Academic Advances of the CTO, Vol. 1, Issue 2, 2017
PhD Thesis Chapter

Stochastic Reliability of a Server under a Random Workload

Rachel Traylor*, Ph.D.

Abstract
Editor's note: This paper is the first chapter from a PhD thesis published in 2016 by Rachel Traylor. We generalize a 2011 model from Cha and Lee that gave a closed form for the survival function of a server under a random workload, and with random service times. In this work, the constant stress assumption of Cha and Lee is relaxed and allowed to be random; that is, every request to the server brings a random stress or workload to the server for the duration of its stay until completion. The efficiency measure of a random stress server is defined as the long run average number of jobs completed per unit time, and provides a way to measure server performance.

Keywords
reliability theory — random traffic — server reliability — probability theory

Contents

Introduction

There are many types of systems which can be dubbed servers, such as a retail checkout counter, a shipping company, a web server, or a customer service hotline. All of these systems have common general behavior. Requests or customers arrive via a stochastic process, the service times vary randomly, and each request stresses the server if only temporarily. A general stochastic model that describes the reliability of such a server can provide the necessary information for optimal resource allocation and efficient task scheduling, leading to significant cost savings for businesses and improved performance metrics[6]. Such topics have been studied in literature for several decades [1, 2, 3, 20].

*Office of the CTO, Dell EMC

Much attention was devoted to reliability principles that model software failures and bug fixes, starting with Jelinski and Moranda in 1972 [11]. The hazard function under this model shows the time between the ith failure and the $i+1$st failure. Littlewood (1980) [16] extended this initial reliability model for software by assuming differences in error size. [12].

These models have been extended into software testing applications [4, 5, 19] and optimal software release times [7, 8, 17, 22]. The explosion of e-commerce and the resulting increase in internet traffic have led to the development of reliability models for Web applications. Heavy traffic can overload and crash a server; thus, various control policies for refreshing content and admission of page requests were created [10, 13, 15, 18, 23].

In particular, Cha and Lee (2011) [9] proposed a stochastic breakdown model for an unreliable web server whose requests arrive at random times according to a nonhomogenous Poisson process and bring a constant stress factor to the system that dissipates upon service completion. The authors provide a fairly general survival function under any service distribution $g_W(w)$, define server efficiency to measure performance, and illustrate a possible admission control policy due to an observed property of the server efficiency under a specific numerical example.

Thus far, no extensions of [9] have been proposed. This work generalizes the model put forth by Cha and Lee in a variety of ways. First, the assumption of constant job stress is relaxed and replaced by a random variable, and a new survival function and efficiency equation are derived. This work, while suitable for IT applications, is general enough for use in almost any industry, including logistics, retail, manufacturing, and engineering systems.

1. Background- Cha and Lee's Model

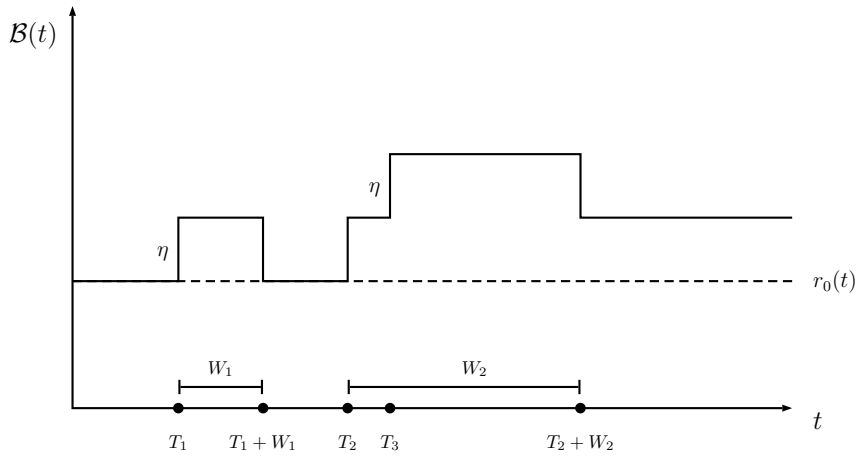

Figure 1. Sample Trajectory of Breakdown Rate Process Under Original Model

System Description and Survival Function

Cha and Lee considered a web server wherein each request arrives via a nonhomogenous Poisson process $\{N(t) : t \geq 0\}$ with intensity function $\lambda(t)$. Each request adds a constant stress η, increasing the breakdown rate for the duration of service. Suppose $r_0(t)$ is the breakdown rate of the idle server. Then the breakdown rate process $\mathscr{B}(t)$ is defined as

$$\mathscr{B}(t) := r_0(t) + \eta \sum_{j=1}^{N(t)} \mathbb{1}(T_j \leq t \leq T_j + W_j)$$

where $N(t)$, $\{T_j\}_{j=1}^{N(t)}$, $\{W_j\}_{j=1}^{N(t)}$ are the random variables that describe the number of arrivals, arrival times, and service times, respectively. It is assumed that $\{T_j\}_{j=1}^{N(t)}$ are independent of each other. Furthermore, the $\{W_j\}_{j=1}^{N(t)} \sim g_W(w)$ are i.i.d. and are mutually independent of all T_j's.

Under these conditions, Cha and Lee proved the following theorem:

Theorem 1. *Suppose that $\{N(t), t \geq 0\}$ is a nonhomogenous Poisson process with intensity function $\lambda(t)$, i.e. $m(t) \equiv \int_0^t \lambda(x)dx$. Assuming the conditional survival function is given by*

$$P\left(Y > t \middle| N(t), \{T_j\}_{j=1}^{N(t)}, \{W_j\}_{j=1}^{N(t)}\right) = \bar{F}_0(t) \exp\left(-\eta \sum_{j=1}^{N(t)} \min(W_j, t - T_j)\right)$$

and $m(t)$ has an inverse, the survival function of Y is given by

$$S_Y(t) = \bar{F}_0(t) \exp\left(-\eta \int_0^t \exp(-\eta w) \bar{G}_W(w) m(t-w) dw\right)$$

and the hazard function of Y, denoted $r(t)$, is given by

$$r(t) = r_0(t) + \eta \int_0^t e^{-\eta w} \bar{G}_W(w) \lambda(t-w) dw$$

Efficiency of the Server

It is natural to develop some measure of server performance. Cha and Lee measure such performance by defining the *efficiency*, ψ, of the web server as the long-run expected number of jobs completed per unit time. That is, with the number of jobs completed as M, the efficiency is defined

$$\psi := \lim_{t \to \infty} \frac{E[M(t)]}{t}$$

Upon breakdown and rebooting, the server is assumed to be 'as good as new', in that performance of the server does not degrade during subsequent reboots. In addition, the model assumes the arrival process after reboot, denoted $\{N^*(t), t \geq 0\}$, is a nonhomogenous Poisson process with the same intensity function $\lambda(t)$ as before, and that $\{N^*(t), t \geq 0\}$ is independent of the arrival process before reboot. In a practical setting, this model assumes no 'bottlenecking' of arrivals occurs in the queue during server downtime that would cause an initial flood to the rebooted server. In addition, the reboot time is assumed to follow a continuous distribution $H(t)$ with expected value ν. This process is a renewal reward process, with the renewal $\{R_n\} = \{M_n\}$, the number of jobs completed. The length of a renewal cycle is $Y_n + H_n$, where Y_n is the length of time the server was operational, and H_n is the time to reboot after a server crash. Then, by [21],

$$\psi = \frac{E[M]}{E[Y] + \nu} \tag{1}$$

where M is the number of jobs completed in a particular renewal cycle, ν is the mean time to reboot of the server, and Y is the length of a particular renewal cycle. Then, using (1), the following closed form of the efficiency of a server under all assumptions of Cha and Lee's model is derived.

