J. Appl. Prob. **48,** 453–466 (2011)
Printed in England
© *Applied Probability Trust* 2011

A STOCHASTIC BREAKDOWN MODEL FOR AN UNRELIABLE WEB SERVER SYSTEM AND AN OPTIMAL ADMISSION CONTROL POLICY

JI HWAN CHA,* *Ewha Womans University*
EUI YONG LEE,** *Sookmyung Women's University*

Abstract

Web servers have to be protected against overload since overload can lead to a server breakdown, which in turn causes high response times and low throughput. In this paper, a stochastic model for breakdowns of server systems due to overload is proposed and an admission control policy which protects Web servers by controlling the amount and rate of work entering the system is studied. Requests from the clients arrive at the server following a nonhomogeneous Poisson process and each requested job takes a random time to be completed. It is assumed that the breakdown rate of the server depends on the number of jobs which are currently being performed by the server. Based on the proposed model, the reliability function and the breakdown rate function of the server system are derived. Furthermore, the long-run expected number of jobs completed per unit time is derived as the efficiency measure, and the optimal admission control policy which maximizes the efficiency will be discussed.

Keywords: Unreliable server; breakdown rate; reliability function; efficiency; optimal admission control policy

2010 Mathematics Subject Classification: Primary 60K10
Secondary 62P30

1. Introduction

The widespread use of Web technologies has already made the Internet an essential channel for mass distribution of information. Nowadays, a popular website can receive more than ten million requests per day, with normal request rates of 12 000 per minute (see Wessels (2001, Chapter 1)). Web servers and Web clients are the fundamental architectural building blocks in the World Wide Web. A Web client is a requester of data (content) and a Web server is the provider of data. A Web server manages and provides the data source while Web browsers send requests to a Web server for specific source data by means of a URL. Upon receipt of a request initiated by a Web client, the Web server then processes the request and sends a response back to the Web client.

The increasing number of Internet users and innovative new services such as e-commerce are placing new demands on Web servers. For example, the increasing growth of e-commerce on the Web means that any server breakdown time that affects the clients being served will result in a corresponding loss of revenue. Thus, it is becoming essential for Web servers to provide steady and stable services in addition to high speed responses. To prevent server overload, the

Received 8 October 2010; revision received 9 February 2011.

* Postal address: Department of Statistics, Ewha Womans University, Seoul, 120-750, Korea.
Email address: jhcha@ewha.ac.kr

** Postal address: Department of Statistics, Sookmyung Women's University, Seoul, 140-742, Korea.
Email address: eylee@sookmyung.ac.kr

https://doi.org/10.1239/jap/1308662638 Published online by Cambridge University Press

amount of work entering the Web server should be controlled. For example, HP's WebQoS triggers rejection of requests once the server starts to be overloaded (see Bhatti and Friedrich (1999)).

In this paper we propose a stochastic breakdown model for a Web server system and discuss the optimal admission control policy which protects Web servers from frequent breakdowns by controlling the amount and rate of work entering the system. Requests from the clients arrive at the server following a nonhomogeneous Poisson process and each requested job takes a random time to be completed. It is assumed that the breakdown rate of the server depends on the number of jobs which is currently being performed by the server. Thus, under the proposed model, the instantaneous susceptibility to breakdown of the server changes in time depending on the number of jobs currently being processed. Once the demand rate exceeds a certain level, to protect the server from high client loads, some requests must be rejected. The optimal rejection rate which maximizes the efficiency of the Web server is also discussed.

Recently, new topics and studies on computers and computer-related systems based on dynamic stochastic modellings are increasingly appearing in the field of applied probability. For example, Agustin and Peña (1999) and Kvam and Peña (2005) developed dynamic reliability models which generalize the famous software reliability model suggested in Jelinski and Moranda (1972). Boland and Ni Chuív (2007) and Barghout (2010) studied software reliability models with imperfect repair/debugging. A list of several related papers also includes Singpurwalla (1991), Boland and Singh (2002), Boland *et al.* (2002), and Agustin (2003). The topic in this study will also extend the range of research in this field (research on computer and computer-related systems) to a new challenging direction.

This paper is organized as follows. In Section 2, a stochastic breakdown model which represents the operating characteristic of the Web server is proposed. Based on it, the distribution function of the time to breakdown and the breakdown rate function of the server are derived. In Section 3, as the efficiency measure of the Web server system, we derive the long-run expected number of jobs completed per unit time. In Section 4, the optimal admission control policy which maximizes the efficiency measure by controlling the amount and rate of work entering the system is discussed. Finally, in Section 5, some concluding remarks are discussed.

2. Stochastic model and breakdown rate function

In this section we consider the stochastic modelling of the Web server breakdowns without consideration of an admission control policy. Let X be the random time to the server breakdown when the client demand rate is given by 0. Thus, server breakdowns in this case may include those caused by technical difficulties, e.g. malfunctions of hardware subsystems or software errors, etc., independently of those caused by the workload. Denote its absolutely continuous cumulative distribution function (CDF) and the corresponding survival function (SF) by $F_0(t) \equiv \mathrm{P}(X \leq t)$ and $\bar{F}_0(t)$, respectively. The *baseline breakdown rate* of the server is then defined by

$$r_0(t) \equiv \frac{f_0(t)}{\bar{F}_0(t)}, \tag{1}$$

where $f_0(t)$ is the probability density function (PDF) of X. Thus, the breakdown rate defined in (1) corresponds to the ordinary failure rate function in reliability theory. Let $\{N(t),\ t \geq 0\}$ be the stochastic counting process describing the arrivals of the client requests in $[0, \infty)$. Denote the arrival times of these client requests as $0 \equiv T_0 < T_1 < T_2 < \cdots$. Upon receipt of a request initiated by a Web client, the Web server then starts to process the request and it takes a random

https://doi.org/10.1239/jap/1308662638 Published online by Cambridge University Press

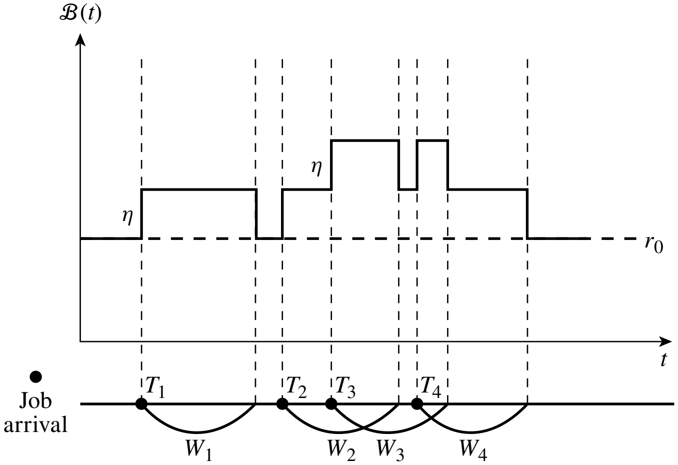

FIGURE 1: The operating characteristics and breakdown rate process.

time for the requested job to be completed. Let $W_j,\ j = 1, 2, \ldots$, be the time to the completion of the jth requested job; we assume that the W_js are independent and identically distributed (i.i.d.) random variables with their common CDF $G_W(w)$, SF $\bar{G}_W(w)$, and corresponding PDF $g_W(w)$. The number of requests that the Web server can process simultaneously is assumed to be sufficiently large.

