International Journal of Statistics and Probability; Vol. 6, No. 2; March 2017
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Reliability of a Clustered-Task Server under Modulated Correlation

Rachel Lunde Traylor[1] & Andrzej Korzeniowski[2]

[1] Dell EMC, Santa Clara, CA

[2] University of Texas at Arlington, Arlington, Texas

Correspondence: Rachel Traylor, Dell EMC, 2421 Mission College Blvd Santa Clara, California, USA.
E-mail: rachel.traylor@dell.com

Received: January 6, 2017 Accepted: February 7, 2017 Online Published: February 21, 2017

doi:10.5539/ijsp.v6n2p93 URL: https://doi.org/10.5539/ijsp.v6n2p93

Abstract

Server resource allocation and traffic management is a large area of research and business concern in order to ensure proper functionality and maintenance procedures. As a result, good server reliability models that can incorporate workload and traffic stress are necessary. This paper generalizes previous dynamic server reliability models for partitioned servers with clustered-task selection by relaxing the assumption that the correlation between channels in the server remain constant. We allow the correlation to vary deterministically with time, or as a function of a random process in discrete or continuous time. The explicit form of the survival function is derived in such cases. Numerical illustrations demonstrate the dangers of erroneously assuming independence among channels, which can lead to costly and unnecessary interventions in the system. In addition, we numerically explore the effects of a variable correlation on the survival function.

Keywords: clustered tasks; server reliability; correlated channels; queuing theory

1. Introduction

Recent years have seen an explosion in the amount of data storage devices and computing resources as well as the need for near constant accessibility, especially as the Internet of Things (IoT) grows. Thus, devices that remain reliable under stress and heavy or inconsistent workload are desirable. Much research has been done on optimal policies to handle spikes in server traffic (Iosup et. al, 2011; Thomas et. al, 2012; Welsh and Culler, 2003), though these are policies to handle overload, and are typically based on some sort of threshold analysis. Other attempts to model and predict server reliability include classification trees (Vishwanath, 2010) and other standard data mining and reliability theory techniques such as Weibull analysis. We propose here an extension of previous work by Cha and Lee (Cha and Lee, 2011), Korzeniowski and Traylor (Korzeniowski and Traylor, 2016), and Traylor (Traylor, 2016) that provides a more analytical and widely applicable solution as opposed to the more traditional data-driven approaches.

1.1 Background

All servers may be viewed as a queue. However, many standard queuing theory assumptions do not mirror reality. For example, the common assumption of Poisson arrivals implies a constant arrival rate, which is unlikely to be the case for most servers. To remedy this, Cha and Lee (Cha and Lee, 2016) proposed a stochastic reliability model for a web server under stress. In their model, customers or jobs arrive to the server via a nonhomogenous Poisson process, which allows the arrival rate to vary over time. Each job brings a constant stress $\eta > 0$ to the server and adds this stress to the hazard function for the duration of its time in the system. The defintion of stress is left to the application, but some examples are memory usage, CPU load, or IOPs. Cha and Lee derived the survival function $S_Y(t) = P(Y > t)$ under the assumption of independent arrival times $\{T_i\}_{i=1}^n$ and i.i.d. service times $\{W_i\}_{i=1}^n \sim G(w)$. Traylor (Traylor, 2016) generalized the model by allowing the workload stress brought by each job to be i.i.d. random variables $\{\mathcal{H}_i\}_{i=1}^n \sim \mathcal{H}$, where the random variable $\mathcal{H}$ may have either a discrete or continuous distribution. Thus, a very general survival function was produced that allowed for a nonconstant arrival rate, any service times distribution $G(w)$, and a workload with any distribution.

Korzeniowski and Traylor (Korzeniowski and Traylor, 2016) applied the random stress model of (Traylor, 2016) to a "partitioned" server with K channels. Each channel is capable of performing a single unique task and has service time distribution $G_k(t), k = 1, ..., K$. Customers still arrive via a nonhomogenous Poisson process with intensity $\lambda(t)$, and upon arrival, select N channels based on the desired tasks each channel performs. The selection is done sequentially, with the customer moving down the channel/task options and either selecting or rejecting it. Thus, channel selection is a Bernoulli random variable $\varepsilon_i, i = 1, ..., K$ with probability of success p. For simplification, each customer adds a constant multiple η of the number of channels selected as a stress factor to the server for the remainder of the customer's time in the system. That is, the stress added by each customer remains ηN until the completion of the last (slowest) task requested, regardless of the completion of other selected tasks, where the sample space for the random variable N is $\{1, 2, ..., K\}$.

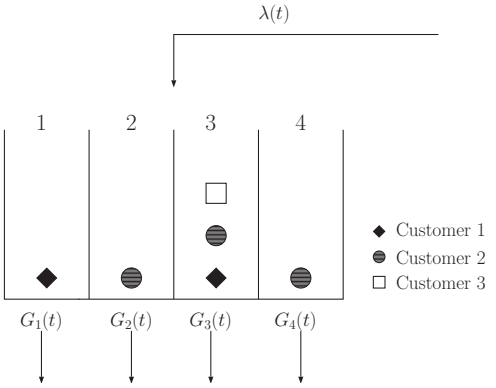


Figure 1. Multichannel Server Illustration

Figure 1 gives an illustration for a four-channel server. $\lambda(t)$ is the intensity of the nonhomogenous Poisson process governing arrivals. Customer 1 selects 2 tasks (Channels 1 and 3), and thus the stress added is 2η, where $\eta > 0$. $G_k(t)$ is the service time distribution of channel k. Let W_{i_k} denote the service time for Customer i at channel k. Then $W_{i_k} \sim G_k(t)$, and the service time for Customer 1 is $\max_{i_k}(W_{i_k})$. The stress 2η remains part of the hazard function until both tasks are completed. Customer 2 selects 3 tasks (Channels 2, 3, and 4), but in this case, Channel 3 has not completed the requested task for Customer 1. Thus, a queue can form in each channel, regardless of the queue length of the other channels. The stress Customer 2 adds to the system is 3η and remains until the final task selected is completed.

