

2G1126 Homework 1

## Estimating Waiting Time at a Server

Michael Ahlberg  
mah@kth.se

Måns Rullgård  
mru@kth.se

Henrik Åhlander  
e99\_hah@e.kth.se

April 5, 2003

## 1 Introduction

We have simulated the performance of a single-threaded web server. As input we used a trace file from a web proxy. The measured parameters were maximum and average delay, maximum and average queue length, and server utilization.

## 2 Experiments and analysis

The trace file contains the arrival time and size of requests for web pages. The time to service a request is taken to depend on the size by the relation

$$time = 0.1 + cost * size,$$

with *time* measured in seconds and *size* in kilobytes. A total of four simulations were performed. Two on the first 5000 entries in the log file, and two on the next 5000. For each set of log entries we ran the simulation once with the parameter *cost* = 0.6, and once with *cost* = 0.3. The results of these simulations are summarized in tables 1 and 2.

Entries	1-5000	5001-10000
Maximum delay in queue / s	14740.578	4700.435
Average delay in queue / s	8576.583	1214.684
Maximum number in queue	3763	3271
Average number in queue	1984.492	1077.766
Server Utilization	1.000	0.997
Delays longer than 0.5 s	1237	1720

Table 1: Cost = 0.6 s / kB

Entries	1-5000	5001-10000
Maximum delay in queue / s	7917.938	3201.071
Average delay in queue / s	4995.462	592.382
Maximum number in queue	1913	2756
Average number in queue	1164.998	718.470
Server Utilization	1.000	0.949
Delays longer than 0.5 s	3177	2152

Table 2: Cost = 0.3 s / kB

The large difference in waiting time between the two data sets prompted us to investigate the cause. We found that among the first 5000 requests, five had a size greater than 1 MB. Of these, two were as large as 30 MB. The largest request among the next 5000 is only 800 kB. This indicates that large requests will block the queue for a long time, causing the queue, and hence waiting times, to grow.

In order to determine more precisely how the performance depends on the size we truncated the request size at several levels and examined the effect. Figure 1 shows the server utilization and figure 2 the average waiting time for different maximum request sizes. Note that the range is truncated in the utilization graph in order for the graph to be more clear.

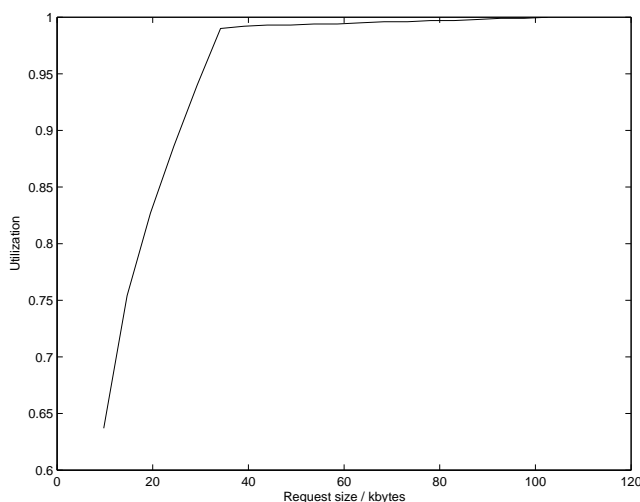


Figure 1: Server utilization vs. request size

### 3 Conclusions

When designing a single-threaded web server it is necessary to analyze how many requests the server is going to handle and the size of them. A few large requests could completely change the performance of the server. To be able to draw the correct conclusions it is very important to use test data that are similar to the data that is expected.

### 4 Who did what?

**Michael** Programming, some analysis.

**Måns** Simulation, analysis, writing.

**Henrik** Analysis, writing.

### 5 Achievements and suggestions

We got an introduction to the basic ideas of trace driven simulation. However, we don't think the project is big enough to be performed by four students. We suggest that the size of the project is increased or the size of the groups changed to two students.

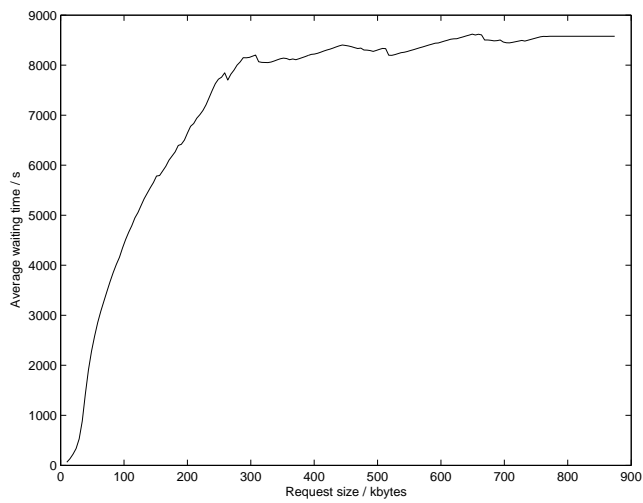


Figure 2: Average waiting time vs. request size