

# Performance evaluation of Web Cache

Pawan Kumar Choudhary and Kishor S. Trivedi  
Duke University, Durham, NC 27708

## 1 Web Cache Example in CSIM

Next we will consider an example of discrete event simulation implemented in CSIM. The example discusses the effect of Web caching on network planning in sense of determining the bandwidth of the access link interconnecting the ISP's subnet with the internet. This is studied by means of simulations [1]. The latency of a browser retrieving files is studied for given traffic characteristics, number of users, bandwidth of access link and cache hit ratio.

We divide our study into two cases. In the first case we assume the arrival process to be Poisson and service times to be generally distributed. In the second case we assume that the arrival process is an ON-OFF process, with sojourn time in ON state being Weibull distributed and in OFF state being Pareto distributed.

This model represents typical network infrastructure interconnecting a subnet with the internet. Browsers access the subnet via Modem, ISDN BRI, Ethernet or Token Ring. When a WWW browser clicks on a hyperlink, several URL requests may be sent from the browser to the web proxy which may be just one cache or a collection of caches. If the proxy has a copy of the requested file and is consistent with the original copy on the remote server, the proxy sends the copy to the browser. This is called a hit. If the proxy does not have a copy of the file that the browser is looking for, the proxy retrieves an original copy from the remote server, sends it to the browser and keeps a copy in the cache. This is called a miss.

In this example we are interested in *file delivering latency* or *mean response time*, which is defined as the time interval from the browser clicking an object to the requested object being displayed on the monitor. Excluding some trivial terms such as the delay of monitor display and cache retrieval and the HTTP interaction within the subnet, the mean response time consists of the following terms:

- Delay on the internet side  $T_1$ : This can be broken into three parts. The delay related with HTTP interaction in the internet, which consists of request sent from the proxy to the server, followed by a response sent back from the server to the proxy. The second term consists of delay on the remote server's subnet which depends on the network speed and load of the remote server. The third term consists of transmission delay and queuing delay on the internet which depends on the speed of routers the TCP connection traversing and the traffic load along the route. In general  $T_1$  is a random variable but for simplicity we assume that it is deterministic.
- $T_2$ : the time of the files sojourning on the subnet and browser's access line, which consists of the transmission delay. If the subnet has high speed network like Fast Ethernet or ATM Switch, the queuing delay in the subnet is negligible compared with the delay on the internet. If browsers use modems to access the ISP's subnet then

$$T_2 = 8Z/speed \quad (1)$$

where  $Z$  is the file length in bytes and  $speed$  is the speed of modem in terms of bits per second. Since file size is treated as a random variable,  $T_2$  is also a random variable.

- $T_3$ : the time of the file sojourning at the entrance of the subnet which consists of transmission delay. It depends on the speed of the access link and size of the file requested.  $T_3$  is given by

