A Probabilistic Approach for Admission Control to Web Servers

Zhengdao Xu Gregor v. Bochmann
School of Information Technology and Engineering
University of Ottawa, P.O.Box 450, Stn. A,
Ottawa, Ontario, Canada, K1N 6N5
zhengdao@cs.toronto.edu bochmann@site.uottawa.ca

Keywords: Admission control; Probabilistic admission control; Quality of Service; Performance oscillations; User differentiation
.

Abstract

We consider that some form of admission control must be applied to a Web server in order to avoid completely unacceptable response times during periods of heavy system overload. In this paper, we show that an on-off approach to admission control introduces oscillations in the server load, which may lead, in particular situations, to unacceptable variations in the server response time. In order to solve this problem, we propose a probabilistic approach to admission. A theoretical model and simulation studies show that these oscillations can be avoided with the probabilistic approach if suitable values are selected for its operating parameters. The most important parameters are the gradual nature of the probability function that determines the user acceptance probability, and the inter-observation period, which determines how frequently the response time of the server is determined. While the oscillations have in general only a small effect on the average response time, we showed that the probabilistic approach has a definite advantage for the distinction between different classes of users that have different priorities for accessing the server.

1. Introduction

With the expansion of the Internet infrastructure, many Internet-based applications have experienced extremely fast growth. Applications such as e-commerce and multimedia applications like video-on-demand require high processing power and/or the transmission of large amounts of data, which is a challenge when the number of users grows beyond expectations. In order to serve a growing user population, various architectures of server pools have been proposed and implemented.

One of the challenges is to provide a scalable system architecture that allows a flexible distribution of the user requests to the different servers in the pool. Among the different approaches to distribute the user requests to the available server resources (for a review, see [Boch 93a]), we consider in this paper a session-based management of user admission, as proposed in [Sale 01], in which a user is allocated to a given server when the user starts a new session with a given application. Several allocation algorithms are described in [Sale 01], which make such allocations based on performance data about the different servers that are obtained through performance monitoring. The objective of these algorithms is to balance the load among the different servers in order to obtain a uniform response time for all active users.

In the case that the number of servers cannot be sufficiently increased in order to adequately serve a growing user population, the system become overloaded. In such a situation, the server response time may increase to unacceptable values. Certain users will therefore abandon their session because of unsatisfactory system response time. In order to avoid an unsatisfied user population, it may be better if the service provider adopts a strategy by which in case of system overload, new user sessions will not be accepted. This means that some form of user admission control will limit the number of active users, and at the same time a maximum response time will be guaranteed to those users that have been accepted. In this spirit, Salem [Sale 02a] has proposed an admission control for differentiated classes of users: A-users have higher priority than B-users for being admitted to the service. This means that, if the response time of the system reaches a first limit, B-users will not be accepted any more while A-users are still admitted. If the response time further increases (because some many A-users want to use the application), then finally new A-users will also be refused when the response time reaches the maximum allowed value.

The admission control approaches mentioned above use an on-off control algorithm with a single threshold which works as follows: When the response time is below the threshold all new users will be accepted, if it is beyond that limit, no new user will be accepted. We note that this decision is based on the measured response time of the server (or the server pool), and this measured value is evaluated over a certain measurement period. We call the interval between two successive measurements the inter-observation time period. It is intuitively evident that, in the case of relatively long inter-observation time periods, the system performance will exhibit an oscillating behaviour, because when the last measurement showed a response time below the threshold all new users will be accepted within the subsequent inter-observation time period. This may lead to a much larger response time at the end of this period. If the next measured value is above the threshold, during the subsequence inter-observation time no new users will be accepted, while a certain number of the previously active users will terminate their session. Therefore the response time will decrease over the subsequent inter-observation time, and the measured value of the response time, at the end of that period, may well lie below the threshold; … and the new users will again be accepted leading to a higher response time, etc.