Theorem 2. *Suppose $\{N(t), t \geq 0\}$ is a nonhomogenous Poisson process with intensity $\lambda(t) \geq 0$. Then the*

efficiency is given by

$$\psi = \frac{1}{\int_0^\infty S_Y(t)dt + \nu}\left[\exp\left(-\int_0^t r_0(x)dx - \int_0^t \lambda(x)dx + a(t) + b(t)\right) \times (r_0(t)a(t) + \eta a(t)b(t))\right]$$

where $a(t) = \int_0^t e^{-\eta v} g_W(v) m(t-v)dv$, $b(t) = \int_0^t e^{-\eta(t-r)} \bar{G}_W(t-r)\lambda(r)dr$, $\bar{G}_W(x) = 1 - \int_0^x g_W(s)ds$, *and* $m(x) = \int_0^x \lambda(s)ds$.

Numerical Example and Control Policies

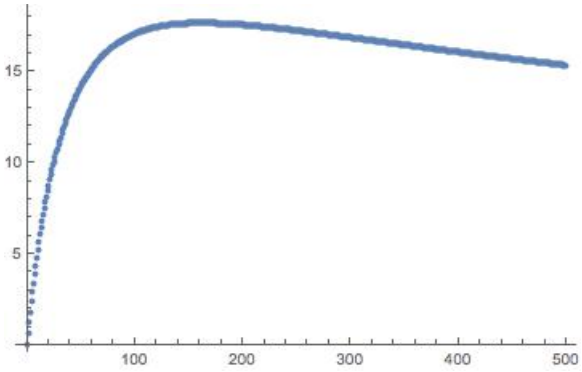

Figure 2. $\psi(\lambda)$ under Rayleigh Service Time Distribution

As an illustrative example, Cha and Lee considered the case when $\lambda(t) \equiv \lambda$, $r_0(t) \equiv r_0 = 0.2$, $\eta = 0.01$, $\nu = 1$, and $g_W(w) = we^{-w^2/2}$ (the PDF of the Rayleigh distribution). As shown in Figure 2, there exists a λ^* such that $\psi(\lambda)$ is maximized. Thus one may implement the obvious optimal control policy for server control to avoid server overload:

(1) If the real time arrival rate $\lambda < \lambda^*$, do not interfere with arrivals.

(2) If $\lambda \geq \lambda^*$, facilitate some appropriate measure of interference.

Examples of interference for a web server in particular include rejection of incoming requests or possible re-routing. Cha and Lee give an interference policy of rejection with probability $1 - \frac{\lambda^*}{\lambda}$. The next section presents a generalization of the above model by relaxing the assumption that the job stress η is constant. This yields a far more generalized model that can encompass random stress with any distribution.

2. Random Stress Reliability Model

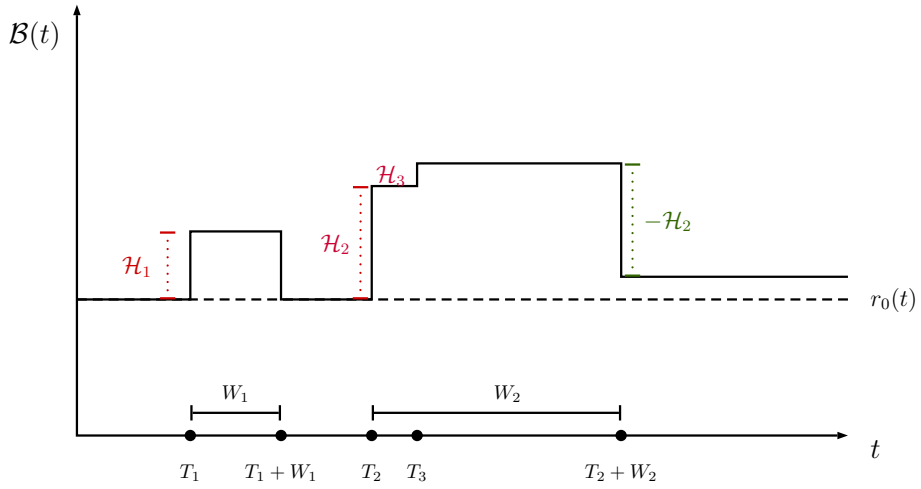

Figure 3. Sample Trajectory of Breakdown Rate Process under Random Stress Model

The basic model here is very similar to Cha and Lee's. The main difference is that we relax the constant stress assumption and allow it to be random. Assume that each job j coming into the server adds a random stress $\mathcal{H}_j$ to the server for the duration of its time in the system. Suppose $\{\mathcal{H}_j\}_{j=1}^{N(t)} \overset{i.i.d.}{\sim} \mathcal{H}$, where WLOG $\mathcal{H}$ is a discrete random variable with a finite sample space $S = \{\eta_i : \eta_i \in \mathbb{R}^+, i = 1,...,m \text{ for } m \in \mathbb{N}\}$ and probability distribution given by

$$P(\mathcal{H} = \eta_i) = p_i, \quad i = 1,...,m$$

The following assumptions from [9] are retained:

(CL1) Requests arrive via a nonhomogenous Poisson Process $\{N(t), t \geq 0\}$ with intensity $\lambda(t)$.

(CL2) Arrival times $\{T_j\}_{j=1}^{N(t)}$ are independent.

(CL3) Service times $\{W_j\}_{j=1}^{N(t)}$ are i.i.d. with pdf $g_W(w)$ and mutually independent of all arrival times.

Then, the random stress breakdown rate (RSBR) process $\mathscr{B}(t)$ is given by

$$\mathscr{B}(t) = r_0(t) + \sum_{j=1}^{N(t)} \mathcal{H}_j \mathbb{1}(T_j < t \leq T_j + W_j), \quad t \geq 0 \tag{2}$$

Compare the sample trajectory shown in Figure 3 to Figure 1. The random stress brought by each request still disappears upon job completion, but the effect on the server is no longer deterministic. Thus, for the same set of arrival times and respective completion times, the realization of the breakdown rate process under the RSBR model has one more element of variation.

Let Y be the random time to breakdown of the web server given the workload from client requests. Let $\mathfrak{T} = \{T_j\}_{j=1}^{N(t)}$, $\mathfrak{W} = \{W_j\}_{j=1}^{N(t)}$, and $\mathfrak{H} = \{\mathcal{H}_j\}_{j=1}^{N(t)}$, with observed values $\mathfrak{t} = \{t_j\}_{j=1}^{N(t)}$, $\mathfrak{w} = \{w_j\}_{j=1}^{N(t)}$, and $\mathfrak{h} = \{\eta_{i_j}\}_{i_j=1}^{N(t)}$. Then the conditional survival function of the server, given the arrival process of client requests ($N(t)$), job stresses ($\mathfrak{H}$), service times ($\mathfrak{W}$), and arrival times ($\mathfrak{T}$) is

$$\begin{aligned} S_{Y|N(t),\mathfrak{T},\mathfrak{W},\mathfrak{H}}(t|n,\mathfrak{t},\mathfrak{w},\mathfrak{h}) &= e^{-\int_0^t \mathscr{B}(x)dx} \\ &= \bar{F}_0(t) e^{-\sum_{j=1}^{N(t)} \mathcal{H}_i \min(W_j, t-T_j)} \end{aligned}$$

where $\overline{F}_0(t) = \exp\left(-\int_0^t r_0(x)dx\right)$.

Survival Function for the RSBR Web Server

Under the RSBR generalization, the survival function of the server is given in the following theorem.

