Dynamic Reliability of a Cluster Server

Andrzej Korzeniowski[1], Rachel Traylor[1]

[1] Department of Mathematics, University of Texas at Arlington

Correspondence: Andrzej Korzeniowski, Department of Mathematics, University of Texas at Arlington, Arlington, TX 76019, USA. Tel: 1-817-272-3595. E-mail: korzeniowski@uta.edu

Abstract

Suppose a single server has K channels, each of which performs a different task. Customers arrive to the server via a nonhomogenous Poisson process with intensity $\lambda(t)$ and select 0 to K tasks for the server to perform. Each channel services the tasks in its queue independently, and the customer's job is complete when the last task selected is complete. The stress to the server is a constant multiple η of the number of tasks selected by each customer, and thus the stress added to the server by each customer is random. Under this model, we provide the survival function for such a server in both the case of independently selected channels and correlated channels. A numerical comparison of expected lifetimes for various arrival rates is given, and the relationship between the dependency of channel selection and expected server lifetime is presented.

Keywords: cluster server, correlated Bernoulli variables, reliability, dynamic server lifetime

1. Introduction

Many single-server queuing models assume each arrival brings one task that must be completed by the server (Gross & Shortle, 2008; Tang, 1997). Common examples include a cashier's register, a hostess stand at a restaurant, or a basic web server. Of particular interest is the reliability of such queuing servers under workload or stress. In particular, Cha and Lee (2011) formulated a dynamic reliability model for a web server wherein customers arrive via a nonhomogenous Poisson process $\{N(t), t \geq 0\}$, the service times are i.i.d. with general distribution $G(w)$, and each task brought by a customer increases the stochastic hazard function by a constant η for the duration of its time in the system. (Cha and Lee, 2011).

Traylor (2015) provided a generalization to the model of Cha and Lee that relaxed the constant stress assumption and allowed for the job stresses to be i.i.d. with probability distribution $\mathcal{H}$. It is assumed that the set of arrival times $\mathfrak{T} = \{T_j\}_{j=1}^{N(t)}$ are mutually independent. The service times W_j, $j = 1, ..., N(t)$ remain i.i.d. and mutually independent of the set of arrival times. It is also assumed that the job stresses $\mathcal{H}_j$, $j = 1, ..., N(t)$ are mutually independent of all arrival times and service times. Under these conditions, Traylor provides the survival function for such a server in the following theorem.

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

$$S_Y(t) = \bar{F}_0(t) \exp\left(-E_{\mathcal{H}}\left[\mathcal{H}\int_0^t e^{-\mathcal{H}w}m(t-w)\bar{G}_W(w)dw\right]\right) \tag{1}$$

where $\bar{F}_0(t) = \exp\left(-\int_0^t r_0(s)ds\right)$, $\bar{G}_w(w) = 1 - G(w)$, $m(t) \equiv E[N(t)] = \int_0^t \lambda(x)dx$. and $r_0(s)$ is the hazard function of the idle server at time s.

Suppose now that this single server is partitioned into K channels, with each channel assigned to a different possible task a customer may choose. Customers may choose 0 to K tasks for the server to complete, and the customer is fully serviced when the last task is complete. The choice of such tasks may be correlated or uncorrelated. We propose a generalization to the random stress model of Traylor (2015) that gives the survival function for such a "clustered-task" server in both the correlated and independent cases. This provides a highly general reliability model for a variety of server types in many different industries. It also allows for a correlated multi-server system to be analyzed in a simple manner.

The paper is organized as follows. Section 2 gives a general description of the server model and its assumptions. Section 3 examines the reliability of the server when the channel selection is independent; correlated channels are discussed in

section 4. Section 5 provides some numerical illustrations of the effect of channel selection dependency on the expected lifetime of the server, and Section 6 gives the conclusion.
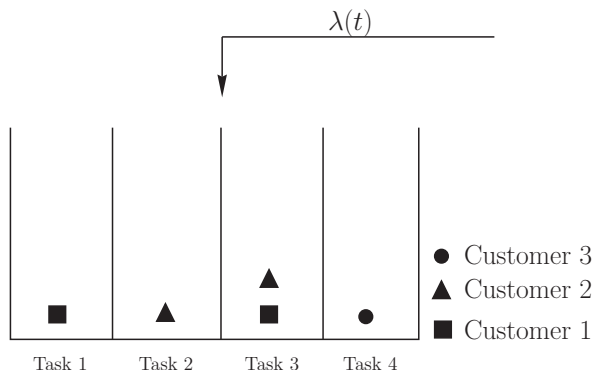

