Computers & Industrial Engineering 110 (2017) 297–306

Contents lists available at ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

Analysis of a two-node closed queueing cyclic network with one non-exponential node

Bacel Maddah [a,*], Muhammad El-Taha [b]

[a] Department of Industrial Engineering and Management, Faculty of Engineering and Architecture, American University of Beirut, Beirut 1107 2020, Lebanon
[b] Department of Mathematics and Statistics, University of Southern Maine, 96 Falmouth Street, Portland, ME 04104-9300, United States

ARTICLE INFO

Article history:
Received 12 October 2016
Received in revised form 10 April 2017
Accepted 5 June 2017
Available online 9 June 2017

Keywords:
Closed queueing networks
Cyclic networks
Machine repair
General repair times
High variability
Birth Processes

ABSTRACT

Consider a two-node closed cyclic queueing network with a total of M customers. The first node consists of $N \leqslant M$ parallel identical servers having exponential service times. Node 2 consists of a single server having a general service time distribution. Waiting space at each node is sufficient to accommodate all M customers.

We derive an exact solution for the steady state system size probabilities. Our approach is based on developing and solving an imbedded Markov chain at Node 2 service completion epochs coupled with a birth process over Node 1 service times. In that sense, our methodology is intuitively appealing and easy to apply. We focus on analytically and numerically investigating the effect of Node 2 service time variability on system performance. We find that closed queueing networks, like the one considered here, are less vulnerable to variability than open networks. However, the effect of variability is still significant to merit considerable attention. Numerical results are presented to illustrate the viability of our method for many service time distributions. As by-products, our work yield new results on birth processes and combinatorial identities, which can be useful in other contexts.

A primary application of this model is the well-known machine repair model where a set of identical machines are attended by a single repairman. Other applications include performance evaluation of manufacturing and computer networks, as well as reliability studies where our model can be easily used to compute system availability.

1. Introduction and literature review

In this article we consider a two-node closed cyclic queueing network with a total of M customers. The first node consists of $N \leqslant M$ parallel identical servers having independent and identically distributed (iid) exponential service times with rate λ each. The second node consists of a single server with iid service requirements $S_i, i = 1, 2, \ldots$, having a general distribution function $B(.)$. Waiting space at each node is sufficient to accommodate all M customers and the service discipline is *FCFS*. Fig. 1 shows the block diagram of the model.

A main motivation of this paper is to develop a methodology for determining the measures of performances of this two-node network model which is easy to understand and feasible to implement. For this purpose, we develop and solve an imbedded Markov chain with states given by the number of customers in the second node at departure epochs. Such an approach is standard in analyzing the general single-server model with Markovian arrivals, $M/G/1$. Moreover, the imbedded Markov chain technique has been successfully applied to study the machine repair model with no spares, $M/G/1//N$, (e.g. Saaty, 1961; Takagi, 1993; van Hoorn, 1984). In our model, this is equivalent to the special case of $M = N$. The fact that M can be larger than N complicates the development of the imbedded Markov chain in our model since the transition rate to the second node is not as smooth as in the case of $M/G/1$ and $M/G/1//N$. This may explain why no other work, that we are aware off, in the literature utilizes an imbedded Markov chain approach for the closed cyclic two-node network model as the one we consider. In this paper, we draw upon results from the theory of birth processes (e.g. Ross, 1997) to facilitate the development of the imbedded Markov chain solution. Our approach is based on the simple intuition that our closed network model is "equivalent" to a $M/G/1$ in certain states and to a $M/G/1//N$ in other states. The resulting structure is intuitively appealing and can be understood and implemented easily with a basic knowledge of Markov chains.

* Corresponding author.
E-mail addresses: bacel.maddah@aub.edu.lb (B. Maddah), el-taha@maine.edu (M. El-Taha).

http://dx.doi.org/10.1016/j.cie.2017.06.007
0360-8352/

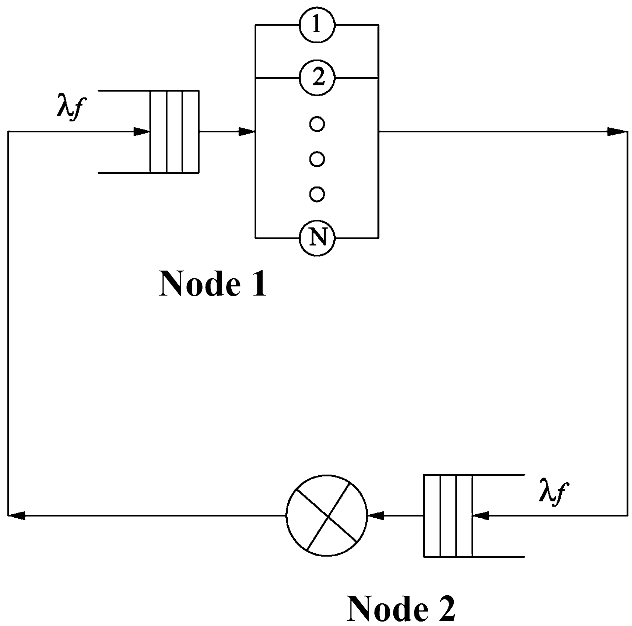

Fig. 1. Block diagram of the model.

Another motivation of this paper is to study the effect of service times variability on system performance within a closed queueing network context. This issue has received attention within an open network framework (e.g. El-Taha & Maddah, 2006; Harchol-Balter, 2002; Whitt, 1999). This attention is fueled, in part, by the detection of "heavy tailed" service times characterized by high variability in several computer and communication systems, especially in Internet traffic (e.g. Crovella & Bestavros, 1997; Fowler, 1997; Harchol-Balter & Downey, 1997; Naldi, 1999). Since our closed network applies in such systems, it is of interest to study how variability affects this model. In this paper, we study the effect of Node 2 service time variability, analytically and numerically. Our main finding is that variability in closed queueing networks leads to system performance deterioration, similar to open networks. However, variability effect is less severe in closed networks due to a "self-correcting" behavior related to the finiteness of population size in closed networks. We formally demonstrate that variability effect is bounded in our network model in the case of Node 2 service times having a hyper-exponential distribution with an infinite variance.

While pursuing our main research objectives, we derive useful by-products, which broaden the contribution of this paper. Specifically, we derive a combinatorial identity involving nested geometric sums. We then utilize this identity to develop a closed-form expression for the transition probabilities of a pure birth process with a birth rate which starts constant (as in a Poisson process) for a small population size, and then decreases linearly (similar to a Yule process) when the population size crosses a threshold. Both of these novel results can be useful in other contexts.

To recap, the main contributions of this paper are (i) developing a simple solution methodology for the two-node closed network model based on an imbedded Markov chain coupled with a birth process; (ii) analyzing the effect of Node 2 service time variability on the performance of the closed network; and (iii) developing novel results on combinatorial identities and birth processes.

A primary application of this model is in machine repair where Node 1 represents a set of N working identical machines that are subject to random failures. Machine times to failure are iid exponential random variables. Node 2 represents a single repairman facility where repair times are iid with an arbitrary (general) distribution. Upon failure, a machine moves to the repair facility where it may start repair immediately, or, otherwise wait in queue until a repair person, attending other failed machines, is available. Once a machine is repaired it is redirected to the production floor. The demand on the system requires that N machines be operational at any given time. However, random failures lead to inevitable delays at the repair facility which tends to degrade the system performance. The machine repair model is widely utilized in the analysis of many systems characterized by a finite population of customers. This model has many applications in manufacturing (see Haque & Armstrong, 2007; Koenigsberg, 1982; Stecke, 1982 for reviews).

One way to guarantee that demand is often met, is to acquire spare machines. That is, the system would own a total of $M = N + Y$ machines, N of which may be deployed to meet the demand while the remaining operational machines, if any, wait in queue at the production facility to resume operation immediately in case of a failure in the deployed machines. Such a model is commonly considered under the assumption of exponential repair and failure times (e.g. Koenigsberg, 1982; Stecke, 1982). The exponential assumption is mostly for analytical tractability. For example, if machines failures are one of two types at random, then the overall repair time distribution is not exponential. Instead, the repair time is a mixture of two random variables, characterized by high variability if the two repair types are sensibly different. In this context, the model is denoted by $M/G/1//N + Y$ when appropriate.

Our network model has other applications as well. In particular, it applies in computer science to the study of limited memory multiprogramming systems where one node represents the central processing unit (CPU) and the other node represents Input/Output (I/O) devices, and a fixed number M of programs is processed continuously at the two nodes in a cyclic fashion (e.g. Allen, 1978, pp. 242–250; Shum, 1980). The assumption of fixed number of programs is justified in heavy traffic since, in this case, a job that finishes processing gets quickly replaced by a new job. In addition, this closed network model has applications in reliability as it can be used to evaluate the availability of a K-out-N system composed of M parts which is functional if at least K parts are operating (e.g. Krishnamoorthy & Ushakumari, 2001; Lam & Zhang, 2000 and references therein). We illustrate this via example in Section 4.

