データベースシステム　115－15
情　報　学　基　礎　49－15
（１９９８．５．１５）

放送型配信機構上での並行処理制御方式「山彦」の補完機構

白田由香利[†‡]　飯沢篤志[†‡]　矢野隆志[‡]
† （株）次世代情報放送システム研究所　{shirota,izw}@ibl.co.jp
‡ （株）リコー　研究開発本部　ソフトウェア研究所　{shirota,izw,tyano}@src.ricoh.co.jp

大規模な情報放送システム、電子図書館システム、ビデオオンデマンド(VOD)システムを構築する場合、利用者からのアクセスを効率良く行わせるためには、データ複製を行うことが有効である。我々はこうした複製間の並行処理制御方式として、放送型配信機構を用いた「山彦」方式を提案し現在評価を行っている。放送では降雨、降雪などの気象条件により受信に失敗する場合があるが、本稿ではこれを通信によって補う「山彦補完機構」について述べる。山彦補完機構では同時に、集計した最新値保有サイト ID の集合を放送配信することにより、システム全体として最新値コピーにかかるコストを小さくできる。

Extension to the Concurrency Control Schemes by Broadcasting Mechanism called ECHO

Yukari Shirota[†‡],　Atsushi Iizawa[†‡],　and Takashi Yano[‡]
† Information Broadcasting Laboratories, Inc.　{shirota,izw}@ibl.co.jp
‡ Software Research Center, R & D Group, Ricoh Company, Ltd.　{shirota,izw,tyano}@src.ricoh.co.jp

When constructing a large-scale information broadcasting system, digital library system, or video on demand (VOD) system, data replication is useful to give users efficient access to the data. For concurrency control of these replica data, we have proposed concurrency control schemes using broadcasting mechanism called ECHO, which is currently in the evaluating phase. While using broadcast is cost effective, one drawback is, in an unlikely event, with changes in the weather such as heavy rain and snowfall, the system could fail to receive the broadcast. To overcome this, we propose an extension to the "ECHO" schemes which can compensate for possible broadcast failure using communication networks. The extend also collects a set of site IDs that possess the latest version and broadcasts them, so that the total cost of the latest version replication will be minimal.

1. はじめに

大規模な情報放送システム[飯沢 97]、電子図書館システム、ビデオオンデマンド(VOD)システムを構築する場合、利用者からのアクセスを効率良く行わせるためには、データ複製を行うことが有効である。しかし超大規模分散データベースシステム（DBS）の場合、複製数も大きくなり、複製間の並行処理制御にかかるコストが増大するという問題がある。我々はその解決策として放送型配信機構を用いた「山彦」方式を提案し評価を続けている[白田 97]。

山彦方式は楽観的制御方式の確認相に放送型配信を利用することを特長とする。システム全体で唯一の「放送局サイト」がグローバルタイムスタンプを利用してコミットの判定を行なう。複製の置かれる多数のサイトは其々「DB サイト」と呼

[†](株)リコーより(株)次世代情報放送システム研究所へ兼任出向中。
The authors are partly on loan from Ricoh Company, Ltd. to Information Broadcasting Laboratories, Inc.

ぶ。DBサイトは、更新を起こす可能性のある「更新サイト」と、放送された情報を受信蓄積するだけで自らは更新は起こさない「受信サイト」の2種類に分類される。

我々が主として想定する応用は情報放送システムであり、配信するデータは雑誌やCD-ROMのような情報のまとまり、情報パッケージを想定している。例えば、都市近郊のタウン情報誌的情報パッケージであれば、内容は、映画や演劇、催し物情報、及びTVの番組表などである。その更新処理としては、映画館の興行予定、時刻表の追加、変更など多種類となり、その更新頻度は小さくない。よって超多地点サイトの複製間での並行処理制御が必要になってくる。

情報放送システムでは、図1に示すような各種のデータベース(DB)を想定している。受信機DBは放送されたコンテンツを家庭で蓄積するものであり、中間DBは放送されるコンテンツをキャッシュすることにより、システム全体としてのアクセス効率を高めるために配置される。素材DBとは番組制作者が提供するDBであり、素材DBでのみ更新処理が発生する。この素材DBは一種類の応用システムにおいて、数百から数千サイトを想定している*。

