マルチメディア通信と分散処理　94－4
分散システム/インターネット運用技術　14－4
（1999．7．15）

日蝕中継におけるWWW分散サーバ群の構築とその有効性

安田 豊　　中山 雅哉
神戸大学 経済経営研究所　　東京大学 情報基盤センター

インターネットにおけるネットワークアプリケーションは一般にサービス能力のスケーラビリティを要求される。分散化はスケーラビリティを確保するための一つの方法であるが、本稿では1997年以来幾度か日蝕インターネット中継のためのWWW分散サーバ群を構築・運用してきたので、まずこれについて報告する。次に1999年2月の日蝕で試みた各分散サーバの処理分担率を動的に調整し得る手法を示し、アクセスログの分析結果からその有用性を検証する。

Effectiveness of Distributed WWW Servers for Solar Eclipse Live

Yutaka Yasuda　　Masaya Nakayama
Research Institute for Economics & Business Administration of Kobe University　　Information Technology Center, the University of Tokyo

Internet application will be required a scalability of it's service performance. Distribution is the one of technique to keep the scalability. This paper shows some distributed WWW servers system case for solar eclipse live from 1997. And the advantage of it is shown by the analyzing the access log of of Feb 1999 eclipse live. It contains the effectiveness evaluation of dynamic adjustment a rate of sharing.

1　はじめに

LIVE! ECLIPSE (以下 LE) は1997年以来、数度の日蝕インターネット中継を実現している非営利団体であり、その活動は主として日蝕の動画像を撮影し、インターネット利用者に生中継することにある。映像の提供には RealVideo などその時点で適用可能なさまざまな技術が試みられてきた。しかしどのような動画中継を行なった場合であっても、全てのユーザにとってアクセスの入口は WWW で提供する情報が元になるため、まず大量のWWWアクセスを処理する必要が生じている。このため、LEでは1997年の最初のインターネット中継から WWW にはミラーサーバを用意して処理の分散化をはかってきた。

毎回の中継で差はあるものの、LEによるインターネット中継に関連する WWW アクセスは極めて短時間に集中する傾向にある。一定のピーク時間を過ぎるとトラフィックは急激に下がり、1台のサーバでの情報提供で十分な量となる。このように、LE 中継の期間のみ実験的に WWW 分散サーバシステムを構築して、正しく機能する事を検証するのに非常に適した題材であると言える。

本稿では、まず、これまで行ってきた日蝕中継の概要と日蝕中継における WWW アクセスの傾向について示す。続いて1999年2月のオーストラリア日蝕で構築した WWW 分散サーバ群の構成と採用した分散方式について概説し、適用結果の分析と評価を行う。そして、最後にこのとき得られたデータを元に WWW 分散サーバ群の有効性について考察する。

2 過去のインターネット中継の概要

LEが行ってきた日蝕のインターネット中継は下記のとおりで、本年2月までに4回行われてきた。

LIVE! ECLIPSE 97 (LE97)
シベリア・モンゴル皆既日食 - 1997.3.9

シベリア・シルカ、モンゴル、日本全国21地点 (部分日蝕) の多地点中継。シルカからのみ動画中継を NVAT (NEC提供) で実施。国内からはWWWによる静止画中継。

WWWサイト数は4。(ミラー含む。以下同。)

LIVE! ECLIPSE 98 (LE98)
ベネズエラ皆既日蝕 - 1998.2.26

中米のマラカイボ (ベネズエラ) とカリブ海のガドループ島から中継。両地点から動画中継をPipeCam (netspace提供) で実施。

WWWサイト数は5。

LIVE! ECLIPSE 98 annular (LE98A)
マレーシア金環日蝕 - 1998.8.22

マレーシアのダヤン島と沖縄(部分日蝕)から中継。マレーシアからの動画中継を RealVideo で実施。沖縄からは静止画中継のみ。

WWWサイト数は9。

LIVE! ECLIPSE 99 annular (LE99A)
オーストラリア金環日蝕 - 1999.2.16

ムレア (西オーストラリア) から中継。動画中継は RealVideo で実施。

WWWサイト数は12。

また、この他にもLE主要メンバーが中心となって、1998年11月の17日前後に獅子座流星群を観測・中継する LIVE! LEONIDS 98 を実施した。この中継では、極大時に合わせて伊豆高原から5時間の RealVideo による動画中継を行い、7つのWWWサイトに対して24時間でトップページビューが26万件、延べ20万人以上に動画を提供することができたが、流星観測では利用者のアクセスは数日間にわたって起き、僅か数分にアクセスが集中する日蝕中継とは処理するべきトラフィックの傾向が大幅に異なるため、今回は考察の対象としていない。