Figure 1. Illustration of Clustered Tasks in a Multichannel Server

2. Model Description

Suppose a server has multiple channels $Q_1, ..., Q_K$, but each channel serves a different type of task. A customer arrives to the server and may select any number from 0 to K tasks for the server to perform. Said customer will select each possible task j with probability p_j. Figure 1 illustrates an example of such a situation in which three customers visit the server and each customer picks a different number and set of tasks at random. A customer is considered fully serviced (i.e. the job is complete) upon completion of the last task belonging to that particular customer.

2.1 Model Assumptions

The following mathematical assumptions are made for the multichannel server with clustered tasks:

(i) Customers arrive to the server with K channels via a nonhomogenous Poisson process (NHPP) with intensity $\lambda(t)$.

(ii) The breakdown rate of the idle server is given by $r_0(t)$.

(iii) Each channel corresponds to a different task the server can perform.

(iv) The selection of each task is a Bernoulli random variable with probability p_k. Thus the number of tasks selected by each customer is a binomial random variable.

(v) The workload stress to the server is a constant multiple η of the number of tasks requested by the customer, i.e. the additional stress is given by ηN, where N is the number of tasks requested.

(vi) The distribution of each channel's service time is given by $G_i(w), i = 1, ..., K$. Since the customer's service is not complete until all requested tasks have finished, the service life distribution for the customers is therefore given by $\max_i G_i(w)$.

Under these assumptions, this model is a special interpretation of the random stress environment developed in Traylor (2015). In this case, the random workload stress is ηN, where N is a binomial random variable, and the service life distribution $G_W(w) = \max_i G_i(w)$, which may be easily obtained through the mathematical properties of order statistics. Two variations are considered in this section: independent channels and correlated channels.

3. Independent Channels

Suppose the selection probabilities for each task in the server are identical, that is, $p_1 = p_2 = \ldots = p_K = p$. Then $N \sim \text{Bin}(K, p)$. Using Theorem 1, the survival function of the multichannel server is given in the following theorem:

Theorem 2 (Survival Function of Multichannel Server with Clustered Tasks and Independent Channels)**.** *Suppose conditions (i)-(vi) above are satisfied. In addition, assume $p_1 = p_2 = \ldots = p_K = p$. Then the survival function of the server is given by*

$$S_Y(t) = \bar{F}_0(t) \exp\left(-K\eta\left[e^{-\eta t}\left(1 - p + pe^{-\eta t}\right)^{K-1} - p(1-p)^{K-1}\right]\int_0^t m(t-w)\bar{G}_W(w)dw\right)$$

where $m(x) = \int_0^x \lambda(s)ds$, $\bar{F}_0(t) = e^{-\int_0^t r_0(s)ds}$, $\bar{G}_W(w) = 1 - G_W(w)$, and $G_W(w) = \max_i G_i(w)$.

Proof. Since $p_1 = \ldots = p_K = p$, the number of tasks selected by any particular customer $N \sim \text{Bin}(K, p)$. Then the random stress distribution $\mathcal{H}$ is given by $\mathcal{H} = \eta N$. Thus

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

In this case,

$$E\left[\mathcal{H} \int_0^t e^{-\mathcal{H}w} m(t-w)\bar{G}_W(w)dw\right]$$

$$= E\left[\eta N \int_0^t e^{-\eta Nw} m(t-w)\bar{G}_W(w)dw\right]$$

$$= \sum_{n=0}^{K}\left[\eta n \int_0^t e^{-\eta nw} m(t-w)\bar{G}_W(w)dw\right] \cdot P(N = n)$$

$$= \sum_{n=0}^{K}\left[\eta n \int_0^t e^{-\eta nw} m(t-w)\bar{G}_W(w)dw\right]\binom{K}{n}p^n(1-p)^{K-n}$$

$$= \eta \int_0^t m(t-w)\bar{G}_W(w)\left(\sum_{n=0}^{K} ne^{-\eta nw}\binom{K}{n}p^n(1-p)^{K-n}\right)dw$$