$$T_3 = 8(Z/bandwidth) \quad (2)$$

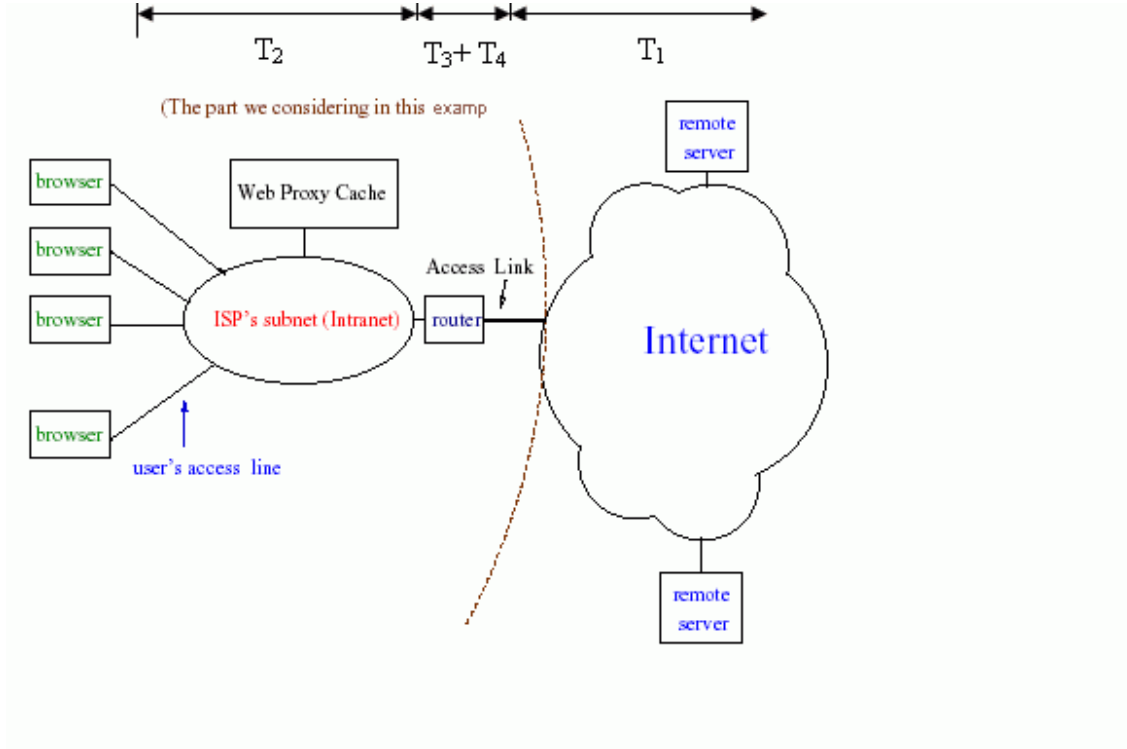


Figure 1: A typical Network Infrastructure Interconnecting an ISP's subnet(Intranet) with the Internet

where *bandwidth* is the bandwidth of the access link in terms of bits per second. Since  $Z$  is random variables,  $T_3$  also becomes random variable.

- $T_4$ : this constitutes the queuing delay due to presence of queue at the access link. Assuming buffer size at the entrance to be infinite, this is given by

$$T_4 = Q/\text{throughput} \quad (\text{queuing delay}) \quad (3)$$

where  $Q$  is the queue length of the entrance buffer in terms of bytes. Since  $Q$  is random variable so is  $T_4$ .

The total latency or response time is given by

$$R = \begin{cases} T_1 + T_2 + T_3 + T_4 & \text{if there is a cache miss;} \\ T_2 & \text{if there is a cache hit.} \end{cases} \quad (4)$$

### 1.1 Case 1

In first case the arrival process is assumed to be Poisson and service time is taken to be generally distributed. This forms M/G/1 queue. This assumption is very coarse and inaccurate, and we will relax this assumption in second case. In both cases three scenarios will be studied:

1. Bandwidth=256kb/s with no web cache
2. Bandwidth=256kb/s with web cache having 50% hit ratio
3. Bandwidth=512kb/s with no web cache