Literature review. There is a vast related literature on two-node networks in the context of the machine repair model and its variants. A recent paper by Haque and Armstrong (2007) presents an extensive review of the related recent literature. Reviews of earlier works are found in Koenigsberg (1982) and Stecke (1982). In the following, we review the literature that is mostly related to this paper. Specifically, we limit the review to articles that consider two-node closed cyclic networks and/or machine repair with non-exponential repair times and spares. To the best of our knowledge, the only article that develops an algorithm that can potentially give exact solution (in terms of system size probabilities) for the $M/G/1//N + Y$ model with general repair distribution is by Gupta and Rao (1996). However, Gupta and Rao (1996) apply the supplementary variable technique, which requires more specialized knowledge compared to our imbedded Markov chain approach. We believe that our solution methodology is easier to implement because it does not require higher order differentiation of the Laplace-Steiltjes transform of the service times density function. In addition, our article is the first to study the effect of variability on system performance in such closed systems. Rao and Gupta (2000) extend the model in Gupta and Rao (1996) by considering warm spare machines (i.e. spare machines that may fail while not in operation). One recent article that also utilizes the supplementary variable technique on the $M/G/1//N$ model, involving no spares, is by Wang, Su, and Yang (2014) who focus on the effect of "imperfect coverage" where a failed machine may not be noticed by the repairman.

Few papers develop exact solution for non-exponential closed systems, especially the $M/G/1//N+Y$ model with a specific distribution of repair times. Wang (1994) studies a $M/G/1//N+Y$ system with an Erlang distribution of repair times, while Wang and Wu (1995) studies a system with two modes of failure each requiring an exponentially distributed repair times, which is equivalent to having a repair time with a hyperexponential distribution. Both models in Wang (1994), Wang and Wu (1995) allow for warm spares. Recently, Jain, Sharma, and Sharma (2008) consider a model similar to Wang and Wu (1995) with additional variations such as balking and reneging. The main contribution of these papers (Jain et al., 2008; Wang, 1994; Wang & Wu, 1995) is developing the system sized probabilities utilizing a two dimensional continuous Markov chain. Our model is general enough to include most of these articles as special cases (excluding features like warm spares, balking, reneging, and features that we do not include in our model).

Some studies in the literature on computer systems assume general processing times at one node in a cyclic network. However, to the best of our knowledge, such studies are limited to having a single server at both nodes. For example, Lewis and Shedler (1971) develop an exact imbedded Markov chain solution for a generalized cyclic queue model with preemption and single-server nodes, where the CPU processes jobs in multiple stages before and after visiting the I/O. Shum (1980) develops an approximation to extend product form solutions to the case of general-service-time single-server queueing networks. Gaver and Shedler (1973) develop a diffusion type approximation for the cyclic queue with a single server at each node. Our model may be useful in similar computer science applications involving multiple parallel processors. Finally, one stream of literature utilize diffusion approximations for general service times and multiple servers at both nodes of the network. The reader interested in diffusion approximations may refer to Jain (1993) and Sivazlian and Wang (1989), Sivazlian and Wang (1990).

The remainder of this paper is organized as follows. In Section 2, we derive an imbedded Markov chain and solve for the imbedded and time-average probabilities of the closed network observed at Node 2 service completion epochs. In Section 3, we analytically investigate the effect of repair time variability on system performance. In Section 4, we present numerical examples using several service (repair) time distributions to illustrate the solution methodology and further discuss, via examples, the effect of repair time variability. Section 5 concludes the paper.

2. The imbedded Markov chain

In this section, we describe a Markov chain imbedded at the service completion instants of the second node. We use ideas from pure birth processes to derive expressions for the one-step transition probabilities of this Markov chain. This allows obtaining the stationary system-size probabilities at departure epochs, and, subsequently, at arbitrary times. Key performance measures are then determined.

Let X_n be the number of customers at Node 2 upon service completion of the n^{th} customer. Then, the number of customers X_n is related to X_{n-1} as follows:

$$X_n = \begin{cases} X_{n-1} + A_n - 1 & \text{if } X_{n-1} > 0 \\ A_n & \text{if } X_{n-1} = 0 \end{cases},$$

where A_n is the number of transitions from Node 1 to Node 2. In the context of the machine repair model, this represents the number of failures during the repair time of the n^{th} machine. Just after the n^{th} customer starts service the number of customers at Node 2 is X_{n-1} and the number of customers at Node 1 is $M - X_{n-1}$. Therefore, the number of transitions A_n depends on X_{n-1} only. This implies that $\{X_n; n = 1, 2, \cdots\}$ is a Markov chain. The one-step transition probabilities of X_n defined as, $p_{ij} = P\{X_n = j | X_{n-1} = i\}$, $i = 0, \ldots, M-1$, $j = 0, \ldots, M-1$, are given by

$$p_{ij} = P\{A_n = j - i + 1 | X_{n-1} = i\}, \text{ if } i \geqslant 1,\ j \geqslant i - 1, \tag{1}$$

and for $i = 0, p_{0j} = p_{1j}$ for all j; otherwise, $p_{ij} = 0$ if $j < i - 1$. Let $\pi_i, i = 0, \ldots, M-1$ be the stationary distribution of the Markov chain X_n.

Our approach is based on using the imbedded Markov chain for solving for the stationary system-size departure epochs probabilities $\{\pi_i, i = 0, \ldots, M-1\}$, $\pi_i = \lim_{n\to\infty} P\{X_n = i\}$. The stationary probabilities are interpreted as

$$\pi_i = \lim_{n\to\infty} P\{n^{th} \text{ departing customer leaves behind } i \text{ customers at Node 2}\}.$$

The stationary distribution $\{\pi_i, i = 0, \ldots, M-1\}$, is evaluated by solving the equation, $\pi P = \pi; \sum_{i=0}^{M-1} \pi_i = 1$, where P is the transition probability matrix, $(P)_{i,j} = p_{ij}$, and $\pi = (\pi_0, \pi_1, \ldots, \pi_{M-1})$.

Now we turn our attention to deriving expression for the one step transition probabilities that can be computed efficiently. Let $\{W(u), 0 \leqslant u \leqslant t\}$ be a pure birth process representing the number of transitions (arrivals) from Node 1 to Node 2 during one service time of length t at Node 2. That $W(u)$ is a pure birth process is due to the fact that the time between two consecutive arrivals (machine failures) is exponential, possibly with different rates. Then, p_{ij} is the probability that $W(u)$ produces $j - i + 1$ births during a random service time at Node 2,

$$p_{ij} = \int_0^\infty P\{W(t) = j + 1 | W(0) = i\} dB(t) = \int_0^\infty Q_{i,j+1}(t) dB(t), \tag{2}$$

where $Q_{ij}(t)$ are the transition probabilities of $W(u)$. The following results are utilized to evaluate $Q_{ij}(t)$.

Lemma 2.1. *For $i = Y, \ldots, M-1$,*

$$Q_{ij}(t) = \binom{M-i}{M-j} e^{-(M-j)\lambda t} \left(1 - e^{-\lambda t}\right)^{j-i};\ j = i, \ldots, M.$$

Proof. With $i \geqslant Y$, the birth rate of $W(u)$ is $\lambda_i = (M - i)\lambda$. For $j = i$, Ross (1997) (Proposition 6.1, p. 314) proves that $Q_{ii}(t) = e^{-\lambda_i t} = e^{-(M-i)\lambda t}$. Therefore, the lemma holds for $i = j$. We complete the proof using induction. Suppose that the lemma is valid for some $j - 1 \geqslant i$. Then, following Ross (1997) (Proposition 6.2 p. 321)

$$\begin{aligned} Q_{ij}(t) &= \lambda_{j-1} e^{-\lambda_j t} \int_0^t e^{\lambda_j s} Q_{i,j-1}(s) ds \\ &= (M - j + 1)\lambda e^{-(M-j)\lambda t} \binom{M-i}{M-j+1} \int_0^t e^{-\lambda s} \left(1 - e^{-\lambda s}\right)^{j-i-1} ds \\ &= \frac{M-j+1}{j-i} \binom{M-i}{M-j+1} e^{-(M-j)\lambda t} \left(1 - e^{-\lambda t}\right)^{j-i} \\ &= \binom{M-i}{M-j} e^{-(M-j)\lambda t} \left(1 - e^{-\lambda t}\right)^{j-i}, \end{aligned}$$

which completes the proof. □

Lemma 2.1 has the following intuitive interpretation in the context of the machine repair model. Since machines fail according to an exponential distribution, starting with $M - i \leqslant N$ machines on the production floor, a machine fails independently of other machines with probability $1-e^{-\lambda t}$. Then, the number of failed machines during a time t is a Binomial random variable with parameters $M-i$ and $(1-e^{-\lambda t})$. With no spares, i.e. $Y=M-N=0$, for the case of the $M/G/1//N$ model, Lemma 2.1 applies for all $i \geqslant 1$. This intuitive argument is commonly utilized in the literature in the analysis of the $M/G/1//N$ model (see, for example, Saaty, 1961, p. 324; van Hoorn, 1984, p. 19). Lemma 2.1 gives a formal argument for this intuition and helps provide a better understanding of the $M/G/1//N+Y$ model. The second result deals with the case when $i \leqslant Y-1$.