Now,

$$\sum_{n=0}^{K} ne^{-\eta nw}\binom{K}{n}p^n(1-p)^{K-n} = \sum_{n=0}^{K} \frac{K!}{(K-n)!n!} ne^{-\eta nw} p^n(1-p)^{K-n}$$

$$= \sum_{n=0}^{K} \frac{K(K-1)!}{(n-1)!(K-1-(n-1))!} e^{-\eta nw} p^n(1-p)^{K-n}$$

$$= \sum_{n=0}^{K} K\binom{K-1}{n-1} e^{-\eta nw} p^n(1-p)^{K-n}$$

Making a change of indices, let $j = n - 1$. Then

$$\sum_{n=0}^{K} K\binom{K-1}{n-1} e^{-\eta nw} p^n(1-p)^{K-n} = K\sum_{j=0}^{K-1} \binom{K-1}{j} p^{j+1}(1-p)^{K-(j+1)} e^{-\eta(j+1)w}$$

Note the above resembles a scaled and shifted moment generating function of a binomial random variable. Let $X \sim \text{Bin}(K-1, p)$. Then

$$K\sum_{j=0}^{K-1} \binom{K-1}{j} p^{j+1}(1-p)^{K-(j+1)} e^{-\eta(j+1)w} = K\left(E\left[e^{-\eta(X+1)t}\right] - P(X = 0)\right)$$

$$= K\left(e^{-\eta t} E\left[e^{-\eta Xt} - p(1-p)^{K-1}\right]\right)$$

$$= K\left(e^{-\eta t}\left[1 - p + pe^{-\eta t}\right]^{K-1} - p(1-p)^{K-1}\right)$$

Thus,

$$S_Y(t) = \bar{F}_0(t)\exp\left(-K\eta\left[e^{-\eta t}\left(1 - p + pe^{-\eta t}\right)^{K-1} - p(1-p)^{K-1}\right]\int_0^t m(t-w)\bar{G}_W(w)dw\right)$$

$\square$

4. Correlated Channels

Now suppose the server tasks are correlated, in that the selection of one particular task may affect the selection of any or all of the other tasks. Thus the channels are a sequence of dependent Bernoulli random variables. Based on ideas from random graphs of Korzeniowski (2003), the construction of dependent Bernoulli random variables was given in Korzeniowski (2013), and is briefly summarized below.

4.1 Dependent Bernoulli Random Variables and the Generalized Binomial Distribution

Korzeniowski (2013) constructed a sequence of dependent Bernoulli random variables using a binary tree that distributes probability mass over dyadic partitions of [0,1]. Let $0 \le \delta \le 1$, $0 < p < 1$, and $q = 1 - p$. Then define the following quantities:

$$
\begin{aligned}
q^+ &:= q + \delta p & p^+ &:= p + \delta q \\
q^- &:= q(1 - \delta) & p^- &:= p(1 - \delta)
\end{aligned}
\tag{2}
$$

The quantities in (2) satisfy the following conditions:

$$
\begin{aligned}
q^+ + p^- &= q^- + p^+ = q + p = 1 \\
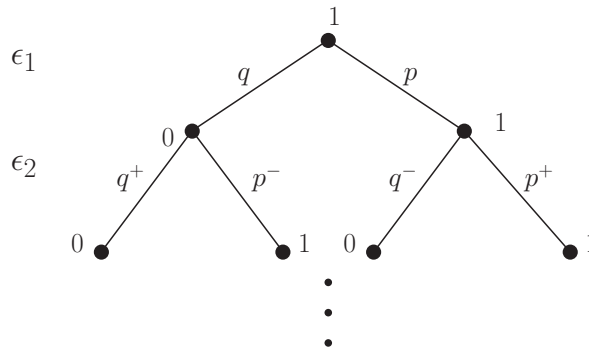qq^+ + pq^- &= q, \quad qp^- + pp^+ = 1
\end{aligned}
\tag{3}
$$

Figure 2. Construction of Dependent Bernoulli Random Variables

Figure 2 shows the construction of the dependent Bernoulli random variables. The following examples using coin flips illustrate the effect of the dependency coefficient δ:

Example 1 ($\delta = 1$). *For $\delta = 1$, $q^+ = q + p = 1$, $q^- = 0$, $p^+ = p + q = 1$, and $p^- = 0$. Supposing the first coin flip $\varepsilon_1 = 1$. Then every successive ε_i will also be 1. Similarly if $\varepsilon_1 = 0$. Thus the result of the first coin flip completely determines the outcomes of all the rest.*