In this study we assume that, during the process of each requested job, the server breakdown rate is additionally increased by a fixed amount η and, as soon as the corresponding requested job is completed, the increased breakdown rate then vanishes. Thus, 'the breakdown rate process' of the server *dynamically responds* to the flow of jobs being processed. The operating characteristics of the Web server and the corresponding breakdown rate process is depicted in Figure 1 when the baseline breakdown rate is given by a constant r_0.

As shown in Figure 1, the breakdown rate process $\mathcal{B}(t)$ can be mathematically defined as

$$\mathcal{B}(t) = r_0(t) + \eta \sum_{j=1}^{N(t)} \mathbf{1}(T_j < t \le T_j + W_j), \qquad t \ge 0,$$

where $\mathcal{B}(t)$ is the level of breakdown rate at time t. Let Y be the random time to the breakdown of the Web server given the workload caused by the requests from Web clients. Then, given the arrival process of the client requests and the job processing times, the conditional survival function of Y is given by

$$\begin{aligned} &\mathrm{P}(Y > t \mid N(t), T_1, T_2, \ldots, T_{N(t)}; W_1, W_2, \ldots, W_{N(t)}) \\ &\quad = \exp\biggl\{-\int_0^t \mathcal{B}(x)\,\mathrm{d}x\biggr\} \\ &\quad = \exp\biggl\{-\int_0^t r_0(x)\,\mathrm{d}x - \eta \sum_{j=1}^{N(t)} \min\{W_j, (t - T_j)\}\biggr\} \\ &\quad = \bar{F}_0(t) \exp\biggl\{-\eta \sum_{j=1}^{N(t)} \min\{W_j, (t - T_j)\}\biggr\}. \end{aligned} \tag{2}$$

https://doi.org/10.1239/jap/1308662638 Published online by Cambridge University Press

The following result gives the survival function and the breakdown rate function of Y.

Theorem 1. *Suppose that $\{N(t),\ t \geq 0\}$ is a nonhomogeneous Poisson process with intensity function $\lambda(t) \geq 0$, i.e. $m(t) \equiv \mathrm{E}[N(t)] = \int_0^t \lambda(x)\,\mathrm{d}x$. Assuming that (2) holds and that $m(t)$ has inverse $m^{-1}(t)$, the survival function of Y is given by*

$$\mathrm{P}(Y > t) = \bar{F}_0(t)\exp\biggl\{-\eta\int_0^t \exp\{-\eta w\}\bar{G}_W(w)m(t-w)\,\mathrm{d}w\biggr\}$$

and the breakdown rate function of Y, denoted as $r(t)$, is given by

$$r(t) = r_0(t) + \eta\int_0^t \exp\{-\eta w\}\bar{G}_W(w)\lambda(t-w)\,\mathrm{d}w.$$

Proof. The unconditional survival function can be obtained by

$$\mathrm{P}(Y > t) = \bar{F}_0(t)\,\mathrm{E}\biggl[\exp\biggl\{-\eta\sum_{j=1}^{N(t)}\min\{W_j, (t-T_j)\}\biggr\}\biggr].$$

Here, the expectation will be obtained by

$$\mathrm{E}\biggl[\exp\biggl\{-\eta\sum_{j=1}^{N(t)}\min\{W_j, (t-T_j)\}\biggr\}\biggr] = \mathrm{E}\biggl[\mathrm{E}\biggl[\exp\biggl\{-\eta\sum_{j=1}^{N(t)}\min\{W_j, (t-T_j)\}\biggr\}\biggm| N(t)\biggr]\biggr].$$

Observe that the joint distribution of $T_1, T_2, \dots, T_n$, given $N(t) = n$, is the same as the joint distribution of order statistics $T'_{(1)} \leq T'_{(2)} \leq \cdots \leq T'_{(n)}$ of i.i.d. random variables $T'_1, T'_2, \dots, T'_n$, where the PDF of the common distribution of the T'_js is given by $\lambda(x)/m(t),\ 0 \leq x \leq t$:

$$(T_1, T_2, \dots, T_n \mid N(t) = n) \overset{\mathrm{D}}{=} (T'_{(1)}, T'_{(2)}, \dots, T'_{(n)}).$$

That is, let $f_{T_1,T_2,\dots,T_n|N(t)}(t_1, t_2, \dots, t_n \mid n)$ and $f_{T'_{(1)},T'_{(2)},\dots,T'_{(n)}}(t_1, t_2, \dots, t_n)$ be the conditional joint PDF of $T_1, T_2, \dots, T_n$ given $N(t) = n$ and the joint PDF of $T'_{(1)}, T'_{(2)}, \dots, T'_{(n)}$, respectively. Then

$$\begin{aligned} f_{T_1,T_2,\dots,T_n\,|\,N(t)}(t_1, t_2, \dots, t_n \mid n) &= f_{T'_{(1)},T'_{(2)},\dots,T'_{(n)}}(t_1, t_2, \dots, t_n) \\ &= n!\prod_{i=1}^{n}\frac{\lambda(t_i)}{m(t)}, \qquad 0 \leq t_1 \leq t_2 \leq \cdots \leq t_n \leq t. \end{aligned}$$