Lemma 2.2. *For $i=1,\ldots,Y-1$,*

$$Q_{ij}(t)=\frac{e^{-N\lambda t}(N\lambda t)^{j-i}}{(j-i)!},\ \text{for } j=i,\ldots Y, \tag{3}$$

$$Q_{ij}(t)=\lambda_{j-1}e^{-\lambda_j t}\int_0^t e^{\lambda_j s}Q_{i,j-1}(s)ds,\ \text{for } j=Y+1,\ldots,M, \tag{4}$$

where $\lambda_j=(M-j)\lambda$.

Proof. The proof follows directly from Proposition 6.2 in Ross (1997) by noting that the birth rates of $W(u)$ are given by $\lambda_i=N\lambda$, if $i<Y$, and $\lambda_i=(M-i)\lambda$, otherwise. The first part of the Lemma follows by using induction, and for $j>Y$ the lemma is restating Ross' Proposition. □

Intuitively, in the context of the machine repair model, Lemma 2.2 states that if the system starts a repair with $i<Y$ machines in the repair facility and ends repair with j machines, then the first $Y-i$ machines fail according to a Poisson process. That is, the system is "equivalent" to an $M/G/1$ queue with arrival rate $N\lambda$. If more than $Y-i$ machines fail, then the last $j-Y$ machines fail according to diminishing failure rates similar to those of an $M/G/1//N$ model with no spares. That is, for $i<Y$, the $M/G/1//N+Y$ model is a mixture of the $M/G/1$ and the $M/G/1//N$ models. Otherwise, for $i \geqslant Y$, the behavior of the model is similar to that of an $M/G/1//N$.

In Lemma 2.2, the evaluation of (4) can be computationally demanding. The following theorem presents a combinatorial identity that allows simplifying the expression in (4).

Theorem 2.3. *(**MTGS**) For Z and k integers, and $S_k=r_1+\cdots+r_k$, we have the following identity*

$$\sum_{r_1=0}^{Z}\sum_{r_2=0}^{Z-r_1}\cdots\sum_{r_k=0}^{Z-S_{k-1}}\prod_{l=1}^{k}\left(\frac{k+1}{l}\right)^{r_l}=(k+1)^{Z+1}\sum_{r=0}^{k}(-1)^{k+r}\binom{k}{r}\times(k+1-r)^{-(Z+1)}.$$

Proof. See Appendix. □

We call the identity in Theorem 2.3 *multidimensional triangular geometric sum* (MTGS) for obvious reasons. Theorem 2.3 leads to the following theorem which provides a computationally efficient closed-form expression for (4).

Theorem 2.4. *For $i=1,\ldots,Y-1$, and $k=1,\ldots,N$, let $S_k=r_1+\cdots+r_k$, then*

$$Q_{i,Y+k}(t)=N^{Y-i}\prod_{l=1}^{k}(N-l+1)\left[\frac{1}{(k-1)!}\sum_{r=0}^{k-1}(-1)^r\binom{k-1}{r}\times(k-r)^{-(Y-i+1)}e^{-(N-k+r)\lambda t}+(-1)^k\sum_{r_1=0}^{Y-i}\sum_{r_2=0}^{Y-i-r_1}\cdots\sum_{r_k=0}^{Y-i-S_{k-1}}\frac{N^{i+S_k-Y}}{\prod_{l=1}^{k}l^{r_l+1}}Q_{i+S_k,Y}(t)\right],$$

where $Q_{i+S_k,Y}(t)=e^{-N\lambda t}\frac{(N\lambda t)^{Y-i-S_k}}{(Y-i-S_k)!}$.

Proof. See Appendix. □

Now, the transition probability matrix $P=[p_{ij}]$ of the Markov chain $\{X_n,n=0,1,\ldots,\}$ is evaluated utilizing (2). For $i\geqslant Y$, it follows from (2) and Lemma 2.1 that

$$p_{ij}=\binom{M-i}{j-i+1}\sum_{r=0}^{j-i+1}(-1)^r\binom{j-i+1}{r}B^*[(M-j-1+r)\lambda],\quad j=i-1,\ldots,M-1, \tag{5}$$

where $B^*(.)$ is the LST of $B(.), B^*(u)=\int_0^\infty e^{-u}dB(u)$. Eq. (5) allows the evaluation of $p_{ij}, i\geqslant Y$ with ease for repair time distributions with closed-form Laplace transforms. For $i<Y$, the evaluation of $p_{ij}, j=i-1,\ldots Y$, is similar to the case of the $M/G/1$ which can be done easily as well. For $i<Y$ and $j>Y, Q_{ij}(t)$ are evaluated using Theorem 2.4, and then p_{ij} are determined. In its current form, Theorem 2.4 allows determining $Q_{ij}(t), i<Y, j>Y$ function of $M/G/1$-type transition probabilities. For instance, it may be seen that, for $i<Y$,

$$Q_{i,Y+1}(t)=N^{Y-i+1}\left[e^{-(N-1)\lambda t}-\sum_{r=0}^{Y-i}N^{i+r-Y}Q_{i+r,Y}(t)\right]$$

$$Q_{i,Y+2}(t)=N^{Y-i+1}(N-1)\left[2^{-(Y-i+1)}e^{-(N-2)\lambda t}-e^{-(N-1)\lambda t}-\sum_{r=0}^{Y-i}N^{i+r-Y}(1-2^{-(r+1)})Q_{i+r,Y}(t)\right].$$

Before proceeding we point out that the MTGS identity in Theorem 2.3 and the birth process result in Theorem 2.4 may be applicable in other context not necessarily involving closed cyclic networks and machine repair. To the best of our knowledge, these two results are novel.

Having determined the transition probability matrix for the Markov chain $\{X_n,n=0,1,\ldots,M-1\}$ the limiting probabilities $\pi_i, i=0,\ldots,M-1$, are then evaluated. However, we are interested in the time-average stationary distribution of the number of customers in at Node 2 at at arbitrary times defined as

$$p_i=\lim_{t\to\infty}P\{X(t)=i\},\ i=0,\ldots,M,$$

where $X(t)$ is the number of customers at Node 2. Moreover we are interested in determining the performance measures of the system. The most important performance measure is the *throughput* or, in the machine repair model, the aggregate failure rate, λ_f. The rate λ_f measures the average traffic flow in the network, as shown in Fig. 1. The following lemma determines the values of λ_f and the p_i's once the π_i's are obtained from the Markov chain analysis.

Lemma 2.5. *Let $E[S]$ be the mean service time, $\lambda_i=N\lambda$ for $i\leqslant Y$ and $\lambda_i=(M-i)\lambda$, otherwise. Then, the throughput, λ_f, is given by*

$$\lambda_f=\frac{N\lambda}{\pi_0+N\lambda E[S]}. \tag{6}$$

The system size probabilities, $p_i,\ i=0,\ldots M$, are given by

$$p_i=\frac{\lambda_f\pi_i}{\lambda_i},\ i=0,\ldots M-1,$$

and $p_M=1-\sum_{i=0}^{M-1}p_i$.

Proof. The probabilities π_i and p_i are related by the following equation (see, for example, El-Taha & Stidham (1999)):

$$\lambda_i p_i=\lambda_f\pi_i,\ i=0,1,\ldots M-1.$$

In addition, Little's law implies that

$$1-p_0=\lambda_f E[S].$$

The result follows upon solving for λ_f and simplifying. □

With λ_f and the p_i's determined from Lemma 2.5, other measures of performance could be determined easily. The mean number of customers at Node 2 (down machines) $E[N_d]$ is given by

$$E[N_d] = \sum_{i=1}^{M} ip_i.$$

Little's law implies that the mean delay at Node 2 is

$$W_q = E[N_d]/\lambda_f - E[S],$$

and the average number of customers at Node 1 (operational machines) is

$$E[N_o] = \frac{\lambda_f}{\lambda}.$$

In Section 4, we evaluate λ_f, W_q, and $E[N_o]$ numerically using several service time distributions.