In general, there are K queues where the service times for each queue are all mutually independent and have service time distribution governed by the channel. The service time distribution at the customer level under this model is therefore $\max_k G_k(t)$, where the maximum may be found by conventional statistical means.

The stress to the server brought by each customer i was denoted in (Traylor, 2016) by $\mathcal{H}_i$. The model of (Korzeniowski and Traylor, 2016) is thus a special case of (Traylor, 2016) with $\mathcal{H}_i = \eta N_i$, where N_i is the random number of tasks selected by Customer i. Since the selection of channels is a sequence of Bernoulli random variables $\{\varepsilon_{i_j}\}_{i_j=1}^{N_i}$, $N_i = \sum_j \varepsilon_{i_j}$, and $N_i \sim N$, where N is binomially distributed.

The assumptions for a multichannel server with clustered tasks are summarized below:

(i) Arrivals follow a nonhomogenous Poisson process (NHPP) with intensity $\lambda(t)$

(ii) The idle server has a baseline hazard function (breakdown rate) given by $r_0(t)$, and the survival function of the idle server is given by $\bar{F}_0(t) = \exp\left(-\int_0^t r_0(s)ds\right)$

(iii) $\{W_{i,k}\}$ are independent across i and k, where k denotes the channel of service and i denotes the customer. That is, the service times are independent variables within the same channels and across channels.

(iv) $\{T_i\}$ are mutually independent and the set is independent of $\{W_{i,k}\}$. In other words, the arrival times are pairwise independent amongst themselves and all mutually independent of the complete set of all service times within and across channels.

(v) $\varepsilon_k \sim$ Bernoulli(p).

(vi) The workload stress to the server brought by customer i is ηN_i, where $N_i = \sum_{k=1}^{K} \varepsilon_k$, i.e. the stress brought to the server by each customer is a constant multiple of the number of channels selected.

(vii) For fixed k, $\{W_{i,k}\}_{i=1}^{N}$ are i.i.d. with cdf G_k. The service times within each channel share the same distribution.

(viii) The service time distribution for job i is given by $G(w) = \max_k G_k(w)$

The survival function of such a system was studied in (Korzeniowski and Traylor, 2016) for both independent channel selection and correlated channel selection. Under correlated channel selection, the selection of channels 2,...,K depend on the selection (or rejection) of channel 1 through the dependency coefficient $\delta \in [0, 1]$. This creates a sequence of dependent Bernoulli random variables via a binary tree that distributes probability mass over dyadic partitions of [0,1] at each level of the tree corresponding to an ε_i, whose construction is detailed in (Korzeniowski, 2013). In this case, each ε_i corresponds to the selection (or rejection) of channel i.

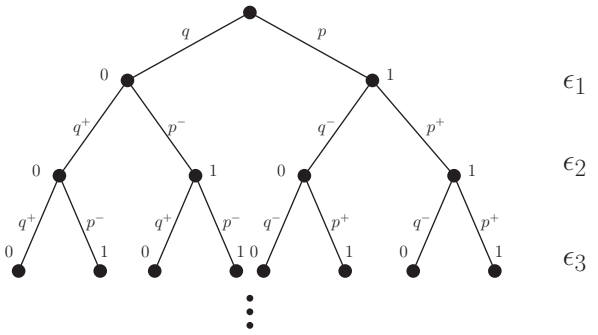


Figure 2. Construction of Dependent Bernoulli Random Variables with Constant Dependency

Korzeniowski (Korzeniowski, 2013) defined the following quantities in his construction for $0 \leq \delta \leq 1$.

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

The quantities above and Figure 2 show how the subsequent Bernoulli variables $\varepsilon_2, \varepsilon_3, ...$ are affected by the outcome of the first Bernoulli variable ε_1. If Task 1 is selected by a customer with probability p, the probability of selecting Task 2, 3,...,K is p^+. Thus, the probability of selecting other tasks is increased by δq. Conversely, if Task 1 is not selected by a customer, the probability of selecting each of Tasks $2, 3, ..., K$ is p^-, and thus the selection probability for other tasks is decreased by δp.

In the extreme cases, for $\delta = 1$, the only random variable is ε_1, as the outcome renders the other channel selections completely deterministic. If $\delta = 1$, $p^+ = 1 = q^+$, and $p^- = q^- = 0$. Therefore, arriving customers under this arrangement will either select all K tasks with probability p, or no tasks (and leaving the server idle) with probability $q = 1 - p$. Conversely, $\delta = 0$ implies that $p = p^+ = p^-$ and $q = q^+ = q^-$ and thus the outcome of ε_1 has no effect at all on the other selection probabilities. Under this condition, the channels are completely independent. Formal construction is given in (Korzeniowski, 2013).

With a sequence of now dependent Bernoulli variables, Korzeniowski introduced the Generalized Binomial Distribution ($GBin(n, p, \delta)$), reproduced in the below theorem.

Theorem 1 (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*

$$\begin{aligned} P(X = k) = q\binom{n-1}{k}(p^-)^k(q^+)^{n-1-k} \\ + p\binom{n-1}{k-1}(p^+)^{k-1}(q^-)^{n-k}. \end{aligned}$$

In addition, Korzeniowski showed that the pairwise correlation between the Bernoulli random variables $\{\varepsilon_i\}_{i=1}^n$ has the form

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

Subsequently, based on the above construction, the survival function of the multichannel server with correlated channel selection was derived in (Korzeniowski and Traylor, 2016).

However, even the assumption in (Korzeniowski and Traylor, 2016) is limiting. It it conceivable that the correlation between channels is nonconstant, and it either varies deterministically as a function of time, or changes randomly. This paper relaxes the constant nature of the dependency coefficient δ from (Korzeniowski and Traylor, 2016) and examines the impact of variable correlation both under a deterministic function of time ($\delta(t)$) and a function of a Markov process ($\delta(X(t))$). This new model allows for the fluctuation of correlation among channel selection either as a known function of time or as random correlation. Section 2.1 gives the extension of Theorem 3 of the survival function derived in (Korzeniowski and Traylor, 2016) to time-dependent correlation between channels. Section 2.2 gives the survival function

of a correlated clustered-task server when the correlation is a function of a discrete time Markov process. Section 2.3 extends Section 2.2 to a continuous time Markovian dependency for a fully generalized survival function. Section 3 explores the effects of the dependency, number of channels, and stress multiplier on the survival function.