山彦方式でのDBの名称と図1に示す情報放送システムでのDB名称の関係を以下に示す(代表サイトについては第3節で述べる)。

山彦方式での名称　　情報放送システムでの名称

- ·放送局サイト ——— 放送局DBサイト
- ·DBサイト
  - 更新サイト ——— 素材DBサイト
    - 代表更新サイト
    - 一般更新サイト
  - 受信サイト ——— 中間DBサイト / 受信機DBサイト
    - 代表受信サイト
    - 一般受信サイト

* 例えば、「タウン情報誌配信システム」と「不動産業者によるアパートマンション情報配信システム」では、別々に数百から数千のサイトが設置される。

複製は更新サイト及び受信サイト上に作られる。データ複製が多数ある場合複製間のデータ一貫性保持が問題となるが、山彦方式では更新サイト上の複製間の一貫性を保つだけでなく、受信サイトでも最新値を受信蓄積できるという利点があった。このように放送には、サイトが超多地点となっても配信コストが変わらない、という利点がある一方、放送配信では降雨、降雪などの気象条件により受信に失敗する可能性がある、という問題がある。本稿ではこれを通信によって補う「山彦補完機構」について述べる。

2．気象条件による放送受信の失敗

衛星放送は、山彦方式で利用する放送機構の主たるものであるが、その降雨に対する信頼性には問題がある。通信の場合、ノードおよびノード間の輻輳および障害が起こった場合の対処の方法として、下位レベルで再送を要求したり、タイムアウト後のリトライするなどの方策をとるが、放送の場合は確認を取るためのアップリンクが無いためである。それを補うため、誤り訂正符号や、同じデータを複数回放送するなどの方策を採るが、衛星放送の周波数は非常に高いため、降雨、降雪、強風、などの気象条件により電波が減衰し受信に影響が出ることがある*。

現在衛星放送では12GHz帯を使っているが2007年からは21GHz帯を世界各国が放送用に利用できるように決められている。降雨による受信への影響は周波数が高くなるに従い大きくなるので、将来21GHzを利用した場合、さらに問題となる*。降雨の影響をなるべく少なくするよう

* 「誤り訂正方式を用いることにより、30Mbps程度の受信ビットレート場合、誤りが起こるのは100年に1回以下の確率になり、事実上誤り訂正後の出力はエラーフリーとみなせる[米田98]」という報告があるが、これは通常の衛星伝走路を経由する間に受ける各種の伝走路ノイズを対象とする評価であり、降雨の影響があった場合は誤り率はさらに大きくなってしまう。

* 降雨の影響は10GHz以上の周波数で大きくなってくる。

に回線設計されてはいるが、1 年に数回降雨の影響による受信障害が発生する可能性はある[長坂 89]。放送衛星の電波への影響としては、降雪時の伝播損失と着雪によるアンテナの利得低下や+強風によるアンテナの受信レベルの低下もある[北城 84]。降雨強度と降雨減衰量の関係及び対策については[山田 97]が詳しい。

降雨の影響はその地域の降雨パターンに依存して変わるので、放送伝送機器はその地域の最悪月の伝送エラー発生率平均値が 99.5%から 99.7%程度になるように伝送機器は設計されている。このため 1 ビットの伝送誤りも許されないデータ配信の場合は、放送配信の誤り訂正符号の他に上位の応用相に対して CRC などを採用し誤り検出を行なっている。

3. 山彦方式で補完したい箇所

山彦方式では、放送と通信の両方を用いている。我々は通信の誤りはないと仮定し◆、山彦方式における DB サイトにおける放送受信の誤りをどのように補完するかについて論じる。

＜用語「代表サイト」の定義＞

以下では、特定のオブジェクトに対して更新が発生した場合、即時に更新伝播を必要とする DB サイトを、そのオブジェクトの「代表サイト」と呼ぶことにする。代表サイト以外の DB サイトは「一般サイト」と呼ぶ。

代表サイトは、そのオブジェクトの複製を管理している DB サイトから選出される。代表サイトの第 1 の役割は一般サイトからの read 要求に応じて、オブジェクトの最新値を read させてやることである。よって受信サイトであってもシステム全体の効率が向上すると判断されれば代表サイトと成りうる。代表サイトの第 2 の役割は、代表サイトの少なくとも一つには各オブジェクトの最新値が存在することを保証することである。これにより、システムから各オブジェクトの最新値がなくなってしまうことを回避する。