Thus,

$$\begin{aligned} &\mathrm{E}\biggl[\exp\biggl\{-\eta\sum_{j=1}^{N(t)}\min\{W_j, (t-T_j)\}\biggr\}\biggm| N(t) = n\biggr] \\ &\quad= \mathrm{E}\biggl[\exp\biggl\{-\eta\sum_{j=1}^{n}\min\{W_j, (t-T'_{(j)})\}\biggr\}\biggr] \\ &\quad= \mathrm{E}\biggl[\exp\biggl\{-\eta\sum_{j=1}^{n}\min\{W_j, (t-T'_j)\}\biggr\}\biggr] \\ &\quad= (\mathrm{E}[\exp\{-\eta\min\{W_1, (t-T'_1)\}\}])^n, \end{aligned} \tag{3}$$

https://doi.org/10.1239/jap/1308662638 Published online by Cambridge University Press

where the second and third equalities hold because $\{W_j,\ j = 1, 2, \dots, n\}$ and $\{T'_j,\ j = 1, 2, \dots, n\}$ are i.i.d. random variables, respectively, and they are independent. In (3), for $w \le t$,

$$
\begin{aligned}
&\mathrm{E}[\exp\{-\eta \min\{W_1, (t - T'_1)\}\} \mid W_1 = w] \\
&\quad = \int_0^{t-w} \exp\{-\eta w\} \frac{\lambda(x)}{m(t)}\,\mathrm{d}x + \int_{t-w}^{t} \exp\{-\eta(t - x)\} \frac{\lambda(x)}{m(t)}\,\mathrm{d}x \\
&\quad = \exp\{-\eta w\} \frac{m(t - w)}{m(t)} + \exp\{-\eta t\} \int_{t-w}^{t} \exp\{\eta x\} \frac{\lambda(x)}{m(t)}\,\mathrm{d}x,
\end{aligned}
$$

and, for $w > t$,

$$
\mathrm{E}[\exp\{-\eta \min\{W_1, (t - T'_1)\}\} \mid W_1 = w] = \exp\{-\eta t\} \int_0^{t} \exp\{\eta x\} \frac{\lambda(x)}{m(t)}\,\mathrm{d}x.
$$

Then

$$
\begin{aligned}
&\mathrm{E}[\exp\{-\eta \min\{W_1, (t - T'_1)\}\}] \\
&\quad = \mathrm{E}[\mathrm{E}[\exp\{-\eta \min\{W_1, (t - T'_1)\}\} \mid W_1]] \\
&\quad = \frac{1}{m(t)} \bigg(\int_0^t \exp\{-\eta w\} m(t - w) g_W(w)\,\mathrm{d}w \\
&\qquad\qquad + \exp\{-\eta t\} \int_0^t \int_{t-w}^{t} \exp\{\eta x\} \lambda(x)\,\mathrm{d}x\, g_W(w)\,\mathrm{d}w \\
&\qquad\qquad + \bar{G}_W(t) \exp\{-\eta t\} \int_0^t \exp\{\eta x\} \lambda(x)\,\mathrm{d}x \bigg) \\
&\quad = \frac{1}{m(t)} \bigg(\int_0^t \exp\{-\eta w\} m(t - w) g_W(w)\,\mathrm{d}w \\
&\qquad\qquad + \exp\{-\eta t\} \int_0^t \exp\{\eta x\} \lambda(x) \int_{t-x}^{t} g_W(w)\,\mathrm{d}w\,\mathrm{d}x \\
&\qquad\qquad + \bar{G}_W(t) \exp\{-\eta t\} \int_0^t \exp\{\eta x\} \lambda(x)\,\mathrm{d}x \bigg) \\
&\quad = \frac{1}{m(t)} \bigg(\int_0^t \exp\{-\eta w\} m(t - w) g_W(w)\,\mathrm{d}w \\
&\qquad\qquad + \exp\{-\eta t\} \int_0^t \exp\{\eta x\} \lambda(x) \bar{G}_W(t - x)\,\mathrm{d}x \bigg) \\
&\quad = \frac{1}{m(t)} \bigg(\int_0^t \exp\{-\eta w\} m(t - w) g_W(w)\,\mathrm{d}w + \int_0^t \exp\{-\eta w\} \lambda(t - w) \bar{G}_W(w)\,\mathrm{d}w \bigg) \\
&\quad = \frac{1}{m(t)} \bigg(\Big[-\exp\{-\eta w\} m(t - w) \bar{G}_W(w) \Big]_0^t \\
&\qquad\qquad - \int_0^t [\eta \exp\{-\eta w\} m(t - w) + \exp\{-\eta w\} \lambda(t - w)] \bar{G}_W(w)\,\mathrm{d}w \\
&\qquad\qquad + \int_0^t \exp\{-\eta w\} \lambda(t - w) \bar{G}_W(w)\,\mathrm{d}w \bigg) \\
&\quad = \frac{1}{m(t)} \bigg(m(t) - \eta \int_0^t \exp\{-\eta w\} m(t - w) \bar{G}_W(w)\,\mathrm{d}w \bigg).
\end{aligned}
$$

https://doi.org/10.1239/jap/1308662638 Published online by Cambridge University Press

Thus, from (3) we have

$$\begin{aligned}
&\mathrm{E}\bigg[\exp\bigg\{-\eta\sum_{j=1}^{N(t)}\min\{W_j,(t-T_j)\}\bigg\}\bigg]\\
&\quad=\sum_{n=0}^{\infty}\bigg\{\frac{1}{m(t)}\bigg(m(t)-\eta\int_0^t\exp\{-\eta w\}\bar{G}_W(w)m(t-w)\,\mathrm{d}w\bigg)\bigg\}^n\frac{(m(t))^n}{n!}\exp\{-m(t)\}\\
&\quad=\exp\bigg\{-\eta\int_0^t\exp\{-\eta w\}\bar{G}_W(w)m(t-w)\,\mathrm{d}w\bigg\}.
\end{aligned}$$

Therefore,

$$\mathrm{P}(Y>t)=\bar{F}_0(t)\exp\bigg\{-\eta\int_0^t\exp\{-\eta w\}\bar{G}_W(w)m(t-w)\,\mathrm{d}w\bigg\}. \tag{4}$$

From (4),

$$\ln\mathrm{P}(Y>t)=-\int_0^t r_0(x)\,\mathrm{d}x-\eta\int_0^t\exp\{-\eta w\}\bar{G}_W(w)m(t-w)\,\mathrm{d}w.$$

Now applying Leibnitz's rule (see, e.g. Casella and Berger (2002, p. 69)), the compound failure rate function $r(t)$ is thus given by

$$r(t)=-\frac{\mathrm{d}}{\mathrm{d}t}\ln\mathrm{P}(Y>t)=r_0(t)+\eta\int_0^t\exp\{-\eta w\}\bar{G}_W(w)\lambda(t-w)\,\mathrm{d}w.$$