Theorem 3 (Survival Function of RSBR Server)**.** *Suppose that jobs arrive to a server according to a nonhomogenous Poisson process* $\{N(t), t \geq 0\}$ *with intensity function* $\lambda(t) \geq 0$ *and* $m(t) \equiv E[N(t)] = \int_0^t \lambda(x)dx$. *Let the arrival times* $\{T_j\}_{j=1}^{N(t)}$ *be independent, and let the service times* $\{W_j\}_{j=1}^{N(t)} g_W(w)$ *be i.i.d. and mutually independent of all arrival times. Assume the random job stresses* $\mathcal{H}_j \sim \mathcal{H}$. *Then*

$$S_Y(t) = \bar{F}_0(t) \exp\left(-E_{\mathcal{H}}\left[\mathcal{H}\int_0^t e^{-\mathcal{H}w} m(t-w)\bar{G}_W(w)dw\right]\right) \tag{3}$$

where $\bar{F}_0(t) = \exp\left(-\int_0^t r_0(s)ds\right)$.

The proof of Theorem 3 can be found in Section 6.

The compound failure rate function $r(t)$ is given by

$$r(t) = -\frac{d}{dt}\ln(S_Y(t)) = r_0(t) + E_{\mathcal{H}}\left[\mathcal{H}\int_0^t e^{-\mathcal{H}w} m(t-w)\bar{G}_W(w)dw\right]$$

See [14] or [21].

3. Efficiency measure of the server under RSBR

Upon server crash, the server must be rebooted. This section gives the server efficiency as defined in [9], but for the RSBR model. The server efficiency given in Theorem 4 holds quite a few similarities to that of [9], except that, like Theorem 3, the expectation must be taken over the random stress variable. The following assumptions from the original model are retained:

(E1) The arrival process after rebooting, $\{N^{rb}(t), t \geq 0\}$, remains a nonhomogenous Poisson process with the same intensity function $\lambda(t), t \geq 0$ as before.

(E2) $\{N^{rb}(t), t \geq 0\}$ is independent of the arrival process of client requests before rebooting. Hence, $\{N^{rb}(t), t \geq 0\} = \{N(t), t \geq 0\}$, since it retains all the same characteristics as before.

(E3) The time to reboot the server follows a continuous distribution $H(t)$ with mean ν.

Recall that $M(t)$ is defined as the total number of jobs completed by the server during the time $(0, t]$. Also, recall the definition of server efficiency from [9]:

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

The efficiency of the server under a random stress environment is given in the following theorem.

Theorem 4 (Server Efficiency under Random Stress Environment)**.** *Suppose that* $\{N(t) : t \geq 0\}$ *is a nonhomogenous Poisson process with intensity function* $\lambda(t), t \geq 0$. *Suppose also the conditions of Theorem 3 and the conditions (E1)-(E3) are met. Then the efficiency of the server is given by*

$$\psi = \frac{1}{\int_0^\infty S_Y(t)dt + \nu}\left\{\int_0^\infty e^{-\int_0^t r_0(x) - \int_0^t \lambda(x)dx + E_{\mathcal{H}}[a(t)+b(t)]}(r_0(t)E_{\mathcal{H}}[a(t)] + E_{\mathcal{H}}[\mathcal{H}a(t)b(t)])dt\right\} \tag{4}$$

where $a(t) = \int_0^t e^{-\mathcal{H}v} g_W(v)m(t-v)dv$ *and* $b(t) = \int_0^t e^{-\mathcal{H}(t-r)}\bar{G}_W(t-r)\lambda(r)dr$.

The proof for Theorem 4 can be found in Section 6.2.

4. Remarks and Implications

Note that $\mathcal{H}$ was assumed discrete, but the proofs of Theorema 3 and 4 are unaffected if $\mathcal{H}$ is continuous. Thus the generality of this model is significantly stronger than in [9]. In practical considerations, these integrals will need to be numerically evaluated.

For certain distributions of $\mathcal{H}$, $S_Y(t)$ has a fairly compact form. Subsequent installments will examine the case where $\mathcal{H}$ has a binomial distribution, formed from both independent and dependent Bernoulli trials. We will also explore the effects of various service life distributions on ψ

5. Conclusion

This paper has given a simple but large generalization of [9]. The random stress model allows for the reality that we likely will not know the stress a request will bring to the server, but may know its distribution. Subsequent installments will further study the properties of the RSBR model and extend it to other single-server models, and networks.

6. Appendix

6.1 Proof of Theorem 3

Proof. Taking the expectation of the conditional survival function in (3),

$$S_Y(t) = \bar{F}_0(t) E\left[\exp\left(-\sum_{j=1}^{N(t)} \mathcal{H}_j \min(W_j, t - T_j)\right)\right]$$

Using the law of total expectation:

$$E\left[e^{-\sum_{j=1}^{N(t)} \mathcal{H}_j \min(W_j, t-T_j)}\right] = E\left[E\left[e^{-\sum_{j=1}^{N(t)} \mathcal{H}_j \min(W_j, t-T_j)} \middle| N(t), \mathfrak{H}\right]\right]$$

Conditioned on $N(t) = n$ and $\mathcal{H}_j = \eta_{i_j}$ for some $i_j \in \{1, ..., m\}$,

$$f_{T_1,\ldots,T_n|N(t),\mathfrak{H}}(t_1 \ldots t_n | n, \mathfrak{h}) = f_{T_1,\ldots,T_n|N(t)}(t_1, \cdots, t_n | n)$$

since the sets $\mathfrak{H}$ and $\mathfrak{T}$ are mutually independent.

Let $T_1', \ldots T_n'$ be i.i.d. random variables with pdf $f(x) = \frac{\lambda(x)}{m(t)}$. By Lemma 1,

$$f_{T_1,\ldots,T_n|N(t)}(t_1, \cdots, t_n | n) = n! \prod_{j=1}^{n} \frac{\lambda(t_j)}{m(t)} \tag{5}$$

for $0 \le t_1 \le \cdots \le t_n \le t$. Then

$$E\left[e^{-\sum_{j=1}^{N(t)} \mathcal{H}_j \min(W_j, t-T_j)} \middle| N(t), \mathfrak{H}\right] = E\left[e^{-\sum_{j=1}^{n} \eta_{i_j} \min(W_j, t-T_j)}\right] \qquad = E\left[e^{-\sum_{j=1}^{n} \eta_{i_{[j]}} \min(W_j, t-T_{[j]})}\right]$$

By Lemma 1 in Section 6.3,

$$\begin{aligned} E\left[e^{-\sum_{j=1}^{n} \eta_{i_{[j]}} \min(W_j, t-T_{[j]})}\right] &= E\left[e^{-\sum_{j=1}^{n} \eta_{i_{j'}} \min(W_j, t-T_{j'})}\right] \\ &= E\left[\prod_{j=1}^{n} e^{-\eta_{i_j} \min(W_j, t-T_{j'})}\right] \\ &= \prod_{j=1}^{n} E\left[e^{-\eta_{i_{j'}} \min(W_j, t-T_{j'})}\right] \end{aligned}$$

The equalities hold because the elements of $\mathfrak{W}$ and $\mathfrak{T}$ are i.i.d., respectively, and are mutually independent. Now, fix j'; then $\eta_{i_{j'}}$ is also fixed. By Lemma 2 (Section 6.3),

$$E\left[e^{-\eta_{i_{j'}}\min(W_j,t-T_{j'})}\right] = \frac{1}{m(t)}\left(m(t) - \eta_{i_{j'}}\int_0^t e^{-\eta_{i_{j'}}w}m(t-w)\bar{G}_W(w)dw\right) \tag{6}$$

Equation (6) is true $\forall\, j$, so

$$E\left[e^{-\sum_{j=1}^{N(t)}\mathcal{H}_j\min(W_j,t-T_j)}\middle| N(t),\mathfrak{H}\right] = \prod_{j=1}^{n}\frac{1}{m(t)}\left(m(t) - \eta_{i_j}\int_0^t e^{-\eta_{i_j}w}m(t-w)\bar{G}_W(w)dw\right) \tag{7}$$

