マルチメディア通信と分散処理 77－6
（1996．7．12）

分散RAID型VODにおけるデータ配置問題について（その2）

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 静岡大学情報学部　　*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分散RAID方式ビデオサーバを提案する。この方式は、サーバ、ディスク台数に比例した性能向上が期待できるため、普及型の安価なサーバによるスケーラブルなシステムの構築が可能になる点が大きな特徴といえる。これまでの報告では、RAID5型データ配置方式について、各縮退モードの性能比較をする事で、ディスク台数にある程度比例した性能向上が観測されている。しかし、理論的な性能比からは若干のずれがあった。その原因の一つとして、各サーバへのアクセスの不均等が観測された為、各クライアントの要求開始時刻の許容時間について考察し、改良を行った。

Data Allocation Problem of Distributed RAID Style VOD(2)

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

*1 Faculty of Information, Shizuoka University　　*2 Mitsubishi Electric Corp.

In this paper, we present a distributed RAID style video server to increase stream supply in VOD systems. This method has a feature that linear performance improvement can be achieved for server/disk number and enables inexpensive servers to build scalable system. In the former report, we tested the system in both normal and degraded mode of RAID5 style data allocation method, and showed linear performance for the disk number. But that result is far from ideal one due to uneven access to disks. We discuss and improve admission interval of starting time of requirement.

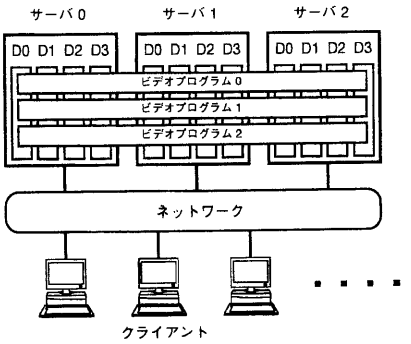

図 1: システムモデル

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、VOD(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。本研究では、通常のネットワークファイル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を供給するような VOD システムを取り扱う。

本研究で取り扱う、「分散 RAID 方式」[2] とは、通常ディスク装置に対して適用される RAID の手法を、複数サーバから構成されるシステムに拡張して適用するものである。つまり、RAID ではディスク装置を複数台並べるのに対して、この方式では安価なファイルサーバを複数台並べ、データを冗長性を持たせて分散配置し、並列に読み出しを行う。RAID と同様、並列アクセスによる高速化、冗長である事による耐故障性向上が期待できる。本研究では特に、分散 RAID 方式 VOD システムを、シミュレーションを通じて考察する。

2 シミュレーションモデル

本研究で扱うシステムモデルを図 1に示す。複数の VOD 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、クライアント共に広く普及している PC 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。

また、本稿におけるシミュレーションは以下の条件の下に行った。

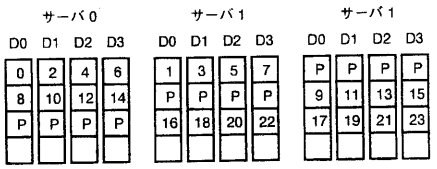

図 2: RAID5 型データ配置方式

サーバ数	3
サーバ当たりのディスク台数	4
データブロックサイズ	$64kb$
ビデオデータレート	$1.2Mbps$
ビデオデータサイズ	24000ブロック
ビデオプログラム本数	54

データブロックとは、ビデオストリームデータを分割して格納、配送する場合の単位である。

3 RAID5 型データ配置方式

本研究では、複数サーバにデータを格納する方法として、RAID5 型データ配置方式を用いる。RAID5 型データ配置方式とは図 2に示すようなデータ配置方式である。

それぞれのデータブロックは、サーバ 0,1,2,0... のように順番に分散して格納される。また、パリティデータは、各サーバに分散して格納されている。

この図の例では、例えばデータブロック 0,1 と対応するパリティの 3 つが組になっている。ここではそのような組のことをパリティグループと呼ぶ。ディスク、又はサーバに故障が起きた場合には、それぞれのパリティグループ内で計算を行う事でデータ回復が行われる。