山彦方式において、放送されたメッセージを受信し損ねる場合としては以下の2通りがある。

(1) 更新サイトが自分宛てのコミット/アボートメッセージの受信に失敗した。

受信に失敗した場合、その更新サイトは来るはずの自分宛てのメッセージを一定時間待つが、正しいメッセージを受け取れないままタイムアウトとなった場合、どのように回復処理をしてよいかという問題が出てくる。もし仮にそのままアボートしてしまう、という方策を取るとすると、受け損ねたメッセージがアボート指示であれば問題ないが、メッセージ内容がコミット指示であった場合、最新値を保持しているサイトがなくなってしまうので問題となる。これを解決するため、タイムアウトした時点で、自分へのコミット許可がおりたか否かをどこかのサイトへ問い合わせることができるようにしたい。

(2) 代表サイトが、他のサイト宛てのコミットメッセージを受信することに失敗した。

代表サイトは常に最新値を保有している必要がある。問題は、「受信に失敗した代表サイトに対して失敗したことをどのように伝えるか」、及び、「どのように最新値を取得させるか」である。

上記問題を解決するため想定したシステム規模は次の通りである。

- 更新サイト数：数 100～数 1000
- 代表サイト数：10～数 10
- 各オブジェクトの最新値を保有するサイト数：　100～数 100

トランザクションの性質としては[白田 97]の評価で利用したものと同じとした。つまり、オブジ

強い雨の場合、放送衛星に使用している 12GHz 帯であれば電波の強さは 1/2～1/3 になり、20GHz 帯であれば、1/5～1/10 となり、30GHz 帯であれば、1/50 程度に弱くなる。周波数が 30GHz 以上になると大気の分子や水蒸気による吸収の影響も大きい[岩崎 85]。

+ 降雪による減衰は 1dB 以下と比較的少ないようであるが、着雪の場合、雪質や着雪状況によっては、受信不能になることがある。

◆ 通信の場合、エラーが発生したとしても双方向の経路があるため、再送処理が容易である。

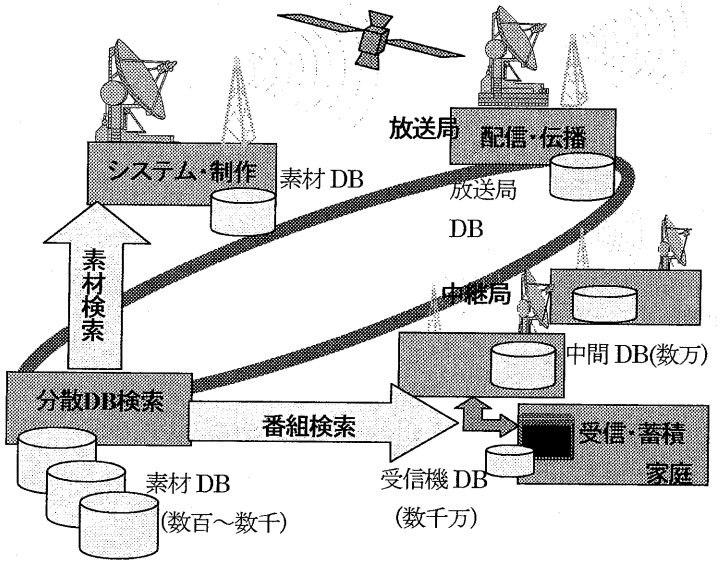

図 1: 情報放送システムで想定している放送型配信機構

ェクトサイズは 5MB、一つのトランザクションの中で各オブジェクトが更新される平均割合は 10%、一つのトランザクションでは平均して、5 個のオブジェクトを read し、2 個のオブジェクトを write する、と仮定した。トランザクションの放送局サイトへの平均到着率 λ は 1.1(sec^{-1})とした。放送受信の誤り発生率平均値は 99.9%と仮定した†。

4．山彦補完機構のための機能拡張

山彦における放送メッセージの受信失敗を補完する機構を「山彦補完機構」と呼ぶ。本節では山彦補完機構のため必要となる機能拡張を説明する。

＜放送局サイトで保持する情報の拡張＞

放送局サイトが格納しておく情報を次のように拡張する。新たに追加した属性は右側に "○"印をつけた。

・ OID(Object Identification)
・ 最終更新タイムスタンプ $t(O_0)$
・ 最終更新サイト ID
・ 最終更新トランザクション ID
・ 最新値保有サイト ID の集合○
・ 代表サイト ID の集合○

