動画配信における負荷分散システムの構築

水越一貴[1] 羽田友和[2] 林本雅之[3] 八代一浩[4] 安藤英俊[1]

[1]山梨大学大学院医学工学総合教育部 [2]株式会社YSK e-com
[3]株式会社日本ネットワークサービス [4]山梨県立大学

概 要

ネットワークインフラの整備が進み，ブロードバンド環境が広く普及したことによってイベント中継が盛んに行われるようになってきている．現地へ行かなくてもイベントを見ることができることから，イベント中継に対する需要は高い．イベント中継を行う場合，配信する側はアクセス集中に備えて複数の配信サーバを用意することになる．地域ネットワークにおいて配信を行う場合，配信サーバは必ずしも同じ場所に設置されるとは限らない．また，配信サーバの性能が同じではない場合もある．このような状況で効率的に配信を行うためには，単純に負荷を分散させるのではなく各配信サーバの負荷に応じた配信量になるように負荷を分散させることが望ましい．本研究では，各配信サーバの負荷を考慮してユーザのアクセスを分散させるシステムを構築した．この負荷分散システムをイベント中継に用いて，その有効性を検討する．

Construction of the Load Balancing System on Media Streaming

Kazutaka MIZUKOSHI[†], Tomokazu HADA[‡],Masayuki HAYASHIMOTO[††]
Kazuhiro YATSUSHIRO[‡‡]and Hidetoshi ANDO[†]

[†]Yamanashi University, [‡]YSK e-com Corporation
[††]NNS Corporation, [‡‡]Yamanashi Prefectural University

Abstract

Maintenance of a network infrastructure progresses, and when broadband spread widely, event relaying is performed increase briskly. Since an event can be seen even if it cannot go to a spot, the demand over such relay is high. When performing such event relaying, the side to distribute will prepare two or more distribution servers in preparation for access concentration. When distributing in a regional network, a distribution server is not necessarily installed in the same place. Moreover, the spec. of a distribution server may not be the same. In order to distribute efficiently in such a situation, it is desirable to distribute load so that load may not be distributed simply and it may become the amount of distribution according to each distribution server. In this research, the system which distributes a user's access in consideration of the load of each distribution server was built. This load balancing system is used for event relaying, and that validity is examined.

1. はじめに

近年，日食中継や祭中継，国体中継，花火中継等多くのイベント中継が各地で行われている．このようなイベント中継は，需要が高くユーザからのアクセスが集中するために複数の配信サーバを設置することが多い．

中継を行う場合，同一地点に配信サーバを設置し，広帯域の回線を用意するのがこれまで一般的だった．しかし，この手法では，配信サーバの設置位置までのネットワーク状態によっては安定的に配信することが難しい．また，広帯域の回線を用意するにしても動画の使用する帯域とユーザ数を考えると限界が生じる．

最近では，ネットワーク上に配信サーバを分散して配置し，ユーザから近い配信サーバへ誘導する広域負荷分散が行われてきている．しかし，配信サーバを分散させて配置することから配信地点ごとに回線帯域が異なったり，配置された配信サーバの性能が必ずしも均一にならなかったりすることがあ

る．その場合，ある配信サーバではユーザアクセスを捌けても違う配信サーバでは捌けなくなる可能性がある．

これらの問題を改善し，地域ネットワークを使ってイベント配信を行うために，本研究ではサーバの負荷を考慮した負荷分散システムを提案する．SNMP（Simple Network Management Protocol）を使って配信サーバの負荷を計測することにより，配信サーバの性能に応じた負荷分散を行うことができる．また，トラフィック量をみることで，使用できる回線帯域を考慮することもできる．

本研究では，動画配信を行う際に有効な負荷分散手法を検討し，その結果から負荷分散システムを構築した．2章では，サーバ負荷分散手法について検討する．3章では，負荷分散システムの構築について述べる．4章では，ライブ中継に構築した負荷分散システムを用いた結果を示す．最後に5章でまとめと今後の課題について述べる．

