ANALYSIS OF WEB TRAFFIC AND USERS' BEHAVIOUR MODELING DURING BUSY HOUR

Dr. Yevgeni Koucheryavy [1], *Andrew Krendzel* [2]

[1]Tampere University of Technology, P.O. Box 553, FIN-33101, Tampere, Finland e-mail: yk@cs.tut.fi

[2]LONIIS (R&D Institute), 11, Warshawskaya str., 196128, St.Petersburg, Russia, e-mail: kren@inser.loniis.spb.su

ABSTRACT

This paper presents results on statistical analysis of World Wide Web server under heavy load. Some empirical digits are presented inside the paper. The main attention paid to the users behaviour, i.e. to dynamic of requests toward the web server. Further, in part 4.2 the special case of 4-state PH-distribution was presented as analytical equations. Than, the parameters for mentioned distribution was computed in order to model interarrival time for users requests due the busy hour.

1. INTRODUCTION

The phenomenal growth in popularity of the WWW has made this service the most growing in communication society. Due to this fact the lifetime of hardware for Internet decreases dramatically (in comparison to the channel-switched hardware). However, the network planning task concerning QoS is still priority of the present studies. We now have the situation when Internet starts migrating from the best-effort network towards QoS support network.

This research was started in order to investigate the model of incoming requests stream for the web server during the busy hour, when the degrading of QoS is normally occurred due to the congestions.

The workload characterization is a very important thing in system and network design. It allows to predict the state of web server under the given traffic conditions and plays a crucial role in designing new network components. Moreover, in some cases (e.g. extremely heavy workload of tested site) it allows to foresee the outlook of the future www servers' workload.

2. WEB SERVER AND LOG COLLECTION

The 16th FIFA's World Cup was held in France since June 10 till July 12, 1998. The Web site for this event named www.france98.com was extremely popular and received about 1 billion requests during the tournament time. Since the April 26 all measured statistics are available in the Internet freely [1].

Actually this site consists of the overall information concerning football World cup, teams, players, matches and upcoming events to have the possibility of surfing through the web for all interested people. The World Cup Web site was installed to on-line at May 6, 1997.

3. BRIEF ACCESS LOG ANALYSIS

The workload of the web server is characterized by the set of access logs collected on a dayly basis from May 1, 1998 till July 23, 1998. Each access log is presented in the Common Log Format [2]. For each request received by web server the following information is stored in form of a log file:

```
remotehost rfc931 authuser [date]
"request" status bytes
```

These fields are defined as follows:

- `remotehost`: the IP address of the client issuing current request;
- `rfc931` : the remote login name of user;
- `authuser` : the user name as which the user has authenticated himself;
- `[date]` : the date and time of the request;
- `"request"` : the request line exactly as it comes from the client;
- `status` : the HTTP response status code returned from the client;
- `bytes` : the content length of the document transferred.

The `request` field includes the method (e.g. GET, HEAD etc.) to be applied for the appropriate request resource, the name of the resource (e.g. index.html) and the protocol version used (e.g. HTTP/1.0).
The overall traffic characteristics, i.e. access logs, are summarised in table 1 [3].

Table 1

Duration	May 1 – July 23, 1998
Total requests	1,352,804,107
Avg. requests per min.	10,796
Total bytes transf. (GB)	4,991
Avg. bytes transf per min	40.8

Thus, totally more than 1.35 billion requests were received by the web server during the collection period, and more than 5 TB of data sent to clients. The site receives about 11,000 request per minute in average.
Despite the vast amount of information available, a lot of interesting and important data is unavailable. For instance, the access logs does not appear to get information on the umber of aborted connections. As a result, the number of bytes transferred reported in Table 1 overestimates the actual traffic data. Moreover, while logs do have a timestamp that records when the server received the request, it has only one second resolution, which is too coarse-grained.

4. 'BUSY HOUR' ANALYSIS

From the beginning of May until the start of World Cup on June 10, the traffic volume was quite light. Beginning on June 10 the volume of traffic grows enormously. The site became very popular. Although the daily traffic volume was quite bursty during the whole World Cup, the traffic volume remained higher than it was at any other time prior to the starting of the each event. The busiest day for the site was June 30 when over 73 million requests were handled. After June 30 the daily traffic volumes began slowly diminishing until the end of the World Cup. Table 2 shows the statistics for busy hour.

Table 2

Duration	Busy hour
Total requests	12,104,059
Avg. requests per min.	201,734
Total bytes transf. (GB)	33.2
Avg. Mbytes transf per min	553

All results obtained in this paper are for the 11:00 – 12:00 of June 30 access logs.