これらの属性情報は、システム全体が格納する全オブジェクトに対して保持する。後述するように放送局サイトへの負荷を分散するためには、代表サイトを限定して数を減らす必要がある。よって応用システムの設計段階で代表サイトは決定され、登録されることとする。

放送局サイトは、コミットを了承した時点で、最終更新タイムスタンプ、及び、最終更新サイト ID、最終更新トランザクション ID を変更し、最新値保有サイト ID の集合をクリアする。

＜アボートメッセージの拡張＞

放送局が配信するアボートメッセージも以下のように拡張する。

abort(サイト ID, トランザクション ID, (OID, 最新値保有サイト ID 集合)のペア値のリスト)

拡張理由は、オブジェクトの最新値保有サイトを公知する役割をアボートメッセージに含めることにより全体の効率化を図るためである。詳細は次節で述べる。

5．山彦補完機構

山彦補完機構とは以下のメッセージの送受信から構成される。

(a) 最終更新タイムスタンプ問合せ
timestamp_query(), timestamp_reply()
(b) 受信更新済報告と最新値保有サイト公知
receive_update_ack(),

† 1 年間に放送受信に影響を与えるほど強い降雨などがあるのは年に数回程度であると予想されるが、一度雨が降り始めると一定時間の間障害が続くので、平均値としては 99.9%とした。

broadcast_latest_sites()

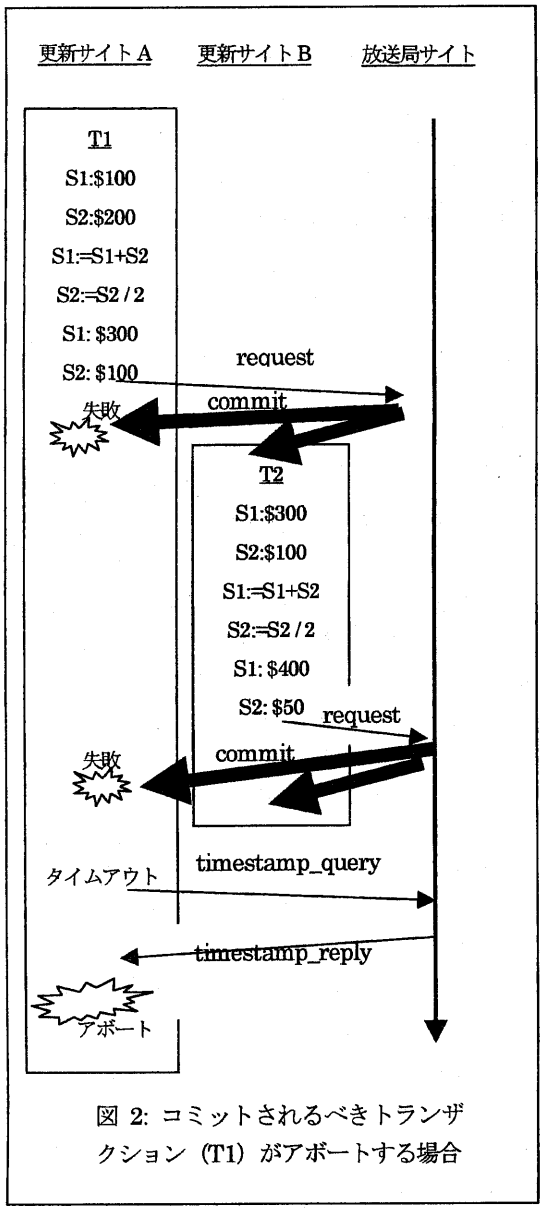

図 2: コミットされるべきトランザクション (T1) がアボートする場合

以下では、これらの内容を説明する。

(a) 最終更新タイムスタンプ問合せメッセージ

コミットメッセージを待ちタイムアウトとなった更新サイトは、放送局サイトに対して各オブジェクトの最終タイムスタンプの値を問い合わせることができるようにする。この問合せ及び返答には通信を用いる。理由は、一度放送によるメッセージの受信に失敗した場合の補完機構としては、通信の方が確実であるためである。最終更新タイムスタンプ問合せメッセージの形式は次の通り。

timestamp_query(OID)

それに対して放送局サイトの返答メッセージの内容は以下の通り。括弧内の引数の値が返ってくる。

timestamp_reply(OID, 最終更新タイムスタンプ, 最終更新サイトID, 最終更新トランザクションID, 最新値保有サイトIDの集合)