The purpose of this paper is to study this oscillation behaviour of Web server admission control algorithms in more detail and to determine how the oscillations can be reduced. For this purpose, the paper proposes a gradual probabilistic admission control, which eliminates the regular oscillations if its parameters are suitably selected. We show in this paper that the impact of these oscillations may have on the average response time of Web servers and how the average response time of Web servers can be limited using the probabilistic approach. However, in the case of differentiated user classes, as discussed in [Sale02a], we show that the distinction between the different user priorities is much more precise when the probabilistic approach to admission control is adopted. We note that our probabilistic admission control corresponds to “proportional control” which is one of the basic control paradigms known within control theory (see for instance in [Ogata]).

After describing in Section 2 the Web server performance simulation model used in this paper, we show some simulation results exhibiting the oscillatory behaviour in the case of on-off admission control. In Section 3 be provide a simplified theoretical model, which is helpful for the understanding of these oscillations. In Section 4 we study the probabilistic approach to admission control. After providing the definition of this approach, we modify the theoretical model in order to show that the probabilistic approach reduces the oscillations. After discussion the impact of the oscillations on the average response time, we present some of our simulation results, which characterize the behaviour of the probabilistic admission control algorithm in the context of stochastic user arrivals and leaving. We also characterize the system parameters and their impact on the reduction of oscillations. Finally we consider several differentiated classes of users with different server access priorities and show that the probabilistic approach has a definite advantage over the on-off approach considered earlier.

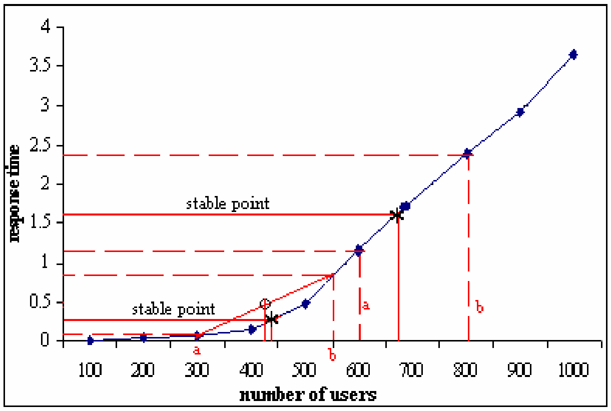
2. A Web server simulation model

Since the performance monitoring aspect and the observation time period is independent of the number of servers in a Web server pool, we consider in the following, for simplicity, that the server pool contains a single server. We have built a simulation model of a Web server, a Web application and a user population. It is assumed that requests for new user sessions arrive randomly, as a Poisson arrival process. Each user session involves a certain number of Web page requests, where the number of pages requested within one session follows an exponential distribution. We use the same model characteristics [Barf 98] which were also used in Salem’s papers, and which are shown in the following table:

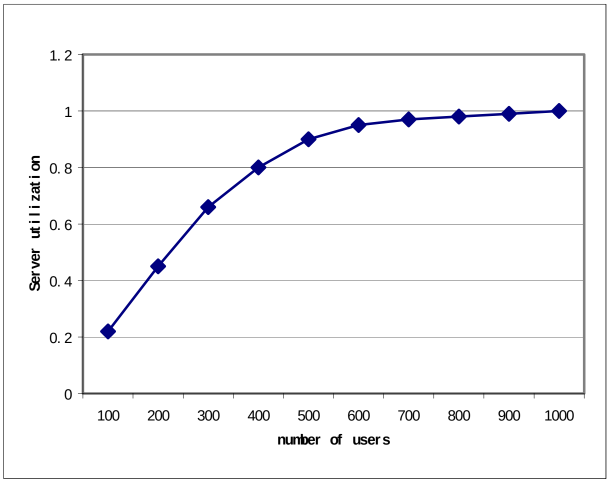
Parameter	Description
Server speed	10^{-6} second/byte
Inter-arrival time between the clients	exponential distribution with a certain mean
Number of pages each client requests	exponential distribution(mean = 36)
Number of embedded objects per page	Pareto distribution (α = 2.43, k=2.3)
Object size (in bytes)	Bounded Pareto distribution (α = 1.25, k = 1800, p =10^8)
Object processing time (in seconds)	Weibull distribution (α = 0.146, β=0.382)
User think time (in seconds)	Pareto distribution (α = 1.5, k=3)

Table 1 Parameters and descriptions of the simulation

By varying the inter-arrival time between the clients, we can vary the average number of users in the system and thereby the server utilization and the average response time. The dependence of the average response time and the average server utilization on the average number of active users in the system is shown in the following figures.