Finally, the expectation of (7) over $N(t),\mathcal{H}_1,...,\mathcal{H}_{N(t)}$ is taken. Denote

$$h_j(t) = m(t) - \eta_{i_j}\int_0^t e^{-\eta_{i_j}w}m(t-w)\bar{G}_W(w)dw$$

Denote $\vec{i} = (i_1,...,i_n)$, $\vec{i}_{-j} = (i_1,...,i_{j-1},i_{j+1},...,i_n)$, $\eta_{\vec{i}} = (\eta_{i_1},...,\eta_{i_n})$, and

$$\sum_{i_n=1}^{m}\cdots\sum_{i_2=1}^{m}\sum_{i_1=1}^{m}(\cdot)P(\mathcal{H}_1=\eta_{i_1})P(\mathcal{H}_2=\eta_{i_2})\cdots P(\mathcal{H}_n=\eta_{i_n}) = \sum_{\vec{i}}(\cdot)P(\mathfrak{H}=\eta_{\vec{i}})$$

Then

$$\begin{aligned}
E\left[\prod_{j=1}^{n}h_j(t)\right] &= \sum_{n=0}^{\infty}\sum_{\vec{i}}\left(\prod_{j=1}^{n}\frac{1}{m(t)}h_j\right)P(\mathcal{H}=\eta_{\vec{i}})P(N(t)=n)\\
&= \sum_{n=0}^{\infty}\frac{1}{m(t)^n}\left(\sum_{i_1=1}^{m}h_1P(\mathcal{H}_1=\eta_{i_1})\right)\sum_{\vec{i}_{-1}}\prod_{j=2}^{n}h_jP(\mathfrak{H}_{-1}=\eta_{\vec{i}_{-1}})P(N(t)=n)\\
&= \sum_{n=0}^{\infty}\frac{1}{m(t)^n}\prod_{j=1}^{n}\left(\sum_{j=1}^{m}h_jP(\mathcal{H}_j=\eta_{i_j})\right)P(N(t)=n)\\
&= \sum_{n=0}^{\infty}\frac{1}{m(t)^n}\prod_{j=1}^{n}\left(\sum_{j=1}^{m}h_jP(\mathcal{H}_j=\eta_{i_j})\right)\frac{m(t)^n}{n!}e^{-m(t)}\\
&= \sum_{n=0}^{\infty}\frac{1}{m(t)^n}\left(\prod_{j=1}^{n}E_{\mathcal{H}_j}[h_j]\right)\frac{m(t)^n}{n!}e^{-m(t)}
\end{aligned}$$

Let $h(t) = m(t) - \mathcal{H}\int_0^t e^{-\mathcal{H}w}m(t-w)\bar{G}_W(w)dw$ Since $\mathcal{H}_j \sim \mathcal{H}$ and i.i.d. $\forall\, j$, $E_{\mathcal{H}_j}[h_j] = E_{\mathcal{H}}[h(t)]$ for all j. Thus

$$\begin{aligned}
&\sum_{n=0}^{\infty}\frac{1}{m(t)^n}\left(\prod_{j=1}^{n}E_{\mathcal{H}_j}[h_j]\right)\frac{m(t)^n}{n!}e^{-m(t)}\\
&\quad= \sum_{n=0}^{\infty}\frac{1}{m(t)^n}\left(E_{\mathcal{H}}[h(t)]\right)^n\frac{m(t)^n}{n!}e^{-m(t)}\\
&\quad= e^{-m(t)}\sum_{n=0}^{\infty}\frac{m(t)^n}{n!}\left(E_{\mathcal{H}}[h(t)]\right)^n\\
&\quad= e^{-m(t)}\exp\left(m(t) - E_{\mathcal{H}}\left[\mathcal{H}\int_0^t e^{-\mathcal{H}w}m(t-w)\bar{G}_W(w)dw\right]\right)\\
&\quad= \exp\left(-E_{\mathcal{H}}\left[\mathcal{H}\int_0^t e^{-\mathcal{H}w}m(t-w)\bar{G}_W(w)dw\right]\right)
\end{aligned}$$

where the third equality uses the Taylor series representation of e^x. □

6.2 Proof of Theorem 4

Proof. From [21] and [9], $\psi = \frac{E[M]}{E[Y]+\nu}$, where Y is the length of time the server is operational during a particular renewal cycle and ν is the mean time to reboot. By [14], $E[Y] = \int_0^\infty S_Y(t)dt$, where $S_Y(t)$ is the unconditional survival function from Theorem 3. Therefore, the completion of the proof relies on deriving $E[M]$.
$M = \sum_{j=1}^{N(Y)} \mathbb{1}(T_j + W_j \leq Y)$ which may be rewritten as

$$M = \sum_{j=1}^{N(Y)} \mathbb{1}(R_j + V_j \leq Y)$$

where $\{(R_j, V_j)\}_{j=1}^{N(Y)}$ may be regarded as a random permutation of $\{(T_j, W_j)\}_{j=1}^{N(Y)}$ due to the mutual independence of $\{T_j\}, \{W_j\}$ and the respective i.i.d nature of both. Therefore,

$$E[M] = E\left[\sum_{j=1}^{N(Y)} \mathbb{1}(R_j + V_j \leq Y)\right]$$

For convenience and clarity, the following notation is introduced:

$$\mathfrak{R} = \{R_1, ..., R_n\}, \qquad \mathfrak{V} = \{V_1, ..., V_n\},$$
$$\mathfrak{H} = \{\mathcal{H}_1, ..., \mathcal{H}_n\}$$

with observed values

$$\mathfrak{r} = \{r_1, ..., r_n\}, \qquad \mathfrak{v} = \{v_1, ..., v_n\},$$
$$\mathfrak{h} = \{\eta_{i_1}, ..., \eta_{i_n}\}$$

By Bayes's Theorem,

$$f_{\mathfrak{R},\mathfrak{V},\mathfrak{H},Y,N}(\mathfrak{r},\mathfrak{v},\mathfrak{h},t,n) = f_{Y|\mathfrak{R},\mathfrak{V},\mathfrak{H},N}(t|\mathfrak{r},\mathfrak{v},\mathfrak{h},n) f_{\mathfrak{R},\mathfrak{V},\mathfrak{H},N}(\mathfrak{r},\mathfrak{v},\mathfrak{h},n)$$

By Lemma 3(Section 6.3), the conditional distribution $f_{Y|\mathfrak{R},\mathfrak{V},\mathfrak{H},N}(t|\mathfrak{r},\mathfrak{v},\mathfrak{h},n)$ is given by

$$f_{Y|\mathfrak{R},\mathfrak{V},\mathfrak{H},N}(t|\mathfrak{r},\mathfrak{v},\mathfrak{h},n) = e^{-\int_0^t r_0(s)ds - \sum_{j=1}^n \eta_{i_j} \min(v_j, t-r_j)} \left(r_0(t) + \sum_{j=1}^n \eta_{i_j} \mathbb{1}(v_j > t - r_j)\right) \tag{8}$$

Since all $\mathcal{H}_j \in \mathfrak{H}$ are i.i.d. and mutually independent of $\mathfrak{R}, \mathfrak{V}$, and N,

$$f_{\mathfrak{R},\mathfrak{V},\mathfrak{H},N}(\mathfrak{r},\mathfrak{v},\mathfrak{h},n) = f_{\mathfrak{R},\mathfrak{V},N}(\mathfrak{r},\mathfrak{v},n) f_{\mathfrak{H}}(\mathfrak{h}) = f_{\mathfrak{R},\mathfrak{V},N}(\mathfrak{r},\mathfrak{v},n) \prod_{j=1}^n P(\mathcal{H}_j = \eta_{i_j})$$