Example 2 ($\delta = 0$). *For $\delta = 0$, $q^+ = q^- = q$, and $p^+ = p^- = p$. Thus, the first coin flip (and all subsequent ones) have no effect on the ones that follow.*

Example 3 ($\delta = \frac{1}{4}$). *Suppose $p = q = \frac{1}{2}$. Then $p^+ = q^+ = \frac{5}{8}$, and $p^- = q^- = \frac{3}{8}$. Then the subsequent outcomes $\varepsilon_i, i \ge 2$ are more likely to match the outcomes of ε_1 than not.*

Now suppose $p = \frac{1}{4}, q = \frac{3}{4}$. Then $p^+ = \frac{7}{16}, p^- = \frac{3}{16}, q^+ = \frac{13}{16}$, and $q^- = \frac{9}{16}$. In this example of an unfair coin, the dependency coefficient δ still attempts to skew the results following the first coin flip in favor of the outcome of ε_1. However, the dependency here heightens the effect of $\varepsilon_1 = 0$ on subsequent flips, and cannot overcome the discrepancy between the probability of success and failure to skew $\varepsilon_i, i \ge 2$ in favor of a 1 following the outcome of $\varepsilon_1 = 1$.

Using these dependent Bernoulli random variables, Korzeniowski (2013) derived a Generalized Binomial Distribution for identically distributed but dependent Bernoulli random variables.

Generalized Binomial Distribution

Let $X = \sum_{i=1}^{n} \varepsilon_i$, where $\varepsilon_i, i = 1, ..., n$ are identically distributed Bernoulli random variables with probability of success p and dependency coefficient δ. Then

$$
P(X = k) = q \binom{n-1}{k} (p^-)^k (q^+)^{n-1-k} + p \binom{n-1}{k-1} (p^+)^{k-1} (q^-)^{n-1-(k-1)}, \quad k = 0, 1, ..., n
\tag{4}
$$

4.2 Survival Function of Correlated Channels in a Cluster Server

Suppose the selection of tasks may be modeled by the dependent Bernoulli random variables given in the previous section. That is, suppose the customer selects Tasks 1,...,K in sequence, and the selection or rejection of Task 1 affects all subsequent tasks by a dependency coefficient δ. From Korzeniowski (2013), the correlation between task selections $\varepsilon_i, \varepsilon_j$ is given by

$$\rho = \text{Cor}(\varepsilon_i, \varepsilon_j) = \begin{cases} \delta, & i = 1; j = 2, ..., K \\ \delta^2, & i \neq j; i, j \geq 2 \end{cases} \tag{5}$$

This illustrates the dependency of Tasks 2,...,K on the outcome of Task 1, and notes that while Tasks 2,...,K are still correlated with each other, the dependency is much lower. In a similar fashion to the independent channel server, the survival function is derived.

Theorem 3. *Suppose conditions (i)-(vi) above are satisfied. In addition, suppose the selection of channels $1, \ldots, K$ are determined by identically distributed Bernoulli random variables with dependency coefficient δ as defined in Korzeniowski (2013). Then the survival function of the server is given by*

$$S_Y(t) = \bar{F}_0(t) \exp\left(-\eta \int_0^t m(t-w)\bar{G}_W(w)\mathscr{S}(w)dw\right) \tag{6}$$

where $m(x) = \int_0^x \lambda(s)ds$, and

$$\mathscr{S}(w) = \sum_{n=0}^{K} e^{-\eta n w} \sum_{j=0}^{K-n-1} \binom{K-1}{n-1, j, K-1-n-j} p^{K-1-j}(1-p)^{j+1}\delta^{K-1-n-j}(1-\delta)^n$$

$$+ \sum_{n=0}^{K} n e^{-\eta n w} \sum_{i=0}^{n-1} \binom{K-1}{K-1-n, i, n-1-i} p^{i+1}(1-p)^{K-n}\delta^{n-1-j}(1-\delta)^{K-n-j}$$