3. The efficiency measure of the Web server

During the operation of the Web server system, if the server breaks down then it is rebooted. Here, after 'rebooting' of the server, the physical state of the server system is assumed to be 'as good as new'. For example, after fixing the malfunctions in subsystems or in installed software, the performance of the information technology (IT) systems (e.g. computers, servers, etc.) could be regarded as the same as before. Furthermore, we assume that the arrival process of the client requests after rebooting, $\{N^*(t),\ t\geq 0\}$, also 'restarts'. Here, the conditions under which the arrival process restarts are (i) the new arrival process $\{N^*(t),\ t\geq 0\}$ after rebooting is a nonhomogeneous Poisson process with the same intensity function $\lambda(t),\ t\geq 0$, and (ii) the arrival process of the client requests $\{N^*(t),\ t\geq 0\}$ is independent of those before rebooting. Physically, this assumption implies that the stochastic pattern of the arrival of client requests is independently repeated in the same manner after each rebooting. If, for example, the intensity is increasing, under this assumption, the increasing pattern is repeated after each rebooting (see case II of Example 1 below). Note that the assumption for 'the restart of the arrival process' is automatically satisfied when the process is a homogeneous Poisson process, i.e. when the intensity function is given by a constant. Thus, the Web server system and its operating characteristics are 'renewed' on each rebooting. The time needed for rebooting of the server is assumed to follow a continuous distribution $H(t)$ with its mean ν.

Let $M(t)$ be the total number of jobs completed by the Web server during $(0,t]$. Then, as the measure of the performance of the Web server, we define the long-run expected number of jobs competed per unit time:

$$\psi\equiv\lim_{t\to\infty}\frac{\mathrm{E}[M(t)]}{t}.$$

https://doi.org/10.1239/jap/1308662638 Published online by Cambridge University Press

In the following, we will call ψ 'the efficiency of the server'. Then, by renewal theory (see, e.g. Ross (1996, Section 3.6)),

$$\psi = \lim_{t\to\infty} \frac{\mathrm{E}[M(t)]}{t} = \frac{\mathrm{E}[M]}{\mathrm{E}[Y]+\nu},$$

where, $\mathrm{E}[Y] = \int_0^\infty \mathrm{P}(Y > t)\,\mathrm{d}t$ and M is the number of jobs completed in the selected renewal cycle.

The following theorem gives the efficiency of the server.

Theorem 2. *Suppose that* $\{N(t),\ t \ge 0\}$ *is a nonhomogeneous Poisson process with intensity function* $\lambda(t) \ge 0$. *Then the efficiency is given by*

$$\begin{aligned}\psi = {}& \frac{1}{(\int_0^\infty \mathrm{P}(Y > t)\,\mathrm{d}t + \nu)}\\ &\times \biggl(\int_0^\infty \biggl[r_0(t)\exp\biggl\{-\int_0^t r_0(x)\,\mathrm{d}x\biggr\}\exp\biggl\{-\int_0^t \lambda(x)\,\mathrm{d}x\biggr\}a(t)\exp\{a(t)+b(t)\}\biggr]\mathrm{d}t\\ &\quad + \int_0^\infty \biggl[\exp\biggl\{-\int_0^t r_0(x)\,\mathrm{d}x\biggr\}\exp\biggl\{-\int_0^t \lambda(x)\,\mathrm{d}x\biggr\}\\ &\qquad\qquad \times \eta a(t)b(t)\exp\{a(t)+b(t)\}\biggr]\mathrm{d}t\biggr),\end{aligned}$$

where

$$a(t) = \int_0^t \exp\{-\eta v\}g_W(v)m(t-v)\,\mathrm{d}v, \qquad b(t) = \int_0^t \exp\{-\eta(t-r)\}\bar{G}_W(t-r)\lambda(r)\,\mathrm{d}r,$$

and

$$a(t)+b(t) = m(t) - \eta\int_0^t \exp\{-\eta v\}\bar{G}_W(v)m(t-v)\,\mathrm{d}v.$$

Proof. We derive $\mathrm{E}[M]$. Observe that

$$M = \sum_{i=1}^{N(Y)} \mathbf{1}(T_i + W_i \le Y),$$

where, by convention, $M \equiv 0$ when $N(Y) = 0$. Note that M can be rewritten as

$$M = \sum_{i=1}^{N(Y)} \mathbf{1}(R_i + V_i \le Y),$$

where $\{(R_i, V_i),\ i = 1, 2, \dots, n\}$ is a 'randomized sample' (that is, a random permutation) of $\{(T_i, W_i),\ i = 1, 2, \dots, n\}$. Thus,

$$\mathrm{E}[M] = \mathrm{E}\biggl[\sum_{i=1}^{N(Y)} \mathbf{1}(R_i + V_i \le Y)\biggr].$$

Now, to obtain the above expectation, we derive $f_{R_1,R_2,\dots,R_n,V_1,V_2,\dots,V_n,Y,N(t)}(r_1, r_2, \dots, r_n, v_1, v_2, \dots, v_n, t, n)$ for $n \ge 1$, using the relation

$$\begin{aligned}&f_{R_1,R_2,\dots,R_n,V_1,V_2,\dots,V_n,Y,N(t)}(r_1, r_2, \dots, r_n, v_1, v_2, \dots, v_n, t, n)\\ &\quad = f_{Y\mid R_1,R_2,\dots,R_n,V_1,V_2,\dots,V_n,N(t)}(t \mid r_1, r_2, \dots, r_n, v_1, v_2, \dots, v_n, n)\\ &\qquad \times f_{R_1,R_2,\dots,R_n,V_1,V_2,\dots,V_n,N(t)}(r_1, r_2, \dots, r_n, v_1, v_2, \dots, v_n, n).\end{aligned}$$

https://doi.org/10.1239/jap/1308662638 Published online by Cambridge University Press

Observe that

$$\begin{aligned}&\mathrm{P}(Y > t \mid N(t) = n,\ R_1 = r_1, \dots, R_n = r_n,\ V_1 = v_1, \dots, V_n = v_n)\\&\quad= \exp\bigg\{-\int_0^t r_0(x)\,\mathrm{d}x - \eta\sum_{j=1}^{n}\min\{v_j, (t - r_j)\}\bigg\}.\end{aligned}$$

Also, note that

$$\begin{aligned}&\mathrm{P}(Y > t + \Delta t \mid N(t) = n,\ R_1 = r_1, \dots, R_n = r_n,\ V_1 = v_1, \dots, V_n = v_n)\\&\quad= \mathrm{P}(Y > t + \Delta t \mid Y > t,\ N(t) = n,\ R_1 = r_1, \dots, R_n = r_n,\ V_1 = v_1, \dots, V_n = v_n)\\&\qquad\times \mathrm{P}(Y > t \mid N(t) = n,\ R_1 = r_1, \dots, R_n = r_n,\ V_1 = v_1, \dots, V_n = v_n).\end{aligned}$$