By Lemma 4(Section 6.3)

$$f_{\mathfrak{R},\mathfrak{V},N}(\mathfrak{r},\mathfrak{v},n) = \frac{1}{n!} \prod_{j=1}^n e^{\int_0^t \lambda(x)dx} \lambda(r_j) g_W(v_j)$$

$$f_{\mathfrak{R},\mathfrak{V},\mathfrak{H},N}(\mathfrak{r},\mathfrak{v},\mathfrak{h},n) = \frac{1}{n!} \prod_{j=1}^n e^{\int_0^t \lambda(x)dx} \lambda(r_j) g_W(v_j) \prod_{j=1}^n P(\mathcal{H}_j = \eta_{i_j}) \tag{9}$$

Finally, by multiplying (8) and (9)

$$f_{\mathfrak{R},\mathfrak{V},\mathfrak{H},Y,N(t)}(\mathfrak{r},\mathfrak{v},\mathfrak{h},t,n) = \frac{e^{-\int_0^t r_0(x)dx-\int_0^t \lambda(x)dx}}{n!}\left[\prod_{j=1}^{n} e^{-\eta_{i_j}\min(v_j,t-r_j)}\lambda(r_j)g_W(v_j)P(\mathcal{H}_j=\eta_{i_j})\right] \times\left[r_0(t)+\sum_{j=1}^{n}\eta_{i_j}\mathbb{1}(v_j>t-r_j)\right]$$

Denote $\mathfrak{v}_{-j}=(v_1,...,v_{j-1},v_{j+1},...,v_n)$, $\mathfrak{r}_{-j}=(r_1,...,r_{j-1},r_{j+1},...,r_n)$.
Let $\int_{\vec{a}}^{\vec{b}} f(\mathbf{x})d\mathbf{x}=\int_a^b\cdots\int_a^b f(x_1,...,x_n)dx_1\cdots dx_n$ for $a,b\in\mathbb{R}$.

$$\begin{aligned}
E[M] &= E\left[\sum_{j=1}^{N(Y)}\mathbb{1}(R_j+V_j\le Y)\right]\\
&= \sum_{n=1}^{\infty}\int_0^t\left[\sum_{j=1}^{n}\sum_{i_j=1}^{m}\int_0^t\int_0^{t-r_j}\int_{\vec{0}}^{\vec{t}}\int_{\vec{0}}^{\vec{\infty}} f_{\mathfrak{R},\mathfrak{V},\mathfrak{H},Y,N(t)}(\mathfrak{r},\mathfrak{v},\mathfrak{h},t,n)d\mathfrak{v}_{-j}d\mathfrak{r}_{-j}dv_jdr_j\right]dt\\
&= \sum_{n=1}^{\infty}\int_0^{\infty}\frac{1}{n!}r_0(t)e^{-\int_0^t r_0(x)dx-\int_0^t\lambda(x)dx}nE_{\mathcal{H}}\left[\int_0^t\int_0^{t-r}e^{-\mathcal{H}v}g_W(v)dv\lambda(r)dr\right]\\
&\quad\times\left\{E_{\mathcal{H}}\left[\int_0^t\int_0^{t-r}e^{-\mathcal{H}v}g_W(v)dv\lambda(r)dr+\int_0^t e^{-\mathcal{H}(t-r)}\bar{G}_W(t-r)\lambda(r)dr\right]\right\}^{n-1}dt\\
&\quad+\sum_{n=1}^{\infty}\int_0^{\infty}\frac{1}{n!}e^{-\int_0^t r_0(x)dx-\int_0^t\lambda(x)dx}\\
&\qquad\times n(n-1)E_{\mathcal{H}}\left[\mathcal{H}\int_0^t\int_0^{t-r}e^{-\mathcal{H}v}g_W(v)dv\lambda(r)dr\int_0^t e^{-\mathcal{H}(t-r)}\bar{G}_W(t-r)\lambda(r)dr\right]\\
&\qquad\times\left\{E_{\mathcal{H}}\left[\int_0^t\int_0^{t-r}e^{-\mathcal{H}v}g_W(v)dv\lambda(r)dr+\int_0^t e^{-\mathcal{H}(t-r)}\bar{G}_W(t-r)\lambda(r)dr\right]\right\}^{n-2}dt
\end{aligned}$$

Let $a(t)=\int_0^t\int_0^{t-r}e^{-\mathcal{H}v}g_W(v)dv\lambda(r)dr$. Through a change of variables, $a(t)=\int_0^t e^{-\mathcal{H}v}g_W(v)m(t-v)dv$. Let $b(t)=\int_0^t e^{-\mathcal{H}(t-r)}\bar{G}_W(t-r)\lambda(r)dr$. Then

$$\begin{aligned}
E[M] &= \sum_{n=1}^{\infty}\int_0^{\infty}\frac{1}{(n-1)!}r_0(t)E_{\mathcal{H}}[a(t)](E_{\mathcal{H}}[a(t)+b(t)])^{n-1}e^{-\int_0^t r_0(x)dx-\int_0^t\lambda(x)dx}dt\\
&\quad+\sum_{n=2}^{\infty}\int_0^{\infty}\frac{1}{(n-2)!}E_{\mathcal{H}}[\mathcal{H}a(t)b(t)](E_{\mathcal{H}}[a(t)+b(t)])^{n-2}dt\\
&= \int_0^{\infty}r_0(t)e^{-\int_0^t r_0(x)-\int_0^t\lambda(x)dx}E_{\mathcal{H}}[a(t)]\left(\sum_{n=1}^{\infty}\frac{1}{(n-1)!}(E_{\mathcal{H}}[a(t)+b(t)])^{n-1}\right)dt\\
&\quad+\int_0^{\infty}e^{-\int_0^t r_0(x)-\int_0^t\lambda(x)dx}E_{\mathcal{H}}[\mathcal{H}a(t)b(t)]\left(\sum_{n=2}^{\infty}\frac{1}{(n-2)!}(E_{\mathcal{H}}[a(t)+b(t)])^{n-2}\right)dt\\
&= \int_0^{\infty}r_0(t)e^{-\int_0^t r_0(x)-\int_0^t\lambda(x)dx}E_{\mathcal{H}}[a(t)]e^{E_{\mathcal{H}}[a(t)+b(t)]}dt\\
&\quad+\int_0^{\infty}e^{-\int_0^t r_0(x)-\int_0^t\lambda(x)dx}E_{\mathcal{H}}[\mathcal{H}a(t)b(t)]e^{E_{\mathcal{H}}[a(t)+b(t)]}dt\\
&= \int_0^{\infty}e^{-\int_0^t r_0(x)-\int_0^t\lambda(x)dx+E_{\mathcal{H}}[a(t)+b(t)]}\left[r_0(t)E_{\mathcal{H}}[a(t)]+E_{\mathcal{H}}[\mathcal{H}a(t)b(t)]\right]dt
\end{aligned}\tag{10}$$

□

6.3 Auxiliary Lemmata

Lemma 1 (Conditional Joint Distribution of Arrival Times). *Let $\{N(t)\}$ be a nonhomgenous Poisson process describing the arrivals of client requests to the web server, and let $T_1,...T_n$ be the arrival times of the client requests. Then, given $N(t)=n$, the conditional joint distribution of $T_1,...T_n$, denoted $f_{T_1,...T_n|N(t)=n}(t_1,...t_n)$ has distribution equal to the joint distribution of the order statistics $T'_{[1]},...T'_{[n]}$, where $T'_1,...T'_n$ are i.i.d. with pdf $f(x)=\frac{\lambda(x)}{m(t)}$. The pdf is given by*

$$f_{T_1,...T_n|N(t)=n}(t_1,...t_n|n)=\frac{n!}{m(t)}\prod_{i=1}^{n}\lambda(t_i), \qquad 0\le t_1\le ...\le t_n\le t$$