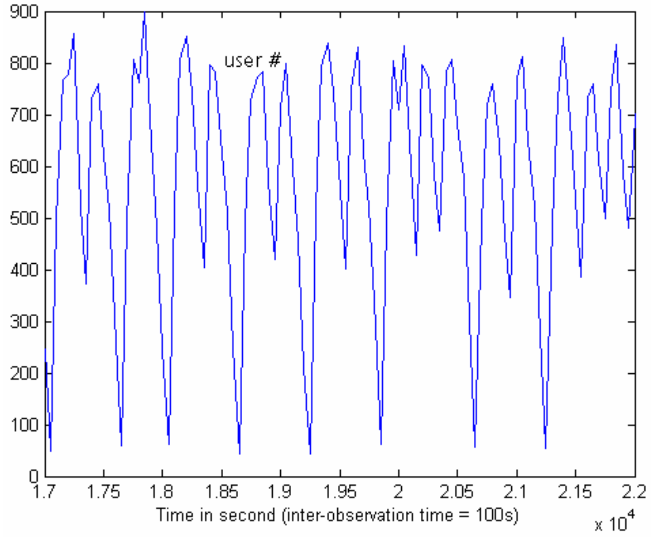
(a)



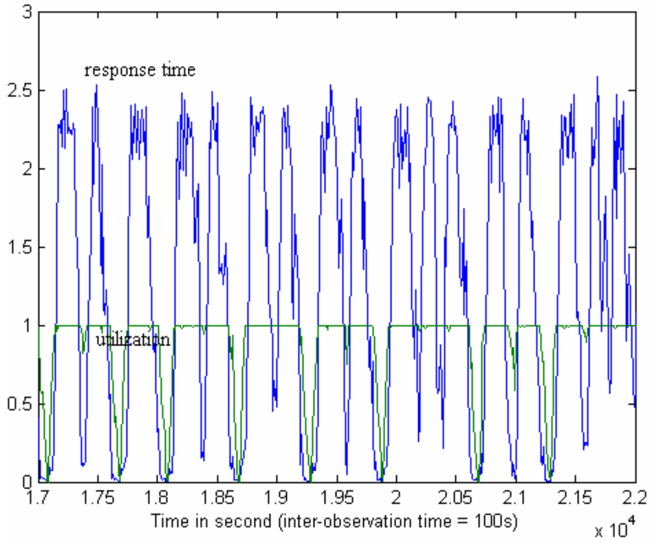
(b)

Figure 1: The response time (a) and server utilization (b) as a function of the number of users in the system

If we introduce into this simulated Web service system the on-off admission control described above, and use the response time threshold of 1.3 seconds, an inter-observation period of 100 seconds, and a client inter-arrival rate of 10 per second, we obtain an oscillatory behavior as shown in Figure 2.



(a)



(b)

Figure 2: Example of oscillations of the number of users (a), response time and server utilization (b)

The figure shows that as the number of users varies between 100 and 800, the response time varies between 0.1s and 2.5s, which is a quite unstable situation, and should better be avoided.

3. A theoretical model of server performance oscillations

To better understand the situation of the oscillations of a Web server system, let us consider a simplified theoretical model as follows. As in the simulation model described in Section 2, we assume that we have a random arrival of new clients following a Poisson process with a given average inter-arrival rate, which we abbreviate by r_a. After being accepted, we assume that the users remain active for a certain time period which (in contrast to the simulation model) is a random variable with an exponential distribution, that is, the probability that an active user terminates its session and leaves the system is a constant, which we abbreviate by p_l.