2. サーバ負荷分散技術

負荷分散技術は，トラフィックを振り分けたり，サーバの負荷に応じて割り振るサーバを変更したりする技術である．その手法としては，DNSによる負荷分散，スイッチによる負荷分散に分別することができる．

2－1．DNSによる負荷分散

DNSによる負荷分散は古典的手法であり，いろいろな場面で使用されている．

・DNSラウンドロビン

ドメイン名に対して複数のIPアドレスを割り当ててDNSサーバにアクセスしてきたユーザに対して順番にIPアドレスを返す手法である．よって得られるIPアドレスは均等になり，均一な負荷分散となる．また，サーバの状態によらず割り当てられるので必ずしも最適な分散が行われるとは言えない．

・TENBIN

TENBIN[2]はDNSを基盤としたサーバ選択システムである．サーバ群の情報を保持し，サーバ群を定期的に調査するシステムからの情報を使ってクライアントからの要求に答える．サーバ群を定期的に調査するシステムは，BGP4[4]の経路情報を集め，それを用いてTENBINがクライアントをネットワーク的に近いサーバへ誘導する．この方式では，負荷分散に経路情報を使っており，サーバの負荷が考慮されていない．

・DNSフィルタ方式

DNSフィルタ方式[5]はクライアントからDNSサーバへの問い合わせを監視する．問い合わせに対する答えが複数ある場合，それらのホストに対して調査を行い，次のクライアントからの問い合わせには最適なホストに接続させる．この方式では，クライアントから見て最も近いサーバに接続できるが，システムがクライアントとローカルDNSの間になくてはならない．

2－2．スイッチによる負荷分散

スイッチによる負荷分散では主にNAT[1]によるアドレス変換やMAT[2]による直接サーバ返答が使われる．

NATでは，サーバ群に対して仮想IPアドレスをつけて負荷分散装置がトラフィックを受け，仮想IPアドレスへの変換を行う．サーバからの返答も負荷分散装置でIPアドレス変換を行いユーザへ送られる．サーバ群の中で負荷が最小のサーバへ振り分ける機能をつけることも出来る．[6]

MATでは，すべてのサーバのループバックインタフェースに同一の仮想IPを付加する．負荷分散装置にはサーバ群のMACアドレステーブルを持たせておき，ユーザアクセスをMACアドレスによってサーバに割り当てる．サーバからの返答には負荷分散装置を介さないで直接ユーザに返答することが可能になる．

これらの方式では，負荷分散装置においてアドレス変換を行う．そのため動画のような実時間で大量のデータが流れる状態では変換による遅延が発生することが考えられる．

3. 負荷分散システムの構築

3－1．システム要求

動画配信において安定で効率的な負荷分散を実現するために以下の条件を満たすようにする．

- 配信サーバごとに利用可能帯域と最大接続数の設定ができるようにする
- 配信サーバの負荷状態を考慮するために，SNMPで配信サーバの負荷状況を取得する
- ユーザはWebページのリンクによって配信サーバにアクセスできるようにする

3－2．構成

負荷分散システムの構成を図1に示す．システムは配信サーバの負荷状態の情報を取得し，分析するプログラムであるLAFS（Load Average Forecast System）とユーザからのアクセスを捌くCGIプログラムであるHADA（High Available Dynamic Access-control）の2つの部分からなる．

---

[1] Network Address Translation
[2] Mac Address Translation

3－3．LAFS

LAFS は SNMP を用いて配信サーバの負荷状態を分析するプログラムである．プログラムは実行速度を考慮して C 言語で記述を行った．LAFS は MAIN 部，FORECAST 部，SNMPGET 部の3つから構成される．(図2)

3－3－1．MAIN 部

MAIN 部では，設定ファイルを読み込み，設定に従ってプロセスを作成する．fork を用いて SNMPGET 部と FORECAST 部のプロセスを作成する．また，2 つの部でデータをやり取りするための共有メモリを作成する．SNMPGET 部については，測定する配信サーバ台数分だけプロセスを作成する．MAIN 部は毎秒プロセス作成を行い，それ以外は sleep で待機する．