Proof. Let $N(t)=n$, and let $0<t_1<...<t_n<t$. Let $h_i, i=1,..,n$ be small enough such that $t_i+h_i<t_{i+1}\ \forall\ i=1,..,n-1$. Denote A_i as the event that the server sees exactly 1 arrival in $[t_i,t_i+h_i]$, and B be the event that no events arrive outside the set
$U:=[0,t_1]\cup[t_1,t_1+h_1]\cup[t_2,t_2+h_2]\cup...\cup[t_n,t_n+h_n]\cup[t_n+h_n,t]$. Then

$$P(t_i\le T_i\le t_i+h_i, i=1,...,n|N(t)=n)=\frac{P(A_1\cap A_2\cap...\cap A_n\cap B)}{P(N(t)=n)}$$

From (**??**) in the preliminaries, $P(N(t)=n)=\frac{e^{-m(t)}m(t)^n}{n!}$. For each $i=1,..,n$

$$\begin{aligned}P(t_i\le T_i\le t_i+h_i)&=P(N(t_i+h_i)-N(t_i)=1)\\&=e^{-(m(t_i+h_i)-m(t_i))}(m(t_i+h_i)-m(t_i))\end{aligned}$$

The next step is the calculation of $P(B)$. The complement of the set U can be broken into the disjoint intervals $[0,t_1],[t_1+h_1,t_2],...[t_i+h_i,t_{i+1}],...[t_n+h_n,t]$ By definition of a NHPP (See, for example, [21]),

$$\begin{aligned}P(B)&=P[N(t)-N(t_n+h_n)=0]P[N(t_1)-N(0)=0]\prod_{i=1}^{n-1}P[N(t_{i+1})-N(t_i+h_i)=0]\\&=e^{-m(t_1)}e^{-(m(t)-m(t_n+h_n))}\prod_{i=1}^{n-1}e^{-(m(t_{i+1})-m(t_i+h_i))}\\&=\exp\left(-\left[m(t)+\sum_{i=1}^{n}m(t_i)-\sum_{i=1}^{n}m(t_i+h_i)\right]\right)\end{aligned}$$

Now, again using the fact that a NHPP has independent increments, and simplifying,

$$\begin{aligned}P(t_i\le T_i\le t_i+h_i, i=1,...,n|N(t)=n)&=\prod_{i=1}^{n}P(t_i\le T_i\le t_i+h_i)\\&=n!\prod_{i=1}^{n}\frac{m(t_i+h_i)-m(t_i)}{m(t)}\end{aligned}$$

Letting $h_i\to 0\ \forall\ i$,

$$f_{T_1,...T_n|N(t)=n}(t_1,...t_n|n)=n!\prod_{i=1}^{n}\frac{\lambda(t_i)}{m(t)}$$

□

Lemma 2 (Expectation of $e^{-\eta_{i_{j'}} \min(W_j, t-T_{j'})}$)**.**

$$E\left[E\left[e^{-\eta_{i_{j'}} \min(W_j, t-T'_j)} \middle| W_j\right]\right] = \frac{1}{m(t)}\left(m(t) - \eta_{i_{j'}} \int_0^t e^{-\eta_{i_{j'}} w} m(t-w)\bar{G}_W(w)dw\right)$$

Proof. Two cases must be considered: (1) $w \leq t$ and (2) $w > t$. For $w \leq t$,

$$\begin{aligned} E\left[e^{-\eta_{i_j} \min(W_j, t-T'_j)} | W_j = w\right] &= \int_0^{t-w} e^{-\eta_{i_j} w} \frac{\lambda(x)}{m(t)} dx + \int_{t-w}^t e^{-\eta_{i_j}(t-x)} \frac{\lambda(x)}{m(t)} dx \\ &= e^{-\eta_{i_j} w} \frac{m(t-w)}{m(t)} + e^{-\eta_{i_j} t} \int_0^t e^{\eta_{i_j} x} \frac{\lambda(x)}{m(t)} dx \end{aligned}$$

For $w > t$,

$$E\left[e^{-\eta_{i_j} \min(W_j, t-T'_j)} | W_j = w\right] = e^{-\eta_{i_j} t} \int_0^t e^{\eta_{i_j} x} \frac{\lambda(x)}{m(t)} dx$$

Therefore,

$$\begin{aligned} E_W\left[e^{-\eta_{i_j} \min(W_j, t-T'_j)}\right] &= E_W\left[E\left[e^{-\eta_{i_j} \min(W_j, t-T'_j)} \middle| W_j = w\right]\right] \\ &= \frac{1}{m(t)}\left(\int_0^t e^{-\eta_{i_j} w} m(t-w) g_W(w) dw\right. \\ &\quad + e^{-\eta_{i_j} t} \int_0^t \int_{t-w}^t e^{\eta_{i_j} x} \lambda(x) dx g_W(w) dw \qquad (*) \\ &\quad \left. + \bar{G}_W(t) e^{-\eta_{i_j} t} \int_0^t e^{\eta_{i_j} x} \lambda(x) dx\right) \end{aligned}$$

Focusing on $(*)$, we make the change of variables $w = t - x$ and change the order of intergration, yielding a new second term.

$$\begin{aligned} &= \frac{1}{m(t)}\left(\int_0^t e^{-\eta_{i_j} w} m(t-w) g_W(w) dw\right. \\ &\quad + e^{-\eta_{i_j} t} \int_0^t e^{\eta_{i_j} x} \lambda(x) \int_{t-x}^t g_W(w) dw dx \\ &\quad \left. + \bar{G}_W(t) e^{-\eta_{i_j} t} \int_0^t e^{\eta_{i_j} x} \lambda(x) dx\right) \end{aligned}$$

Combining the second and third terms:

$$\begin{aligned} &= \frac{1}{m(t)}\left(\int_0^t e^{-\eta_{i_j} w} m(t-w) g_W(w) dw\right. \\ &\quad \left. + e^{-\eta_{i_j} t} \int_0^t e^{\eta_{i_j} x} \lambda(x) \bar{G}_w(t-x) dx\right) \end{aligned}$$

Changing variables again in the second term, using $w = t - x$, we get

$$\begin{aligned} &= \frac{1}{m(t)}\left(\int_0^t e^{-\eta_{i_j} w} m(t-w) g_W(w) dw\right. \\ &\quad \left. + \int_0^t e^{-\eta_{i_j} w} \lambda(t-w) \bar{G}_w(w) dw\right) \end{aligned}$$

Integrating the first term by parts, we get

$$
\begin{aligned}
&= \frac{1}{m(t)} \Bigg(\left[-e^{-\eta_{i_j} w} m(t-w) \bar{G}_w(w) \right] \Big|_0^t \\
&\quad - \int_0^t (\eta_{i_j} e^{-\eta_{i_j} w} m(t-w) + e^{-\eta w} \lambda(t-w)) \bar{G}_W(w) dw \\
&\quad + \int_0^t e^{-\eta_{i_j} w} \lambda(t-w) \bar{G}_w(w) dw \Bigg) \\
&= \frac{1}{m(t)} \left(m(t) - \eta_{i_j} \int_0^t e^{-\eta_{i_j} w} m(t-w) \bar{G}_W(w) dw \right)
\end{aligned}
$$

□

Lemma 3 (Conditional distribution of Renewal Cycle Length)**.** *The conditional distribution* $f_{Y|\mathfrak{R},\mathfrak{V},\mathfrak{H},N}(t|\mathfrak{r},\mathfrak{v},\mathfrak{h},n)$ *is given by*

$$
\begin{aligned}
f_{Y|\mathfrak{R},\mathfrak{V},\mathfrak{H},N}(t|\mathfrak{r},\mathfrak{v},\mathfrak{h},n) &= \exp\left(- \int_0^t r_0(s) ds - \sum_{j=1}^n \eta_{i_j} \min(v_j, t - r_j) \right) \\
&\quad \times \left(r_0(t) + \sum_{j=1}^n \eta_{i_j} \mathbb{1}(v_j > t - r_j) \right)
\end{aligned}
$$

Proof. Denote the condition $C = \{\mathfrak{R} = \mathfrak{r}, \mathfrak{V} = \mathfrak{v}, \mathfrak{H} = \mathfrak{h}, N(t) = n\}$. Then

$$
f_{Y|C}(t|c) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \left(P(Y > t | C = c) - P(Y > t + \Delta t | C = c) \right) \tag{11}
$$

From (3),

$$
P(Y > t | C = c) = \exp\left(- \int_0^t r_0(s) ds - \sum_{j=1}^n \eta_j^{i_j} \min(v_j, t - r_j) \right)
$$

Recall that $\mathcal{B}(s) = r_0(s) + \sum_{j=1}^{N(t)} \mathbb{1}(R_j \le s \le R_j + V_j)$. We now derive $P(Y > t + \Delta t | C = c)$.