If we use y for the number of active users in the system, we can establish a differential equation, which governs the evolution of the number of users in the system. In the absence of admission control, we have the following differential equation:

$dy = r_a * dt – y *p_l \, dt$ or $dy / dt = – y *p_l + r_a$

where dy is the small change of the number of users within a small time interval dt. We note that we have ignored here the statistical fluctuations that are introduced due to the random arrival and leaving processes. However, this equation holds in the limiting case of a very large number of users where the statistical fluctuations become small compared with the total number of users in the system. If we solve this equation, we get the following formula which describes the number of active users in the system as a function of the time (where y_0 is the initial number of users at t = 0) :

$$y = f_{up}(t) = r_a/p_l + (y_0 - r_a/p_l)* e^{-p_l t}$$

To compute the number of users leaving the system when no new clients are accepted because of the admission control, we can use the same differential equation with no new arrivals (r_a = 0). We then get the following equation for the evolution of the number of active users (where again y_0 is the initial number of users at t = 0) :

$y = f_{down}(t) = y_0*e^{-Pl\ t}$

Equipped with these equations, we now address the problem of the oscillation of the number of users in the system when the on-off approach to admission control is used. We assume here that a threshold on the number of active users is given, and the actual number of active users is determined not continually, but only after a given inter-observation time period. If the so determined user number is below the threshold, all new clients will be accepted within the next inter-observation period; if it is above the threshold, no new client will be accepted during this period.

Figure 3 shows the oscillations for an example, using a threshold of 450 users, an inter-observation period T = 100 seconds, r_a = 10 users/second, and p_l = 0.01 per second. As shown in the figure, the number of users follows alternatively the model functions $f_{up}(t)$ and $f_{down}(t)$ and the switch-over occurs when the actual number of users becomes available at the end of each inter-observation period.

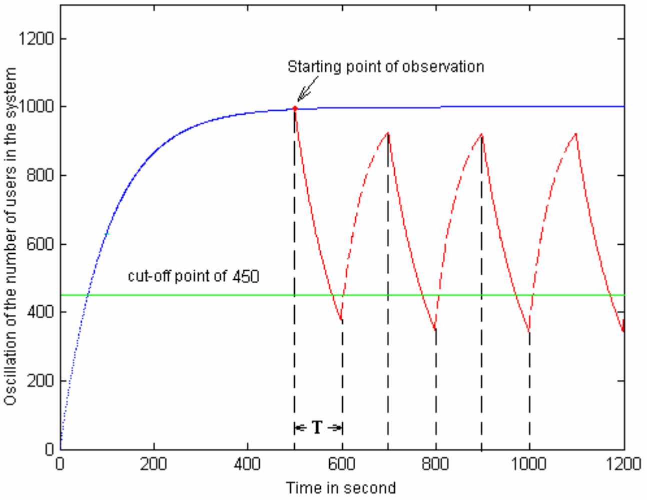


Figure 3. Theoretical model of oscillation

The simulation study also shows that the asymptotic amplitude of this oscillation (that is, the amplitude in the limit when time goes to infinity) does not depend on the starting point of the oscillation and can be calculated as

$$A = r_a/p_l * (1-e^{-p_l*T})/(1+e^{-p_l*T}) .$$

However, this formula only holds when the period of the oscillation is equal to twice the inter-observation period, as shown in the figure above. We note that longer oscillation period may occur when the threshold is very low (it may take several inter-observation periods before the number of users returns below the threshold) or very high (it may take more than one inter-observation period before the number of users gets above the threshold).

4. Probabilistic admission control

4.1. Probabilistic admission control: Definition

In our proposed probabilistic approach to admission control, to avoid system oscillation, a new client is admitted with a certain probability. This probability is equal to one when the server load is low, and it gradually becomes zero when the server load becomes very high. Since the server response time is what counts for the user, and since we assume that the server response time can be measured at regular intervals, we postulate that the acceptance probability for a new client is a function of the last measured response time of the server, which we abbreviate by r. In this paper, we assume that the function P is piecewise linear, and has the following form:

$$P_{a,b}(r) = \begin{cases} 1 \ (\text{if } r < a) \\ (r\text{-}a)/(b\text{-}a) \ (\text{if } a < r < b) \\ 0 \ (\text{if } r > b) \end{cases}$$

where a and b are two constants which indicate at which response time to start partial rejection and at which response time all users will be rejected, respectively.