3－3－2．SNMPGET 部

SNMPGET 部は，MAIN 部によって配信サーバの台数分だけプロセスが作成される．MAIN 部から受け取った設定にある MIB[3]情報を配信サーバから SNMPGET Request を用いて取得する．取得したデータを共有メモリに書き込んだ後に終了する．

3－3－3．FORECAST 部

FORECAST 部は，MAIN 部によって毎秒 1 プロセスが作成される．共有メモリから SNMPGET 部が取得したデータを取り出し，予測トラフィックと負荷状況を計算する．その結果を HADA に UDP を使って送信した後に終了する．

3－3－4．配信サーバから取得するデータ

配信サーバの負荷状況やトラフィックを計算するために次の MIB 情報を取得する．

出力バイト数　ifOutOctets
1. 3. 6. 1. 2. 1. 2. 2. 1. 16
TCP コネクション数　tcpCurrEstab
1. 3. 6. 1. 2. 1. 6. 9
CPU 負荷率　hrProcessorLoad
1. 3. 6. 1. 2. 1. 25. 3. 3. 1. 2
メモリ全体量　hrStorageSize
1. 3. 6. 1. 2. 1. 25. 2. 3. 1. 5
メモリ使用量　hrStorageUsed
1. 3. 6. 1. 2. 1. 25. 2. 3. 1. 6

3－3－5．HADA へ送信するデータ

送信するデータを図3に示す．図3の塊が配信サーバ1台分のデータになり，これを配信サーバの台数分だけリストにする．

リストの順番は負荷状況の低い順とする．取得したデータから計算した負荷状況を示す Weight が 0 以下の配信サーバは，ネットワークまたはサーバ自身に障害があるとして送信対象としない．よってリストは配信サーバ全台数分にならないこともある．

[3] Management Information Base

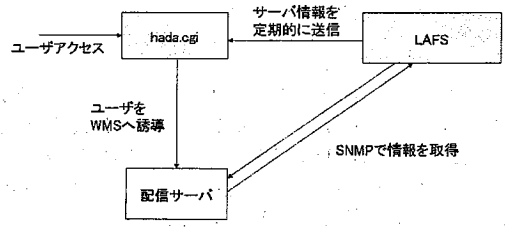

図1．負荷分散システムの構成

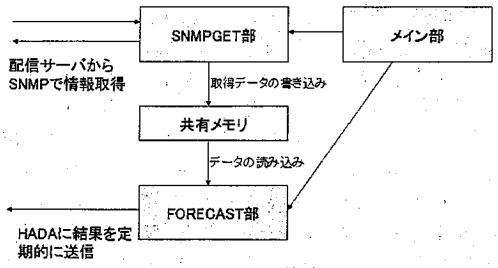

図2．LAFS の構成

配信サーバの IPアドレス (32bits)	配信サーバの netmask (32bits)	配信サーバの 予測Traffic (32bits)	配信サーバの connection数 (32bits)

図3．送信するデータ

3－3－6．負荷の計算

・予測トラフィックの計算
現在のトラフィックと TCP コネクション数から次の接続を受けるとトラフィックがどのように変化するかを予測する．
OO＝ 新しく取得した ifOutOctets
OO'＝ 前回に取得した ifOutOctets
CE＝tcpCurrEstab
ET ＝ 前回取得してから新しく取得できるまでの経過時間(秒)
・秒単位の平均トラフィック
TA＝(OO - OO') / ET
・予測トラフィック
FT＝TA＋(TA / CE)

・Weight の計算
各配信サーバの負荷状況 (Weight) を数値化する．
FT＝ 予測トラフィック
PL＝hrProcessorLoad
SS＝hrStorageSize
SU＝hrStorageUsed
Weight＝FT＊(PL＊2＋(SU / SS＊100))
メモリよりも CPU 負荷に重みを置くため PL を2倍にしている．
この Weight によって HADA に送信するデータの

配信サーバ優先順位を決める.

