RMONを用いたWebトラフィック解析

4P－6

丹野 秀和　関 義長
NEC インターネット技術研究所

1. はじめに

近年の Web 利用者の増加に伴い、インターネットトラフィックの大部分を HTTP トラフィックが占めるようになってきた。このような中、HTTP トラフィックの解析を行い、URL アクセス分布等の傾向を的確に把握する要求が高まってきている。

一方で、RMON-MIB を実装した機器を用いたネットワークトラフィック監視手法が広く用いられるようになってきた。RMON-MIB の機能のうち、フィルタ機能、パケットキャプチャ機能を組み合わせることにより、ある特定の条件にマッチしたパケットをキャプチャし、解析を行うことが可能となる。

本報告では、RMON プローブ（以下、プローブ）のフィルタ機能、キャプチャ機能を用いた Web トラフィック測定・解析手法について述べる。測定の際、プローブのバッファサイズの問題で、全ての HTTP パケットをキャプチャすることが困難なため、ネットワーク上の一部の HTTP パケットをサンプリング収集する手法を用いた。そして、測定結果より URL アクセス数分布等の解析を行った。

2. 測定手法

2.1 測定システム

図 1 に測定システムの概略図を示す。測定にはインターフェースを複数持つプローブを使用し、HTTP トラフィックの測定を行うネットワークと、キャプチャしたパケットデータを制御用 PC（以下、PC）に転送するネットワークとに分けた。PC 上では HTTP トラフィック測定用に作成したツールを動作させ、プローブの制御とプローブがキャプチャしたパケットデータのダウンロードと保存を行った。

2.2 測定方法

今回の測定では、HTTP Get リクエスト（以下、リクエスト）と HTTP レスポンス（以下、レスポンス）の測定を行った。プローブが一つしかなく同時に収集できないので、リクエスト、レスポンスのそれぞれにマッチするフィルタを別々に設定し、毎時 00 分から 30 分まではリクエストのフィルタ、毎時 30 分から 60 分まではレスポンスのフィルタを用いてキャプチャを行った。

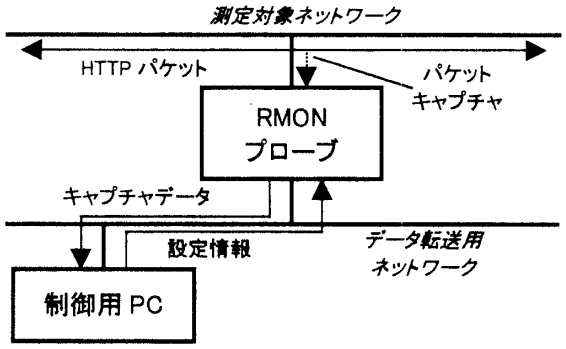

図 1. 測定システムの概略図

キャプチャは以下に示すサイクルを繰り返し行うようにし、ネットワーク上を流れる一部の HTTP パケットをキャプチャするようにした。

① プローブ上でキャプチャを開始
② 一定個数収集したら、キャプチャを停止
③ キャプチャしたパケットを PC にダウンロード
④ PC 上でパケットを解析、結果をログに追加

ログには、フィルタに一致したパケット数、キャプチャしたパケット数、URL 毎のアクセス数、サーバ毎の Content-Length フィールドの平均値等を記録し、30 分毎にフィルタを変更する際にファイルに保存するようにした。

3. 測定結果の解析

3.1 測定の概要

今回の測定は、HTTP トラフィックが全トラフィックの 90%程度を占めるネットワークで行った。プローブにフィルタを設定する際、IP、TCP の可変ヘッダに対応するためにフィルタを複数設定する必要があるが、今回測定対象としたネットワークでは HTTP トラフィックの 99%以上が IP ヘッダ、TCP ヘッダとも最小サイズの 20 バイトであったため、

The Web Traffic Analysis Using RMON
Hidekazu Tanno,　Yoshinaga Seki
NEC Corporation