The advantage of this approach over the on-off approach to admission control is that we can reject or accept user gradually. By choosing a proper inter-observation time, we hope to avoid the performance oscillations, as discussed in the following.

4.2. A theoretical model for oscillations with probabilistic admission control

For our theoretical investigation of probabilistic admission control, we use again a simplified model where the admission control decision is based on the number of users in the system, and not the response time. The only difference with the theoretical model described in Section 3 is that here we use a linear probability function $P_{a,\ b}$ (y) to control admission. In this context, a and b are numbers of users rather than response times. To model the number of users, we have the following differential equations:

$dy = r_a * P*dt – y *p_l\ dt$ for $y \in [a, b]$ ①

$dy = – y *p_l\ dt$ for $y > b$ ②

$dy = r_a *dt – y *p_l\ dt$ for $a > y$ ③

It is clear that the solutions for ② and ③ are $f_{down}(t)$ and $f_{up}(t)$ respectively, as discussed in Section 3. By solving the differential equation ①, we get

$$y = b*r_a/(r_a + p_l*(b\text{-}a)) + c * e^{-(r_a + p_l*(b\text{-}a))*t/(b\text{-}a)}$$

, and

$$c = (a\text{-}b*r_a/(r_a+p_l*(b\text{-}a)))\ /\ (1\text{-}a*p_l/r_a)^{(r_a/p_l/(b\text{-}a) + 1)}$$

.

From this solution, when t → ∞ , the second term in y will vanish, thus y reaches what we call stable point and is expressed as $y_{stable} = b*r_a/(r_a + p_l * (b\text{-}a))$. When the number of users y reaches the stable point, the oscillation will finally vanish. It is not difficult to proof that if $p_l < r_a$, $a < y_{stable} <b$, and this stable point is located between a and b depending on the workload. If the workload is really high ($r_a/p_l \to \infty$) , y_{stable} = b; the stable point will be at point b. As an example we show in Figure 4 the case where a = 200, b = 800 and the other settings are the same as before (r_a = 10, p_l = 0.01, inter-observation time T=100s). As shown in the figure, within 10 inter-observation time periods, the oscillation is totally reduced and the system reaches a stable point at $y_{stable} = b*r_a/(r_a + p_l * (b\text{-}a)) = 500$.

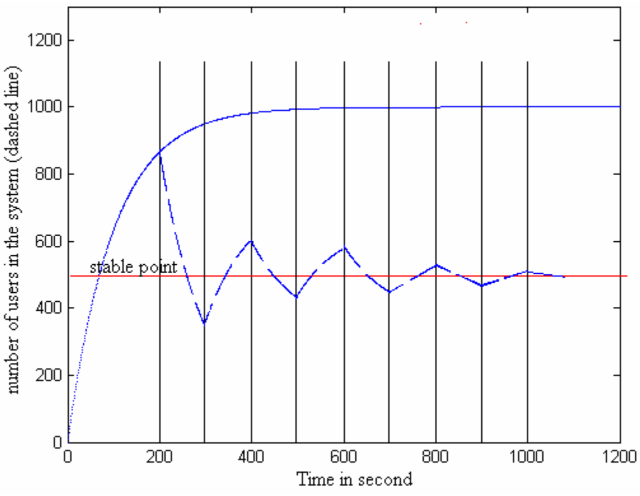


Figure 4. The oscillation gets stable at the stable point with P(200, 800)

4.3. The impact of oscillations on the average response time

Let us take a look at Figure 1(a) again. If we assume that with an on-off admission control the number of users oscillates between the extreme values of 300 and 550 users, we may as an alternative adopt probabilistic admission control with the values of a = 300 and b = 550; and as shown in the above figure, the oscillation will vanish and a stable point will be reached with y_{stable} = 440. The average response time in the case of probabilistic admission control will therefore equal the corresponding point on the curve of Figure 1(a), namely around 0.25 seconds. In the case of the on-off admission control, the response time will oscillate between the values corresponding to the number of user values 300 and 550, that is, between 0.1 and 0.9 seconds, as shown in the figure. If we assume that the number of users is either 300 or 550 and has not any intermediate value (which is not true), then the average response time in the case of on-off admission control will be the average of 0.1 and 0.9, that is 0.5 seconds. This is 0.25 seconds worse than the response time with probabilistic admission control. In reality, the advantage of the probabilistic admission control will be smaller, since the oscillations also cover intermediate user numbers between 300 and 550.