3－4．HADA

HADA は，ユーザからのアクセスを受ける CGI プログラムである．一般的に CGI で使われる Perl で記述した．HADA は LAFS から定期的に送られてくる各サーバの優先度と予測トラフィック，コネクション数を基にユーザを配信サーバに振り分ける．各配信サーバの使用可能帯域と接続数上限の情報を使い，負荷が低い配信サーバでも使わないようにすることができる．

HADA のフローチャートを図4に示す．HADA は，LAFS からサーバの優先度とそれぞれのサーバの予測トラフィック及び TCP コネクション数を毎秒受け取る．ユーザアクセスがあった時点でもっとも優先度が高いサーバの予測トラフィックと利用可能帯域，TCP コネクション数と接続数上限をそれぞれ比較し，上回っていなければそのサーバに配信を割り当てる．割り当てられた配信サーバへ誘導できるように Web ページを生成する．もし，上回っていた場合は次に優先度の高いサーバについて同様の比較を行う．

4．　負荷分散システムの運用

構築した負荷分散システムを山梨県市川大門町で行われた第 17 回神明の花火大会のインターネットライブ中継において使用した．

この中継は，山梨県地域情報ネットワーク相互接続機構[4]によって行われたものである．

2005 年 8 月 7 日 19 時 30 分から 21 時の間中継が行われた．中継では，Microsoft 社が提供している Windows Media[5]を使い行った．

4－1．構成

全体のネットワーク構成を図5に示す．撮影した映像は WME（Windows Media Encoder）でエンコードしてから無線 LAN にて中継地点まで送信する．中継地点からは ADSL 回線を使い山梨県の商用地域 IX である BeX-J[6]に配置した WMS-D へ送信する．WMS-D は配信元になる WMS（Windows Media Server）であり，再配信用の WMS へのみ配信を行う．ユーザからのアクセスは負荷分散サーバを経由して再配信サーバが受ける．

WMS からの配信はすべてユニキャストで行った．また，mms，http，rtsp によるアクセスを有効にし，配信レートは 500kbps とした．配信サーバとして合計 6 台の WMS サーバを用意した．6 台のサーバのうち WMS 1 から WMS 5 までを地域 IX に接続している ISP 等に分散配置した．WMS 6 は一般家庭内に配置し，B フレッツによりアクセスできるようにした．それぞれのサーバのスペックは表1のようにそれぞれ異なっている．

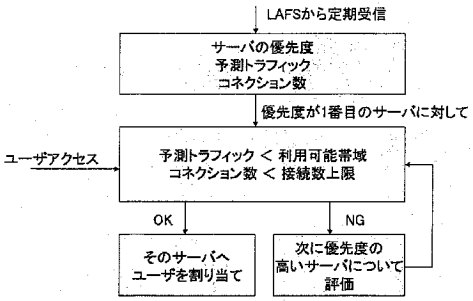

図4．HADA のフローチャート

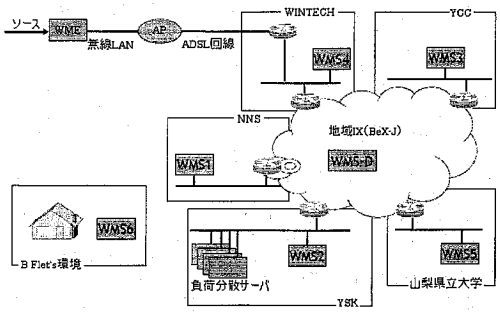

図5．全体のネットワーク構成

配信のための負荷分散システムは図6のように構成した．負荷分散サーバは4台用意し，それぞれに HADA と LAFS を設定した．　負荷分散サーバへのアクセスは DNS ラウンドロビンを使って分散するようにした．

負荷分散サーバは 6 台の WMS に対して LAFS を使ってデータ収集をする．負荷分散サーバでは，LAFS からのデータに基づき HADA で作成した Web 画面によって各 WMS へ導く．その際，HADA は設定された各 WMS に対する利用可能帯域と接続上限を考慮する．（表2）その後は配信サーバが直接ユーザに配信を行う．