Here,

$$\begin{aligned}&\mathrm{P}(Y > t + \Delta t \mid Y > t,\ N(t) = n,\ R_1 = r_1, \dots, R_n = r_n,\ V_1 = v_1, \dots, V_n = v_n)\\&\quad= 1 - \mathrm{P}(t < Y \le t + \Delta t \mid Y > t,\ N(t) = n,\ R_1 = r_1, \dots, R_n = r_n,\\&\qquad\qquad V_1 = v_1, \dots, V_n = v_n)\\&\quad= 1 - \bigg\{\bigg(r_0(t) + \eta\sum_{j=1}^{n}\mathbf{1}(v_j > t - r_j)\bigg)\Delta t + \eta\varepsilon_0(\Delta t)(\lambda(t)\Delta t + o(\Delta t))\\&\qquad\qquad + \eta\varepsilon_0(\Delta t)\sum_{n=2}^{\infty} n\frac{(m(t+\Delta t) - m(t))^n}{n!}\exp\{-(m(t+\Delta t) - m(t))\}\bigg\}\\&\quad= 1 - \bigg\{\bigg(r_0(t) + \eta\sum_{j=1}^{n}\mathbf{1}(v_j > t - r_j)\bigg)\Delta t + \eta\varepsilon_0(\Delta t)(\lambda(t)\Delta t + o(\Delta t))\\&\qquad\qquad + \eta\varepsilon_0(\Delta t)[(m(t+\Delta t) - m(t))\\&\qquad\qquad\qquad - (m(t+\Delta t) - m(t))\exp\{-(m(t+\Delta t) - m(t))\}]\bigg\}\\&\quad= 1 - \bigg(r_0(t) + \eta\sum_{j=1}^{n}\mathbf{1}(v_j > t - r_j)\bigg)\Delta t + o(\Delta t),\end{aligned}$$

where $\varepsilon_0(t)$ represents any function which satisfies $\lim_{t\to 0}\varepsilon_0(t) = 0$. Therefore,

$$\begin{aligned}&f_{Y \mid R_1, R_2, \dots, R_n, V_1, V_2, \dots, V_n, N(t)}(t \mid r_1, r_2, \dots, r_n, v_1, v_2, \dots, v_n, n)\\&\quad= \lim_{\Delta t\to 0}\frac{1}{\Delta t}[\mathrm{P}(Y > t \mid N(t) = n,\ R_1 = r_1, \dots, R_n = r_n,\ V_1 = v_1, \dots, V_n = v_n)\\&\qquad\qquad - \mathrm{P}(Y > t + \Delta t \mid N(t) = n,\ R_1 = r_1, \dots, R_n = r_n,\\&\qquad\qquad\qquad V_1 = v_1, \dots, V_n = v_n)]\\&\quad= \bigg(r_0(t) + \eta\sum_{j=1}^{n}\mathbf{1}(v_j > t - r_j)\bigg)\exp\bigg\{-\int_0^t r_0(x)\,\mathrm{d}x - \eta\sum_{j=1}^{n}\min\{v_j, (t - r_j)\}\bigg\}.\end{aligned} \tag{5}$$

https://doi.org/10.1239/jap/1308662638 Published online by Cambridge University Press

On the other hand, if we let $(w_1, w_2, \dots, w_n)$ be the realizations of $(W_1, W_2, \dots, W_n)$ and $(r_{(1)}, r_{(2)}, \dots, r_{(n)})$ be the ordered vector of $(r_1, r_2, \dots, r_n)$, then

$$\begin{aligned}
&f_{R_1,R_2,\dots,R_n,V_1,V_2,\dots,V_n,N(t)}(r_1, r_2, \dots, r_n, v_1, v_2, \dots, v_n, n)\\
&\quad= \frac{1}{n!}\lambda(r_{(1)}) \exp\biggl\{-\int_0^{r_{(1)}} \lambda(x)\,\mathrm{d}x\biggr\}\lambda(r_{(2)}) \exp\biggl\{-\int_{r_{(1)}}^{r_{(2)}} \lambda(x)\,\mathrm{d}x\biggr\}\cdots\\
&\qquad\times \lambda(r_{(n)}) \exp\biggl\{-\int_{r_{(n-1)}}^{r_{(n)}} \lambda(x)\,\mathrm{d}x\biggr\} \exp\biggl\{-\int_{r_{(n)}}^{t} \lambda(x)\,\mathrm{d}x\biggr\} \prod_{j=1}^{n} g_W(v_{i_j})\\
&\quad= \frac{1}{n!}\prod_{j=1}^{n}(\lambda(r_j)g_W(v_j)) \exp\biggl\{-\int_0^t \lambda(x)\,\mathrm{d}x\biggr\},
\end{aligned}\tag{6}$$

where v_{i_j} is the element in $\{v_1, v_2, \dots, v_n\}$ which corresponds to w_j, $j = 1, 2, \dots, n$. Thus, by multiplying equations (5) and (6), for $n \ge 1$,

$$\begin{aligned}
&f_{R_1,R_2,\dots,R_n,V_1,V_2,\dots,V_n,Y,N(t)}(r_1, r_2, \dots, r_n, v_1, v_2, \dots, v_n, t, n)\\
&\quad= \frac{1}{n!}\biggl(r_0(t) + \eta \sum_{j=1}^{n} \mathbf{1}(v_j > t - r_j)\biggr) \exp\biggl\{-\int_0^t r_0(x)\,\mathrm{d}x - \eta \sum_{j=1}^{n} \min\{v_j, (t - r_j)\}\biggr\}\\
&\qquad\times \prod_{j=1}^{n}(\lambda(r_j)g_W(v_j)) \exp\biggl\{-\int_0^t \lambda(x)\,\mathrm{d}x\biggr\}\\
&\quad= \frac{1}{n!}r_0(t) \exp\biggl\{-\int_0^t r_0(x)\,\mathrm{d}x\biggr\} \exp\biggl\{-\int_0^t \lambda(x)\,\mathrm{d}x\biggr\}\\
&\qquad\times \prod_{j=1}^{n} \exp\{-\eta \min\{v_j, (t - r_j)\}\} \prod_{j=1}^{n}(\lambda(r_j)g_W(v_j))\\
&\qquad+ \frac{1}{n!}\biggl\{\eta \sum_{j=1}^{n} \mathbf{1}(v_j > t - r_j)\biggr\} \prod_{j=1}^{n} \exp\{-\eta \min\{v_j, (t - r_j)\}\} \prod_{j=1}^{n}(\lambda(r_j)g_W(v_j))\\
&\qquad\times \exp\biggl\{-\int_0^t r_0(x)\,\mathrm{d}x\biggr\} \exp\biggl\{-\int_0^t \lambda(x)\,\mathrm{d}x\biggr\}.
\end{aligned}$$