We conclude that the impact of the oscillations on the average response time is normally not very big. In fact, it becomes negligible if the range of the oscillation covers only the linear part of the curve of Figure 1(a), for instance between the number of users 600 and 800, as indicated on the figure.

4.4. Simulation study of oscillations with probabilistic admission control

We have studied the oscillations occurring with probabilistic admission control within the Web server model described in Section 2. The main conclusion is that although the nature of the oscillations follows in general the characteristics of the theoretic model described above, it is important to note that the oscillations will not disappear completely, as indicated in the asymptotic behavior of the theoretical model shown in Figure 4. In fact, certain amplitude of oscillation remains which is presumably introduced by the stochastic nature of the arrival and leaving of users. The amplitude of this remaining oscillation decreases with smaller inter-observation periods and with more gradual probability functions. Finally, it has no recognizable oscillation period any more, which means that the oscillations become random, reflecting the stochastic nature of the user arrival and leaving pattern.

In the following we show some results of our simulations with three different probability functions $P_{0.1,\ 2.5}(r)$, $P_{0.7,\ 1.9}(r)$ and $P_{1.3,\ 1.3}(r)$, which are shown in the Figure 5. Notice that function $P_{1.3,\ 1.3}(r)$ represents exactly the on-off approach, as used in [Sale 02a].

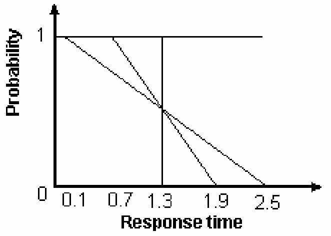


Figure 5. Probability functions

We measured the period and amplitude of the oscillations of the number of users in the system for different probability functions with different inter-observation times when the inter-arrival time equals 0.1seconds, as shown in Figure 6, 7:

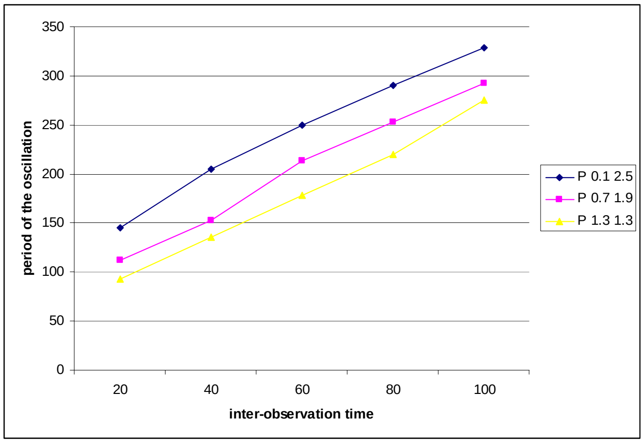


Figure 6. The period of the oscillations for different probability functions

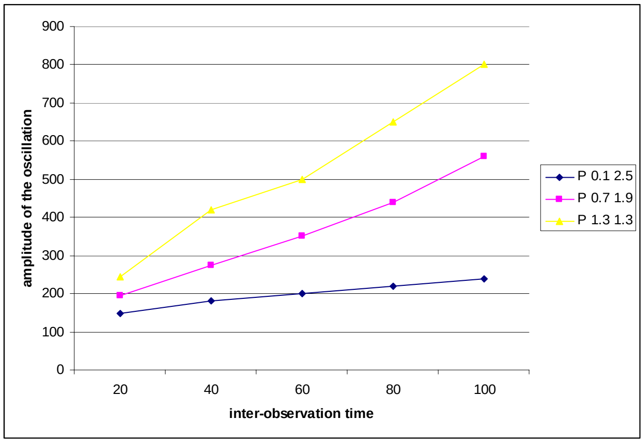


Figure 7. The amplitude of the oscillations for different probability functions

We can see from the above figures, for a given inter-observation time period, the more gradual the probability function is (like $P_{0.1, 2.5}(r)$) the larger the period of the oscillation will be. A more gradual probability function also leads to smaller amplitudes, that is, a less severe oscillation. This means that the oscillation is somewhat dampened out by using a gradual probability function.