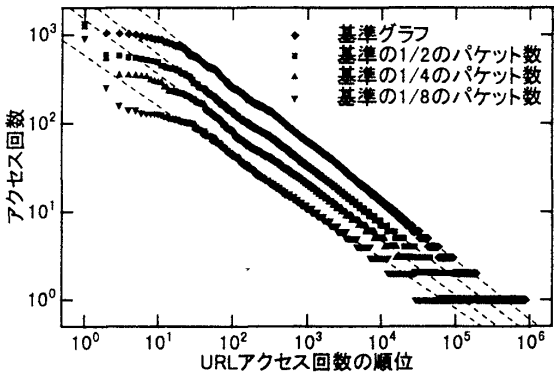

図 2. URL 毎のアクセス回数分布

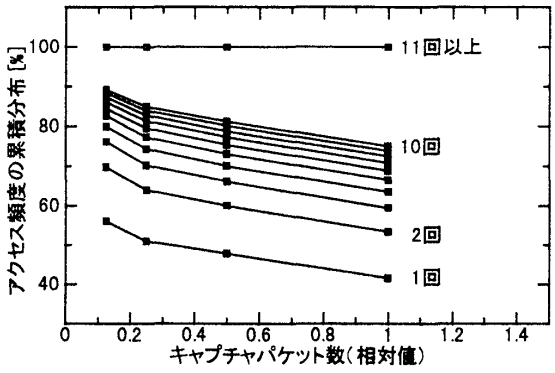

図 3. アクセス頻度別トラフィック分布

フィルタはこの最小サイズに合わせたもののみ設定した。また、パケットキャプチャは 1 サイクルで 100 パケット収集するように設定した。上記の条件で測定を行ったところ、フィルタにマッチしたパケットの 5〜7%をキャプチャすることができた。

3.2 URL アクセス分布の解析

図2にキャプチャしたリクエストパケットのURL別アクセス数分布を示す。横軸は各 URL に対するアクセス数の順位を表し、縦軸はその URL のアクセス数を表している。横軸、縦軸とも対数軸でプロットすると、測定値がある直線上に分布することが分かる。この結果は、数理言語学におけるジップの法則に対応づけることができる[1]。ジップの法則とは、fを単語の使用頻度、rを使用度数の多い方から振った順位とし、k、Cを定数とするとそれらが

$$f \cdot r^k = C \qquad (1)$$

という式に従うという法則である。図 2 の破線は、測定値が(1)式に従うと仮定して最小二乗法により求めた直線である。

解析するパケット数を変化させてプロットを行うと、グラフの切片（(1)式の C）は変化するが、傾き（(1)式の k）はほとんど変化しないことが分かる。この結果は、サンプリング収集した結果より全トラフィックの傾向を推定する際に、図 2 のグラフを平行移動することによって推定することが可能であることを示唆している。

3.3 キャッシュによるトラフィック低減の推定

図 3 は、収集したパケットを URL 毎のアクセス回数別に分類したときに、各アクセス回数に分類されたパケットが全収集パケットの何パーセントを占めるかを表したグラフである。図 3 の横軸はキャプチャパケット数の相対値を表している（図 2 の基準グラフのパケット数を 1 とした)。図 3 より、パケット数が増えると全データに占める低アクセス頻度の割合が減少する傾向が見られる。この傾向がさらにパケット数を増やした際にも当てはまると仮定すると、ネットワーク上の HTTP トラフィック全体のアクセス頻度分布を推定することが可能となる。

上記の推定により、例えばネットワーク間の HTTP トラフィックに対してキャッシュを導入する際に、事前にその効果を知ることが可能となる。キャッシュによりアクセス回数 n 回のトラフィックは $1/n$ に減少すると仮定することにより、図 3 の結果からキャッシュを導入した際の HTTP トラフィックの低減の度合を推定することが可能となる。

4. おわりに

本報告では、RMON プローブを用いて HTTP パケットをサンプリング収集する手法を提案し、実際に測定を行った結果から URL アクセス数分布等の解析を行った結果について報告した。そして、ジップの法則の適用等による、測定結果からの全トラフィックの状態推定の可能性について述べた。今後は異なった環境で同様の測定を行うなどして、この測定手法の有効性、妥当性についてさらなる検討を行う必要がある。

参考文献

[1] 吉田健一：“WWW 用分散キャッシュ構成の検討”, コンピュータソフトウェア, Vol.15, No.5, p38-p48 (1998)