3. Effect of high variability

In this section, we analytically investigate the effect of Node 2 service time (repair time) variability on system performance. The following lemma develops bounds on the throughput which are function of the first moment of service times only.

Lemma 3.1. *The aggregate failure rate, λ_f, satisfies*

$$\frac{1}{1/(N\lambda) + E[S]} \leqslant \lambda_f \leqslant \frac{1}{E[S]}.$$

Proof. Follows directly from (6) by noting that $0 \leqslant \pi_0 \leqslant 1$. □

Lemma 3.1 indicates some insensitivity to service time variability especially if $N\lambda$ is large. Moreover, this lemma shows that even an infinite variance of Node 2 service times (repair times) will not completely deteriorate the system performance since the aggregate throughput (failure rate) is always above the lower bound in the lemma. This is in contrast with open systems where, in heavy traffic, service times with infinite variance lead to infinite delay (e.g. Scheller-Wolf & Sigman, 1997). We elaborate more on this with a sharper result in Lemma 3.2.

We present a result when Node 2 service (repair) time has infinite variance. Consider a service/repair time having a two phase balanced hyper-exponential distribution, H_2. The H_2 distribution, a mixture of two exponential distribution, has a density function of the form

$$b(t) = 2p^2\mu e^{-2p\mu t} + 2(1-p)^2\mu e^{-2(1-p)\mu t}, (0 \leqslant p \leqslant \frac{1}{2}). \tag{7}$$

The variance of this distribution is infinite when $p \to 0$. The following lemma describes the asymptotic performance of our model under H_2 service times with infinite variance.

Lemma 3.2. *Suppose that Node 2 has a H_2 service time distribution function with parameters p and μ as given in (7). As $p \to 0$, the system throughput approaches $\frac{N\lambda}{\pi_0^{M,2\mu} + N\lambda E[S]}$, where $\pi_0^{M,2\mu}$ is the probability of having zero customers at Node 2 at a departure instant in a $M/M/1//N+Y$ system with the same throughput (failure rate) but with twice the service rate as the $M/H_2/1//N+Y$ system.*

Proof. Recall that $P = [p_{ij}]$ is the transition matrix of our imbedded Markov chain. Then, it follows from (2) and (7) that $P = pP^{M,2p\mu} + (1-p)P^{M,2(1-p)\mu}$, where $P^{M,2p\mu}$ and $P^{M,2(1-p)\mu}$ are the transition matrices of the imbedded Markov chains of the $M/M/1//N+Y$ systems with failure rates $2p\mu$ and $2(1-p)\mu$ respectively. Denote by $\pi(p)$ the system size departure epochs probabilities as a function of p. In the following we find $\lim_{p\to 0}\pi(p)$. Note that $\pi(p) = \pi(p)[pP^{M,2p\mu} + (1-p)P^{M,2(1-p)\mu}]$. Then, $\lim_{p\to 0}\pi(p) = \lim_{p\to 0}\pi(p)P^{M,2\mu}$. (Observe that the $\lim_{p\to 0} pP^{M,2p\mu} = \mathbf{0}$, where $\mathbf{0}$ is a matrix with all elements equal zero, because all elements of $P^{M,2p\mu}$ are bounded between 0 and 1.) The uniqueness of the limiting probabilities corresponding to $P^{M,2\mu}$ implies that $\lim_{p\to 0}\pi(p) = \pi^{M,2\mu}$, where $\pi^{M,2\mu}$ are the limiting probabilities corresponding to $P^{M,2\mu}$. Utilizing (6) completes the proof. □

The result in Lemma 3.2 may appear counter-intuitive at first as it indicates that extremely high service time variability will not completely damage the system performance. (As aforementioned, this is not the case in open systems.) However, Lemma 3.2 can be understood by thinking of the basic characteristic of the system at hand. The system has a finite population of customers circulating between two-nodes (in the machine repair model, the system owns a *finite* number of machines). When high variability of service (repair) times leads to a long queue at the repair facility, the arrival rate to the second node (repair sub-system) diminishes (as the number of operational machines decreases), which allows fast clearing of the queue. The system, in this sense, is self-correcting. Shum (1980) makes a similar point about insensitivity to extremely high variability. However, Shum (1980) does not provide analytical (or numerical) evidence.

More specifically, Lemma 3.2 is believed to develop a tighter lower bound than Lemma 3.1 on the throughput of the $M/H_2/1//N+Y$ system under infinite variance. This is reasonable since in an alternate system with twice the repair rate, such as the one in the lemma, we expect that the probability of having an empty system increases (i.e. $\pi_0 \leqslant \pi_0^{M,2\mu}$). This indicates that the limit in the lemma is indeed a lower bound on λ_f. In Section 4, we observe numerically that the λ_f converges to this bound as the variance of repair times approaches infinity.

4. Applications

In this section, we present several examples to illustrate the application of our model and draw useful insights. In Example 1, we illustrates the application of the model for the case of deterministic service times. In Example 2, we numerically evaluate the system performance for many distributions and interpret the results focusing on the effect of service time variability. In Example 3, we discuss the effect of high variability in detail utilizing numerical results on the hyper-exponential distribution. In Example 4 we explain how can our model be used to assess system reliability. Finally, in Example 5, we explain how the model can be amended to handle a case of group arrivals.

Example 1. Consider a deterministic service time with value b. Consider a system with four machines and with a demand of two machines. That is, $M = 4, N = 2$ and $Y = 2$. Lemma 2.2 and Theorem 2.4 imply that

$$Q_{11}(t) = e^{-2\lambda t},\ Q_{12}(t) = 2\lambda te^{-\lambda t},$$
$$Q_{13}(t) = 4e^{-\lambda t} - 4\lambda te^{-2\lambda t} - 4e^{-2\lambda t},\ \text{and}$$
$$Q_{14}(t) = 1 - 4e^{-\lambda t} + 2\lambda te^{-2\lambda t} + 3e^{-2\lambda t}.$$

Lemma 2.1 implies that

$$Q_{22} = e^{-2\lambda t},\ Q_{23} = 2e^{-\lambda t}(1 - e^{-\lambda t}),\ Q_{24}(t) = (1 - e^{-\lambda t})^2,$$
$$Q_{33}(t) = e^{-\lambda t},\ \text{and} Q_{34}(t) = 1 - e^{-\lambda t}.$$

Note that $Q_{ij}(t) = 0$ for all values of i and j not shown above. With deterministic service times of duration b, Eq. (1) implies that the transition probabilities matrix of the Markov chain is given by

$$P = \begin{pmatrix} e^{-2\lambda b} & 2\lambda b e^{-2\lambda b} & 4e^{-\lambda b} - 4\lambda b e^{-2\lambda b} - 4e^{-2\lambda b} & 1 - 4e^{-\lambda b} + 2\lambda b e^{-\lambda b} + 3be^{-2\lambda b} \\ e^{-2\lambda b} & 2\lambda b e^{-2\lambda b} & 4e^{-\lambda b} - 4\lambda b e^{-2\lambda b} - 4e^{-2\lambda b} & 1 - 4e^{-\lambda b} + 2\lambda b e^{-\lambda b} + 3be^{-2\lambda b} \\ 0 & e^{-2\lambda b} & 2e^{-\lambda b}(1 - e^{-\lambda b}) & (1 - e^{-\lambda b})^2 \\ 0 & 0 & e^{-\lambda b} & 1 - e^{-\lambda b} \end{pmatrix}$$

Example 2. In this example, we consider the following repair time distributions: Deterministic, uniform, exponential, and the two phase hyper-exponential (H_2) distribution discussed in Section 3. We fix the mean of the distribution to 1.0, i.e. $\mu = 1$, for all four distributions. For the $H_2, p = 0.1$ and the uniform distribution has support $(0, 2)$. Let $\rho = \lambda/\mu$ be the offered load at the repair server. Applying our solution methodology using the four distributions for the system with $M = 4$ customers, $N = 2$ and $Y = 2$, we obtain the results in Tables 1 and 2.

To verify our numerical computations, results with exponential service/repair times were compared with similar values obtained from the classical approach based on a birth-death process. Results from both approaches were exactly the same on all cases as expected. For other distributions, we utilized simulation to verify the results. All results were valid when compared with simulation outputs with an average run length of five millions repair completions.

Gupta and Rao (1996) obtain similar results using the supplementary variable technique. Their approach results in an algorithm that requires as input the Laplace Stieltjes transform (LST) of the Node 2 service time distribution and their $(j+1)^{\text{st}}$ derivatives for $j = 0, 1, \ldots, Y-2$. This is a significant requirement when the number of spares is relatively large. In contrast, our developed method does not need higher order derivatives as input. Moreover, our method is more efficient computationally as our computations show.