Let $A_k = \{N(t + \Delta t) - N(t) = k\}$ be the event such that k requests arrived between t and $t + \Delta t$. From the definition of a nonhomogenous Poisson process [21],

$$
\begin{aligned}
P(A_0) &= 1 - \lambda(t)\Delta t + o(\Delta t) = e^{-(m(t+\Delta t) - m(t))} \\
P(A_1) &= \lambda(t)\Delta t + o(\Delta t) = (m(t + \Delta t) - m(t)) e^{-(m(t+\Delta t) - m(t))} \\
&\vdots \\
P(A_k) &= \frac{(m(t + \Delta t) - m(t))^k}{k!} e^{-(m(t+\Delta t) - m(t))}
\end{aligned}
$$

for $k \ge 2$. By the Law of Total Probability,

$$
P(Y > t + \Delta t | C = c) = \sum_{k=0}^{\infty} P(Y > t + \Delta t | C \cap A_k) P(A_k) \tag{12}
$$

We look at each fixed k and show that the terms for $k \geq 2$ are of $o(\Delta k)$. For $k = 0$,

$$\begin{aligned} P(Y > t + \Delta t | C \cap A_0) P(A_0) &= e^{-\int_0^{t+\Delta t} \mathcal{B}(s)ds}(1 - \lambda(t)\Delta t + o(\Delta t)) \\ &= e^{-\int_0^{t} \mathcal{B}(s)ds} e^{-\int_t^{t+\Delta t} \mathcal{B}(s)ds}(1 - \lambda(t)\Delta t + o(\Delta t)) \end{aligned} \tag{13}$$

$$\begin{aligned} \int_t^{t+\Delta t} \mathcal{B}(s)ds &= \int_t^{t+\Delta t} r_0(s)ds + \left(\sum_{j=1}^{n} \eta_j^{i_j} \mathbb{1}(v_j > t - r_j) \right) \Delta t \\ &= r_0(t) + o(\Delta t) + \left(\sum_{j=1}^{n} \eta_j^{i_j} \mathbb{1}(v_j > t - r_j) \right) \Delta t \end{aligned}$$

Using the fact that $e^{-x+o(x)} = 1 - x + o(x)$ for small x,

$$e^{-\int_t^{t+\Delta t} \mathcal{B}(s)ds} = 1 - \left(r_0(t) + \sum_{j=1}^{n} \eta_j^{i_j} \mathbb{1}(v_j > t - r_j) \right) \Delta t + o(\Delta t)$$

Substituting into (13),

$$\begin{aligned} P(Y > t + \Delta t | C \cap A_0) P(A_0) &= \left[e^{-\int_0^{t} \mathcal{B}(s)ds} 1 - \left(r_0(t) + \sum_{j=1}^{n} \eta_j^{i_j} \mathbb{1}(v_j > t - r_j) \right) \Delta t + o(\Delta t) \right] \\ &\quad [1 - \lambda(t)\Delta t + o(\Delta t)] \\ &= e^{-\int_0^{t} \mathcal{B}(s)ds} \left(1 - [r_0(t) + \sum_{j=1}^{n} \eta_j^{i_j} \mathbb{1}_{v_j > t - r_j}] \Delta t - \lambda(t)\Delta t + o(\Delta t) \right) \end{aligned} \tag{14}$$

Now, for $k = 1$, we must contend with the arrival of one request in $[t, t + \Delta t]$ in addition to the n fixed arrivals given by C. Call this arrival t_1, with corresponding time to completion w_1. We then have two cases:

(1) $t_1 + w_1 < t + \Delta t$, that is, the request arrives and is serviced before $t + \Delta t$, or

(2) $t_1 + w_1 > t + \Delta t$. In this case, the service time is greater than Δt.

With both of these cases, the failure rate increases by an $\eta_{t_1}^{n_{i_{t_1}}}$ for a time smaller than Δt, which we will denote as $\Delta^1(t)$. Then

$$\Delta^1(t) = \begin{cases} w_1, & t_1 + w_1 < t + \Delta t \\ t + \Delta t - t_1, & t_1 + w_1 > t + \Delta t \end{cases}$$

In what follows we apply previous arguments to calculate $P(Y > t + \Delta t | C \cap A_1) P(A_1)$. Now,

$$\exp\left(-\int_t^{t+\Delta t} \mathcal{B}(s)ds \right) = \exp\left(-r_0(t) + \left[\sum_{j=1}^{n} \eta_j^{i_j} \mathbb{1}(v_j > t - r_j) \right] \Delta t + \eta_{t_1}^{i_{t_1}} \Delta^1 t + o(\Delta t) \right)$$

The above simplifies to

$$\exp\left(-\int_t^{t+\Delta t} \mathcal{B}(s)ds \right) = 1 - [r_0(t) + \sum_{j=1}^{n} \eta_j^{i_j} \mathbb{1}(v_j > t - r_j)] \Delta t + \eta_{t_1}^{i_{t_1}} \Delta^1 t + o(\Delta t)$$

Now,

$$P(Y > t+\Delta t|C\cap A_1)P(A_1) = e^{-\int_0^t \mathcal{B}(s)ds}\left(\left[1-\left[r_0(t)+\sum_{j=1}^{n}\eta_j^{i_j}\mathbb{1}(v_j > t-r_j)\right]\Delta t\right.\right.$$
$$\left.\left.+\eta_{t_1}^{i_{t_1}}\Delta^1 t + o(\Delta t)\right][\lambda(t)\Delta t + o(\Delta t)]\right)$$
$$= e^{-\int_0^t \mathcal{B}(s)ds}(\lambda(t)\Delta t + o(\Delta t)) \tag{15}$$

Combining (15) with (14),

$$\sum_{k=0}^{1} P(Y > t+\Delta t|C\cap A_k)P(A_k) = e^{-\int_0^t \mathcal{B}(s)ds}\left(1-(r_0(t)+\sum_{j=1}^{n}\eta_j^{i_j}\mathbb{1}_{v_j>t-r_j})\Delta t\right)+o(\Delta t) \tag{16}$$

For $k \geq 2$, we will now show that the contribution to (12) is negligible. Using similar notation established in the case of $k = 1$,

$$P(Y > t+\Delta t|C\cap A_k) = e^{-\int_0^t \mathcal{B}(s)ds}\left[1-[(r_0(t)+\sum_{j=1}^{n}\eta_j^{i_j}\mathbb{1}_{v_j>t-r_j})\Delta t+\sum_{l=1}^{k}\eta_{t_l}^{i_{t_l}}\Delta^l t]+o(\Delta t)\right]$$