本研究では、ディスク 1 台の故障とサーバ 1 台の故障をシミュレートする。それぞれをディスク縮退、サーバ縮退と呼ぶ。全てのサーバ、ディスクが正常に動作している場合を含めて、システムには 3 つの状態が存在する事になる。

RAID5 型のデータ配置方式には、ディスク縮退時の性能に問題があった。パリティグループの位置が固定されている為、あるディスクが故障すると、対応したの 2 台のディスクの負荷のみが増加するのである。そこで、[4] では、RAID5 型のデータ配置方法の改良を行った。

あるサーバ内のディスクに注目した場合、このような 4 つのデータの並びが存在する（図 3）。RAID5 型において、このデータの並びは、常にデータブロック番号が左から昇順に並ぶようになっている。このデータの並びを図 4のように 4 通り考えた。この 4 通りの並びを 3 台のサーバ間で組

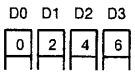

図 3: RAID5 型のディスク群

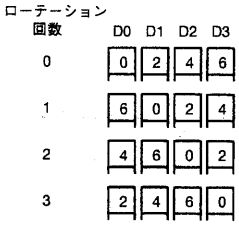

図 4: ローテーションパターン

み合わせる事で、パリティグループの固定を避け、RAID5 型の欠点であるディスク縮退時の性能を向上させる事ができる。RAID5 改良型のデータ配置方法を図 5に示す。

4 スケーラビリティ

本研究で提案する「分散 RAID 方式」では、大きな特徴として、サーバ、ディスク台数に比例した性能向上が可能な為、普及型の安価なサーバによるスケーラブルなシステムの構築が可能になる点が挙げられる。本研究のシミュレーションでは、システムはディスクネックになっているため、それぞれの縮退モードにおいては、稼働ディスク台数に比例した性能差が観測できるはずである。実際、[4] においては、正常、サーバ縮退、ディスク縮退の 3 つの縮退モードについて、その稼働ディスク台数にある程度比例したビデオストリーム配信性能が観測された。（図 6、表 1）。しかし、理論的な性能比からは若干のずれがあった。本稿では特に、この理論的な性能比を実現するために実験、考察を行う。

5 理論的な性能比

本稿では、理論的な性能比の実現を目的とする。しかし、本稿で採用する RAID 5 改良型のデータ配置方式 [4] においては、理論的な性能比は単純なディスクの台数比とは若干異なるものとなる。

この方式においては、ディスク故障に相当する負荷は、その故障ディスクを持たないサーバのみで負担することになる。このデータ配置方法では、パリティグループはそれぞれ異なるサーバに格納されており、故障ディスクの負荷を、同じサーバに含まれるディスクが受け持つ事は無いからである。

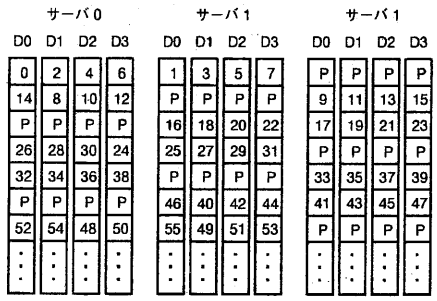

図 5: RAID5 改良型データ配置方式

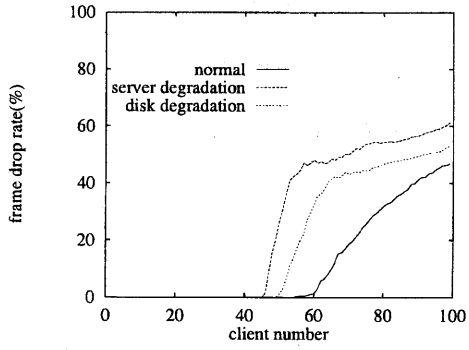

図 6: 性能比較：分散型

る。例えば、本シミュレーションにおいては、故障により増加した 1 台分の負荷を、残り 11 台のディスクで均等に負担する事は出来ず、故障ディスクを含まないサーバ内のディスク、つまり 8 台のディスクでのみ負担する事になる。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。