4.1. Traffic

As far as we have access logs which include the `status` field, the successful / unsuccessful status of file transfer could be determined. That's why it's possible to use two kinds of statistics for the detailed analysis: i) successful file transfer; ii) unsuccessful file transfer. In this paper we concentrate only on the sequence, which consists of the file transfer for both types, i.e. traffic pattern. It is quite important to know a lot about traffic features from the network designing point of view.
The main parameters of mentioned traffic portion are summarized in Table 3.

Table 3

Mean, bytes	Std	Max, MB	Min, Bytes
17,458	75.32	2.35	3,850

It should be noted that 4 years ago the self-similarity feature of Internet traffic has been proposed in [4]. For a given statistic we found a high degree of self-similarity (a Hurst parameter value of $H \approx 0.85$), which is directly corresponds to the results obtained in [4] where complete information on explaining of world wide web traffic self-similarity was presented.

4.2. Analysis and modeling of users behaviour

4.2.1. Users behaviour

One of the most significant indicator characterizing the web server behavior is the load produced by users (requests for the information). In order to model web traffic load we should know the parameters of incoming request sequence, particularly the interarrival rate.
The main parameters of incoming load are summarized in Table 4.

Table 4

Mean	Std	Max	Min
3,150	1.78	3,520	3,950

Further, in next paragraph we'll introduce the method of incoming requests sequence modelling based on PH-distribution.

4.2.2. PH distribution

In order to analyze sophisticated systems with non-renewal input, we shall introduce the phase-type Markov renewal process [5,7]. In a continuous-state Markov chain with k transient states and (k+1)st absorbing state, suppose that upon entering the absorbing state, the process instantaneously jumps to transient state j, j = 1, 2, … , k with probability of f_j.

The PH-distribution is defined as the inter-visit time distribution to the absorbing state, and characterized by (f, G), where G is the transition rate matrix among the transient states, which is an irreducible $k \times k$ matrix. The row vector $\overset{\nu}{f}$ with component f_j called the initial probability vector. The PH-distribution is said to be in phase j if the underlying Markov process is in state j. The PH-distribution includes hyper-exponential, Erlangian, exponentional distribution, etc.

Following to [5,6] the probability distribution $F(\cdot)$ of the phase type is given by $F(x) = 1 - f^T e^{Gx} I,\ x \geq 0$,

where $\sum_{j=1}^{m} f_j \leq 1,\ f_j \geq 0,\ j = \overline{1,m}$ and I denotes identity matrix of the appropriate size with diagonal components equal to 1.

Let's define the special case of 4-state PH-distribution. In this case PH process having Poisson arrival rate λ_j in phase j, j = 1, 2, 3, 4, which appears alternately with exponentially distributed lifetime with mean r_{ij}^{-1}, where i is outgoing state. This is characterized by (R, Λ) where R is the infinitesimal generator of the underlying Markov chain and Λ the arrival rate matrix, defined by:

$$R = \begin{bmatrix} -(r_{12}+r_{13}+r_{14}) & r_{12} & r_{13} & r_{14} \\ r_{21} & -(r_{21}+r_{23}+r_{24}) & r_{23} & r_{24} \\ r_{31} & r_{32} & -(r_{31}+r_{32}+r_{34}) & r_{34} \\ r_{41} & r_{42} & r_{43} & -(r_{41}+r_{42}+r_{43}) \end{bmatrix}$$

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{bmatrix}.$$

The transient probabilities are defined as follows:

$$\theta_{ij} = \begin{cases} 1 + G_{ij} / \nu_i, & i = j \\ G_{ij} / \nu_i, & i \neq j \end{cases}$$

$$\sum_{j=1}^{m} \theta_{ij} \leq 1,\ \theta_{ij} \geq 0,\ i, j = \overline{1,m}$$ and

$$\theta_{i0} = 1 - \sum_{j=1}^{m} \theta_{ij}\ ;$$

where $\nu_i \geq -G_{ii}$, $i = \overline{1,m}$, and $G = R - \Lambda$.