Furthermore, for a given probability function, the period and the amplitude of the oscillation depend on the inter-observation period. The smaller the inter-observation period, the smaller the amplitude and the smaller the period of the oscillation.

5. Probabilistic admission control for differentiated user classes

As mentioned in the introduction, one may consider different classes of users with different priorities for access to the server. Salem [Sale 02a] considered such a model. The results of his simulations using an on-off approach to access control showed that when the response time increases and surpasses the threshold applying to the low-priority B-users, these B-users are not completely refused from accessing the system. This is presumably due to the statistical (and possibly oscillation) nature of the response time values obtained after the consecutive inter-observation periods. As we will show below, the use of the probabilistic admission control will effectively improve the discrimination between class A and B users. The reason for this improvement is presumably the reduction of the oscillations due to the probabilistic approach.

Here we consider performance of our probabilistic approach for two differentiated user groups, user group A with higher priority and user group B with lower priority. We use two different probability functions for the two groups:

A-Users:	$P_{1.05, 1.55}(r)$
B-Users:	$P_{0.55, 1.05}(r)$

Since group A are the users with higher priority, also called “elite group”, they should have access priority over group B users. Therefore we start to reject A-users only after the response time exceeds 1.05s, at which moment all requests from B-users have already been turned down.

To compare the probabilistic approach with the on-off approach using two different thresholds for the two groups (0.8s for group B and 1.3s for group A), we test the user acceptance percentage over a variety of customer arrival rates (inter-arrival time ranging from 0.7s to 0.05s, and equal number of A and B users). Using an inter-observation time of 10s in order to reduce the oscillations, we obtain the result shown in Figure 8.

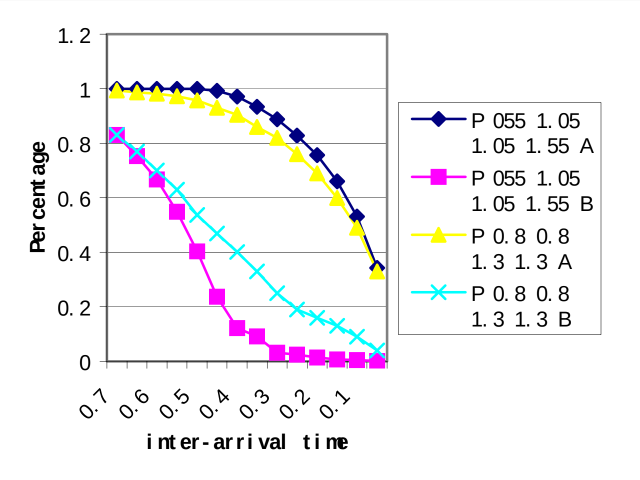


Figure 8. Percentage of users accepted

The probabilistic approach (identified by the probability function P 0.55 1.05 1.05 1.55, meaning $P_{1.05, 1.55}(r)$ for A-user and $P_{0.55, 1.05}(r)$ for B-user) enforces a clear privilege of A-users over B-users in terms of the acceptance probability (percentage). While both groups enjoy the same response time provided by the system, the probability of accepting A-users is substantially higher than for B-users, especially for higher user arrival rates. For on-off admission control (identified by the probability function P 0.8 0.8 1.3 1.3), the priority of A-users over B-users is not so clearly identified. This must be due to the oscillation of the response time, which is induced by the on-off approach, which undermines the service differentiation between the two groups of users.