RAID 5 改良型のデータ配置方式における、各縮退モード間の理論的な性能比は以下の計算により求まる。

まず以下を定義する。

S：サーバ台数
D：サーバあたりのディスク台数
P：性能比率
x：故障ディスクを含むサーバ内のディスクの稼働率
y：その他のディスクの稼働率

正常時の性能を $S * D$ とした場合の性能比率 P 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

データ配置方式	ストリーム数	性能比
正常	58.87	12.00
サーバ縮退	45.75	10.2
ディスク縮退	49.93	9.36

表 1: 性能比較：分散型

る。

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

また、稼働率は1を越える事が無く、また$y > x$である事から$y=1$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 正常時：サーバ縮退時：ディスク縮退時で 12 : 8 : 10.67 となる。

前回までのシミュレーションによって観測された性能比は 1 : 0.78 : 0.85 = 12 : 9.36 : 10.2 であり、厳密には理論比に達していない。

6 アクセスの不均等

本研究のような分散システムにおいて、期待通りの性能が観測されない場合、考え得る最も大きな原因は、各ノードが均等に稼働していない事である。我々は、当システムにおいてもやはり同様の現象が起きていると考えた。そこで、各サーバ、ディスクへのアクセスの状況を観察する為、それぞれの平均待ち行列長を測定した。

6.1 ディスクへのアクセス

ここで問題とするアクセスの不均等には、クライアント数に関するものと、時間的なものの2つの問題があると考えた。具体的には、クライアント数がある程度以上の場合には、それ以下の場合と比べ過大なアクセスの集中がおきている、または、ある時刻を境に急激なアクセスの集中が起きているのではないか、という2つの問題である。そこで、我々は全ディスク平均待ち行列長、またその時間的推移を観察した。

図7では、特に急激なアクセスの集中、偏りは見られない。全ディスク平均待ち行列長が1を越

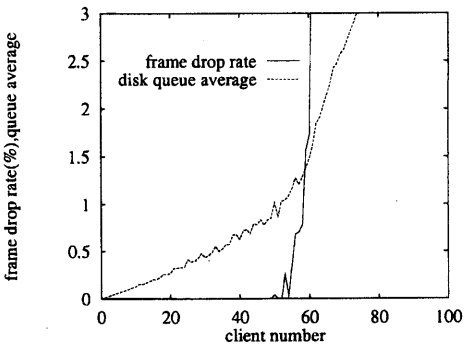

図 7: 全ディスク平均待ち行列長 対 コマ落ち率

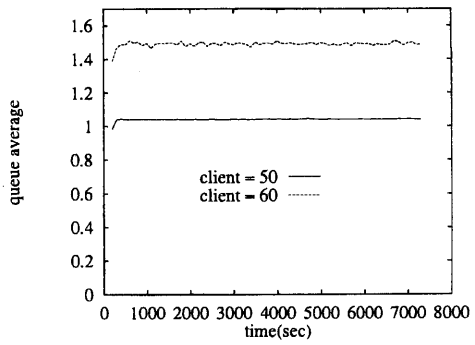

図 8: 全ディスク平均待ち行列長の時間的推移

えるクライアント数と、コマ落ち率の立ち上がるクライアント数が非常によく対応していることからも、妥当な結果であると考えられる。また、図8は、全ディスク平均待ち行列長が1を越える前後のクライアント数についての、アクセスの時間的推移を示している。このグラフからは、シミュレーション時間全体を通じて平均待ち行列長はほぼ一定のレベルに保たれ、雪崩現象に類する現象はおきていない事が分かる。ディスクへのアクセスに関しては、ここでは特に不均等の原因となる要素は観測できなかった。

6.2 サーバへのアクセス

つづいて同様に、サーバの平均待ち行列長と時間的推移について観測を行った。

図10については、ディスクの場合と同様、シミュレーション時間を通じて平均待ち行列長はほぼ一定のレベルに保たれている。しかし、コマ落ち率と平均待ち行列長に関しては、クライアント数50の場合、その他のクライアント数と比べ、