トランザクションを T とする。T は自分がコミットすべきか否かを以下のように判定する。但し、更新した各オブジェクトを Oi とし、Tがアクセスした Oi のタイムスタンプを t(Oi, T)、放送局サイトが保持する主複製オブジェクトのタイムスタンプを $t(O_0)$と表現する。タイムスタンプ値は順次増大するとする。

```
for oid in (T が write した全オブジェクトの OID){
  timestamp_query(oid)を放送局サイトに送る;
  timestamp_reply()を受信する;
  if ( (O0 の最終更新トランザクション ID ==
    T のトランザクション ID)      &&
    (O0 の最終更新サイト ID ==
    T の行われたサイト ID)      &&
    (t(Oi, T) < t(O0)) )
  then return("コミット", t(O0))
  else {
   最新値保有サイト ID 集合を保存;
   continue;
  }
}
return("アボート");
```

つまり、T 自身が更新したという痕跡が少なくとも一つあればコミットとなる。しかし、コミットを許可されたにもかかわらず、他のトランザクションが次々と更新を行い、結果として自分の更新の痕跡が一つも無い場合がある（放送受信に失敗する頻度自体が年間数回程度なので、現実にこのようなことが起こる確率は非常に少ない）。例えば、図 2 では、T1 が timestamp_query で問い

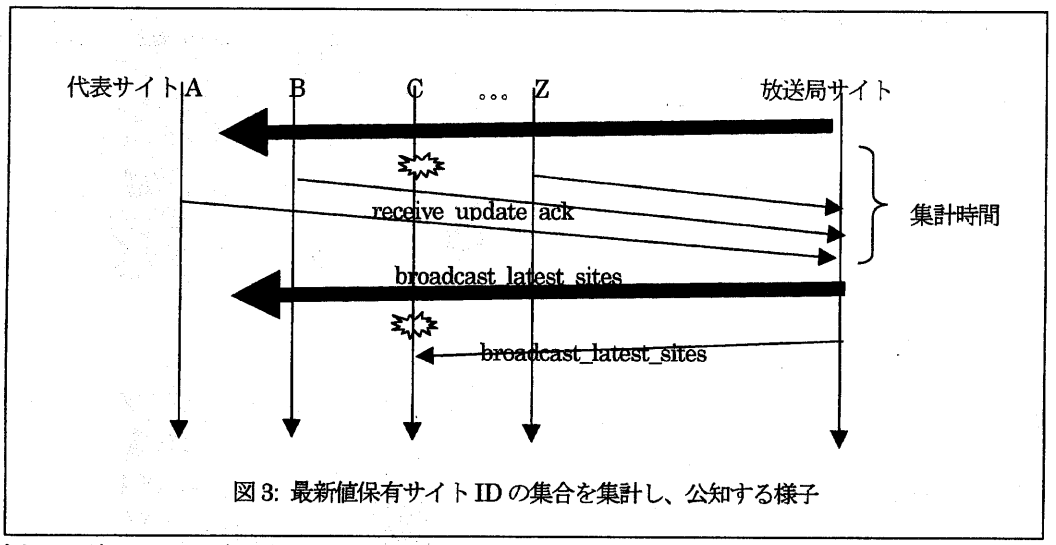

図 3: 最新値保有サイト ID の集合を集計し、公知する様子

合わせる前に T2 が更新を行っている。この場合においても、T1 はアボートするが正しい最新値は放送されたコミットメッセージによって少なくとも T2 には継承されているので、問題は無い。

判定結果がコミットであった場合 T は、返り値として得られた t (O_0)を、T が更新した全オブジェクトのタイムスタンプ値として書込みを行なう。これは、一つのトランザクション中で複数のオブジェクトが更新された場合、同じタイムスタンプ値が使われるからである♥。

放送局サイトではオブジェクトの最新値は保存しないので、アボートとなった T はリトライの準備のため、どこかのサイトからオブジェクト最新値をコピーする必要がある。放送受信に成功した場合は abort()により、放送受信に失敗した場合は timestamp_query により、最新値保有サイト ID 集合を知らされる。T はその ID 集合の中からアクセスコスト最小のサイトを選択し、そこからコピーする‡。