2. Survival Function of a Cluster Server under Nonconstant Dependency

2.1 Dependency as a Function of Time

Let Y be the random time to server breakdown. Server breakdown may be defined in any engineering-applicable way, but here we will define server breakdown from the perspective of the customer, in that the server can no longer provide any kind of service to the customer, even if the hardware still functions. The survival function of the server under constant δ is given in Theorem 3 of (Korzeniowski and Traylor, 2016) by

$$S_Y(t) = \bar{F}_0(t) \exp\left(-\eta \int_0^t m(t-w)\bar{G}_W(w)\mathscr{S}(w)dw\right) \tag{2}$$

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

$$\mathscr{S}(w) = \sum_{n=0}^{K} ne^{-\eta n w}\beta^n \sum_{j=0}^{K-n-1} \binom{K-1}{n, j, K-1-n-j} p^{K-1-j}q^{j+1}\delta^{K-1-n-j} + \sum_{n=0}^{K} ne^{\eta n w}q^{K-n} \sum_{i=0}^{n-1} \binom{K-1}{n-1-i, i, K-n} p^{i+1}\delta^{n-1-i}\beta^{K-n-i} \tag{3}$$

where $q = 1 - p$, and $\beta = 1 - \delta$. Now, let $\delta : \mathbb{R}^+ \to [0, 1]$ be a right continuous function of running time t. Then the survival function of a natural generalization of (2) and (3) when $\delta \to \delta(s)$ and $\mathscr{S}(w) \to \mathscr{S}(w, s), 0 \le s \le t$.

Examples of deterministic dependency functions are sinusoidal or periodic functions, though the only requirement is that it be bounded between [0,1]. Situations in which some element of seasonality in dependency is known or estimated benefit greatly from this generalization. This seasonality is separate from any cyclic or seasonal component in arrival times or arrival rate, which is accounted for in the nonconstant intensity $\lambda(t)$ of the Poisson arrival process. More complex situations exist in which the customers may arrive under heavy (or light) traffic, but the stress they impart on the server is not the same for different instances of a particular arrival rate because the customers are interacting with the server differently.

For example, suppose the stream of traffic to the server is constant. The previous model assumed the channel selection probabilities remained constant over time, and thus the server was stressed according to a static probability distribution. There may be times when customers are more likely to select additional tasks if the first task is selected, tending toward an "all or nothing" selection approach, but this phenomenon is not permanent. Thus, allowing δ to vary with time encompasses these more complex scenarios.

Another interesting generalization lies in allowing the dependency coefficient to be governed by a random process, notably a Markov process. We first investigate a discrete time Markov process and then a continuous time process.

2.2 Markovian Correlation Structure

Let $\{X_n\}_{n\in\mathbb{N}}$ be a discrete Markov chain with discrete state space $S = \{s_1, ..., s_m\}$ where $0 \le s_1 < s_2 < \ldots < s_m \le 1$. Since the values of the Markov chain $\{X_n\}$ in $\delta(X_n)$ are immaterial, we assume $X_n \in \{1, 2, ..., m\}$. Furthermore, we embed X_n into the process with right continuous realizations on $[0, t]$ as follows:

$$X(s) = X_n, \qquad n\Delta \le s < (n+1)\Delta \qquad \text{for } n \in \mathbb{N}, \quad 0 \le n\Delta \le t.$$

In other words, transitions occur at times $n\Delta$ with Δ as a fixed interval length, and $\lceil t/\Delta \rceil$ is the number of transitions in the interval $[0, t]$. Then we have that $\delta(X(s)) = \delta(X_n) = \delta(i) \equiv \delta_i$ for $n\Delta \le s < (n+1)\Delta$ and $X_n = i$, where $\delta : \{1, 2, ..., m\} \to [0, 1]$.

Let $\mathbb{P}$ be the $m \times m$ transition matrix with transition probabilities p_{ij} of an irreducible Markov chain with a unique steady state distribution $\pi = (\pi_1, \pi_2, \ldots, \pi_m)$. WLOG, we assume the initial distribution to be π.

As an explicit example, suppose we have three possible values for δ : $0, 1/2$, and 1. Then the underlying Markov state space is given by $S = \{0, 1, 2\}$. The Markov process $X(s)$ moves among the three states in discrete time steps of size Δ according to the transition matrix $\mathbb{P}$. We require $\mathbb{P}$ be irreducible with a unique steady state distribution $\pi = (\pi_1, \pi_2, \pi_3)$, and WLOG assume the initial distribution to choose the first δ is this stationary distribution π. Then we have that $\delta(X(s)) \in \{0, 1/2, 1\}$ according to the value of $X(s)$ at time s. This value of δ remains until the next transition time of the Markov chain.

Under these assumptions, the following theorem for the survival function of the clustered task server is given below.

Theorem 2. *Consider $X(s)$ above and assume* (i)-(vii) *from Section 1.1 hold. Then the survival function of the server is given by*

$$S_Y(t) = \bar{F}_0(t) \exp\left(-\eta \int_0^t m(t-w)\bar{G}_W(w)\left[\sum_{i=1}^{m} \pi_i(\mathscr{S}_1(w,0) + \mathscr{S}_2(w,0))\right]\right) \tag{4}$$

where $\mathscr{S}_1(w,s) = q(K-1)\left(e^{-\eta w}p^-\right)\left(e^{-\eta w}p^- + q^+\right)^{K-2}$,
$\mathscr{S}_2(w,s) = pe^{-\eta w}\left[(K-1)\left(p^+e^{-\eta w}\right)\left(p^+e^{-\eta w} + q^-\right)^{K-2} + \left(p^+e^{-\eta w} + q^-\right)^{K-1}\right]$

and

$$p^+ = p + \delta(X(s))q \qquad p^- = p(1 - \delta(X(s)))$$
$$q^+ = q + \delta(X(s))p \qquad q^- = q(1 - \delta(X(s)))$$