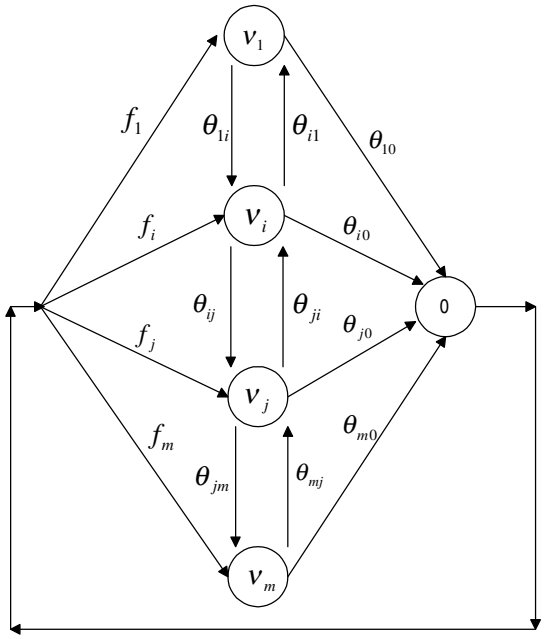

Fig.1 The state diagram for PH process.

Let's take a look at the case when $\nu_i = -G_{ii}$

It leads to:

$\nu_1 = r_{12} + r_{13} + r_{14} + \lambda_1$

$\nu_2 = r_{21} + r_{23} + r_{24} + \lambda_2$

$\nu_3 = r_{31} + r_{32} + r_{34} + \lambda_3$

$\nu_4 = r_{41} + r_{42} + r_{43} + \lambda_4$

By using these equations we can obtain analytical one for transient probabilities θ_{ij} ($i = \overline{1,m}$, $j = \overline{0,m}$).

Hence, for adequate traffic modeling by using 4-state PH-distribution the following parameters should be computed: $r_{ij}, i, j = \overline{1,4}, i \neq j$, $\lambda_j, j = \overline{1,4}$ and row vector $\overset{\nu}{f}$.

Table 5.

r_{12}	r_{13}	r_{14}	r_{21}	r_{23}	r_{24}	r_{31}	r_{32}	r_{34}	r_{41}	r_{42}	r_{43}	λ_1	λ_2	λ_3	λ_4
2.3	2.4	2.4	1.1	1.3	1.1	3.0	3.0	3.3	0.7	0.8	0.7	3239	3151	3111	3098

4.2.3. Parameters computing

For the Markov chain generating the process we need to divide the arrival process into scenes. Using 'Scene determining' method presented in [8] we can find the boundaries, which depend only on a few statistical parameters. These parameters should be available by simply scanning the size of incoming amount of requests per measurement time slot.

In mentioned book the following algorithm was suggested. Let A_i denotes the size of incoming amount of cells i and n is the scene number. The scene number is denoted by s and coefficient of variation c of a sample $\{x_k : k = 1, \dots, M\}$ is defined in accordance with:

$$c = \frac{\frac{1}{M}\sum_{k=1}^{M} x_k}{\frac{1}{M}\left\{\sum_{k=1}^{M} x_k^2 - \frac{1}{M}\left[\sum_{k=1}^{M} x_k\right]^2\right\}}$$

(i) Set n = 1 and s = 1. Set current scene boundary $b_{left}(s) = 1$.

(ii) Increment n by 1. Compute the coefficient of variation c_{new} of the sequence $A_{b_{left}(s)}$ to A_n.

(iii) Increment n by 1. Set $c_{new} = c_{old}$. Compute the coefficient of variation c_{new} of the sequence $A_{b_{left}(s)}$ to A_n.

(a) If $|c_{new} - c_{old}| \cdot (n - b_{left(s)} + 1) > \varepsilon$, set the right scene boundary $b_{right}(s) = n - 1$ and the left scene boundary of the new scene $b_{left}(s+1) = n$. Increment s by 1 and go to step (ii).

(b) If the above does not hold, go to step (iii).

Iterating this algorithm over the whole requests incoming sequence we get a series of scene boundary pairs. The value ε limits the amount of variation that is tolerated for a one scene. If adding a new measurement to the current scene increases or decreases the variation too much, a new scene is assumed to start at this measurement number.

Parameters presented in Table 4 were obtained on a base of the above algorithm. It's computed for 4-state PH-distribution.

It should be noted that for the absolute determination of the PH process we need to have the values of phase appearance probabilities, i.e. row vector $\vec{f}$ (see Fig.1). The Table 5 gives us these values.

Table 6

f_1	f_2	f_3	f_4
0.4	0.2	0.35	0.15

5. RESULTS AND DISCUSSION

To obtain the results on performance of presented model we have used the stochastic simulation. Further, we have checked that for empirical and simulated processes the first and second order statistics are match, i.e. marginal distribution and autocorrelation functions (ACF). For the model presented above at Fig.2-4 we show how it matches with experimental requests sequence (see Tables 5 & 6).

