能を確保できる利点もある。

RAID5 型は、RAID4 型では 1 つのサーバがパリティ専用サーバとなるのに対し、パリティをすべてのディスクに分散して格納する方式である。すべてのサーバ、ディスクが稼働するので、RAID4 型に比べ有利である（図 3）。

4 性能評価

人間が実際にビデオ映像を見た場合、コマ落ち率が 1 %を越えるとコマ落ちに気付くようになり、さらに 5 %では「かなり目立つ」といった印象で

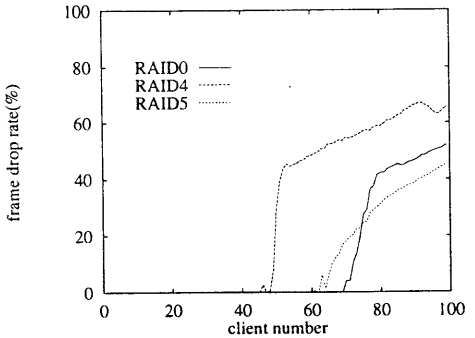

図 4: 正常時のコマ落ち率

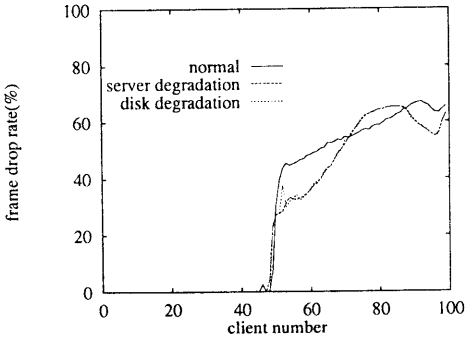

図 5: 性能比較： RAID4 型

データ配置方式	可能最大ストリーム数
RAID0 型	69.23
RAID4 型	45.38
RAID5 型	62.16

表 1: 正常時の可能最大ストリーム数

縮退モード	可能最大ストリーム数
正常	45.38
サーバ縮退	47.23
ディスク縮退	48.04

表 2: 性能比較： RAID4 型

ある。そこで以下ではコマ落ち率が 1 %に達した時点でのビデオストリーム供給数を可能最大ストリーム数とし、これを性能評価の基準とする。

正常時における各方式の性能比較を行なった結果を（図 4）、（表 1）に示す。グラフにおける縦軸はコマ落ち率（%）、横軸はビデオストリーム供給数である。RAID0, 4, 5 型の可能最大ストリーム数の比は 1 : 0.66 : 0.90 = 3 : 1.98 : 2.70 となり、稼働ディスク台数の比 12 : 8 : 12 = 3 : 2 : 3 とほぼ一致することから、それぞれ期待通りの性能であると言える。

次に、RAID4 型について、正常時、縮退時の性能比較を行なった。結果を（図 5）、（表 2）に示す。正常時、サーバ縮退時、ディスク縮退時の可能最大ストリーム数の比は 1 : 1.04 : 1.06 である。これも、RAID4 型の特徴である正常時も縮退時も性能が大きく変化しない、という特性と合致する。

最後に、RAID5 型について、正常時、縮退時の性能比較を行なった。結果を（図 6）、（表 3）に示す。可能最大クライアント数の比は 1 : 0.77 : 0.77 = 12 : 9.24 : 9.24 である。稼働ディスク台数の比は 12 : 8 : 11 であるので、結果はこれに比例していない。特にディスク縮退時については、可能最大ストリーム数と稼働ディスク台数の比は 9.24 : 11 と大きなものとなっている。この結果ついては以下の節で考察を加える。

5 ディスク縮退時の動きについて

我々は、正常時とディスク縮退時の性能比が 12 : 11 にならない原因として、ディスク縮退の場合、何らかの原因で 11 台のディスクが均等に稼働していない、つまり特定のディスクにアクセスが集中しているのではないかと考えた。どのディスクであっても、稼働率が 1 を越えるものが存在する場合には、データブロックの配送の遅れが生じ、結果としてコマ落ちを引き起こすことになると考えられるからである。そこでディスク縮退時の各ディスクの負荷を調べた（図 7）。