Proof. As in the proof of Theorem 3 of (Korzeniowski and Traylor, 2016), with $\mathcal{H} = \eta N$ and $N \sim GBin(K, p, \delta)$, we have

$$S_Y(t) = \bar{F}_0(t) \exp\left(-E\left[\mathcal{H} \int_0^t e^{-\mathcal{H}w} m(t-w)\bar{G}_W(w)dw\right]\right) \tag{5}$$

The expectation in this case is given by

$$E\left[\mathcal{H} \int_0^t e^{-\mathcal{H}w} m(t-w)\bar{G}_W(w)dw\right] = E_X\left[E_N\left[\eta N \int_0^t e^{-\eta N w} m(t-w)\bar{G}_W(w)dw|X\right]\right]$$

Now, focusing on the internal conditional expectation,

$$\begin{aligned} E_N\left[\eta N \int_0^t e^{-\eta N w} m(t-w)\bar{G}_W(w)dw|X\right] &= \sum_{k=0}^{K} \eta k \left[\int_0^t e^{-\eta k w} m(t-w)\bar{G}_W(w)dw\right] P(N = n|X(0) = i) \\ &= \sum_{k=0}^{K} \eta k \left[\int_0^t e^{-\eta k w} m(t-w)\bar{G}_W(w)dw\right] q\binom{K-1}{k}(p^-)^k(q^+)^{K-1-k} \\ &\quad + \sum_{k=0}^{K} \eta k \left[\int_0^t e^{-\eta k w} m(t-w)\bar{G}_W(w)dw\right] p\binom{K-1}{k-1}(p^+)^{k-1}(q^-)^{K-k} \end{aligned}$$

because $X(s)$ is stationary with $X(0) \sim \pi$. Focusing on the first term,

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

Now, denoting the sum above as $\mathscr{S}_1(w,s)$, we have the following:

$$\begin{aligned} \mathscr{S}_1(w,s) &= \sum_{k=0}^{K} q\binom{K-1}{k} k \left(e^{-\eta w}p^-\right)^k (q^+)^{K-1-k} \\ &= \sum_{k=1}^{K} q\binom{K-1}{k} k \left(e^{-\eta w}p^-\right)^k (q^+)^{K-1-k} \\ &= q(K-1)\left(e^{-\eta w}p^-\right)\left(e^{-\eta w}p^- + q^+\right)^{K-2} \end{aligned}$$

and thus the first term is given by

$$\eta \int_0^t m(t-w)\bar{G}_W(w)\left[\mathscr{S}_1(w,s)\right]dw. \tag{6}$$

Moving to the second term,

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

Define $\mathscr{S}_2(w,s) := pe^{-\eta w}\sum_{k=0}^{K}\binom{K-1}{k-1}k\left(e^{-\eta w}p^-\right)^{k-1}(q^+)^{K-k}$. Making use of the fact that $k\binom{K-1}{k-1} = (k-1)\binom{K-1}{k-1} + \binom{K-1}{k-1}$,

$$\begin{aligned}\mathscr{S}_2(w,s) &= pe^{-\eta w}\left[\sum_{k=0}^{K}(k-1)\binom{K-1}{k-1}\left(e^{-\eta w}p^-\right)^{k-1}(q^+)^{K-k} + \sum_{k=0}^{K}\binom{K-1}{k-1}\left(e^{-\eta w}p^-\right)^{k-1}(q^+)^{K-k}\right]\\ &= pe^{-\eta w}\left[(K-1)\left(p^+e^{-\eta w}\right)\left(p^+e^{-\eta w}+q^-\right)^{K-2} + \left(p^+e^{-\eta w}+q^-\right)^{K-1}\right]\end{aligned}$$

yielding the second term as

$$\eta \int_0^t m(t-w)\bar{G}_W(w)\left[\mathscr{S}_2(w,s)\right]dw \tag{7}$$

Thus, combining (6) and (7) yields the conditional expectation

$$E_N\left[\eta N\int_0^t e^{-\eta N w}m(t-w)\bar{G}_W(w)dw|X\right] = \eta\int_0^t m(t-w)\bar{G}_W(w)(\mathscr{S}_1(w,s)+\mathscr{S}_2(w,s)dw)$$

Taking the expectation over X and inserting into (5) completes the proof.

□

Special cases of discrete-time Markovian dependency are:

(I) A 2 state Markov chain with state space $X_n \in \{0,1\}$, where $\delta(0) = 0$ and $\delta(1) = 1$, with transition matrix $\mathbb{P} = \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix}$, where $0 < p < 1, 0 < q < 1$. (Note: p,q here are not the same as the probability of channel selection given earlier.)

The channel selection random variables $\varepsilon_i, i \geq 2$ will either be completely dependent on the outcome of ε_1, or completely independent.

(II) A finite state Markov chain with $m \times m$ transition matrix $\mathbb{P}$ where each row has exactly one 1 in each row and exactly one 1 in each column. This corresponds to a deterministic dynamic (periodic) $\delta(X_n)$ cycling through values $\delta(i) \equiv \delta_i, i = 1,2,...,m$. Since $\mathbb{P}$ is doubly stochastic in this case, $\pi = (\frac{1}{m},\ldots,\frac{1}{m})$ is the uniform distribution.

2.3 Continuous Time Markovian Dependency Structure

We relax the discrete time Markov assumption underneath δ one step further and allow for the Markov chain to evolve in continuous time. We still require the state space (and thus the infinitesimal rate matrix) to be finite.

Consider the dependency structure $\delta(X(s))$, $0 \leq s \leq t$, where $X(s) \in \{1,2,...,m\}$ is a continuous time Markov process, assumed to have unique stationary distribution π. For example, a birth-death process may be used here. Specifically, for the infinitesimal rate matrix that characterizes the Markov process,

$$Q = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ q_{m1} & q_{m2} & \cdots & q_{mm} \end{bmatrix}$$

with $\pi Q = \pi$, where $\pi = (\pi_1,...,\pi_m)$ is the stationary initial distribution. Thus, $X(s)$ is stationary. Under this structure, we have that the survival function is identical to that of Theorem 2.