In Tables 1 and 2, the effect of congestion as reflected by the offered load and service time variability is clear. For example, with deterministic service times when the offered load, ρ increases from 0.4 to 0.8 the average waiting time in the repair facility, W_q, increases from 0.784 to 1.703, by a factor of approximately 2. With other distributions, similar increase in the delay is observed. Similarly, the mean number of Node 1 customers (up machines operating,) $E[N_o]$, decreases as ρ increases (i.e. the failure rate increases). The throughput, λ_f as a measure of traffic intensity (λ_f/μ) also increases as ρ increases.

The variability of the four considered distributions increases according to their order in the tables. It can be observed that the system performance diminishes as one goes down in Tables 1 and 2. For example, in Table 1, λ_f decreases from 0.732 to 0.623, by approximately 15%, when the service/repair time distribution is shifted from deterministic with no variability to hyper-exponential with high variability. Similar effects are observed with other performance measures in the tables. Although the effect of variability (on the throughput for example) is not modest, one may still argue, by observing the results above that the system exhibits certain degree of "insensitivity" toward service time variability compared to open models (e.g., El-Taha & Maddah, 2006). This is not surprising given our analytical results in Section 3.

Example 3. In this example, we further investigate the effect of service time variability on system performance. We consider the hyper-exponential distribution and vary the squared coefficient of variation of service times, C_S^2, which is equal to the ratio of the variance to the square of the mean.

It is clear from Table 3, that system performance declines as variability increases. However, the decline happens at a decreasing rate. That is, as variability increases from "moderate" to "high" (say lines 1 and 3 of Table 3) the system performance deteriorates by a somewhat significant amount. On the other hand, when variability is increased from "high" to "extremely high" the system performance is almost insensitive (compare lines 3 and 4). Additional results, not shown here, indicate that λ_f decreases by very modest amounts as C_S^2 increases beyond the values in line 4 of Table 3. In fact, the lowest values of λ_f we observe as C_S^2 increases are 0.560 for $\rho = 0.4$ and 0.812 for $\rho = 0.8$. These values correspond to the lower bound derived in Lemma 3.2. These findings can be also related to the insensitivity and self-correcting behavior of the model relative to variability discussed in Section 3.

We finally caution the reader to the values of the "fraction of time the Node 2 server is idle", p_0, in Table 3. As it is observed in Table 3, p_0 increases as variability increases. This may also seem counter intuitive. However, this could be understood by noting that the H_2 type of variability is due to a very small number of jobs causing most of the delay (as it is case with heavy tailed distributions). Whenever a large job is getting processed, the

Table 1
The model with $M = 4, N = 2, Y = 2$, and $\rho = 0.4$.

Distribution	p_0	p_1	p_2	p_3	p_4	λ_f	W_q	$E[N_o]$
Deterministic	0.268	0.329	0.254	0.129	0.021	0.732	0.784	1.83
Uniform	0.288	0.290	0.237	0.151	0.035	0.712	0.903	1.78
Exponential	0.317	0.253	0.203	0.162	0.065	0.683	1.065	1.708
Hyper-exponential	0.377	0.218	0.136	0.098	0.172	0.623	1.359	1.558

Table 2
The Model with $M = 4, N = 2, Y = 2$, and $\rho = 0.8$.

Distribution	p_0	p_1	p_2	p_3	p_4	λ_f	W_q	$E[N_o]$
Deterministic	0.024	0.093	0.274	0.439	0.170	0.976	1.703	1.221
Uniform	0.048	0.113	0.237	0.393	0.209	0.952	1.733	1.190
Exponential	0.080	0.128	0.204	0.327	0.261	0.920	1.785	1.15
Hyper-exponential	0.136	0.142	0.159	0.207	0.356	0.864	1.899	1.080

Table 3
Effect of Variability on the Model with $M = 4, N = 2, Y = 2, S \sim H_2$.

H_2		$\rho = 0.4$				$\rho = 0.8$			
p	C_S^2	λ_f	p_0	W_q	$E[N_o]$	λ_f	p_0	W_q	$E[N_o]$
0.05	9.256	0.604	0.396	1.488	1.511	0.842	0.158	1.951	1.503
0.02	24.510	0.578	0.422	1.643	1.445	0.825	0.175	1.993	1.032
0.005	99.503	0.565	0.435	1.724	1.413	0.815	0.185	2.019	1.019
0.001	499.501	0.561	0.439	1.749	1.403	0.813	0.187	2.026	1.016

queue length reaches its maximum value of $N + Y - 1$ for a significant amount of time. This, however, happens rarely. At other instants, the Node 2 queue is mostly idle.

Example 4. In this example, we model Node 2 as a repair facility and the N servers at Node 1 as system components subject to failures and Y spare components. We consider the availability of K-out-N redundant system with Y spares, exponential failure times and general repair times. In this system, at least K out of the N components must be working for the system to function. The key measure of performance here is to obtain the system availability given by $\sum_{i=0}^{M-K} p_i$. The one-step transition probabilities for this system are computed in the same way as our model except that here the state space is reduced to $\{0, 1, \ldots, M - K\}$.

For example, consider a system with constant repair time b, five total components/machines, with a demand of three machines. The system functions if two or more components/machines are operating. That is, $N = 3, K = 2$ and $Y = 2$. In this case, the one-step transition probabilities are given by

$$P = \begin{pmatrix} e^{-3\lambda b} & 3\lambda b e^{-3\lambda b} & 9e^{-2\lambda b} - 9e^{-3\lambda b} - 9\lambda b e^{-3\lambda b} & 1 + 6\lambda b e^{-3\lambda b} + 8e^{-3\lambda b} - 9e^{-2\lambda b} \\ e^{-3\lambda b} & 3\lambda b e^{-3\lambda b} & 9e^{-2\lambda b} - 9e^{-3\lambda b} - 9\lambda b e^{-3\lambda b} & 1 + 6\lambda b e^{-3\lambda b} + 8e^{-3\lambda b} - 9e^{-2\lambda b} \\ 0 & e^{-3\lambda b} & 3e^{-2\lambda b}(1 - e^{-\lambda b}) & (1 - e^{-\lambda b})^2 (1 - e^{-\lambda b} + 3e^{-2\lambda b}) \\ 0 & 0 & e^{-2\lambda b} & 1 - e^{-2\lambda b} \end{pmatrix}$$

Then, the system availability can be obtained by solving for the system size departure epochs probabilities, π_i and the system size probabilities, p_i, $i = 0, 1, 2, 3$, as discussed in Section 2.

Example 5. It is generally difficult to analyze group arrivals in closed networks. To the best of our knowledge, the few works in the literature on this utilize involved diffusion approximations (e.g. Jain, 1993). This example illustrates how our model can (exactly) accommodate a special type of group arrivals resulting from a clustered arrangement of customers at Node 1. To simplify the presentation, the discussion is in the context of machine repair. Consider that machines are organized in k clusters having c machines each. Suppose also that the demand requires l clusters to be operational and that up to $q = k - l$ clusters can act as spares. When a machine in a cluster fails, according to an exponential time with rate λ_m, then the whole cluster is sent to repair. Before that repair actually starts, the repairman needs the identify which machine in the cluster has failed. This requires an inspection time I. After inspection, the failed machine is repaired for a time R. The other $c - 1$ operational machines in the cluster wait in the repair shop during the repair time. Upon repair completion, the fixed machine rejoins its cluster which is deployed back to the production floor.

With some thought, if one considers a cluster as an individual customer in this queueing system, it can be seen that this system can be analyzed with the base $M/G/1//N + Y$ model described in this paper with $M = k, N = l$, and $Y = q$, a repair time $S = I + R$, and a failure rate $\lambda = c\lambda_m$. The only additional complexity here is in figuring out the distribution of the convolution $S = I + R$. However, since reasonable distributions for I are either exponential or deterministic, the convolution exercise is not expected to be too hard.[1]

5. Conclusion

In this paper, we develop an imbedded Markov chain solution for two-node closed cyclic network with one exponential multi-server node and one general single server node. A primary example of such a system is the machine repair model with general repair times and spares. Our solution methodology is based on exploiting properties of a birth process representing arrivals to the single-server system. This approach is simple to understand and apply. The wide range of applications of this closed model and the abundance of non-exponential service times in many systems make our model quite useful. The computational effort required by our approach is somewhat modest especially for small systems.

We investigate, analytically and numerically, the effect of service time variability on system performance. We find that because the population size is finite, the effect of variability is limited, in the sense that very high variability is not much worse than high or moderate variability. The question of reducing variability in queueing networks have received some attention recently (see, for example, El-Taha & Maddah, 2006; Harchol-Balter, 2002; Whitt, 1999). However, all works have been limited to open queueing networks. Our findings here are encouraging from a modeling point of view as it implies that closed systems are robust to variability so that rough, back of the envelope, approximations may be utilized with reasonable accuracy. On the other hand, the insensitivity result may discourage variance reduction alternatives. The conclusion is that decisions for applying variability reduction schemes in closed networks, as opposed to open networks, should be considered with caution to avoid costly alternatives.