From this,

$$\begin{aligned}
\mathrm{E}[M] &= \mathrm{E}\biggl[\sum_{i=1}^{N(Y)} \mathbf{1}(R_i + V_i \le Y)\biggr]\\
&= \sum_{n=1}^{\infty} \int_0^{\infty} \biggl[\sum_{i=1}^{n} \int_0^t \int_0^{t-r_i} \int_0^t \cdots \int_0^t \int_0^{\infty} \cdots\\
&\qquad\qquad \times \int_0^{\infty} f_{R_1,R_2,\dots,R_n,V_1,V_2,\dots,V_n,Y,N(t)}(r_1, r_2, \dots, r_n, v_1, v_2, \dots, v_n, t, n)\\
&\qquad\qquad \times \mathrm{d}v_1 \cdots \mathrm{d}v_{i-1}\,\mathrm{d}v_{i+1} \cdots \mathrm{d}v_n\,\mathrm{d}r_1 \cdots \mathrm{d}r_{i-1}\,\mathrm{d}r_{i+1} \cdots \mathrm{d}r_n\,\mathrm{d}v_i\,\mathrm{d}r_i\biggr]\,\mathrm{d}t
\end{aligned}$$

https://doi.org/10.1239/jap/1308662638 Published online by Cambridge University Press

$$
\begin{aligned}
&= \sum_{n=1}^{\infty} \int_0^{\infty} \biggl[\frac{1}{n!} r_0(t) \exp\biggl\{ - \int_0^t r_0(x)\,\mathrm{d}x \biggr\} \exp\biggl\{ - \int_0^t \lambda(x)\,\mathrm{d}x \biggr\} \\
&\qquad\qquad \times n \biggl[\int_0^t \int_0^{t-r} \exp\{-\eta v\} g_W(v)\,\mathrm{d}v \lambda(r)\,\mathrm{d}r \biggr] \\
&\qquad\qquad \times \biggl[\int_0^t \int_0^{t-r} \exp\{-\eta v\} g_W(v)\,\mathrm{d}v \lambda(r)\,\mathrm{d}r \\
&\qquad\qquad\quad + \int_0^t \exp\{-\eta(t-r)\} \bar{G}_W(t-r) \lambda(r)\,\mathrm{d}r \biggr]^{n-1} \biggr]\,\mathrm{d}t \\
&\quad + \sum_{n=1}^{\infty} \int_0^{\infty} \biggl[\frac{1}{n!} \exp\biggl\{ - \int_0^t r_0(x)\,\mathrm{d}x \biggr\} \exp\biggl\{ - \int_0^t \lambda(x)\,\mathrm{d}x \biggr\} \\
&\qquad\qquad \times n(n-1)\eta \biggl[\int_0^t \int_0^{t-r} \exp\{-\eta v\} g_W(v)\,\mathrm{d}v \lambda(r)\,\mathrm{d}r \biggr] \\
&\qquad\qquad \times \biggl[\int_0^t \exp\{-\eta(t-r)\} \bar{G}_W(t-r) \lambda(r)\,\mathrm{d}r \biggr] \\
&\qquad\qquad \times \biggl[\int_0^t \int_0^{t-r} \exp\{-\eta v\} g_W(v)\,\mathrm{d}v \lambda(r)\,\mathrm{d}r \\
&\qquad\qquad\quad + \int_0^t \exp\{-\eta(t-r)\} \bar{G}_W(t-r) \lambda(r)\,\mathrm{d}r \biggr]^{n-2} \biggr]\,\mathrm{d}t \\
&= \sum_{n=1}^{\infty} \int_0^{\infty} \biggl[\frac{1}{(n-1)!} r_0(t) \exp\biggl\{ - \int_0^t r_0(x)\,\mathrm{d}x \biggr\} \\
&\qquad\qquad \times \exp\biggl\{ - \int_0^t \lambda(x)\,\mathrm{d}x \biggr\} a(t)[a(t)+b(t)]^{n-1} \biggr]\,\mathrm{d}t \\
&\quad + \sum_{n=2}^{\infty} \int_0^{\infty} \biggl[\frac{1}{(n-2)!} \exp\biggl\{ - \int_0^t r_0(x)\,\mathrm{d}x \biggr\} \exp\biggl\{ - \int_0^t \lambda(x)\,\mathrm{d}x \biggr\} \\
&\qquad\qquad \times \eta a(t) b(t) [a(t)+b(t)]^{n-2} \biggr],
\end{aligned}
$$

where

$$
a(t) = \int_0^t \int_0^{t-r} \exp\{-\eta v\} g_W(v)\,\mathrm{d}v \lambda(r)\,\mathrm{d}r = \int_0^t \exp\{-\eta v\} g_W(v) m(t-v)\,\mathrm{d}v,
$$

and

$$
b(t) = \int_0^t \exp\{-\eta(t-r)\} \bar{G}_W(t-r) \lambda(r)\,\mathrm{d}r.
$$

Thus, we now have the desired result.

4. The optimal admission control: illustrative examples

In this section we discuss the problem of determining the optimal admission control policy which maximizes the efficiency of the Web server by considering two illustrative examples. We consider the cases when the intensity of the nonhomogeneous Poisson process is given by

https://doi.org/10.1239/jap/1308662638 Published online by Cambridge University Press

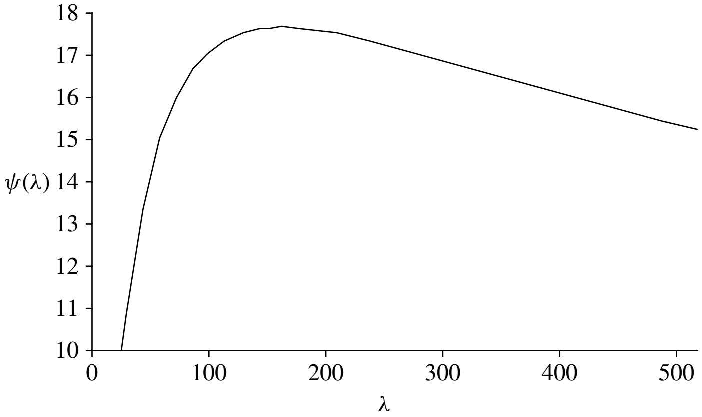

FIGURE 2: The long-run average number of jobs per unit time as a function of λ.

a constant. Then, as mentioned before, the condition for the restart of the arrival process in the previous section is automatically satisfied.