2.4 Steady-State v. Realized Survival Functions

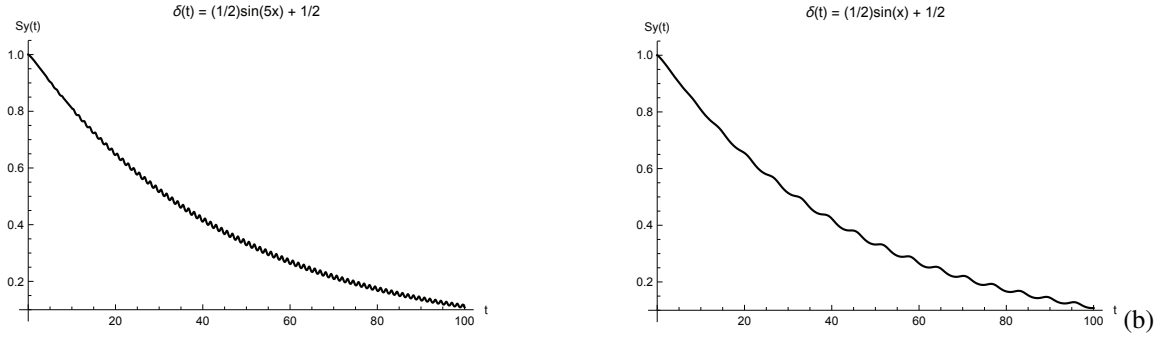
The result give in Theorem 2 is obtained by taking the expectation over the random process X. This yields the"steady state" survival function. In practice, one would observe a particular realization of the process X, and would see a "realized" survival function with jumps. See Figure 6. The monotonicity seems to be lost in the survival function, supposedly in violation of reliability principles. However, a closer look shows this is not the case.

For each fixed δ, a survival function exists, as in (Korzeniowski and Traylor, 2016). The state space S for a particular Markov chain X corresponds to a discrete set of possible values of δ, and therefore a family of survival functions, one for each δ. The trajectory of the Markov chain through the state space shifts among the family of survival functions, corresponding to the observed state at a particular time t.

3. Illustrations and Numerical Studies

The following section gives some numerical illustrations that show how the various components of the survival function interact. In general, the effects of δ on the survival function manifest in an intuitive way based on the underlying function that governs the value of δ.

3.1 Temporal Dependency Effects



(a) $\delta(t) = 1/2\sin(5x) + 1/2$ (b) $\delta(t) = 1/2\sin(x) + 1/2$

Figure 3. Survival Functions under Sinusoidal Temporal Dependency

Figure 3 illustrates the survival function for a clustered task server under two different deterministic temporal dependency functions. In both cases, the departure from monotonicity exactly parallels the underlying temporal dependency defined. The "baseline" survival function is given by $S_{Y,B}(t) = \bar{F}_0(t) = e^{-\int_0^t r_0(s)ds}$, where $r_0(s)$ is the hazard function of the idle server.

In order to isolate the effects of the temporal dependency on the survival function, the simulations in Figure 3 used $r_0 \equiv 0.01$, stress multiplier $\eta = 0.1$, arrival rate $\lambda(t) \equiv \lambda = 1$, and $K = 3$ channels, and a simple exponential customer service time distribution of $\bar{G}_W(w) = e^{-w}$. Both temporal dependency functions are sinusoidal, and are given by

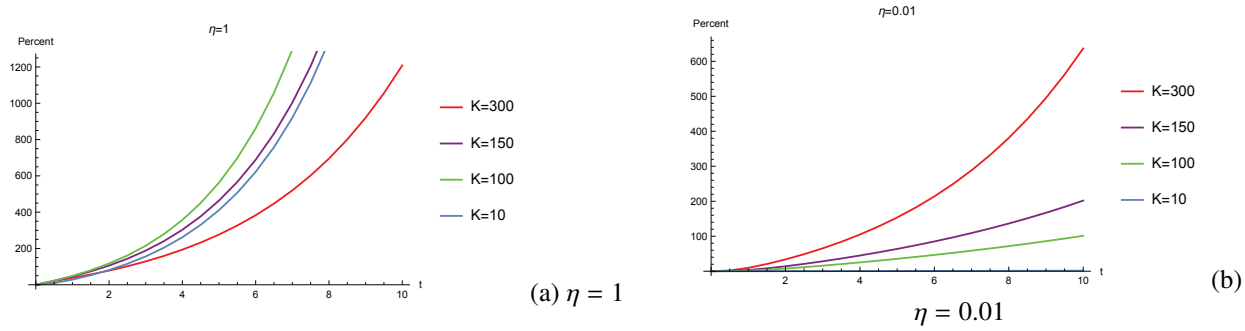
$$\delta_1(t) = \frac{1}{2}\sin(5t) + \frac{1}{2} \text{ and } \delta_2(t) = \frac{1}{2}\sin(t) + \frac{1}{2}$$

The amplitude and vertical shifts ensure both functions have the full range of [0,1]. To illustrate how prominent the effect of $\delta(t)$ is on the survival function, the frequency was varied between 1 and 5. The effect of the workload on the server survival function is given by

$\exp\left(-\eta \int_0^t m(t-w)\bar{G}_W(w)\mathscr{S}(w)dw\right)$ where $\mathscr{S}(w)$ is given in Section 2.1. Thus, this term becomes exponential decay in a polynomial of trigonometric functions and thus oscillatory.

The survival function loses its traditional monotonicity due to the dynamic nature of the channel dependency, with an explanation analogous to that of Section 2.4. This model now has the ability to reflect the effects of customer choice and behavior in addition to the number of customer arrivals.