この図から、ある特定の 2 台のディスクに負荷が集中していることが分かる。つまり、この 2 つのディスクに格納されているデータブロックに関しては、クライアントへのデータの到着が著しく遅れることになり、結果としてシステム全体のデータ配信性能を引き下げていると考えられる。

特定のディスクに負荷が集中する原因は以下に述べる通りである。まず次のように定義する。

d_n : データブロック n
g : パリティグループの番号
p_g : パリティグループ g のパリティデータ
S : サーバ台数
D : ディスク台数

但し、$g = [\frac{n}{S-1}]$

パリティグループとは、あるパリティデータを作成する場合に XOR の要素とされるすべてのデータブロックと、そのパリティデータからなる集合

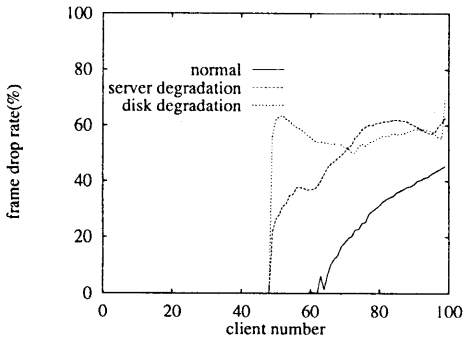

図 6: 性能比較： RAID5 型

データ配置方式	最大可能クライアント数
正常	62.16
サーバ縮退	48.04
ディスク縮退	48.02

表 3: 性能比較： RAID5 型

である。

ここでは本シミュレーションの仮定にならい、サーバ台数が 3 台、サーバ当たりのディスク台数 4 台の場合を考える。この場合、例えばデータブロック d_2, d_3 は共にパリティグループ 1 に属し、p_1 がこれに対応するパリティである。つまり、故障により d_2 が読み出せない場合には d_3 と p_1 を XOR することにより d_2 を復元する。正常時に d_2, d_3 を読み出す代りに故障時には d_3, p_1 の 2 つを読み出すということである。

RAID5 型では、そのデータ配置のアルゴリズムから、結果として、同じパリティグループに属するデータブロックは各サーバで同じ番号のディスクに格納される。この例ならば、d_2, d_3, p_1 はそれぞれサーバ 0, 1, 2 のディスク 1 に格納される。

つまり、正常時であればサーバ 0, 1 のディスク 1 に対してアクセスが行なわれるのに対し、サーバ 0 のディスク 1 が故障した場合にはサーバ 1, 2 のディスク 1 に対してアクセスが行なわれる。

パリティデータがどのサーバに格納されるかはそれぞれのパリティグループによって異なる。パリティが格納されるサーバを s_p とすると、s_p は以下のように決定される。

$$s_p = (g\%D)\%S$$

この例ならば、パリティグループ 0 ～ 3 ではサーバ 2 に、4 ～ 7 ではサーバ 1 に、8 ～ 11 ではサーバ 0 に格納される。

サーバ 0 のディスク 1 が故障した場合には、パリティがサーバ 0 に格納されるパリティグループ

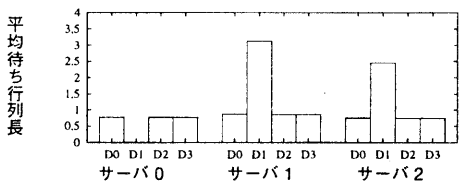

図 7: 各ディスクの負荷

については正常時と変わらずサーバ 1, 2 のディスク 1 へアクセスが行なわれ、その他のグループではそれぞれサーバ 1, 2 のディスク 1 へのアクセスが行なわれる。結果としてサーバ 0 のディスク 1 が故障した場合にはサーバ 1、2 のディスク 1 にアクセスが集中する。つまり、サーバ 0 のディスク 1 が故障した場合には、それに相当する負荷はサーバ 1, 2 のディスク 1 だけが負担することになる。

以上の現象は、RAID5 型の「同じパリティグループに属するデータブロックは各サーバで同じ位置のディスクに格納される」特性のために起きるものといえる。

6 パリティセットの分散

以上の考察から、RAID5 型においてディスク縮退時の可能最大ストリーム数を引き上げるためには、パリティセットを各サーバで異なる位置のディスクに格納する必要があると考えられる。ここで、パリティセットを各サーバで異なる位置のディスクに格納することを「パリティセットの分散」と呼ぶ。