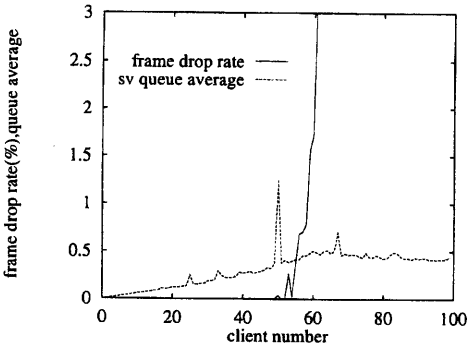

図 9: 全サーバ平均待ち行列長 対 コマ落ち率

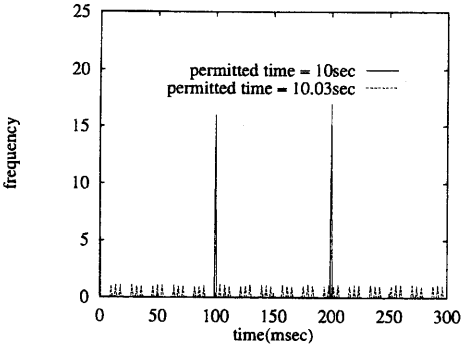

図 11: 要求開始時刻の分布

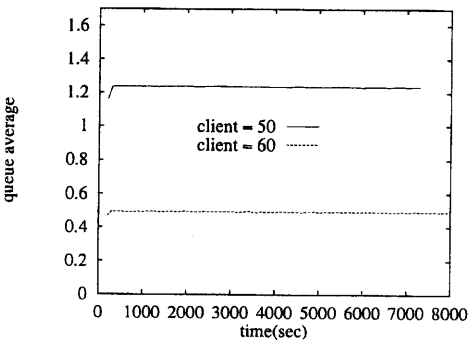

図 10: 全サーバ平均待ち行列長の時間的推移

アクセスが集中する現象が生じた（図 9）。

考えられる要因を以下に述べる。本研究のシミュレーションシステムでは、各クライアントの要求の開始時刻をずらすように設定している。これは、全てのクライアントが同時に要求を開始すると、特定のサーバ、ディスクに非常に偏ったアクセスが行われるので、これを避ける為である。具体的には、要求開始時刻のずれの許容時間をパラメータで設定し、それぞれのクライアントのプログラムが、許容時間内で均等に開始されるようになっている。つまり、各クライアントの要求開始時刻は、（許容時間 / クライアント数）づつずらされている。もちろん、この許容時間は大きければ大きい程システムには有利だが、余り大きくすると、各クライアントが同時にビデオ映像を放映する時間が短くなり、シミュレーションの意味が無くなる。現状では、この許容時間を 10 秒に設定している。クライアント数が 50 の場合では、それぞれの要求開始時刻は 10/50 秒 =0.2 秒づつずらされる事になる。

ここで問題となるのは、クライアントのアルゴリズムがある周期に従っている事である。つまり、データブロックのサイズが 64KB であり、これが 0.3 秒分の映像に相当する為、クライアントのアルゴリズムは 0.3 秒の周期に従う事になる。各クライアントの要求開始が 0.2 秒づつずらされた場合、クライアントのアルゴリズムの各周期で見ると、それぞれの周期の開始から 0 秒, 0.1 秒, 0.2 秒のそれぞれの時点にクライアントの要求が集中する形になる（図 11）。一般的には、（許容時間 / クライアント数）が、クライアントアルゴリズムの周期との間に、比較的小さな公倍数を持つ場合に、このような現象が起こると考えられる。

6.3 アクセス不均等の解決

この問題の解決の為には、要求開始の許容時間について、本来の意味と、クライアントのアルゴリズムの周期（0.3 秒）との 2 つの面について設定する必要がある。今回は、許容間隔を、本来の 10 秒に、クライアントアルゴリズムの 1 周期にあたる 0.3 秒を加えた設定として解決を試みた。変更前とくらべ、クライアントの要求が平均化されている事が分かる。（図 11）。