表1．各サーバのスペック

サーバ	CPU	Memory	HDD
負荷分散サーバ1	P3 Xeon 600MHz	512MB	40GB
負荷分散サーバ2	P3 800MHz	128MB	17GB
負荷分散サーバ3	P3 800MHz	256MB	10GB
負荷分散サーバ4	Celeron 700MHz	512MB	28GB
WME	Athlon XP 2800+	1.0GB	40GB
WMS-D	P3 1.0GHz	1.0GB	40GB
WMS1	Xeon 3.2GHz	1.0GB	40GB
WMS2	P3 1GHz	1.37GB	17GB
WMS3	P3 1GHz	512MB	10GB
WMS4	Xeon 2.4GHz(x2)	1.5GB	80GBx2
WMS5	Celeron 2.6GHz	1.0GB	20GB
WMS6	P4 2.66GHz	512MB	40GB

[4] N@VEL http://www.navel-y.jp/
[5] http://www.microsoft.com/japan/windows/windowsmedia/
[6] http://www.bex-j.net/

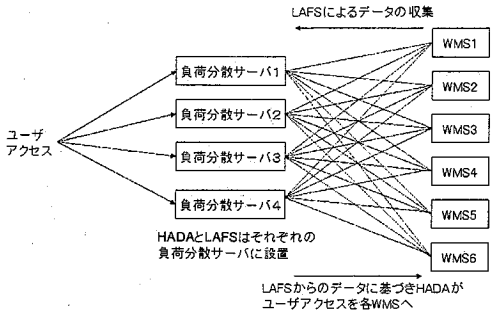

図6. 負荷分散システムの構成

4－2．結果と考察

花火大会が始まる30分前の午後7時から終了する9時すぎまで構築した負荷分散システムを用いて動画の配信を行った．開始時間の午後7時30分には，多くのアクセスが集まった．しかし，HADAの設定ミスにより，LAFSからのサーバリストの優先度が1位のものしか使わないようになっていた．そのため3分の1程度しかユーザアクセスを捌くことができなかった．午後8時ごろ気がつき設定を直した後は，アクセスしてきたユーザをすべて捌くことができるようになった．（図7）

表2. 各WMSの利用可能帯域と接続上限

配信サーバ	利用可能帯域	接続数上限
WMS1	15Mbps	27
WMS2	80Mbps	142
WMS3	25Mbps	42
WMS4	40Mbps	72
WMS5	80Mbps	142
WMS6	設定なし	22

各サーバのトラフィック（図8）を見ると配信レートである500kbps×コネクション数よりもはるかに少ない量しか流れていない．これには，WMSサーバがユーザへ配信する際に自動で配信レートを落としていることが考えられる．しかし，WMS6では想定トラフィック通りの値が計測されている．このことからもネットワークの状態によって配信レートが変動していると考えられる．また，WMS2とWMS6のトラフィックが途中から同値なのはLAFSプログラムの変数処理部分で桁溢れを起こしていたのが原因だった．

コネクション数（図9）においてWMS1が常時50以上のコネクションを張っていることがわかる．WMS1には配信コネクション以外の接続があったためこのようになってしまった．tcpCurrEstabを

