

WWW データアクセスパターンを考慮したキャッシングシステム

5 T - 5

西川記史* 細川貴史*
*(株)日立製作所
システム開発研究所

吉田健一** 辻 洋*
**(株)日立製作所
基礎研究所

1. はじめに

WWW 等ネットワークサービスの急速な普及に伴い、トラフィックの増加に伴うレスポンスの低下が問題となってきた。この問題を解決するためキャッシングプロキシが利用されているが、従来のキャッシングシステム¹⁾では、キャッシュ容量が十分でない場合ではヒット率が低い、大規模なネットワークにおいては個々のマシンの性能によるボトルネック等の問題がある。そこで我々は、WWW アクセスパターンが Zipf の法則に従うことに基づき²⁾、キャッシュ対象のページを選択・分散管理するキャッシュシステムを開発した。

本キャッシュシステムは、アクセスログを定期的に分析し、アクセス頻度の高いサイトのページのみをキャッシュする選択機能、及びマシンへのアクセスを均等に行うための分散機能を有する。本発表ではそのアルゴリズムと構成について述べる。

2. キャッシングシステムの設計

現在 Harvest/squid 等主なキャッシングプロキシでは、expire 情報を用いたキャッシュ置き換えを行っている。前述の知見を用いた解析結果によれば²⁾、これらの方式ではディスク資源を有効に用いることはできない。

またキャッシュシステムの運用を考慮した場合、既存のキャッシュシステムとの共存、及び複数サーバの集中管理も重要であり、以下のような構成とした:

- (1) LRU を用いたキャッシュ置き換えアルゴリズムの採用、
- (2) パラメタによるキャッシュシステムの制御、
- (3) キャッシュとは独立した分析プログラムによるパラメタ自動生成。

3. アルゴリズムと構成

3.1 選択機能

キャッシングプロキシは通常リクエストに対するログを出力する。分析プログラムは定期的にそのアクセスログを収集し、サイト毎にアクセス回数を集計する。その後、上位 N% (N はユーザ指定) のサイト名を抽出し、それを新たなキャッシングパラメタとしてキャッシュサーバに設定する(図 3.1)。キャッシュプログラムは、定期的にパラメタが更新されていないかどうかをチェックし、更新されている場合は、パラメタを置き換える。

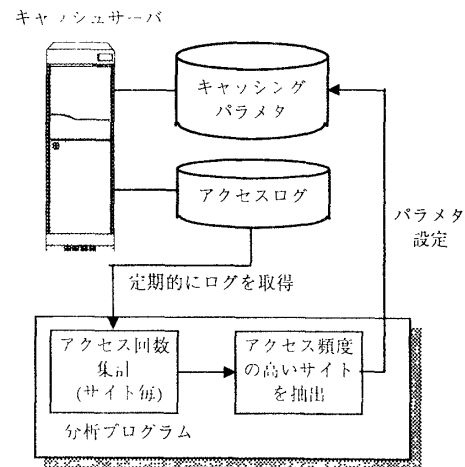


図 3.1 アクセスログ分析と選択

Caching Systems based on WWW Data Access Patterns
Norifumi NISHIKAWA* Takafumi HOSOKAWA*
Ken'ichi YOSHIDA** Hiroshi TSUJI*
* Systems Development Laboratory, Hitachi, Ltd.
** Advanced Research Laboratory, Hitachi, Ltd.

本キャッシュシステムでは、キャッシュ置き換えのアルゴリズムとしてLRUを用いている。キャッシングパラメタが分析プログラムにより変更されると、キャッシュシステムは、パラメタから消去されたサイトのページをLRUの先頭に移動しないようになり、結果としてこれらのサイトのページは新たなキャッシュ対象のページがキャッシュされるに連れ、キャッシュから削除される。

3.2 分散機能

本キャッシュにおける分散処理方式は、アクセスが集中する上位サーバ(ネットワーク側に配置される)のキャッシュを分散させるものである。分析プログラムは、上位サーバからアクセスログを取得し、上位サーバへのアクセス頻度が均等になるような分散パラメタを生成し、下位サーバに配布する。下位サーバは、この分散パラメタに基づき、アクセス先毎に決められた上位サーバを経由してページを取得する。この結果、上位サーバでは下位サーバのパラメタに従ってキャッシュの内容が分散される。

分析プログラムによるパラメタの生成は、まずアクセス頻度の高いサイトとそのアクセス回数を取得し、次に現在の分散の状態(どの上位サーバにどのサイトのページがキャッシュされているか)から、キャッシュ対象サイトでなくなったサイトを削除し、新たにキャッシュ対象となったサイトを、各サーバ毎のアクセス回数が等しくなるように分配する。

このようにパラメタを生成することにより、アクセス頻度の高いサイトが変わった場合でも、各上位サーバがキャッシュするサイトの変更が最小限に抑えられ、キャッシュされたページを有効に利用することが可能となる。また、下位サーバはどの上位サーバにどの

ページがキャッシュされているかを知っているため、無駄な通信を抑えることが可能である。

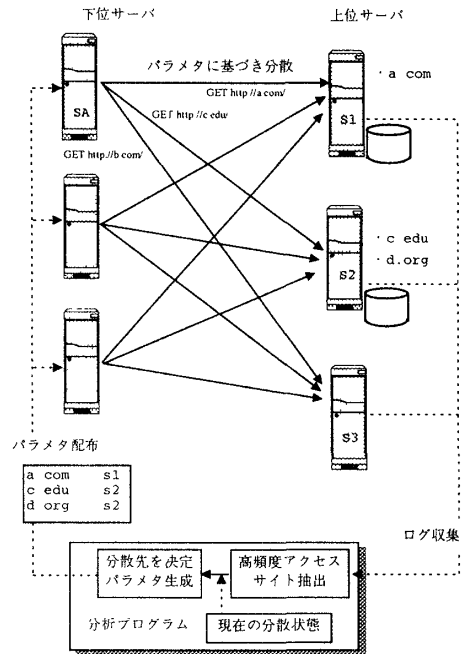


図 3.2 分散キャッシュ

4. まとめ

本稿では、WWW アクセスパターンが Zip の法則に従うことに基づき、選択及び分散を行うキャッシュの構成を示した。キャッシュの制御にパラメタを用い、またパラメタの生成に外部プログラムを用いることにより、既存のキャッシュシステムとの連携、及び複数キャッシングプロキシの集中管理が可能となった。

参考文献

- 1) Chankhunthod, A., etc., *A Hierarchical Internet Object Cache*, USENIX96, pp.153-163, 1996
- 2) 西川他, WWW トラフィック分析と分散キャッシュ、分散システム運用技術、pp.7-12、1997