Finally, we point out that our solution approach based on analyzing birth processes may be applicable to other non-product form queueing networks having one or more nodes with exponential service times. Moreover, the specific birth process studied in this paper, having constant birth rates followed by linearly diminishing rates, could be applicable to other contexts (perhaps biolog-

[1] If a cluster inspection stops when the failed machine is identified, then I is likely to be exponentially distributed. Otherwise, if the inspection requires checking all the machines in a cluster, with, perhaps, some minor maintenance performed, then I might be assumed deterministic.

ical systems). Incidentally, we obtain an interesting combinatorial result (see Theorem 2.3) while analyzing this birth process. This combinatorial identity could also be of use to other researchers.

Appendix A. Closed form for $Q_{i,Y+k}, i \leqslant Y, k = 1, \ldots, N-1$

In order to prove Theorem 2.4, we need the following supporting results.

Fact 1. $\int s^Z e^{as} ds = e^{as}\sum_{r=0}^{Z}(-1)^r \frac{Z!}{(Z-r)!}\frac{s^{Z-r}}{a^{r+1}}$, for all a real and Z integer. See, for example, Gradshteyn and Ryzhik (1994, p. 112).

Proof of Theorem 2.3. The proof is by collapsing the $k-$ *order* sum to the left hand side of the identity into sums of smaller order. First collapse the sum into two sums of order $k-1$ as follows:

$$\begin{aligned}
&\sum_{r_1=0}^{Z}\sum_{r_2=0}^{Z-r_1}\cdots\sum_{r_k=0}^{Z-S_{k-1}}\prod_{l=1}^{k}\left(\frac{k+1}{l}\right)^{r_l}\\
&\quad= \sum_{r_1=0}^{Z}\sum_{r_2=0}^{Z-r_1}\cdots\sum_{r_k=0}^{Z-S_{k-1}}\prod_{l=1}^{k-1}\left(\frac{k+1}{l}\right)^{r_l}\left(\frac{k+1}{k}\right)^{r_k}\\
&\quad= k\sum_{r_1=0}^{Z}\sum_{r_2=0}^{Z-r_1}\cdots\sum_{r_{k-1}=0}^{Z-S_{k-2}}\prod_{l=1}^{k-1}\left(\frac{k+1}{l}\right)^{r_l}\left[\left(\frac{k+1}{k}\right)^{Z-S_{k-1}+1}-1\right]\\
&\quad= k(k+1)^{Z+1}k^{-(Z+1)}\sum_{r_1=0}^{Z}\sum_{r_2=0}^{Z-r_1}\cdots\sum_{r_{k-1}=0}^{Z-S_{k-2}}\prod_{l=1}^{k-1}\left(\frac{k}{l}\right)^{r_l}\\
&\qquad - k\sum_{r_1=0}^{Z}\sum_{r_2=0}^{Z-r_1}\cdots\sum_{r_{k-1}=0}^{Z-S_{k-2}}\prod_{l=1}^{k-1}\left(\frac{k+1}{l}\right)^{r_l}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sum_{r_1=0}^{Z}\sum_{r_2=0}^{Z-r_1}\cdots\sum_{r_k=0}^{Z-S_{k-1}}\prod_{l=1}^{k}\left(\frac{k+1}{l}\right)^{r_l}\Big/(k+1)^{Z+1} = \binom{k}{1}\sum_{r=0}^{1}(-1)^{r+1}\binom{1}{r}\\
&\quad\times(k+1-r)^{-(Z+1)}\sum_{r_1=0}^{Z}\sum_{r_2=0}^{Z-r_1}\cdots\sum_{r_{k-1}=0}^{Z-S_{k-2}}\prod_{l=1}^{k-1}\left(\frac{k+1-r}{l}\right)^{r_l}.
\end{aligned}$$

By induction assume that

$$\begin{aligned}
&\sum_{r_1=0}^{Z}\sum_{r_2=0}^{Z-r_1}\cdots\sum_{r_k=0}^{Z-S_{k-1}}\prod_{l=1}^{k}\left(\frac{k+1}{l}\right)^{r_l}\Big/(k+1)^{Z+1} = \binom{k}{v}\sum_{r=0}^{v}(-1)^{r+v}\binom{v}{r}\\
&\quad\times(k+1-r)^{-(Z+1)}\sum_{r_1=0}^{Z}\sum_{r_2=0}^{Z-r_1}\cdots\sum_{r_{k-v}=0}^{Z-S_{k-v-1}}\prod_{l=1}^{k-v}\left(\frac{k+1-r}{l}\right)^{r_l},
\end{aligned}$$

for $v = 1, \ldots, w < k$. In the following, we prove that this identity holds for $v = w+1$

$$\begin{aligned}
&\binom{k}{w}\sum_{r=0}^{w}(-1)^{r+w}\binom{w}{r}(k+1-r)^{-(Z+1)}\sum_{r_1=0}^{Z}\cdots\sum_{r_{k-w}=0}^{Z-S_{k-w-1}}\prod_{l=1}^{k-w}\left(\frac{k+1-r}{l}\right)^{r_l}\\
&= \binom{k}{w}\sum_{r=0}^{w}\Bigg\{(-1)^{r+w}\binom{w}{r}(k+1-r)^{-(Z+1)}\\
&\quad\times\sum_{r_1=0}^{Z}\cdots\sum_{r_{k-w-1}=0}^{Z-S_{k-w-2}}\prod_{l=1}^{k-w-1}\left(\frac{k+1-r}{l}\right)^{r_l}(k-w)\frac{\left(\frac{k+1-r}{k-w}\right)^{Z-S_{k-w-1}+1}-1}{w-r+1}\Bigg\}\\
&= \binom{k}{w+1}\Bigg\{\left[\sum_{r=0}^{w}(-1)^{r+w}\binom{w+1}{r}\right](k-w)^{-(Z+1)}\sum_{r_1=0}^{Z}\cdots\sum_{r_{k-w-1}=0}^{Z-S_{k-w-2}}\prod_{l=1}^{k-w-1}\\
&\quad\times\left(\frac{k-w}{l}\right)^{r_l}+\sum_{r=0}^{w}(-1)^{r+w+1}\binom{w+1}{r}(k+1-r)^{-(Z+1)}\\
&\quad\times\sum_{r_1=0}^{Z}\cdots\sum_{r_{k-w-1}=0}^{Z-S_{k-w-2}}\prod_{l=1}^{k-w-1}\left(\frac{k+1-r}{l}\right)^{r_l}\Bigg\},
\end{aligned}$$

where the term in brackets above is equal to 1 based on the identity

$$\sum_{r=0}^{w}(-1)^{r}\binom{w+1}{r} = (-1)^{w}\binom{w}{w},$$

see, for example, Feller (1968, p. 64). Upon simplifying and rearranging, it follows that

$$\begin{aligned}
&\sum_{r_1=0}^{Z}\sum_{r_2=0}^{Z-r_1}\cdots\sum_{r_k=0}^{Z-S_{k-1}}\prod_{l=1}^{k}\left(\frac{k+1}{l}\right)^{r_l}\Big/(k+1)^{Z+1} = \binom{k}{w+1}\sum_{r=0}^{w+1}(-1)^{r+w}\\
&\quad\times\binom{w+1}{r}(k+1-r)^{-(Z+1)}\sum_{r_1=0}^{Z}\sum_{r_2=0}^{Z-r_1}\cdots\sum_{r_{k-w-1}=0}^{Z-S_{k-w-2}}\prod_{l=1}^{k-w-1}\left(\frac{k+1-r}{l}\right)^{r_l},
\end{aligned}$$

which completes the inductive proof. For $v = k-1$, the inductive proof implies

$$\begin{aligned}
&\sum_{r_1=0}^{Z}\sum_{r_2=0}^{Z-r_1}\cdots\sum_{r_k=0}^{Z-S_{k-1}}\prod_{l=1}^{k}\left(\frac{k+1}{l}\right)^{r_l}\Big/(k+1)^{Z+1}\\
&= \binom{k}{k-1}\sum_{r=0}^{k-1}(-1)^{r+k-1}\binom{k-1}{r}(k+1-r)^{-(Z+1)}\sum_{r_1=0}^{Z}(k+1-r)^{r_1}\\
&= k\sum_{r=0}^{k-1}(-1)^{r+k-1}\binom{k-1}{r}(k+1-r)^{-(Z+1)}\frac{(k+1-r)^{Z+1}-1}{k-r}\\
&= \sum_{r=0}^{k-1}(-1)^{r+k-1}\binom{k}{r}+\sum_{r=0}^{k-1}(-1)^{r+k}\binom{k}{r}(k+1-r)^{-(Z+1)}\\
&= 1+\sum_{r=0}^{k-1}(-1)^{r+k}\binom{k}{r}(k+1-r)^{-(Z+1)},
\end{aligned}$$

where the last equality also follows from the identity in Feller. Upon rearrangement,

$$\begin{aligned}
\sum_{r_1=0}^{Z}\sum_{r_2=0}^{Z-r_1}\cdots\sum_{r_k=0}^{Z-S_{k-1}}\prod_{l=1}^{k}\left(\frac{k+1}{l}\right)^{r_l} &= (k+1)^{Z+1}\sum_{r=0}^{k}(-1)^{r+k}\\
&\quad\times\binom{k}{r}(k+1-r)^{-(Z+1)},
\end{aligned}$$

which completes the proof. □

The following lemma is another supporting result we need.