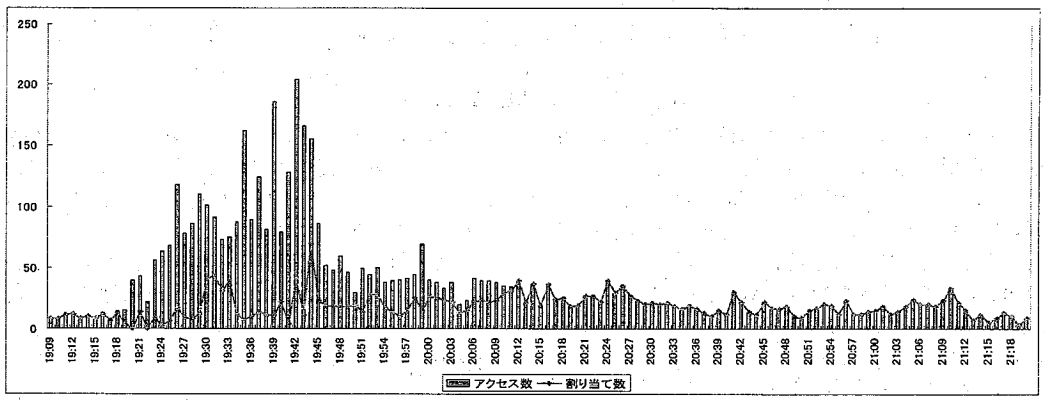

図7. HADAによるユーザアクセスの割り当て数

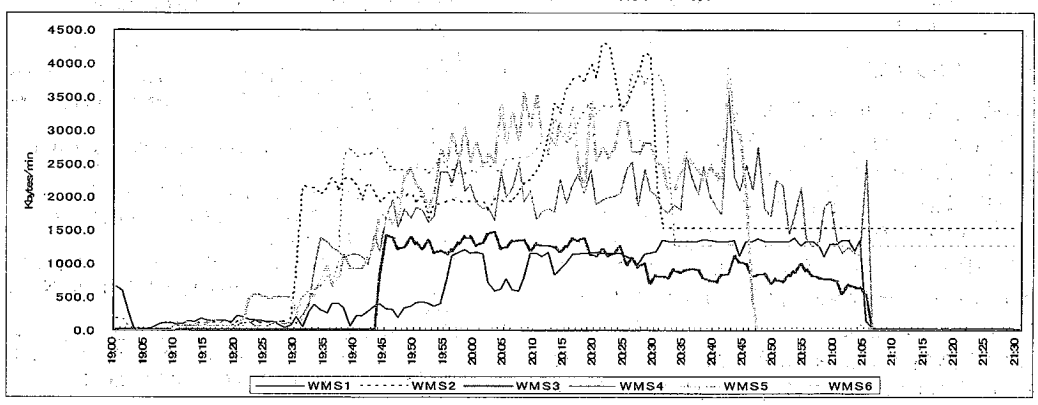

図8. LAFSによる各WMSのトラフィック

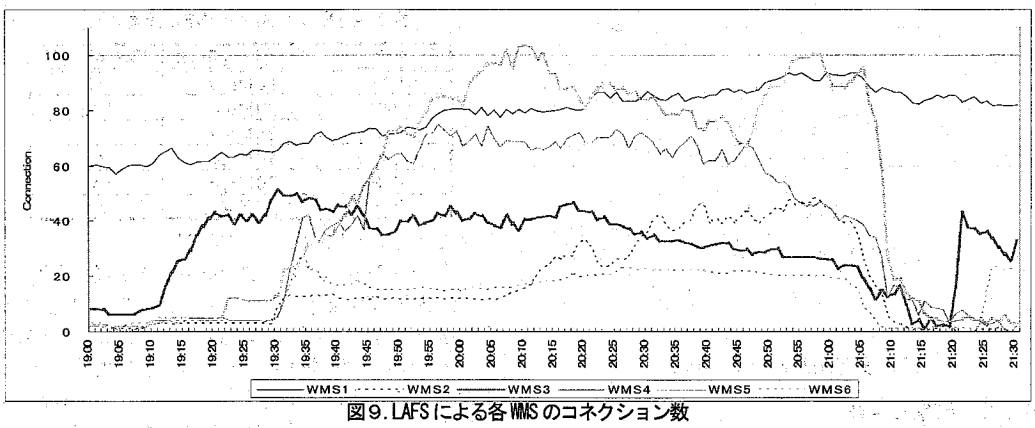

図９. LAFS による各 WMS のコネクション数

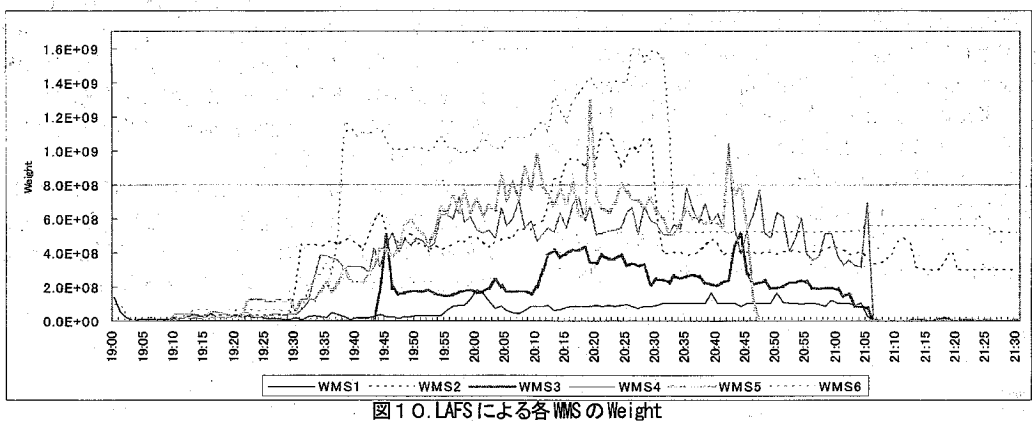

図１０. LAFS による各 WMS の Weight

用いるとコネクション数を WMS 全体で見てしまうため配信以外の接続との判別がつかない.

B フレッツ接続の WMS６では，接続数上限まで捌くことが可能であった．負荷状況や利用帯域から接続数を多くしても対応可能だと考えられる.

配信サーバの優先度を示す Weight では，優先度が高くなっている部分でその WMS のコネクション数が増加している傾向が見える．（図１０）しかし，実際の割り当てでは HADA によって利用可能帯域と接続上限が考慮されるので必ずしも優先度順になっていない.

５．　おわりに

本研究では複数の動画配信サーバを用いて配信する際にサーバ負荷，利用可能帯域を考慮してアクセスを分散させるための負荷分散サーバを構築した．また実際にこの負荷分散システムを使用してライブ中継を行った．前半のもっともアクセスが集中した時間において設定ミスによりうまく動作しなかったが，その後の時間では配信サーバに対する分散ができたといえる.