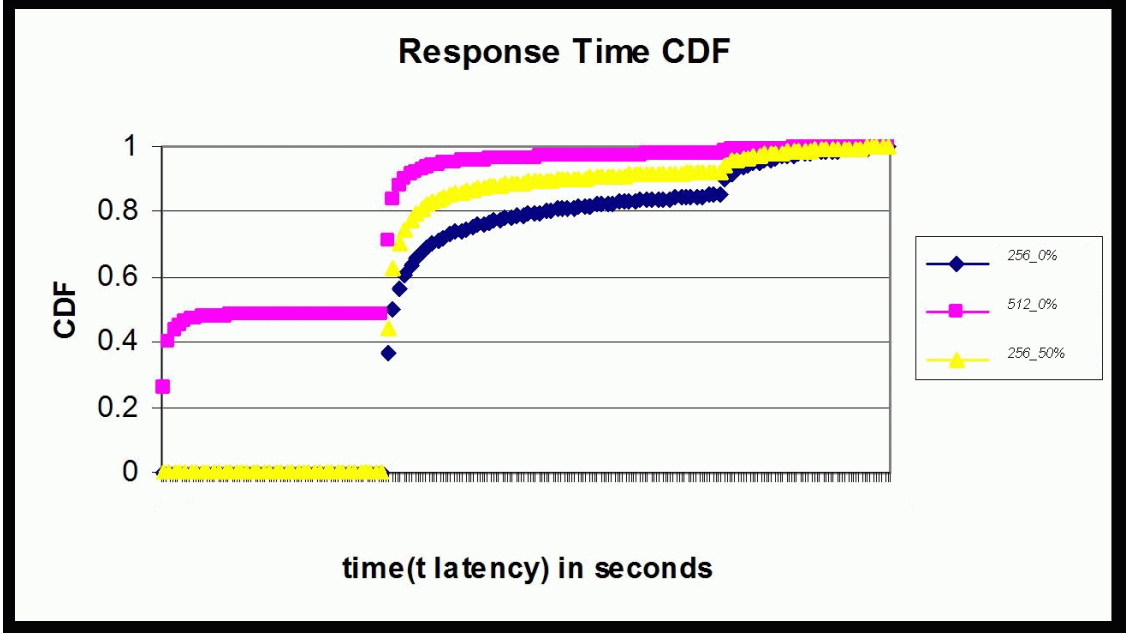


Figure 2: CDF of Response time for the three cases when arrival is Poisson process

Figure (2) compares the Cumulative Distribution function (CDF) for the response time  $R$ .

In this figure we see that two cases of bandwidth 512kb/s with no web cache and 256kb/s with web cache having hit ratio of 50% have nearly the same CDF. Infact doubling the bandwidth is not so effective as compared to having a web cache with respect to decreasing retrieval latency. Note that 500 browsers are logging on the internet simultaneously approaching the real situation.

Note that this forms an M/G/1 queue when there is a miss. So mean response time is given by

$$E[R_{overall}] = (1 - h)(E[T_1] + E[T_2] + E[T_3] + E[T_4]) + h(E[T_2]) \quad (5)$$

where  $h$  is the hit ratio.

$$E[R_{overall}] = (1 - h)(E[T_1] + E[T_2] + E[T_3] + E[R_{M/G/1}]) + h(E[T_2]) \quad (6)$$

Taking special case, we take  $E[T_2] = E[Z]/speed$  to be 4 seconds and delay on the internet side to be constant 4 seconds. The transmission delay on router side is taken to be 3 seconds. Also the queue on the access link (router) is taken M/D/1. The mean response time for an M/D/1 queue is given by

$$E[R]_{M/D/1} = \frac{1}{\lambda} \left( \rho + \frac{\rho^2}{2(1 - \rho)} \right) \quad (7)$$

where  $\lambda$  is average arrival rate and  $\rho = \lambda E[B]$ ,  $E[B]$  is average service time. Now if there is no web cache, it becomes special case of Equation 5 where  $h = 0$ . Using this formula and simulation in CSIM we calculated the mean response time and found that they are quite close. See Table (1). In this arrival rate  $\lambda$  has been taken to be 0.1 and service time is taken to be deterministic 3 seconds.

## 1.2 Case 2

In this case we will assume arrival process to be ON-OFF process. In this the sojourn time in ON period is determined by Weibull distribution. The ON periods are initiated by user's clicks on the hypertext links.