Lemma 5.1. *Let* $\tilde{Q}_{iYj}(t) \triangleq \lambda_j e^{-\lambda_{j+1}t}\int_0^t e^{\lambda_{j+1}s}Q_{iY}(s)ds$, $i = 1, \ldots, Y-1, j = Y+2, \ldots M-1$. *Then,*

$$\begin{aligned}
\tilde{Q}_{iYj}(t) = (M-j)\Bigg[&\frac{N^{Y-i}}{(j-Y+1)^{Y-i+1}}e^{-(M-j-1)\lambda t}\\
&-\sum_{r=0}^{Y-i}\frac{N^{r}}{(j-Y+1)^{r+1}}Q_{r+i,Y}(t)\Bigg].
\end{aligned}$$

Proof. Utilizing Fact 1,

$$\begin{aligned}
\tilde{Q}_{iYj}(t) &= (M-j)\lambda e^{-(M-j-1)\lambda t}\int_0^t e^{(M-j-1)\lambda s}e^{-N\lambda s}\frac{(N\lambda s)^{Y-i}}{(Y-i)!}ds\\
&= \frac{(M-j)\lambda(N\lambda)^{Y-i}}{(Y-i)!}e^{-(M-j-1)\lambda t}\int_0^t e^{(Y-j-1)\lambda s}s^{Y-i}ds\\
&= \frac{(M-j)\lambda(N\lambda)^{Y-i}}{(Y-i)!}e^{-(M-j-1)\lambda t}\\
&\quad\times\left[e^{(Y-j-1)\lambda s}\sum_{r=0}^{Y-i}(-1)^{r}\frac{(Y-i)!}{(Y-i-r)!}\frac{s^{Y-i-r}}{(\lambda(Y-j-1))^{r+1}}\right]_0^t\\
&= \frac{(M-j)\lambda(N\lambda)^{Y-i}}{(Y-i)!}e^{-(M-j-1)\lambda t}
\end{aligned}$$

$$\times\left[e^{(Y-j-1)\lambda s}\sum_{r=0}^{Y-i}(-1)^r\frac{(Y-i)!}{(Y-i-r)!}\frac{s^{Y-i-r}}{(-1)^{r+1}(\lambda(j-Y+1))^{r+1}}\right]_0^t$$

$$=(M-j)\left[\frac{\lambda(N\lambda)^{Y-i}}{\lambda^{Y-i+1}(-Y+j+1)^{Y-i+1}}e^{-(M-j-1)\lambda t}-\sum_{r=0}^{Y-i}\frac{(N\lambda t)^{Y-i-r}}{(Y-i-r)!}e^{-N\lambda t}\frac{(N\lambda)^r\lambda}{\lambda^{r+1}(j-Y+1)^{r+1}}\right]$$

$$=(M-j)\left[\frac{N^{Y-i}}{(j-Y+1)^{Y-i+1}}e^{-(M-j-1)\lambda t}-\sum_{r=0}^{Y-i}\frac{N^r}{(j-Y+1)^{r+1}}Q_{i+r,Y}\right],$$

which completes the proof. □

Proof of Theorem 2.4. For $k=1$, Fact 1 implies

$$Q_{i,Y+1}(t)=\lambda_Y e^{-\lambda_{Y+1}t}\int_0^t e^{\lambda_{Y+1}s}Q_{iY}(s)ds$$

$$=N\lambda e^{-(N-1)\lambda t}\int_0^t e^{(N-1)\lambda s}e^{-N\lambda s}\frac{(N\lambda s)^{Y-i}}{(Y-i)!}ds$$

$$=\frac{(N\lambda)^{Y-i+1}}{(Y-i)!}e^{-(N-1)\lambda t}\int_0^t s^{Y-i}e^{-\lambda s}ds$$

$$=N^{Y-i+1}e^{-(N-1)\lambda t}-\sum_{r=0}^{Y-i}\frac{(N\lambda)^{r+1}}{\lambda^{r+1}}\left(\frac{(N\lambda t)^{Y-i-r}}{(Y-i-r)!}e^{-N\lambda t}\right)$$

$$=N^{Y-i+1}\left[e^{-(N-1)\lambda t}-\sum_{r=0}^{Y-i}N^{i+r-Y}Q_{i+r,Y}(t)\right].$$

Therefore, the theorem holds for $k=1$. By induction assume the theorem holds for $Q_{i,Y+k}(t)$. In the following we prove that it also holds for $Q_{i,Y+k+1}(t)$.

$$Q_{i,Y+k+1}(t)=\lambda_{Y+k}e^{-\lambda_{Y+k+1}t}\int_0^t e^{\lambda_{Y+k+1}s}$$

$$Q_{i,Y+k}(s)ds=\left(\prod_{l=1}^{k}(N-l+1)\right)N^{Y-i}$$

$$\times\left[\frac{(N-k)\lambda}{(k-1)!}e^{-(N-k-1)\lambda t}\sum_{r=0}^{k-1}(-1)^r\binom{k-1}{r}(k-r)^{-(Y-i+1)}\right.$$

$$\times\int_0^t e^{-(r+1)\lambda s}ds+(-1)^k\sum_{r_1=0}^{Y-i}\sum_{r_2=0}^{Y-i-r_1}\cdots\sum_{r_k=0}^{Y-i-S_{k-1}}\frac{N^{i+S_k-Y}}{\prod_{l=1}^{k}l^{r_l+1}}$$

$$\left.\times\tilde{Q}_{i+S_k,Y,Y+k}(t)\right].$$

Note that

$$\frac{(N-k)\lambda}{(k-1)!}e^{-(N-k-1)\lambda t}\sum_{r=0}^{k-1}(-1)^r\binom{k-1}{r}(k-r)^{-(Y-i+1)}\int_0^t e^{-(r+1)\lambda s}ds$$

$$=(N-k)e^{-(N-k-1)\lambda t}\sum_{r=0}^{k-1}(-1)^r\frac{(k-r)^{-(Y-i+1)}}{(r+1)!(k-1-r)!}\left(1-e^{-(r+1)\lambda t}\right)$$

$$=(N-k)\left(e^{-(N-k-1)\lambda t}\sum_{r=0}^{k-1}(-1)^r\frac{(k-r)^{-(Y-i+1)}}{(r+1)!(k-1-r)!}-\sum_{r=0}^{k-1}(-1)^r\frac{(k-r)^{-(Y-i+1)}}{(r+1)!(k-1-r)!}e^{-(N-k+r)\lambda t}\right)$$

$$=(N-k)\left(e^{-(N-k-1)\lambda t}\sum_{r=0}^{k-1}(-1)^r\frac{(k-r)^{-(Y-i+1)}}{(r+1)!(k-1-r)!}+\sum_{r=1}^{k}(-1)^r\frac{(k-r+1)^{-(Y-i+1)}}{r!(k-r)!}e^{-(N-k+r-1)\lambda t}\right).$$