We have run the same simulation for the probabilistic approach with various inter-observation times (10s, 60s, 100s) as shown in Figure 9. We find that the priority of the A-users over B-user is more clearly exhibited for shorter inter-observation periods. This is just as expected, since a shorter inter-observation period means a closer monitoring of the system, thus smaller oscillations.

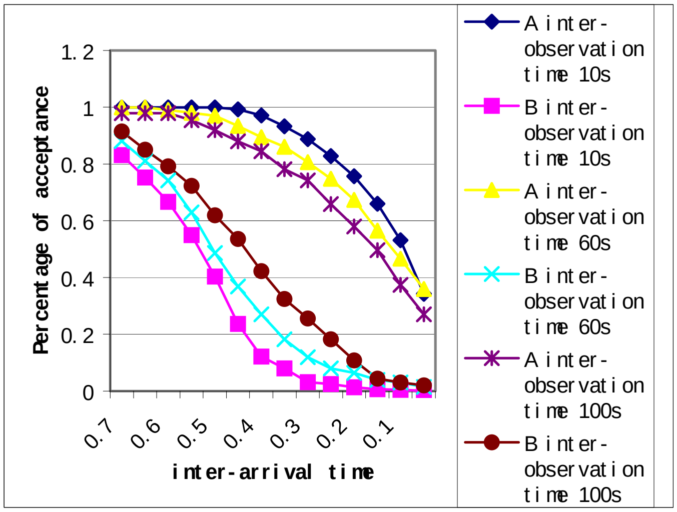


Figure 9. Percentage of users accepted with probabilistic approach for different inter-observation time

6. Conclusions

We consider in this paper that some form of admission control must be applied to a Web server (or any other kind of server for that matter) in order to avoid completely unacceptable response times during periods of heavy system overload. We have shown that an on-off approach to admission control introduces oscillations in the server load, which may lead, in particular situations, to unacceptable variations in the server response time. In order to solve this problem, we proposed a probabilistic approach to the admission. Our theoretical investigations and simulation studies have shown that these oscillations can be avoided with the probabilistic approach if suitable values are selected for its operating parameters. The most important parameters are the gradual nature of the probability function that determines the user acceptance probability, and the inter-observation period, which determines how frequently the response time of the server is determined. While the oscillations have in general only a small effect on the average response time, we showed that the probabilistic approach has a definite advantage for the distinction between different classes of users that have different priorities for accessing the server.

7. References

[Sale 01] M.-V. Mohamed-Salem, J. W. Wong and G. v. Bochmann, *A scalable load-sharing architecture for distributed applications*, Proc. 9th IEEE Conference on Software, Telecommunications and Computer Networks, SoftCom 2001, October 2001, pp. 747-755.

[Sale 02a] M.-V. M. Salem, G. v. Bochmann and J. W. Wong, *Server selection for differentiated classes of users*, Int. Symp. on Performance Evaluation of Computer and Telecommunication Systems (SPECTS 2002), San Diego, July 2002, pp.

[Ogata] Katsuhiko Ogata, "Modern Control Engineering", Third Edition Prentice Hall 1997

[Barf 98] Paul Barford and Marck Crovella, "Generating Representative Web Workload for Network and Server Performance Evaluation", in Proceeding of the 1998 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, pp. 151-160, July 1998.

[Boch 03a] G. v. Bochmann, J. W. Wong, T. C. Lau, D. Bourne, D. Evans, B. Kerhervé, M. V. Salem and H. Ye, *Scalability of Web-based electronic commerce systems*, IEEE Communications Magazine, July 2003.

Biography

Zhengdao Xu He received his B.S. degree in the Department of Computer Science from the Zhejiang University, China, and his M.S. degree in the School of Information Technology and Engineering (SITE) from the University of Ottawa. He is now a Ph.D student in the Department of Computer Science, University of Toronto. His current research interests include the middlware systems and pervasive computing.

Gregor v. Bochmann He is a professor at the School of Information Technology and Engineering at the University of Ottawa since January 1998. Previously, he was professor at the University of Montreal for 25 years. He is a fellow of the IEEE and ACM and a member of the Royal Society of Canada. His present work is aimed at methodologies for the design, implementation and testing of communication protocols and distributed systems.