3 利用者のアクセス傾向と問題点

これまでの経験から、日蝕中継特有のアクセス傾向があることやそれに伴う問題点が幾つか明らかになった。

3.1 短いピークタイムとトラブル対策

日蝕では、太陽が完全に欠けた状態でいる時間はせいぜい数分である。このため図1に示すように毎回数分から数十分という非常に短い時間にアクセスが集中することになる。過去の全日蝕において同様の傾向が見られた [1]。LEでは、その名が示すようにライブ中継であることを重視しているため、このピークタイムの発生は不可避である。

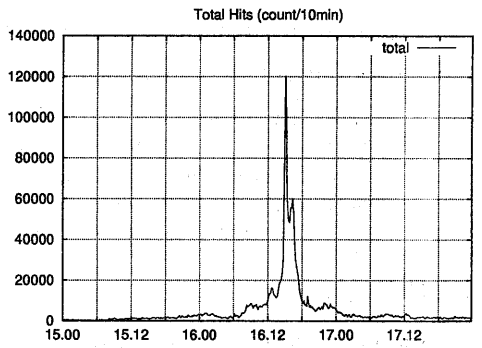

図 1: 時間的集中傾向 (横軸は DD.HH)

ピークタイムの短かさは各サーバに予測困難な問題を発生させる。過去の中継ではピークタイムに入ると短時間のうちに異なる原因で幾つかのサイトのサービスが停止するケースが散見された。代表的なものを以下に示す。

- 過負荷によってサービスが停止した。(LE99A)
- 動画データを提供しているサイトと同居していたために、回線の飽和によってWWWサービスができなくなった。(LE98)
- 動画データを提供しているサイトと同居していたために、ミラー元に対するデータ参照がtimeoutするサイトが幾つか出た。(LE98A)

- ログが急激に増えたため、ディスクが溢れてシステムが停止した。(LE99A)
- ログが急激に増えたため、普段は使われなかったディスク上のBad Blockをさわってしまい、システムが停止した。(LE97)

これらの事例から、システムに対する直接の過負荷以外にも、トラフィック増大をきっかけにサービスが停止する場合があることがわかる。徐々にトラフィックが増える場合はチューニングも含めて対処する余地があるが、日蝕サービスでは全ての事象が短時間で発生するため、ほとんどの場合適切な処置を施すことは出来ない。

3.2 要求されるモデル

まず各サーバを不安定にする要因は事前に排除することが重要なのは言うまでもないが、完全なサーバを作れないこともまた明らかである。すなわち各サーバの安定性は、それほど高い品質を確保できないと仮定するべきである。

また、LEは非営利の活動であり、ミラーに利用するサーバは無償で提供されたものである。そのため用意される各サーバのハードウェア性能、OSや接続回線の利用可能帯域を総合したサービス能力には大きな差がある。

本稿ではこの状況を以下のようにとらえる。

1. 各サイトの最大サービス能力には差がある。
2. 各サイトのサービス能力は不安定であり、停止する可能性もある。

このような不揃いなサーバを用いながら、全体としてWWWサービス能力を維持できる分散サーバ群を構築し、その有効性の検証を試みる。

4 LE99A分散サーバ群の構成

4.1 アクセス振り分けの方法

WWWアクセスのミラーサイトへの分散機構は、LE99Aまでは特に何も用意しておらず、LEトップページである `http://www.solar-eclipse.org/` 上に、 `Mirror Site #1, #2, #3` などのリンクを単に列挙しただけであった。これだけでもかなり均等に負荷分散することが継続的なログ解析から判っている[1]が、LE99Aでは列挙するにはサーバ数が増え過ぎたため、新たな分散機構を試みることとした。