Proof. By Theorem 1,

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

Similar to the proof of Theorem 2, $\mathcal{H} = \eta X$, where this time X has the generalized binomial distribution given in (4). Then

$$E\left[\mathcal{H} \int_0^t e^{-\mathcal{H}w}m(t-w)\bar{G}_W(w)dw\right]$$

$$= \sum_{x=0}^{K} \left[\eta x \int_0^t e^{-\eta x w}m(t-w)\bar{G}_W(w)dw\right] P(X = x)$$

$$= \sum_{x=0}^{K} \eta x \left[\int_0^t e^{-\eta x w}m(t-w)\bar{G}_W(w)dw\right]\left[q\binom{K-1}{x}(p^-)^x(q^+)^{K-1-x}\right]$$

$$+ \sum_{x=0}^{K} \eta x \left[\int_0^t e^{-\eta x w}m(t-w)\bar{G}_W(w)dw\right]\left[p\binom{K-1}{x-1}(p^+)^{x-1}(q^-)^{K-x}\right]$$

$$= \eta \int_0^t m(t-w)\bar{G}_W(w)(\mathscr{S}_1(w) + \mathscr{S}_2(w))dw$$

where $\mathscr{S}_1(w) = \sum_{x=0}^{K} x e^{-\eta x w} q\binom{K-1}{x}(p^-)^x(q^+)^{K-1-x}$
and $\mathscr{S}_2(w) = \sum_{x=0}^{K} x e^{-\eta x w} p\binom{K-1}{x-1}(p^+)^{x-1}(q^-)^{K-x}$. Using the definitions given in (2),

$$\mathscr{S}_1(w) = \sum_{x=0}^{K} x e^{-\eta x w}(1-p)\binom{K-1}{x}(p-\delta p)^x(1-p+\delta p)^{K-1-n}$$

$$= \sum_{x=0}^{K} x e^{-\eta x w}(1-p)\binom{K-1}{x}p^x(1-\delta)^x \sum_{j=0}^{K-1-x} \binom{K-1-x}{j}(1-p)^j(\delta p)^{K-1-x-j}$$

Now, $x\binom{K-1}{x}\binom{K-1-x}{j} = \frac{(K-1)!}{(x-1)!j!(K-1-x-j)!} = \binom{K-1}{x-1,j,K-1-x-j}$. Then

$$\mathscr{S}_1(w) = \sum_{x=0}^{K} e^{-\eta x w} \sum_{j=0}^{K-x-1} \binom{K-1}{x-1,\, j,\, K-1-x-j}(1-p)^{j+1}(1-\delta)^x \delta^{K-1-x-j} p^{K-1-j}$$

Similarly,

$$\mathscr{S}_2(w) = \sum_{x=0}^{K} x e^{-\eta x w} p \binom{K-1}{x-1}(p+\delta(1-p))^{x-1}((1-p)(1-\delta))^{K-x}$$

$$= \sum_{x=0}^{K} x e^{-\eta x w} p(1-\delta)^{K-x}(1-p)^{K-x} \sum_{i=0}^{x-1} \binom{x-1}{i} p^i (1-\delta)^i \delta^{x-1-i}$$

$$= \sum_{x=0}^{K} x e^{-\eta x w} \sum_{i=0}^{x-1} x \binom{K-1}{K-1-x,\, i,\, x-1-i} p^{i+1} \delta^{x-1-i}(1-\delta)^{K-x+i}(1-p)^{K-x}$$

Clearly $\mathscr{S}(w) = \mathscr{S}_1(w) + \mathscr{S}_2(w)$ □

5. Numerical Illustrations

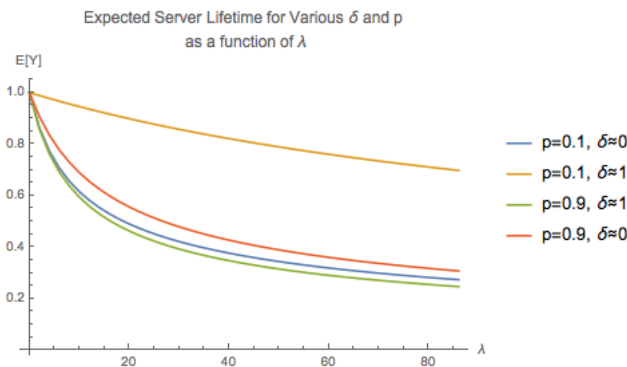### 5.1 Expected Server Lifetime

Figure 3. Expected Server Lifetime for Various Selection Probabilities
and Dependency Coefficients

To measure the effects of the probability of task selection p and the dependency coefficient δ, we look at the expected server lifetime, given by $E[Y] = \int_0^\infty S_Y(t)$ as a function of the arrival rate λ. Figure 3 shows the expected server lifetime for $K = 3$ channels under two different selection probabilities ($p = 0.1, 0.9$) and dependency coefficients ($\delta = 0, 1$). In addition, $\eta = r_0 = 1$. Mathematica was used for calculations.