3.2 Dependency v. Independency - when can independence be assumed?



(a) $\eta = 1$ (b) $\eta = 0.01$

Figure 4. Percent Difference in Independent and Dependent Channel Survival Functions

Table 1. Comparison of Binomial and Generalized Binomial PDF

	0	1	2	3	4	5
$P(N_D = n)$	0.490075	0.00985075	7.45×10^{-5}	7.45×10^{-5}	0.00985075	0.490075
$P(N_I = n)$	0.03125	0.15625	0.3125	0.3125	0.15625	0.03125

In this section we study the effect of δ alone on the survival function. In many practical applications, approximations are desired, and the natural approximation to correlated channels is to simply assume independence. We wish to establish conditions under which this approximation is acceptable. Numerical experiments were performed for various channel counts and values of η, shown in Figure 4. In these numerical experiments, we let $\lambda = 1, r_0 = 0.01$, $p = 1/2$, and $\bar{G}_W(w) = e^{-w}$, and $\delta \in \{1/100, 99/100\}$, approximating near independence and near dependence of the other channel selection probabilities on ε_1.

We investigate the effect of dependency by calculating the long-term survival function under two different Markov chains. The long-term survival function is given by Theorem 2. By taking the expectation over the Markov process, X, the resulting survival function is the "steady state" survival function. The two transition matrices for the comparison are given by

$$\mathbb{P}_1 = \begin{bmatrix} 1/100 & 99/100 \\ 1/100 & 99/100 \end{bmatrix} \quad \mathbb{P}_2 = \begin{bmatrix} 99/100 & 1/100 \\ 99/100 & 1/100 \end{bmatrix}$$

with stationary distributions $\pi_1 = (1/100, 99/100)$, and $\pi_2 = (99/100, 1/100)$. Thus, π_1 weights in favor of high channel selection dependency, and π_2 weights toward independent channel selection. For each η and channel count K, the percent difference between the two survival functions under π_1 and π_2 is calculated and shown in Figure 4. Denote $S_{Y,1}(t)$ as the survival function under π_1, and $S_{Y,2}$ the survival function under π_2. Then the percent difference is given by

$$\%\text{difference}\ (t) = 100\left(\frac{S_{Y,1}(t) - S_{Y,2}(t)}{S_{Y,2}(t)}\right)$$

The results in Figure 4 show that, regardless of the value of K or η, $S_{Y,2}(t) < S_{Y,1}(t)$. That is, independent channels have a smaller probability of survival than dependent channels. At first, the results seem counterintuitive, but under a totally dependent channel selection structure, we have that either all will be selected, or none will be selected.

Table 1 compares the PDFs for the essentially independent $GBin(N_I = 5, 1/2, 1/100)$ and $GBin(N_D = 5, 1/2, 99/100)$. Examining the expectation in (5), we see that almost half of the weight is on a term that evaluates to 0 in the dependent case, whereas only 3% of the weight belongs to 0 in the independent case. In the independent case, the probability of at least one channel being selected by any particular customer is 0.9678, compared to a .509925 probability of the same event in the dependent case. More instances of a channel selections by customers imply a less idle server, and thus greater stress. The independent channel system is only idle around 3% of the time, whereas the dependent channel system is idle almost half the time. Thus, we see why the independent case actually stresses the system more and produces a lower survival function than the dependent case.

The effect of the channel count on the difference between dependent and independent channels depends on the value of η. For η small, the effect of dependent and independent channel selection on the survival function depends heavily on K. See Figure 4b. For small K, the difference between almost independent and essentially dependent channel selection is negligible. But as K increases, we see that the difference between dependent and independent channels increases dramatically in both time and by K. At no point do the percent difference functions cross for $\eta = 0.01$.

For larger η, ($\eta \geq 1$), Figure 4a shows that the channel count has little effect on the percent difference over time between independent and dependent channel selection probabilities. However, the magnitude of the percent difference for all simulated channel counts is almost double that of the small η and $K = 300$ instance in Figure 4b. Thus, we see that only in very special circumstances can one ignore the dependency of channel selection and estimate the survival function with the more simplistic Binomial distribution for channel selection probabilities. In fact, since the survival function under independent channel selection is always below that of the dependent channel selection model, erroneous use of assumption of independence in channel selection would result in vastly overestimating the probability of failure. Policies designed around this overestimation can lead to suboptimal resource allocation and unnecessary traffic intervention protocols.

The next section isolates the channel effects alone on the survival function under Markovian channel dependency.

3.3 Channel Effects

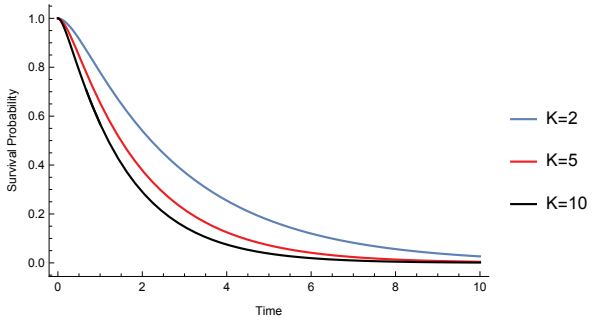


Figure 5. Channel Effect Comparison of Survival Functions

We investigate the effect of the number of channels alone on the survival function. We fix $\lambda = \eta = 1$, $r_0 \equiv 0.01$, $\bar{G}_W(w) = e^{-w}$ $p = 3/4$, and use a 3 state Markov chain with $\delta \in \{1/10, 1/2, 9/10\}$ and transition matrix

$$\mathbb{P} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 2/3 & 1/3 \\ 1/3 & 0 & 2/3 \end{bmatrix}$$

with steady state distribution $\pi = (1/4, 1/4, 1/2)$. Figure 5 shows the survival function for the survival function for $K \in 2, 5, 10$. Intuitively, an increase in the number of channels decreases the survival function, as seen in Figure 5. A larger number of possible tasks available for selection means that customer are more likely to select larger numbers of tasks upon server visitation, and thus the server stress increases. The most notable manifestation is the derivative of the survival function and its change as a result of an increase in channel counts. For large K, the survival function decreases sharply in a very short period of time.