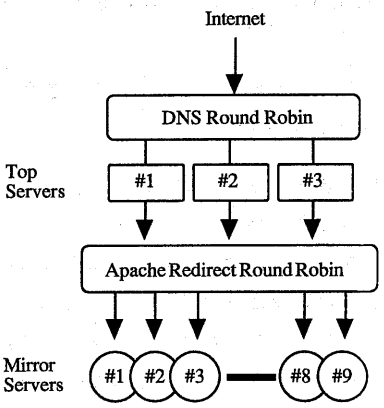

図 2: LE99A サーバ配置

まずLE99AではWWWサーバを図2のように配置した。InternetからのLEトップページへのアクセスは、まずDNSによるRound Robinによって3台のTop Serverに振り分けられる。そこからピーク時に最もアクセスが集中する動画情報へのリンクを含む部分にアクセスすると、Top ServerでWWWサーバソフトとして利用されているApacheのRound Robinされた Redirectの機能[2]によって、多数用意されたMirror Serverに動的に振り分けられる。こうしてユーザはいずれかのMirror Serverに自動的に誘導され、それ以降はずっと特定のMirrorサーバを対象にWWWページの閲覧を行うことになる。

4.2 二種類の Round Robin

この分散機構の基本は Apache Redirect の Round Robin機能を用いたアクセス振り分けにある。しかし鍵となるRound Robinを行うApacheサーバが一台しかない場合は、この一台がなんらかの理由でダウンした時に全体が機能しなくなる危険があるため、これを回避するためにDNS Round Robinによる複数のTop Serverを設けてApache Redirect Round Robinを行うサーバに冗長度を

与えたものである。

全体を DNS Round Robin で構成しなかったのは、Apache Redirect の以下の利点を重視したためである。

1. Round Robin する要素に重みづけを設定できる
2. 重みづけ情報や、リストからの排除の設定が即時に反映される。

利用できる各サイトの WWW サーバの能力にはかなりの開きがあるため、能力の高いサイトに対して多くのアクセスを誘導する重みづけが設定でき、また随時変更できることは有用である。特にダウンしてしまったサーバは速やかに排除しなくてはならないため、これらの設定変更は運用中に随時行なわれ、即時に反映される必要がある。

DNS による Round Robin での重みづけについては過去に WIDE プロジェクトの IAA 実験 [3] において適用された例があるが、DNS の特性から即時反映を可能にする運用は難しいと判断した。

4.3 重みづけの方法

重みづけの初期値はマシン能力や利用可能な回線の帯域、および事前の HTTP によるファイル転送能力の測定結果から推定した値とした。

この HTTP による転送能力は 50KBytes のファイルを GET したときの転送レートで測定する。測定の正確さを高めるために、ネットワーク上の経路の異なる数箇所から一定時間ごとに数回の転送を行い、その平均と Timeout 率を求める。

このサービス能力の測定 (以降単に能力測定) は LE 中継の間継続して行われ、図 3のように記録される。このサービス能力の変化に応じて重みづけの値は動的に変更される。

```
date  time |      sun-i|       OKI|       ksu|
02/16 14:07|0   212995|0   761870|0   222301|
02/16 14:22|0   178394|0   945171|0   216971|
02/16 14:37|0   111805|0   541453|X    68619|
```

図 3: 転送レート記録 (bps, X は Timeout あり)

5 評価

ここでは、LE99A における WWW 分散サーバの運用結果を各サーバのアクセスログを元に分析・評価する。

5.1 Top Server における分散

図 4 に Top Server のピーク時トラフィック変化を示す。横軸は 1999 年 2 月 16 日の時刻、縦軸はログから得たアクセス数で、Top Server ごとにプロットしている。

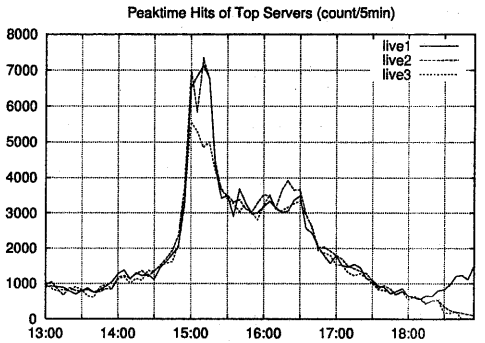

図 4: ピーク時の Top Server アクセス

この図から、細かい部分も合わせて、三つのサーバに均等にアクセスが分配されていることがわかる。15:15 前後のピーク時における能力測定では live3 サービス能力の低下が見られていたが、アクセスログの結果もそれを裏付けている。