For $\delta = 0$, $p = p^+ = p^-$ and $q = q^+ = q^-$, and hence the channels are uncorrelated and the selection of tasks $1, ..., K$ are independent Bernoulli random variables. When $\delta = 1$, the selection decision made at channel 1 completely determines the subsequent task selections. Thus, when $\delta = 1$ and $p = 0.1$, it is highly likely that the first task will not be selected and thus no others will be selected. This results in a server that is expected to remain fairly idle, even as the arrival intensity increases. Thus, the expected lifetime changes very little compared to a completely idle server ($\lambda = 0$).

On the opposite end, for $p = 0.9, \delta = 1$, we again have a perfectly correlated set of channels, but the selection probability for Task 1 is very high. Thus, 0.9λ customers per unit time will select all tasks, and 0.1λ customers per unit time will select no tasks. This will result in a high server workload, and the expected lifetime decreases sharply with increasing λ.
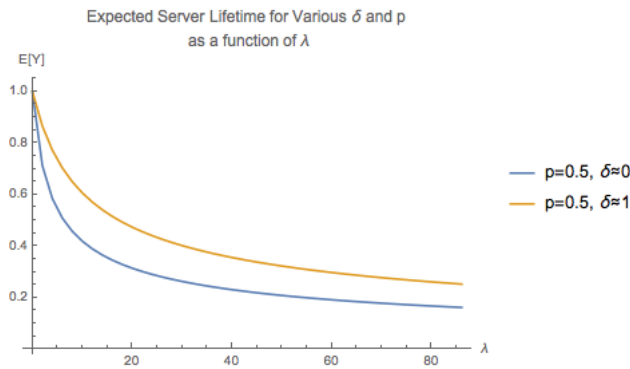
Figure 4. Expected Server Lifetime for $p = 0.5$ and
Various Dependency Coefficients

Figure 4 compares the effect of the dependency coefficient δ when the selection probability of Task 1 is $p = 0.5$. A completely correlated system with $p = 0.5$ has a significantly longer expected lifetime for a given arrival rate λ than independent channels.

6. Conclusion

This paper provides a reliability model for a multichannel clustered-task server under very general assumptions. Requests arrive via a nonhomogenous Poisson process and consist of the random selection of 0 to K tasks in sequence. A request is considered serviced when the last task in that request is completed. Since each channel has a generic service distribution $G_i(w)$, the service distribution of the requests is given by $\max_i G_i(w)$. The stress to the server brought by each request is a constant multiple η of the number of tasks selected, and thus is random with either a binomial distribution or a generalized binomial distribution. The survival function in both cases was formulated, and a numerical comparison of the correlated and uncorrelated cases was done using the expected server lifetime as a metric.

References

Cha, K. H., & Lee, E. Y. (2011). A Stochastic Breakdown Model for an Unreliable Web Server System and an Optimal Admission Control Policy. *J. Appl. Probab., 48*(2), 453-466. http://dx.doi.org/10.1239/jap/1308662638.

Gross, D., & Shortle, J. (2008). *Fundamentals of Queuing Theory*. John Wiley and Sons, Hoboken, New Jersey.

Korzeniowski, A. (2013). On Correlated Random Graphs. *Journal of Probability and Statistical Science, 11*. 43-58.

Korzeniowski, A. (2003). On Universal Representations of Random Graphs. *Annals of Combinatorics, 7*. 299-313.

Tang, Y. H. (1997).A single-server M/G/1 queueing system subject to breakdownssome reliability and queueing problems. *Microelectronics Reliability, 37*. http://dx.doi.org/10.1016/S0026-2714(96)00018-2

Traylor, R. (2015). *A Stochastic Reliability Model of a Server under a Random Workload*. Preprint. Available at http://arxiv.org/abs/1511.04130

Traylor, R. (2015). *Reliability of Series and Parallel Systems with Correlated Nodes*. Preprint. Available at http://arxiv.org/abs/1601.03630

Copyrights