Here we use quantile-quantile plot in order to analyse of matching the empirical distribution (obtained from access logs) and simulated one (using 4 state PH-distribution). If we have a function $q_{(X_1,\dots)(Y_1,\dots)}(\cdot)$ and if the graph of this function fits the graph of the $u(x) = x$ function well, then the tested distribution are quite similar.

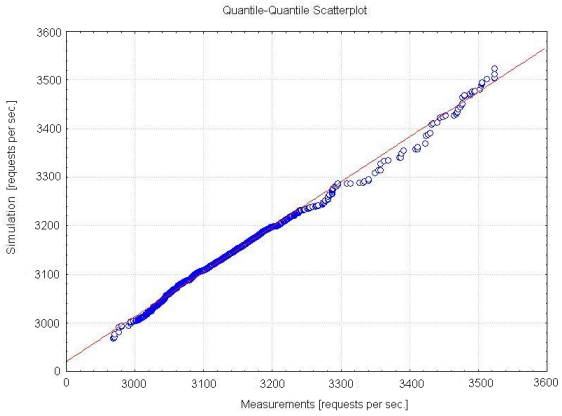

Fig. 2. Quantile-quantile plot

The simulation results indicate that the present method is suitable for modeling the stream of request to web server during busy hour.

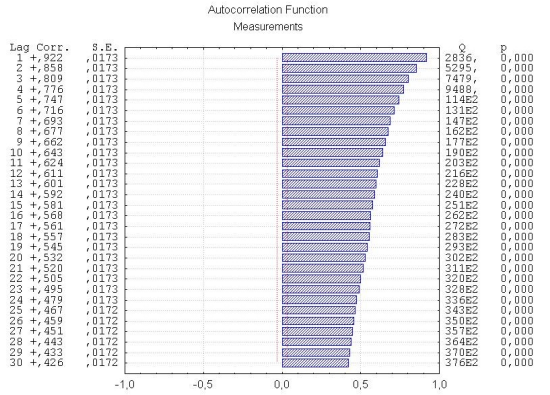

Fig. 3 Autocorrelation function for measured data.

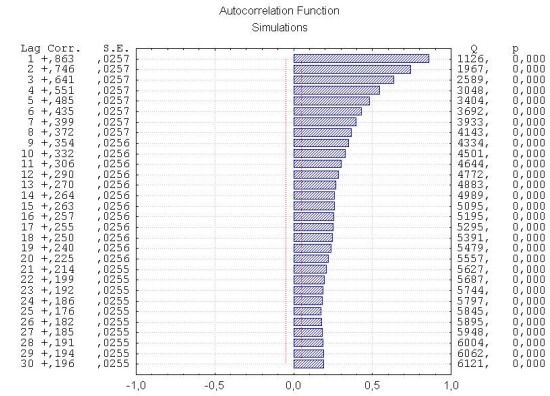

Fig.4 Autocorrelation function for simulated data

However, the usage of any second order statistical method two processes that have exactly the same distribution functions and ACFs can produce completely different queueing response. Such example was constructed in [9]. We will consider it in our future work.

REFERENCES

[1] The Internet Traffic Archive http://ita.ee.lbl.gov/html/contrib/WorldCup.html

[2] W3C Web page, *"Logging in W3C httpd"*, http://www.w3.org/Daemon/User/Config/Logging.html

[3] M. Arlitt T. Jin, "*Workload Characterization of the 1998 World Cup Web Site*", report of Hewlett-Packard Laboratories, September 23, 1999, p.90.

[4] M. Crovella and A.Bestavors, "*Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes",* IEEE/ACM Transactions on Networking, Vol. 5, No. 6, pp. 835-846, December 1997.

[5] Basharin G.P., Bocharov P.P., Kogan Y.• . *"Queuing analysis"*, Moscow, Nauka, 1989, 336 p.

[6] Neuts M. *"Matrix-Geometric solutions in stochastic models"*, Dover publications, NY, 1994, 332 p.

[7] Akimaru H, Kawashima K. *"Teletraffic. Theory and Applications"*, Springer-Verlag, 1993, 226 p.

[8] Roberts J., Mocci U., Virtamo J. *"Broadband Network Teletraffic. Final Report of Action COST 242"*, Springer, 1996, 586 p.

[9] Jelenkovic P.R. *"The effect of multiple time scales and subexponentiality on behavior of a broadband network multiplexer"*, PhD thesis, Columbia University, 1996.