Example 1. Let $r_0(t) = r_0 = 0.2$, $\eta = 0.01$, $g_W(w) = w \exp\{-0.5w^2\}$, $w \geq 0$, and $\nu = 1.0$.

Case I. Nowadays, as the Web businesses are mostly global and worldwide, the connections by Web clients to the server are made from all over the world and, in this situation, the arrival rate would be mostly 'time independent'. In this case, a constant intensity function can be adopted.

Now we assume that $\lambda(t) = \lambda$ for all $t \geq 0$. Then, for this case, the long-run average number of jobs per unit time as a function of λ, $\psi(\lambda)$, is given in Figure 2.

As can be seen from Figure 2, there exists a λ^* which maximizes $\psi(\lambda)$. Therefore, the optimal admission control policy is as follows:

(i) if $\lambda < \lambda^*$ then do not apply any control policy;

(ii) If $\lambda \geq \lambda^*$ then, on each arrival request, randomly reject it with probability $1 - \lambda^*/\lambda$, and accept it with probability λ^*/λ.

Case II. Practically, in some situations, the arrival rate may have an increasing trend after booting of the Web server, and then it stays at an almost constant level. Thus, it would be meaningful to consider the case of increasing $\lambda(t)$: $\lambda(t) = 400(1 - \mathrm{e}^{-10t})$, $t \geq 0$. In almost all cases, the rejection policy is triggered when the arrival rate exceeds a certain threshold value so that the processing rate should not increase anymore and be preserved as a constant level, i.e. as the threshold level (see Bhatti and Friedrich (1999) and Voigt and Gunningberg (2001)). Let the threshold value be λ_{c}. Then in this case, the long-run average number of jobs per unit time as a function of λ_{c}, $\psi(\lambda_{\mathrm{c}})$, is given in Figure 3.

Observe that, in this case, it is sufficient to search for the optimal λ_{c}^* which maximizes $\psi(\lambda_{\mathrm{c}})$ in the interval $[0, 400]$ as the maximum level (supremum) of $\lambda(t)$ is given by 400. The threshold values which exceed this level would yield the same control policy as that with $\lambda_{\mathrm{c}} = 400$ (i.e. do not apply any control policy). In this example, as can be seen from Figure 3, there exists a $\lambda_{\mathrm{c}}^* < 400$ which maximizes $\psi(\lambda_{\mathrm{c}})$.

Example 2. As discussed before, to protect the server from high client loads, some requests must be rejected. Sometimes requests from Web clients can be classified into the following two

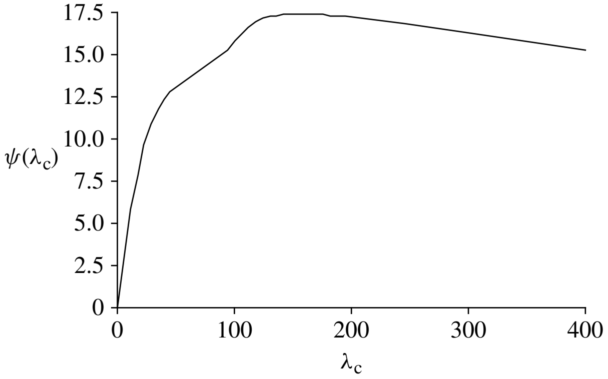

FIGURE 3: The long-run average number of jobs per unit time as a function of λ_c.

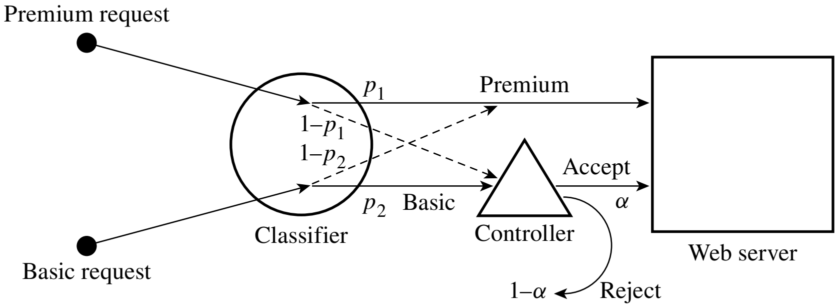

FIGURE 4: The flow diagram.

categories based on preliminary information, such as the client IP address, HTTP cookies, the URL request type, or filename path, which are obtained on the arrivals of the requests: (i) the premium request that should be processed with high priority (e.g. requests from premium clients) and (ii) the basic requests. In this situation, when applying the admission control policy, basic requests rather than premium requests should be rejected (see Bhatti and Friedrich (1999)).

Let $\{N_{\mathrm{P}}(t),\ t \geq t\}$ and $\{N_{\mathrm{B}}(t),\ t \geq t\}$ be two independent homogeneous Poisson processes with constant intensities λ_1 and λ_2, which represent the arrival processes of the premium requests and basic requests, respectively. On the arrivals of requests, they are classified by the classifier. Assume that misclassifications can occur at the classification stage and that there are two types of misclassification: (i) a premium request is misclassified into a basic request (type I), (ii) a basic request is misclassified into a premium request (type II). The probability of a type-I misclassification is $1 - p_1$ and that of a type-II misclassification is $1 - p_2$. Additionally, assume that the random times for the completions of the premium and basic jobs follow the same distribution, and the rewards obtained from the successful completion of the premium and basic jobs are given by κ_1 and κ_2, $\kappa_1 > \kappa_2 > 0$, respectively.

If necessary, the rejection policy is applied only to jobs classified into basic jobs, and jobs classified into premium jobs are processed without rejection. Thus, after the classification stage, the jobs classified into basic jobs are randomly rejected with rejection probability $1 - \alpha$. The flow diagram for the whole process is depicted in Figure 4.

https://doi.org/10.1239/jap/1308662638 Published online by Cambridge University Press

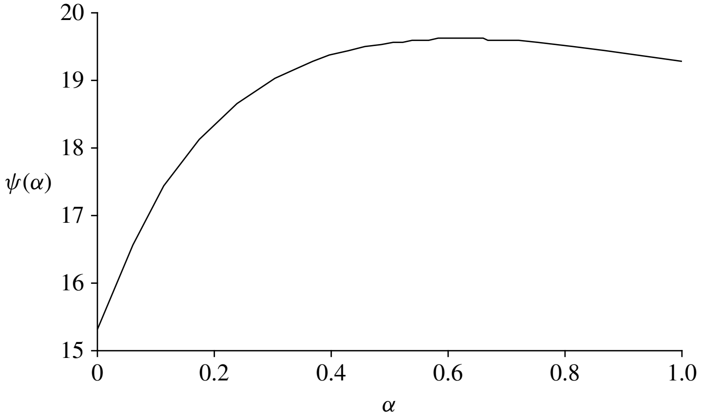

FIGURE 5: The long-run average reward per unit time for $r_0 = 0.2$ and $\eta = 0.01$.