以下で、パリティセット分散の一方式を提案し、その性能測定を行なう。

6.1 分散のアルゴリズム

本方式では、パリティセットの分散を、RAID5 型のデータ配置アルゴリズムを大きく変更することなく行なうために、データ配置アルゴリズムにさらに処理を付け加える形で分散を行なった。図は、すべてサーバ 3 台、サーバ当たりのディスク 4 台の場合の例である。

まず、各サーバ内のディスク群に着目する。すると、RAID5 型では必ず左から昇順にデータブロックが並ぶ（図 8）。本方式では、この各サーバ内のディスク群におけるデータブロックの並びにローテーション処理を加える。ローテーションの回数を r とすると、r は以下のようになる。

$$0 \leq r \leq D-1$$

つまり、全くローテーションをしないものから D-1 回のローテーションを行なうものまで、D 種類

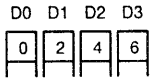

図 8: RAID5 型のディスク群

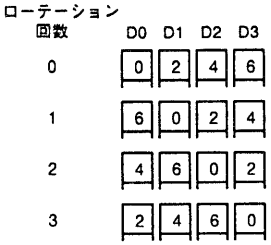

図 9: ローテーションパターン

の並びを作ることができる（図 9）。このローテーションパターンを各サーバで別々に設定する。

各サーバのローテーションパターンの組合せにより、縮退時のアクセスの仕方が決定される。組合せ数は D^S 個であり、本方式ではこれらを順に並べることでパリティグループの分散を行なった（図 10、11）。

6.2 性能比較

RAID5 型で、パリティグループの分散を行なうものの各ディスクの負荷を（図 12）、性能比較を（図 13）、（表 4）に示した。

今回のパリティセット分散の方式では、結果的に、ディスク故障に相当する負荷は、その故障ディスクを持つサーバ以外の 2 つのサーバで負担することになる。そのため正常時とディスク縮退時の性能比は 12 : 11 とはならない。正常時を $S*D$ とした場合の性能比は、ディスク 1 台分の仕事を、それを含まないサーバのディスクで負担すると考え、以下のようになる。

P：性能比
x：故障ディスクを含むサーバの稼働率
y：その他のサーバの稼働率

$$\begin{cases} P = x*(D-1)+y*(S-1)*D \\ y = x*(1+\frac{1}{(S-1)*D}) \end{cases}$$

上の節でも述べたように、どのディスクであっても、稼働率が 1 を越えるものが存在する場合には、データブロックの配送の遅れが生じ、結果として

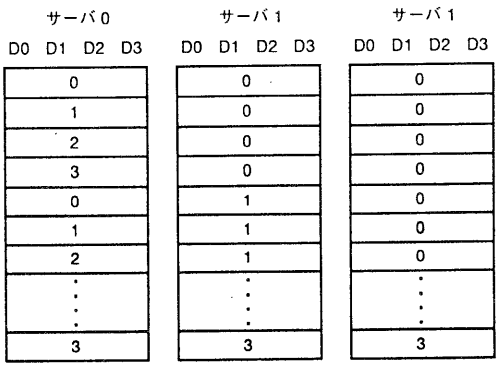

図 10: 本方式：ローテーションパターンの組合わせ

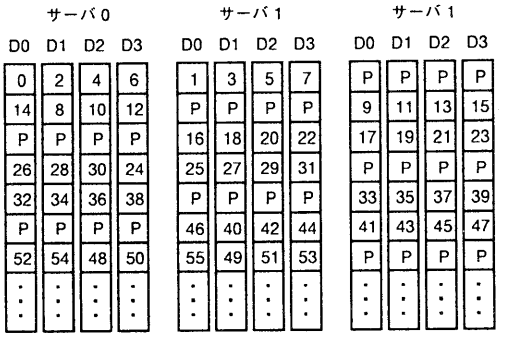

図 11: 本方式：データ配置

コマ落ちを引き起こすことになる。その意味において $\max(x,y)=1$ とおくと、

$$y = 1 = x*(1+\frac{1}{(S-1)*D})$$
$$x = \frac{(S-1)*D}{(S-1)*D+1}$$
$$P = x*(D-1)+y*(S-1)*D = \frac{(S-1)*S*D^2}{(S-1)*D+1}$$