＜ローカルな一貫性の保持について＞

上記のように最新値保有 DB サイトから最新値をコピーする場合、ローカルな一貫性が崩れる場合がある。例えば、最新値保有 DB サイトから最新値をコピーをしている間に、他の DB サイトで更新が起こり、もはやそのコピー値が最新値ではなくなった場合である。

山彦では、read-only トランザクションに対しては一貫性チェックなしで、ローカルに実行してしまうという方式をとっていた。よって、コピーしているうちに情報が古くなっても問題にならない。Read-only トランザクションにおいても、最新値を read したい、という場合は、更新トランザクション同様に放送局サイトでタイムスタンプによる確認を受ければよい。

上記 timestamp_query の効果は以下の通り。

・自分宛てのコミット/アボートメッセージの受信に失敗したトランザクションでも自分がコミットしてよいか判断できる。

・オブジェクトの最新値保有サイト ID の集合を取得できる。

(b) 受信更新済報告メッセージと最新値保有サイト公知メッセージ

各メッセージの形式は以下の通り。

・受信更新済報告

receive_update_ack(自分のサイト ID , (OID, t(O_0))のペア値のリスト)

♥ 山彦方式では、トランザクション中の全ての更新タイムスタンプを同じ時刻に設定する、としている。

‡ アクセスコストの計算方式は、サイトの CPU 負荷や通信経路の混雑度情報の収集などに基づいて行われるが、その情報収集方式、計算方式は今後の研究課題である。

・最新値保有サイト公知

broadcast_latest_sites((OID, t (O_0), 最新値保有サイトIDの集合) の3つ組リスト)

山彦方式において、放送局サイトは以下のような手順で最新値保有サイトIDの集合を収集する(図3参照)。但し以下ではコミットメッセージはDBサイトA宛てであると仮定し説明する。

(1) 放送局サイトがDBサイトA宛てのコミットメッセージを放送する。放送局サイトは、コミットと判定した時点で最新値保有サイトID集合の値をクリアする§。

(2) このコミットメッセージを受信したDBサイトのうち、代表サイト及び他のサイトがオブジェクトの最新値をreadしにきてもよいと思うサイトはreceive_update_ackを放送局サイトに通信で送る。代表サイトは受信したら必ずreceive_update_ackを返さねばならない。

(3) 放送局サイトでは、DBサイトから通信で送られてくるreceive_update_ackを収集し、自分のDB上の「最新値保有サイトID集合」に付加する。放送局サイトは収集の時間を予め決めておき、制限時間内に届いたメッセージだけを集計し、次のアクションに移る。しかし、締め切りに遅れた受信更新済報告メッセージも、DBには追記されていく。

(4) 放送局サイトは、集計に間に合った最新値保有サイトID集合を放送で配信する。このメッセージがbroadcast_latest_sitesである。

(5) 放送局サイトは、Tにより更新されたオブジェクト其々に対して以下を行う。
そのオブジェクトの代表サイトID集合のうち、最新値保有サイトID集合に含まれていないサイトID集合に対して、broadcast_latest_sitesを通信で送る。

§ この時点ではまだ最新値保有サイトID集合にサイトAを挿入してはいけない。理由はサイトAがコミットメッセージの受信に失敗し、最悪の場合アボートする可能性があるからである。

最新値をシステムのどこかに保持するためには、最低一つの代表サイトからreceive_update_ackが届くまでは、放送局サイトは最新値を保持していることとする。最悪の場合として、リクエストを依頼した更新サイトもコミットメッセージを受信し損ねている可能性があるからである。その時は放送局サイトは再度同じコミットメッセージを放送する。代表サイトに限定した理由は一般サイトに比べて信頼性が高いからである。

上記の最新値保有サイトIDの公知による効果は以下の通り。

- どこかのDBサイトから最新値をコピーしようとしている場合、複数の最新値保有サイトIDが放送配信されるので、その中からアクセスコストが最小であるDBサイトを選ぶことが可能となる。
- システム中代表サイトのうち少なくとも一つのサイトには各オブジェクトの最新値が蓄積されていることを保証できる。
- 代表サイトが最新値を受信し損ねても、再度、放送局サイトが代表サイトに対して通信で送信してくれる。その結果、代表サイトは「自分が最新値の放送受信に失敗したこと」及び「どこに最新値があるのか」を知ることが可能となる。