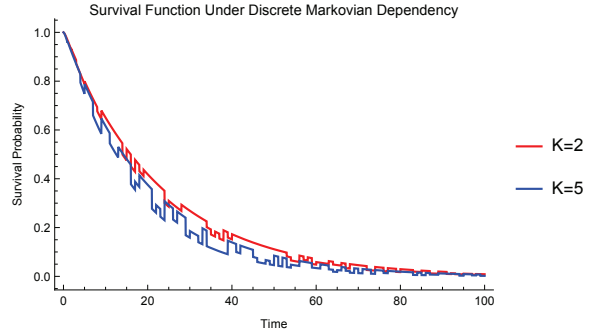


Figure 6. Conditional Survival Function for Various Channels

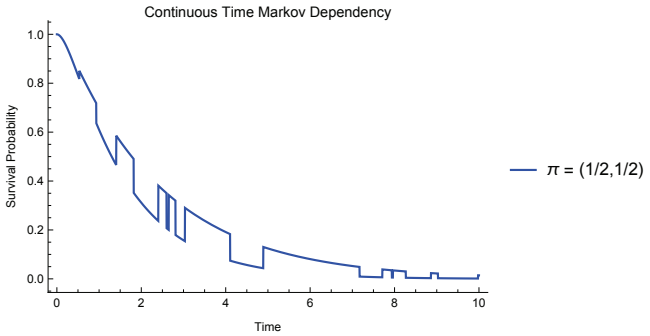
As discussed in Section 2.4, taking the expectation over X in Theorem 2 yields the "steady state" survival function.

In practice, when δ is a function of a random process, an observer of the server will see the survival function under a particular realization of that random process. So, of interest is the survival function under a specific realization of the Markov process. For this example, we show how specific realizations of the Markov process above manifest for $K = 2$ and $K = 5$. The transition matrix remains the same as above, and the first state transition is selected according to the steady state distribution of the Markov chain π. As expected, the survival function for $K = 5$ is below that for $K = 2$, but now the effect of the state transitions can be observed. The variability in the survival function is much higher for $K = 5$; that is, the survival function may jump more dramatically than for $K = 2$. This implies that there is a fair amount of variability around the steady state expected survival function as K increases. Thus, in designing possible admission control or other traffic management policies, one should not rely only on the steady state function given in Theorem 2.

Denote Ω as the set of survival functions under all possible trajectories of the underlying Markov chain. From Section 3.2, the largest the survival function can be for a given K is under complete dependency, i.e. for $\delta \equiv 1 \; \forall \; t \in \mathbb{R}^+$, and the lower bound is under the independent channel model, i.e. $\delta \equiv 0 \; \forall \; t \in \mathbb{R}^+$. Let $\mathscr{D}$ denote the possible values of δ that correspond to the Markov state space S. Ω is bounded by survival functions of constant δ, where the upper bound is given by (2) for $\delta = \max(\delta : \delta \in \mathscr{D})$, and the lower bound is give by (2) for $\delta = \min(\delta : \delta \in \mathscr{D})$. This provides strict bounds for the survival function for a given K and $\mathscr{D}$ (assuming all other quantites besides t are also fixed). A significant advantage of this model is that one may incorporate variability in the channel dependency without resorting to more traditional statistical means that provide mere estimates. Rather than a mean function with a corresponding variance to use in confidence bands, the bounds presented here are strict, so the probability of a particular trajectory straying outside the given bounds for the specific Markov chain model is precisely 0.

As a final illustration, the survival function for a specific trajectory of a continuous time Markov chain is presented in the next section.

3.4 Continuous Time Markov Dependency



Figureb 7. Conditional Survival Function under Continuous Time Markov Switching Process

Figure 7 illustrates a survival function conditioned on a realization of a continuous time Markov dependency structure. In this illustration, we have a two state Markov chain with $\delta \in \{1/100, 99/100\}$, rate matrix $\mathbb{Q} = \begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix}$ and stationary distribution $\pi = (1/2, 1/2)$. The channel count is fixed at $K = 5$, and the channel selection probability is $p = 1/2$. The customer service time distribution remains exponential with $\bar{G}_W(w) = e^{-w}$, and $\lambda = \eta = 1$. Under a continuous time Markov chain, the entries of the transition rate matrix $\mathbb{Q}$ describe the exponential rates at which the process departs state i and arrives at state j. In this particular example, the transition probabilities are a function of time, and the probability of transitioning from state 0 to state 1 is given by $P(t) = e^{-2t}$. Thus, the times spent in each state are now random, as opposed to the discrete time Markov chain where transitions occur at multiples of a specified time interval. In this example, we illustrate a switching process, where the channel selection switches from almost independent to almost dependent, but the switching times are random according to an exponential distribution with rate parameter 2. This switching process represents the most "wild" the behavior of the survival function can get, switching between the extremes of possible behavior for a particular channel count K.

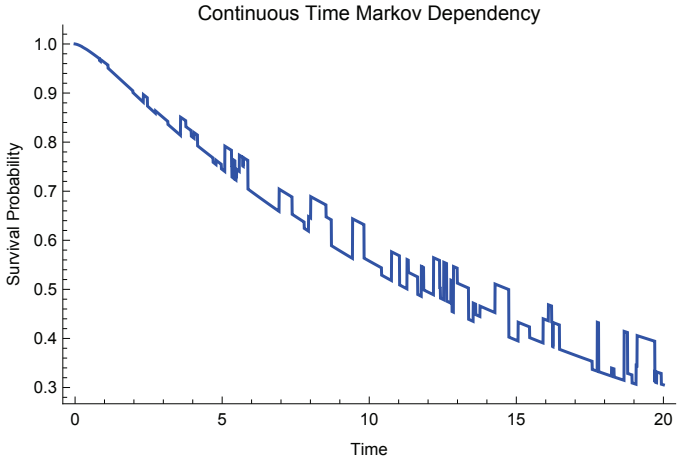


Figure 8. Conditional Survival Function under Continuous Time Markov Chain

To show a continuous time random walk between states, we let the set of possible dependencies $\mathscr{D} = \{1/100, 1/4, 1/2, 3/4, 99/100\}$, which yields a 5 state Markov chain. The transition rate matrix is given by

$$\mathbb{Q} = \begin{bmatrix} -4 & 1 & 1 & 1 & 1 \\ 2 & -5 & 1 & 1 & 1 \\ 1 & 2 & -5 & 1 & 1 \\ 1 & 3 & 2 & -7 & 1 \\ 1 & 1 & 1 & 1 & -4 \end{bmatrix}$$

All other factors remain the same as for the switching process above. The illustration is given in Figure 8.