Now, we see that $\forall\, k \geq 2$,

$$P(Y > t+\Delta t|C\cap A_k) \leq e^{-\int_0^t \mathcal{B}(s)ds}$$

and hence

$$\sum_{k=2}^{\infty} P(Y > t+\Delta t|C\cap A_k)P(A_k) \leq e^{-\int_0^t \mathcal{B}(s)ds}\sum_{k=2}^{\infty}P(A_k)$$
$$= e^{-\int_0^t \mathcal{B}(s)ds}(1-P(A_0)-P(A_1))$$
$$= e^{-\int_0^t \mathcal{B}(s)ds}o(\Delta t)$$

Therefore,

$$P(Y > t+\Delta t|C) = e^{-\int_0^t \mathcal{B}(s)ds}\left(1-(r_0(t)+\sum_{j=1}^{n}\eta_j^{i_j}\mathbb{1}_{v_j>t-r_j})\Delta t\right)+o(\Delta t)$$

Now, we see that

$$\frac{1}{\Delta t}(P(Y > t|C=c) - P(Y > t+\Delta t|C=c)) = e^{-\int_0^1 \mathcal{B}(s)ds}\left(r_0(t)+\sum_{j=1}^{n}\eta_j^{i_j}\mathbb{1}_{v_j>t-r_j}+\frac{o(\Delta t)}{\Delta t}\right)$$

Then, letting $\Delta t \to 0$,

$$f_{Y|\mathfrak{R},\mathfrak{V},\mathfrak{H},N}(t|\mathfrak{r},\mathfrak{v},\mathfrak{h},n) = \left\{\exp\left(-\int_0^t r_0(s)ds - \sum_{j=1}^{n}\eta_j^{i_j}\min(v_j, t-r_j)\right)\right\}\left(r_0(t)+\sum_{j=1}^{n}\eta_j^{i_j}\mathbb{1}_{v_j>t-r_j}\right)$$

□

Lemma 4 (Joint PDF of $\mathfrak{R}, \mathfrak{V}, N$)**.**

$$f_{\mathfrak{R},\mathfrak{V},N(t)}(\mathfrak{r},\mathfrak{v},n) = \frac{1}{n!}\exp\left(-\int_0^t \lambda(s)ds\right)\prod_{j=1}^{n}\lambda(r_j)g_W(v_j)$$

Proof. Let $\mathcal{W} = (W_1,...,W_n)$ be the service times under the condition that $N(t) = n$, and let $\mathcal{W} = \sqsupseteq = (w_1,...,w_n)$. Let $(r_{[1]},...,r_{[n]})$ be the ordered vector of $\mathfrak{r}$. There are $n!$ possible orderings of $\mathfrak{r}$. Now, $P(r_{[i]} \leq R_{[i]} \leq r_{[i]} + h_i)$ for some $h_i > 0, i = 1,...,n$ is given by

$$P(r_{[i]} \leq R_{[i]} \leq r_{[i]} + h_i) = e^{-(m(r_{[i]}+h_i)-m(r_{[i]})}(m(r_{[i]} + h_i) - m(r_{[i]}))$$

We may see that the joint distribution of $\mathfrak{R}$ is identical to the distribution of the order statistics of $\mathfrak{R}$:

$$f_{\mathfrak{R},N(t)}(\mathfrak{r},n) = \frac{1}{n!}\prod_{j=1}^{n}\lambda(r_j)\exp\left(-\int_0^t \lambda(s)ds\right)$$

$\mathfrak{R}, \mathfrak{V}$ are mutually independent. Therefore,

$$f_{\mathfrak{R},\mathfrak{V},N(t)}(\mathfrak{r},\mathfrak{v},n) = \frac{1}{n!}\exp\left(-\int)_0^t\lambda(s)ds\right)\prod_{j=1}^{n}\lambda(r_j)g_W(v_j)$$

□

Acknowledgments

The author is grateful to her thesis advisor, Dr. Andrzej Korzeniowski at the University of Texas at Arlington

References

[1] Robert R. Abernethy. *The New Weibull Handbook: Reliability and Statistical Analysis for Predicting Life, Safety, Supportability, Risk, Cost and Warranty Claims*. Dr. Robert. Abernethy, 2006.

[2] Arnold O. Allen. *Probability, Statistics, and Queuing Theory with Computer Science Appliations, 2nd edition*. Academic Press, Inc., 1990.

[3] Fred Douglis Mark Chamness Guanlin Lu Darren Sawyer Surendar Chandra Windsor Hu Ao Ma, Rachel Traylor. Raidshield: Characterizing, monitoring, and proactively protecting against disk failures. *ACM Transactions on Storage*, 11, 2015.

[4] M.A. Augustin and E.A. Pena. A dynamic competing risks model. *Probability in the Engineering and Informational Sciences*, 13:333–358, 1999.

[5] M. Bargout. Predicting software reliability using an imperfect debugging jelinski moranda nonhomogenous poisson process model. *Model Assisted Statistical Applications*, 5:31–41, 2010.

[6] N. Bhatti and R. Freidrich. Web server support for tiered services. *IEEE Network*, 13:64–71, 1999.

[7] P.J. Boland and Ni Chuiv. Optimal times for software release when repar is imperfect. *Statistical and Probability Letters*, 77:1176–1184, 2007.

[8] P.J. Boland and H. Singh. Determining the optimal time for software in the geometric poisson reliability model. *International Journal of Reliability and Quality Safety Engineering*, 9:201–213, 2002.

[9] Ki Hwan Cha and Eui Yong Lee. A stochastic breakdown model for an unreliable web server system and an optimal admission control policy. *Journal of Applied Probability*, 48(2):453–466, 2011.

[10] David Culler and Matt Welsh. Adaptive overload control for busy internet servers. In *In Proceedings of the 4th conference on USENIX Symposium on Internet Technologies and Systems*, volume 4, 2003.

[11] Z. Jelinski and P. Moranda. Software reliability research. *Statistical Computer Performance Evaluation*, pages 465–484, 1972.

[12] Stephen H. Kan. *Metrics and Models in Software Quality Engineering*. Addison-Wesley, 2003.

[13] Nigel Thomas Katja Gilly, Carlos Juiz and Ramon Puigjaner. Adaptive admission control algorithm in a qos-aware web system. *Journal of Information Sciences*, 199:58–77, 2012.

[14] Lawrence M. Leemis. *Reliability: Probabilistic Models and Statistical Methods, 2nd edition*. Prentice-Hall, 2009.

[15] Y. Ling and J. Mi. An optimal trade-off between content freshness and refresh cost. *Journal of Applied Probability*, 41:721–734, 2004.

[16] B. Littlewood. The littlewood-verrall model for software reliability compared with some rivals. *Journal of Systems and Software*, 1:251–258, 1980.

[17] J. Mi. Age replacement policy and optimal work size. *Journal of Applied Probability*, 39:296–311, 2002.

[18] Alexandru Iosup Nezih Yigitbasi, Ozan Sonmez and Dick Epema. Performance evaluation of overload control in multi-cluster grids. In *Proceedings of the 2011 IEEE/ACM 12th annual Conference on Grid Computing*, pages 173–180, 2011.

[19] H. Singh P.J. Boland and B. Cukic. Stochastic orders in partition and random testing of software. *Journal of Applied Probability*, 39:555–565, 2002.

[20] Marvin Rausand and Arnljot Høyland. *System Reliability Theory: Models, Statistical Methods, and Applications, 2nd edition*. John Wiley and Sons, 2004.

[21] Sheldon Ross. *Stochastic Processes, 2nd edition*.

[22] N.D. Singpurwalla. Determining an optimal time interval for testing and debugging software. *IEEE Transactions on Software Engineering*, 17:313–319, 1991.

[23] Chuan Ye and Haining Wang. Profit-aware overload protection in e-commerce websites. *Journal of Network and Computer Applications*, 32:347–356, 2009.