6. 放送局サイトへの負荷集中に対する考察

山彦方式では放送局サイトへの負荷集中がシステムのボトルネックとなるため、負荷を分散させるようにシステムを設計する必要がある。以下、その仕組みについて考察する。

6.1 代表サイトの限定

放送局サイトの蓄積コストとメッセージ通信負荷を軽減するためには、代表サイトの登録を制限すること、及びreceive_update_ackを返すDBサイトを制限する仕組みが必要である。

第3節で想定した環境で放送局サイトでの最新値保有サイトIDの集合の蓄積コストを考えてみる。DBサイトIDを4Byteし、代表サイトに

は最新値保有フラグ 1bit を其々付けるとする。代表サイト数 30、最新値保有サイト数 300 の場合、一つのオブジェクト 5MB に対して 4B×270＝1,080B であり、約 4,630 対 1 となる。正味のコンテンツに対してこの程度に小さい容量であることが望ましいと考える。

6.2 最新更新タイムスタンプ問合せの発生頻度

コミット/アボートメッセージの受信失敗に対する補完機構として timestamp_query の利用を第 4 節で我々は提案した。別の案として考えられるものが「コミット/アボートメッセージを放送と通信の両方で送る」という案である。両者を比較してみる。

放送局サイトへのトランザクションの平均到着率を 1.1[sec^{-1}] とすると、1.1 × 60 × 60 × 365=1,445,400。つまり年間約 145 万回のトランザクションが到着することになる。受信に失敗し、timestamp_replay()を発行するのはそのうちの 0.1%とすると、約 1,450 回程度である。また降雨は一定時間続くため、障害はバースト的に派生すると予想されるので、実際に障害を受けるトランザクションの数はさらに小さいと考えられる。0.1%の失敗のために毎回通信メッセージを送るのは得策ではないと判断し、我々は補完機構として上述したような timestamp_reply を採用した。

6.3 最新値保有サイト ID の集合の取得方法

山彦補完機構では、broadcast_latest_sites によって、どのサイトでも受信しようと思えば最新値保有サイト ID の集合を取得することができる。放送局サイトの負荷分散のためには、予め自分が read しそうなオブジェクトの最新値保有サイト ID の集合を自分の DB に蓄積しておき、まずはそこに記してあるサイトのどこかにアクセスし最新値をコピーすることが望ましい。

これでも失敗した場合でも、すぐには timestamp_query を放送局サイトに送らないようにする。DB サイト間に階層化構造を作り、まずは親の DB サイトに timestamp_query を通信で送信することにする。階層化構造は親ノードになるに従い、多くの記憶空間と CPU を最新値保有サイトの取得のために割いているとする。

7. まとめ

山彦補完機構について述べた。その第 1 の利点は、放送されたメッセージを受信し損ねた場合でも通信によりコミット承認情報、及び、最新値保有サイト情報を知ることができること、第 2 に、最新値保有サイト ID の情報を集計し、放送配信することで、多くのサイトが最新値保有サイト情報を知り、最新値が必要になった場合のコピーの効率化が図れること、である。

我々は山彦補完方式の発展として、集計した最新値保有サイトの情報から最適な複製配置を計算し放送によって指示する方式を研究している。

参考文献

[飯沢 97] 飯沢篤志、浅田一繁、白田由香利: 「情報放送のための超大規模分散データベースシステム」、情報処理学会研究会報告：夏のデータベースワークショップ'97、97-DBS-113-44、1997。

[岩崎 85] 岩崎昇三 監修、久保庄二 著：「衛星通信」、オーム社、 1985。

[北城 84] 北城幹雄 編：「放送ニューメディアの受信技術 衛星放送/文字放送」、1984 年、電子技術出版。

[白田 97] 白田由香利、飯沢篤志、矢野隆志: 「放送型配信機構上での並行処理制御方式:山彦」、アドバンスト・データベース・シンポジウム(ADBS)'97 論文集、pp. 15-22、東京、1997。

[長坂 89] 長坂 進夫、小野義一、渡辺詔二、小山賢二:「現代テレビ・放送技術」、オーム社、1989。

[米田 98] 米田泰司 他：「ディレク・ティービー用ディジタル STB」、Matsushita Technical Journal、Vol. 44、No. 1、pp. 27-35、1998 年 2 月。

[山田 97] 山田宰：「デジタル放送のさらなる発展をめざして」、NHK 技研 R&D、No. 46、pp. 17-40、1997 年 8 月。