ピークを過ぎて負荷の心配がなくなったため、18:30 頃に Top Server を live1 だけにする処置を DNS 設定に加えた。この結果はほぼ即時に反映されはじめている。

図 5 は DNS のリクエスト量の変化を示す。横軸は 1999 年 2 月の日付と時刻、縦軸は DNS のログから得たアクセス数である。DNS サーバは 2 台用意したが、そのどちらにもほぼ均等にリクエストが来ていることがわかる。

全 WWW サーバのログによると、LE99A ではピークタイムを含む 24 時間で 60ヶ国以上、ピークの 1 時間で 10000 以上の IP アドレスからのアクセスがあった。LE99A では www.solar-eclipse.org

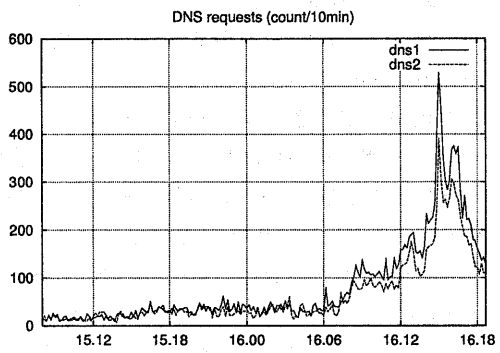

図 5: DNS リクエスト数 (横軸は DD.HH)

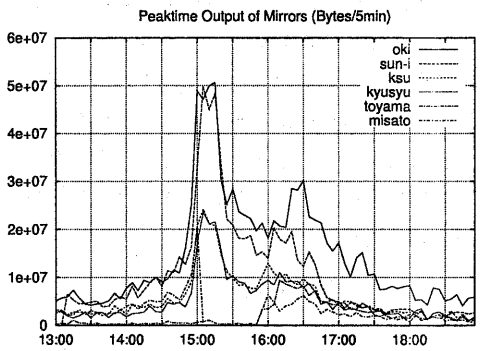

図 6: ピーク時の Mirror Server 転送量

と www.live-eclipse.org の 2 ドメイン名を使い分けたため、これを統一して更に問い合わせを減らすことも可能であるが、そうせずとも DNS の処理量は問題にならないほど少ないことがわかる。

5.2 Mirror Server における分散

グラフ上に oki, sun-i で示される 2 つの Mirror Server には重み 3 が、ksu, kyusyu, toyama, misato については重み 1 が初期値として設定された。Mirror Server は全部で 9 つあったが残り 3 つのうち、まず一つは他の重み 1 のサイト並に能力が出ていたと推測されるがレンタルホストであったためログが回収できなかった。次の一つは常に能力が出ず、あとの一つは早々にサーバが過負荷でダウンしてしまい、Round Robin リストから外した。これらの理由により、Mirror Server のグラフはすべて 6 サーバでプロットしている。

図 6 に Mirror Server のピーク時トラフィック変化を示す。横軸は同じく 1999 年 2 月 16 日の時間、縦軸はログから得た単位時間あたりのバイト転送量を示しており、Mirror Server ごとにプロットしている。

細かな部分も含めて、6 つのサーバに重みづけに応じてアクセスが分配されていることがわかる。このカーブはまた Top Server のそれとほぼ同じ形である。

5.3 動的な設定変更

misato は 15:10 頃にログの増大でディスクが一杯になりサービスが停止したため Round Robin リストから外したが、その後処置を施して重み 1 で 16:10 頃にリストに加えた。toyama は設定の失敗から Mirror に参加したのが遅く、15:50 頃からリストに加えられた。図 6 にはこれらの設定変更がほとんど即時に反映されていることが示されている。

また、sun-i は 15:30 頃から転送能力の測定値が不安定になった。図 7 に、代表的な三つの Mirror Server のサービス量の変化を抜き出してみる。