3.5 Expected Lifetime: Effects of Dependency

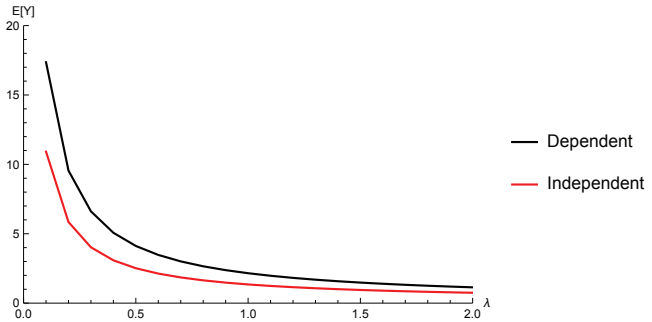


Figure 9. Expected Lifetime Comparison between Dependent and Independent Channels

Another common measure of reliability performance is given by the expected lifetime. Let Y be the random time to failure of the server. Then the expected lifetime of the server is given by

$$E[Y] = \int_0^\infty S_Y(t)dt$$

Figure 9 gives the expected lifetime for a clustered task server as a function of arrival rate λ for $K = 100$ channels and $\eta = 0.01$. The survival functions integrated are given in Section 3.2. As λ increases, both expected lifetimes naturally tend to 0, but the expected lifetimes under an independent channel structure are much lower than that for the dependent channel structure. For example, assuming a constant arrival rate $\lambda = 0.1$, the expected lifetime (in generic units of time *tu*) under essentially dependent channels is 17.36 *tu*, and the expected lifetime of a server with independent channels is 10.91 *tu*. This means that we expect the server to fail (or need rebooting) due to workload every 17.36 *tu* or 10.91*tu* for dependent and independent channels, respectively. For illustration, if we take the time units to be days, we would expect to have handle 21 outages per year from a dependent channel server, and 33 outages per year from an independent channel server, which is a 57% increase in the number of outages per year.

The next section briefly investigates other possible dependency structures and the impact on the survival function of the clustered-task server.

4. Conclusion

This paper has described a generalization of the work on clustered-task server reliability with correlated channels by Korzeniowski and Traylor in which the dependency structure was relaxed from constant δ to a function of time ($\delta(t)$) and a function of both a continuous time and discrete time finite state Markov process ($\delta(X(s))$). In both cases, the survival function was given in closed form.

Numerical illustrations were given in Section 3 to explore the various effects of dependency, number of channels, and the Markov process itself on the survival function. In all cases, the underlying function governing the dependency manifested in the shape of the survival function. Section 3.1 showed that, for deterministic temporal dependency, the sinusoidal function shape is directly apparent in the survival function, causing a loss of the typical monotonicity of the survival function as it oscillates between the minimum and maximum values of $\delta(t)$. Section 3.2 explored the actual effect of dependency on the survival function under a discrete time Markov switching process for $\delta = 1/100$ and $\delta = 99/100$. Regardless of the number of channels, independent task selection always produces a smaller survival probability. However, the difference depends on the value of the stress multiplier η and the number of channels K. For larger η, the difference is stark. As t increases, the percent difference between the survival functions under independent and dependent channels exceeds 100%. For small η and small K, the difference is negligible, and thus under this special case, one may assume independence even under the reality of complete dependence.

Section 3.3 examined the effects of the channel count alone on the steady-state survival function. The results were intuitive, with greater channel counts decreasing the survival function. Of greater interest is the survival function conditioned on a specific Markov realization. In this way, the variability was examined and strict bounds were derived. In general, the steady-state form of the survival function given in Theorem 2 should not be used alone when creating admission and routing/control policies. Section 3.4 illustrated the survival function for various trajectories of a continuous time Markov chain.

Section 3.5 examined the effects of dependency on the expected lifetime of a clustered-task server. The difference in expected lifetime between servers with dependent and independent channels seems less stark as a function of arrival rate, but calculations that show the number of expected outages reveal a vastly different level of reliability for dependent channels.

The models and simulations given illustrate a novel approach to modeling server reliability. Dependency among task selection is common, and we have provided a closed form of the survival function that incorporates the dependency. This allows for more robust and accurate policies to be derived based on this model, as more complex behaviors may now be rigorously accounted for.

References

Cha, K., & Lee, E. (2011). A stochastic breakdown model for an unreliable web server. *Journal of Applied Probability 48*(2), 453C466.

Iosup, A., Yigitbasi, N., Sonmez, O., & Epema, D. (2011). Performance evaluation of overload control in multi-cluster grids. *Proceedings of the 2011 IEEE/ACM 12th Annual Conference on Grid Computing*, 173-180. https://doi.org/10.1109/Grid.2011.30

Korzeniowski, A. On correlated random graphs. (2013). *Journal of Probability and Statistical Science 11*, 43-58.

Korzeniowski, A., & Traylor, R. (2016). Dynamic reliability of a cluster server. *International Journal of Statistics and Probability 5*(4), 45-51. https://doi.org/10.5539/ijsp.v5n4p45

Thomas, N., Gilly, K., Juiz, C., & Puigjaner, R. (2012). Adaptive admission control algorithm in a qos-aware web system. *Journal of Information Sciences 199*, 58-77. https://doi.org/10.1016/j.ins.2012.02.018

Traylor, R. (2016). Stochastic Reliability Models for a General Server and Related Networks. Phd thesis, University of Texas at Arlington.

Vishwanath, K. V., & Nagappan, N. (2010). Characterizing cloud computing hardware reliability. *Proceedings of the 1st ACM symposium on Cloud computing*, 193-204. https://doi.org/10.1145/1807128.1807161

Welsh, M., and Culler, D. (2003). Adaptive overload control for busy internet servers. *Proceedings of the 4th conference on USENIX Symposium on Internet Technologies and Systems*, *4*.