この改良の結果を図 12に示す。サーバの平均待ち行列は、かなり平均化されたといえる。しかし、その付近のコマ落ち率が、改良前に比べ振動している様子が観察できる。

また、各縮退モードでのコマ落ち率は（図 13、表 2）のようになった。このように、サーバ負荷はある程度平均化されたが、、コマ落ち率については予想されたような効果は上がらなかった。

サーバの平均待ち行列長がある程度平均化された事は確かに確認出来るので、その意味では各クライアントの要求開始時刻の許容時間の設定の変更は成功したと言える。しかし、それでも各サーバへのアクセスの不均等が起きている。また、この設定変更によって、変更前には見られなかったコマ落ち率の振動を引き起こしている。各クライ

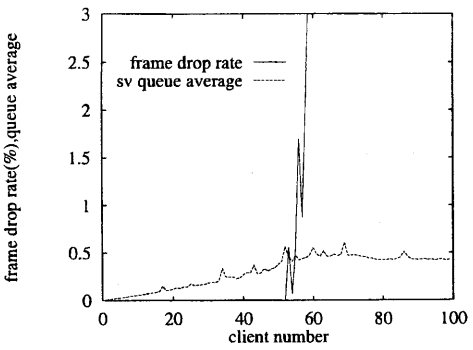

図 12: 全サーバ平均待ち行列長 対 コマ落ち率

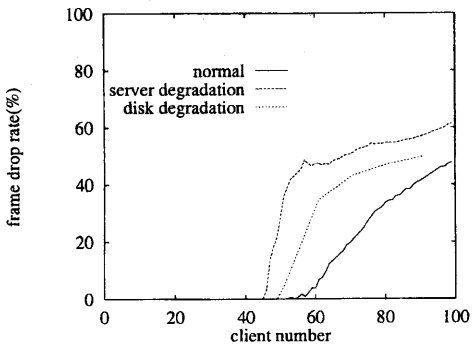

図 13: 性能比較：許容時間 =10.03sec

アントの要求開始時刻は平均化されているので、これらの現象の原因は、クライアントプログラムが開始された以降の部分にあると考えられる。1つには、各クライアントの要求開始時刻が理想的に平均化されている為、逆に、あるクライアントが何らかの原因によって、要求開始当初の周期からずれるような場合に対して、アクセスの不均等が連鎖的に起こりやすくなっている事が考えられる。また、このように考えた場合、各クライアントのアルゴリズムが、0.3 秒の周期に比較的正確に従う事自体が問題と考える事も出来る。

この点については、要求開始時間間隔による影響をあまり受けない事を目指したシステムを考案中である。現状では、各クライアントは、ビデオストリームデータの各データブロックについて、あくまで正確な時刻に映像出力を行おうとする。しかし、各データブロックの出力時刻がある程度ずれても、実際に人間が映像を通して見る場合においては、さほどの問題は起こらないとの考え方も出来る。そこで、各クライアントがそれぞれのデータブロックについて、ある程度の出力時間のずれを許容できるようなアルゴリズムを開発中である。このアルゴリズムにより、本稿で観測されたようなアクセスの集中は起こりにくくなると考えている。

データ配置方式	ストリーム数	性能比
正常	55.22	12.00
サーバ縮退	45.28	9.84
ディスク縮退	49.60	10.78

表 2: 性能比較：許容時間 =10.03sec

7 むすび

本稿では、シミュレーションで観測されたサーバ負荷の不均等をなくすために、各クライアントの要求開始時刻の許容時間の設定を変更して実験をおこなった。結果、サーバ負荷の不均等は軽減されたが、システムの性能向上には結び付かなかった。このことは、各クライアントのアルゴリズムに1つの原因があるとも考えられる。そのため、現在新たなクライアントのアルゴリズムを開発中である。

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] 清水、中村、峯村、山口、渡辺、水野：“分散 RAID 型 V.O.D. におけるデータ配置問題について”、情報処理学会研究報告 96-DPS-75,p61-66,Mar.1996.

[5] 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.

[6] 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