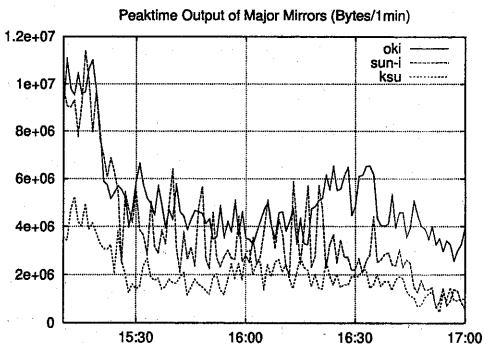

図 7: sun-i サイト不安定時

今回は sun-i のなんらかの資源が飽和したと考えて、重みづけの配分を変更した。まず 16:20 頃に管理者から余裕ありと報告された oki の重みを 3 から 5 に上げた。それでも能力測定値が安定し

なかったため、16:40 頃に sun-i の重みづけを 1 に下げた。これらの処置によって sun-i のサービス量は安定し、その後も他の重み 1 の他サイトと同量のサービスを提供し続けることが出来た。

次に図 8に、各 Mirror Server の全体に対する処理の分担率を示す。このグラフでは重みづけの効果がより明確に現れている。

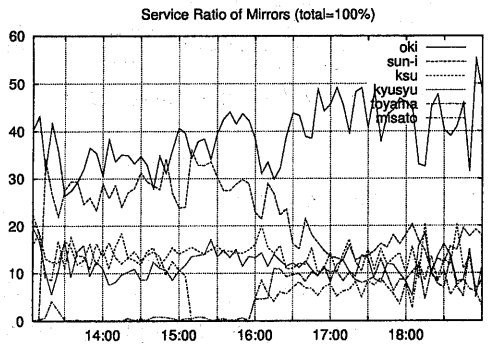

図 8: 動的な重みづけ変更の効果

横軸は同じく時間、縦軸はログから得た単位時間あたりのバイト転送量における全 Mirror に対する分担率 (%) である。特に misato, toyama, oki, sun-i に対して行われた動的な対処の効果について注目されたい。特に不安定になった sun-i の重みを 16:40 頃に 3 から 1 に変更した結果、重み 1 のサイトの並びに速やかに移行し、安定して機能し続けたことがわかる。

これらの動的な操作は WWW のアクセスログを見て行ったのではなく、複数地点から継続的に能力測定を実施し、その結果を元に行っていた。すなわぎ、アクセスログの分析結果はその測定結果がおおよそ正しく、処置が適正であったことを裏付けていることになる。この様に、外部からの測定結果を元にして、リアルタイムに状況に合わせた分担率の最適化を行うことができたことになる。

6 まとめ

分散化はスケーラビリティを確保するための手法の一つであるが、安定で均質な要素サーバと回線環境を揃えられるとは限らない。能力に差のあるサーバを集めて処理の分担率を最適化し、そこで全体として有効に機能させる機構が求められる。その意味で LE における分散サーバ群は典型的なモデルであると言える。

本稿では能力差のある要素サーバを用いた WWW 分散サーバ群において、各サーバの状況に合わせて動的に処理分担率を変更するための手法を示した。それに基づく WWW 分散サーバ群を LE99A において構築・運用した結果、各サーバ単体の処理量を大きく越える最大約 5.6Mbps の処理量を実現できた。そこでは各サイトの過負荷やサーバ停止などに合わせて動的な処理分担率の最適化を試み、その有効性を実証した。そのために必要なサービス能力の推定が、外部から HTTP による転送能力を測ることによって可能であることも明らかになった。

1999 年 8 月にはヨーロッパ・アジアを横断する皆既日蝕 (LE99) があり、そこでは本稿で示した分散サーバシステムの規模をより大きくして運用を試みる予定である。そこでは経路や言語など、ユーザの属性に応じたサーバへの誘導も試みたいと考えている。

なお、最新の LE の情報は以下の URL から入手できる。`http://www.solar-eclipse.org/`

謝辞

梅本肇氏、尾久土正己委員長をはじめとする、LIVE! ECLIPSE 実行委員会の諸氏およびすべての LIVE! ECLIPSE の協力団体、個人に深く感謝します。

参考文献

[1] 安田豊, “LIVE! ECLIPSE ログ分析”
`http://www.rieb.kobe-u.ac.jp/~yasuda/LE/index-j.html`

[2] Apache Project, “RewriteMap”
`http://www.apache.org/`

[3] WIDE プロジェクト, “IAA システムについて”
`http://www.iaa.wide.ad.jp/IAA-1998/System/system.html`