つまり、正常時とディスク縮退時の性能比は $S*D : \frac{(S-1)*S*D^2}{(S-1)*D+1}$ となる。これに $S=3, D=4$ を代入すると $P=\frac{32}{3}=10.67$ となる。つまりこの方式での理論的な性能比は 12 : 8 : 10.67 となる。

シミュレーションの結果、可能最大ストリーム数の比は $1 : 0.78 : 0.85 = 12 : 9.36 : 10.2$ である。パリティセットを分散しない方式にくらべ、ディスク縮退の性能が向上していると言えるが、理論的な性能比には達していない。このことについては今後の研究の課題である。

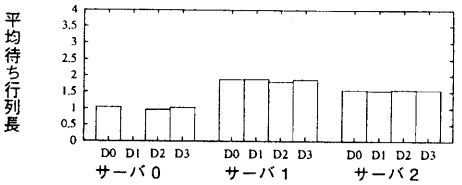

図 12: 分散型の各ディスクの負荷

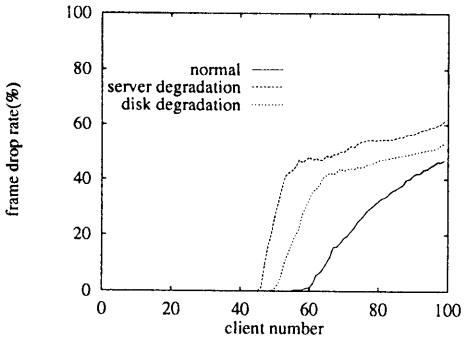

図 13: 性能比較：分散型

また、故障ディスクを含むサーバにも、故障に相当する負荷を分担させることが理想的ではあるが、そのためにはこれまでの分散 RAID 方式のデータブロック配置方式の範疇では不可能であり、新たな配置法が必要になると考えている。

7 むすび

本稿では分散 RAID 方式ビデオサーバの提案を行ない、RAID0、4、5 型の各方式の、正常時、サーバ縮退時、ディスク縮退時のそれぞれにおいてのシミュレーション評価を行なった。その結果、RAID0、4 方式では満足な評価が得られたが、RAID5 型については、特にディスク縮退時に期待した評価が得られなかった。原因として、これまでの分散 RAID 方式のデータ配置法では、パリティセットがそれぞれ特定のディスクに固定されるという問題があることが分かった。そこでこの問題を解決する 1 つの方法として、RAID5 型のデータブロック配置方式に、各サーバ内でローテーション処理を加える方法を提案した。評価の結果、理論値には達しないまでも、性能が向上することが観測された。

データ配置方式	可能最大ストリーム数
正常	58.87
サーバ縮退	45.75
ディスク縮退	49.93

表 4: 性能比較：分散型

参考文献

[1] 中村、山口、峯村、渡辺、水野：“ビデオストリーム配信性能の一検証”、情報処理学会研究報告 95-DPS-72,p.37-42,sep.1995.

[2] 中村、峯村、山口、清水、渡辺、水野：“分散 RAID 方式ビデオサーバー”、情報処理学会研究報告 95-DPS-72,p123-128,Dec.1995.

[3] 中村、峯村、山口、清水、渡辺、水野：“分散 RAID 方式ビデオサーバー（その 2）”、情報処理学会研究報告 96-DPS-74,p227-232,Jan.1996.

[4] Fouad A.Tobagi, Joseph Pang, Randall Baird, Mark Gang: “Streaming RAID - A Disk Array Management System For Video Files”, ACM Multimedia 93 Proceedings, 1993.8.1-6,p393-400.

[5] D.James Gemmuel, Harrick M.Vin, Dilip D.Kandlur, P.Venkat Rangan, Lawrence A.Rowe: “Multimedia Storage Servers:A Tutorial”, IEEE computer, May 1995, p40-49