今後の課題として，配信の開始時と割り当てる時をより近づける仕組みを考える必要がある．今回の手法ではユーザにアクセスする配信サーバを考慮してから Web 画面を生成している．しかし，アクセスしてくるユーザは必ずしもすぐに配信サーバへアクセスするとは限らない.

また，配信サーバの負荷がどの変数に依存するかをさらに検討することで，負荷分散の精度をより高くすることができると考えられる.

参考文献

[1] 馬場始三・山口英：DNS を用いた広域負荷分散の実装，情報処理学会研究報告 1998-DSM-9，Vol.1998，No.36，pp.37-42（1998.5）.

[2] http://www.tenbin.org/

[3] 安田豊・中山雅哉：日蝕中継における WWW 分散サーバ群の構築とその有効性，情報処理学会研究報告 1999-DSM-14，Vol1999，No.56，pp.19-24（1999.7）

[4] Rekhter, Y. and Li, T.: A Border Gateway Protocol 4 (BGP-4), RFC1771 (1995).

[5] 横田裕恩・木村成伴・海老原義彦：DNS フィルタ方式によるミラーサーバ選択法の提案と実装，情報処理学会論文誌 Vol.44，No.3，pp.682-691（2003.3）.

[6] 井上博之・山口英：NAT による WWW サーバの負荷分散機構の実装，情報処理学会研究報告 1996-DPS-78，Vol1996，No.95，pp.19-24（1996.9）

[7] 水越一貴・牧野晋・林英輔：通信トラフィック監視システムの試作とバーストラフィックの検出，情報処理学会研究報告 2004-DSM-34，Vol.2004，No.77，pp.31-36（2004.7）.