Table 1: Comparison of Analytical calculation vs. Simulation

	Analytical(Mean Response Time)	Simulation(Mean Response Time)
No Web Cache	14.5	14.82460
Web Cache 50% hit ratio,	9.275	9.4238

Table 2: Parameters for Simulation model for second case

ON Period	Weibull( $k = 0.9, \theta = e^{4.4}$ )
OFF Period	Pareto( $k = 60, \alpha = 0.5$ )
Interarrival Period	During ON Weibull( $k = 0.5, \theta = 1.5$ )
File Size	Pareto( $k = 300, \alpha = 1.3$ )

The ON period is found to follow Weibull distribution whose pdf is given by

$$p_{on}(x) = \frac{k}{\theta} \left(\frac{x}{\theta}\right)^{k-1} \exp\left(-\left(\frac{x}{\theta}\right)^k\right) \text{ with } k=0.77 \text{ to } 0.91 \text{ and } \theta=e^{4.4} \text{ to } e^{4.6} \quad (8)$$

During the OFF period, no request is generated. The duration of OFF period follows a Pareto distribution whose pdf is given by

$$p_{off}(x) = \frac{\alpha k^\alpha}{x^{\alpha+1}} \text{ with } \alpha=0.58 \text{ to } 0.9, k=60 \quad (9)$$

File size of the request is also Pareto distributed with  $\alpha=1.1$  to  $1.3$  and  $k$  is determined by mean length of files.

$$k = (\alpha - 1) \frac{E[file]}{\alpha} \quad (10)$$

In this equation (4) will be used to determine mean response time of the system.

For our model parameters from Table 2 has been taken

From this graph we can see that a Web Cache with 50% hit ratio is more effective than doubling the bandwidth. The simulation has been done in CSIM

## References

- [1] Hairong Sun, Xinyu Zang, and Kishor S. Trivedi. The effect of web caching on network planning, September 1999.

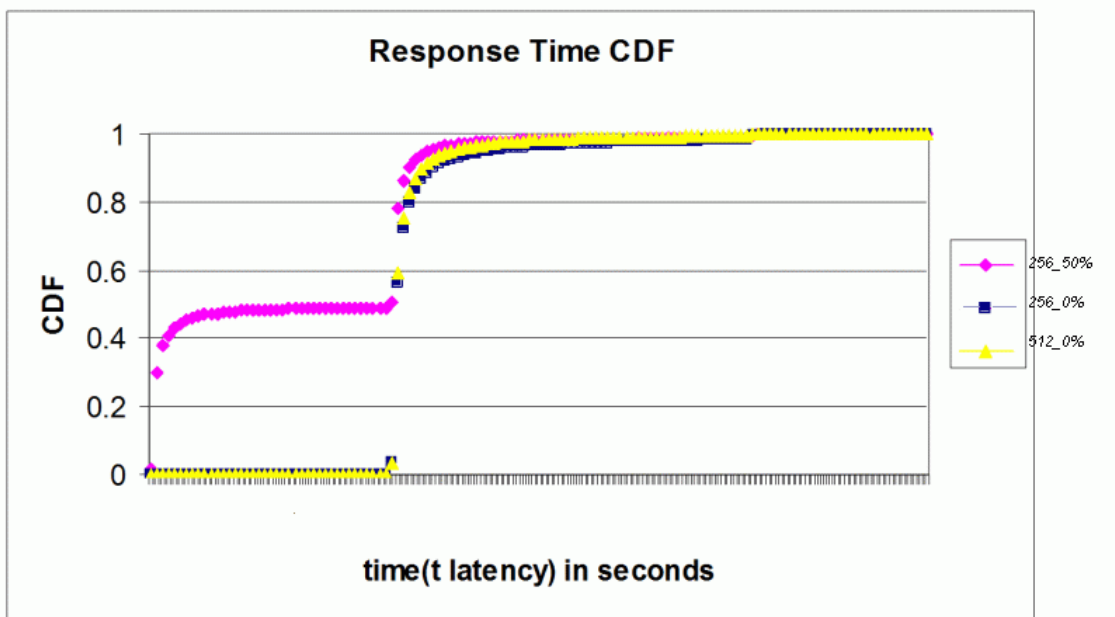


Figure 3: CDF of Response time for the three cases when arrival is ON-OFF process