When the acceptance rate is given by α, the total arrival rate of the requests after the admission control process is given by $[\lambda_1 p_1 + \lambda_2(1 - p_2)] + \alpha[\lambda_1(1 - p_1) + \lambda_2 p_2]$. Furthermore, when a requested job is performed by the Web Server after the classification and screening stage, it is a premium job with probability

$$\beta \equiv \frac{\lambda_1 p_1 + \alpha\lambda_1(1 - p_1)}{[\lambda_1 p_1 + \lambda_2(1 - p_2)] + \alpha[\lambda_1(1 - p_1) + \lambda_2 p_2]},$$

and is a basic job with probability $1 - \beta$. If we now define the efficiency $\psi(\alpha)$ as the long-run average reward per unit time (as a function of α), then

$$\psi(\alpha) = \frac{(\beta\kappa_1 + (1 - \beta)\kappa_2)\,\mathrm{E}[M]}{\mathrm{E}[Y] + \nu},$$

where $\mathrm{E}[M]$ and $\mathrm{E}[Y]$ are obtained by setting

$$\lambda(t) = [\lambda_1 p_1 + \lambda_2(1 - p_2)] + \alpha[\lambda_1(1 - p_1) + \lambda_2 p_2] \quad \text{for all } t \geq 0.$$

Then, when $r_0(t) = r_0 = 0.2$, $\eta = 0.01$, $g_W(w) = w\exp\{-0.5w^2\}$, $w \geq 0$, $\nu = 1.0$, $\lambda_1 = 15$, $\lambda_2 = 150$, $p_1 = p_2 = 0.95$, $\kappa_1 = 2.0$, and $\kappa_1 = 1.0$, the efficiency function $\psi(\alpha)$ is given in Figure 5.

As can be seen from Figure 5, there exists a unique optimal acceptance rate α^*.

5. Concluding remarks

Recently, as the IT industry rapidly develops, new topics and studies on the testing and operations of software, computers, and computer-related systems that exist for the purpose of data, information, and knowledge processing are accordingly increasingly appearing in the field of applied probability (see, e.g. Boland *et al.* (2002), Mi (2002), Ling and Mi (2004), and Agustin (2003)). In this paper we proposed and considered another new topic in the field: the problem of determining the optimal admission control policy which maximizes the efficiency of the Web server. In order to model the operating characteristics of the Web server system, a stochastic model was proposed so that the breakdown rate process of the server *dynamically responds* to the flow of the jobs being processed. Based on the proposed model, the breakdown

https://doi.org/10.1239/jap/1308662638 Published online by Cambridge University Press

rate and the efficiency of the Web server were derived. Considering two case studies, the optimal admission control policy which maximizes the efficiency measure by controlling the amount and rate of work entering the system was discussed. The topic in this study could be expanded, and the range of the study could be extended by considering and including various practical aspects. Some helpful references could be Voigt and Gunningberg (2001) and Voigt *et al.* (2001). To the authors' knowledge, this kind of topic has not been discussed in the literature based on a stochastic and probabilistic approach. Therefore, the topic discussed in this paper would stimulate further related studies in the field.

Acknowledgements

The authors would like to thank the anonymous referee for very helpful and careful comments and suggestions, which have considerably improved the presentation of this paper. The work of the first author was supported by the Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2009-0093827). The work of the second author was supported by the Basic Science Research Program through the NRF funded by the Ministry of Education, Science and Technology (2010-0023928).

References

Agustin, M. A. (2003). Statistical properties of a system reliability estimator using the Littlewoood software reliability model. *J. Appl. Prob.* **40,** 766–778.

Agustin, M. A. and Peña, E. A. (1999). A dynamic competing risks model. *Prob. Eng. Inf. Sci.* **13,** 333–358.

Barghout, M. (2010). Predicting software reliability using an imperfect debugging Jelinski Moranda non-homogeneous Poisson process model. *Model Assisted Statist. Appl.* **5,** 31–41.

Bhatti, N. and Friedrich, R. (1999). Web server support for tiered services. *IEEE Network* **13,** 64–71.

Boland, P. J. and Ni Chuív. N. (2007). Optimal times for software release when repair is imperfect. *Statist. Prob. Lett.* **77,** 1176–1184.

Boland, P. J. and Singh, H. (2002). Determining the optimal release time for software in the geometric Poisson reliability model. *Internat. J. Reliab. Quality Safety Eng.* **9,** 201–213.

Boland, P. J., Singh, H. and Cukic, B. (2002). Stochastic orders in partition and random testing of software. *J. Appl. Prob.* **39,** 555–565.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*, 2nd edn. Duxbury, CA.

Jelinski, Z. and Moranda, P. (1972). Software reliability research. In *Statistical Computer Performance Evaluation*, ed. W. Freiberger, Academic Press, New York, pp. 465–484.

Kvam, P. H. and Peña, E. A. (2005). Estimating load-sharing properties in a dynamic reliability system. *J. Amer. Statist. Assoc.* **100,** 262–272.

Ling, Y. and Mi, J. (2004). An optimal trade-off between content freshness and refresh cost. *J. Appl. Prob.* **41,** 721–734.

Mi, J. (2002). Age-replacement policy and optimal work size. *J. Appl. Prob.* **39,** 296–311.

Ross, S. M. (1996). *Stochastic Processes*, 2nd edn. John Wiley, New York.

Singpurwalla, N. D. (1991). Determining an optimal time interval for testing and debugging software. *IEEE Trans Software Eng.* **17,** 313–319.

Voigt, T. and Gunningberg, P. (2001). Kernel-based control of persistent web server connections. In *ACM SIGMETRICS Performance Evaluation Review*, Association for Computing Machinery, New York, pp. 20–25.

Voigt, T., Tewari, R., Freimuth, D. and Mehra, A. (2001). Kernel mechanisms for service differentiation in overloaded web servers. In *Proc. General Track: 2002 USENIX Ann. Tech. Conf.*, ed. Y. Park, pp. 189–202.

Wessels, D. (2001). *Web Caching*. O'Reilly, Sebastopol, CA.

https://doi.org/10.1239/jap/1308662638 Published online by Cambridge University Press