Moreover, Lemma 5.1 implies that

$$(-1)^k\sum_{r_1=0}^{Y-i}\sum_{r_2=0}^{Y-i-r_1}\cdots\sum_{r_k=0}^{Y-i-S_{k-1}}\frac{N^{i+S_k-Y}}{\prod_{l=1}^{k}l^{r_l+1}}\tilde{Q}_{i+S_k,Y,Y+k}(t)$$

$$=e^{-(N-k-1)\lambda t}(-1)^k(N-k)\sum_{r_1=0}^{Y-i}\sum_{r_2=0}^{Y-i-r_1}\cdots\sum_{r_k=0}^{Y-i-S_{k-1}}\frac{1}{\left(\prod_{l=1}^{k}l^{r_l+1}\right)(k+1)^{Y-i-S_{k-1}+1}}$$

$$-(-1)^k(N-k)\sum_{r_1=0}^{Y-i}\sum_{r_2=0}^{Y-i-r_1}\cdots\sum_{r_k=0}^{Y-i-S_{k-1}}\frac{N^{i+S_k-Y}}{\prod_{l=1}^{k}l^{r_l+1}}$$

$$\times\sum_{r=0}^{Y-i-r_1-\cdots-r_k}\frac{N^r}{(k+1)^{r+1}}Q_{r+i+S_k,Y}(t)$$

$$=\frac{e^{-(N-k-1)\lambda t}(-1)^k(N-k)}{k!(k+1)^{Y-i+1}}\sum_{r_1=0}^{Y-i}\sum_{r_2=0}^{Y-i-r_1}\cdots\sum_{r_k=0}^{Y-i-S_{k-1}}\prod_{l=1}^{k}\left(\frac{k+1}{l}\right)^{r_l}$$

$$+(-1)^{k+1}(N-k)\sum_{r_1=0}^{Y-i}\sum_{r_2=0}^{Y-i-r_1}\cdots\sum_{r_{k+1}=0}^{Y-i-S_k}\frac{N^{i+S_{k+1}-Y}}{\prod_{l=1}^{k+1}l^{r_l+1}}Q_{i+S_{k+1},Y}(t).$$

Therefore,

$$\frac{Q_{i,Y+k+1}(t)}{\left(\prod_{l=1}^{k+1}(N-l+1)\right)N^{Y-i}}$$

$$=e^{-(N-k-1)\lambda t}\left\{\sum_{r=0}^{k-1}(-1)^r\frac{(k-r)^{-(Y-i+1)}}{(r+1)!(k-1-r)!}+\frac{(-1)^k}{k!(k+1)^{(Y-i+1)}}\sum_{r_1=0}^{Y-i}\sum_{r_2=0}^{Y-i-r_1}\cdots\sum_{r_k=0}^{Y-i-S_{k-1}}\prod_{l=1}^{k}\left(\frac{k+1}{l}\right)^{r_l}\right\}$$

$$+\sum_{r=1}^{k}(-1)^r\frac{(k-r+1)^{-(Y-i+1)}}{r!(k-r)!}e^{-(N-k+r-1)\lambda t}$$

$$+(-1)^{k+1}\sum_{r_1=0}^{Y-i}\sum_{r_2=0}^{Y-i-r_1}\cdots\sum_{r_{k+1}=0}^{Y-i-S_k}\frac{N^{i+S_{k+1}-Y}}{\prod_{l=1}^{k+1}l^{r_l+1}}Q_{i+S_{k+1},Y}(t).$$

Theorem 2.3 implies that the term in curly braces is equal to

$$\sum_{r=0}^{k-1}(-1)^r\frac{(k-r)^{-(Y-i+1)}}{(r+1)!(k-1-r)!}+\sum_{r=0}^{k}(-1)^r\frac{(k+1-r)^{-(Y-i+1)}}{r!(k-r)!}$$

$$=\sum_{r=0}^{k-1}(-1)^r\frac{(k-r)^{-(Y-i+1)}}{(r+1)!(k-1-r)!}+\sum_{r=1}^{k}(-1)^r\frac{(k+1-r)^{-(Y-i+1)}}{r!(k-r)!}+\frac{(k+1)^{-(Y-i+1)}}{k!}=\sum_{r=0}^{k-1}(-1)^r\frac{(k-r)^{-(Y-i+1)}}{(r+1)!(k-1-r)!}$$

$$+\sum_{r=0}^{k-1}(-1)^{r+1}\frac{(k-r)^{-(Y-i+1)}}{(r+1)!(k-r-1)!}+\frac{(k+1)^{-(Y-i+1)}}{k!}$$

$$=\frac{(k+1)^{-(Y-i+1)}}{k!}.$$

This completes the proof. □

References

Allen, A. O. (1978). *Probability, statistics, and queueing theory with computer science applications*. New York: Academic Press.

Crovella, M. E., & Bestavros, A. (1997). Self-similarity in World Wide Web traffic: Evidence and possible causes. *IEEE/ACM Transactions on Networking, 5*, 835–846.

El-Taha, M., & Maddah, B. (2006). Allocation of service time in a multi-server system. *Management Science, 52*, 623–637.

El-Taha, M., & Stidham, S. Jr., (1999). *Sample-path analysis of queueing systems*. Boston: Kluwer Academic Publishing.

Feller, W. (1968). *An introduction to probability theory and its applications* (3rd ed.). NY: John Wiley and Sons.

Fowler, T.B. (1997). A short tutorial on fractiles and internet traffic, The Telecommunications Review, 10: 1–14, Mitretek Systems, McLean, VA.

Gaver, P., & Shedler, G. S. (1973). Processor utilization in multiprogramming systems via diffusion approximations. *Operations Research, 21*, 569–576.

Gradshteyn, I. S., & Ryzhik, I. M. (1994). *Table of integrals, series, and products* (5th ed.). Boston: Academic Press.

Gupta, U. C., & Rao, T. S. S. (1996). On the M/G/1 machine interference model with spares. *European Journal of Operations Research, 89*, 164–171.

Haque, L., & Armstrong, M. J. (2007). A survey of the machine interference problem. *European Journal of Operations Research, 179*, 469–482.

Harchol-Balter, M. (2002). Task assignment with unknown duration. *Journal of the ACM, 49*, 260–288.

Harchol-Balter, M., & Downey, A. B. (1997). Exploiting process lifetime distributions for dynamic load balancing. *ACM Transactions on Computer Systems, 15*, 253–285.

Jain, M. (1993). Diffusion equation for $GI^X/G/r$ machine interference problem with spare machines. *Microelectronics Reliability, 33*, 1415–1418.

Jain, M., Sharma, G. C., & Sharma, R. (2008). Performance modeling of state dependent system with mixed standbys and two modes of failure. *Applied Mathematical Modelling, 32*, 712–724.

Koenigsberg, E. (1982). Twenty five years of cyclic queues and closed queue networks: A review. *Journal of the Operational Research Society, 33*, 605–619.

Krishnamoorthy, A., & Ushakumari, P. V. (2001). k-out-of-n: G system with repair: the D-policy. *Computers & Operations Research, 28*, 973–981.

Lam, Y., & Zhang, Y. L. (2000). Repairable consecutive-k-out-of-n: F system with markov dependence. *Naval Research Logistics, 47*, 18–39.

Lewis, P. A. W., & Shedler, G. S. (1971). A cyclic-queue model of system overhead in multiprogrammed computer systems. *Journal of ACM, 18*, 199–220.

Naldi, M. (1999). Measurement-based modeling of internet dial-up access connections. *Computer Networks, 31*, 2381–2390.

Rao, T. S. S., & Gupta, U. C. (2000). Performance modeling of the M/G/1 machine repair problem with cold-, warm-, and hot-standbys. *Computers & Industrial Engineering, 38*, 251–267.

Ross, S. M. (1997). *Introduction to probability models* (6th ed.). New York: Academic Press.

Saaty, T. (1961). *Elements of queueing theory with applications*. New York: Dover Publications.

Scheller-Wolf, A., & Sigman, K. (1997). Delay moments for GI/G/s queues. *Queueing Systems: Theory and Applications, 25*, 77–95.

Shum, A. (1980). *Queueing models for computer systems with general service time distributions*. New York: Garland Publishing Company.

Sivazlian, B. D., & Wang, K. H. (1989). System characteristics and economic analysis of the $G/G/R$ machine repair problem with warm standbys using diffusion approximation. *Microelectronics Reliability, 29*, 829–848.

Sivazlian, B. D., & Wang, K. H. (1990). Diffusion approximation to the $G/G/R$ machine repair problem with warm standby spares. *Naval Research Logistics, 37*, 753–772.

Stecke, K. (1982). Machine interference: The assignment of machines to operators in handbook of industrial engineering. In G. Salvendy (Ed.). New York: John Wiley and Sons.

Takagi, H. (1993). *Queueing analysis, a foundation for performance evaluation. Volume 2: Finite systems*. Amsterdam: North Holland.

van Hoorn, M. (1984). *Algorithms and approximations for queueing systems*. Amsterdam: Center for Mathematics and Computer Science.

Wang, K. H. (1994). Comparative analysis for the $M/E_k/1$ machine repair problem with spares. *Computers & Industrial Engineering, 26*, 765–774.

Wang, K. H., Su, J. H., & Yang, D.-Y. (2014). Analysis and optimization of an $M/G/1$ machine repair problem with multiple imperfect coverage. *Applied Mathematics and Computation, 242*, 590–600. 242.

Wang, K. H., & Wu, J. D. (1995). Cost analysis of the $M/M/R$ machine repair problem with spares and two modes of failure. *Journal of the Operational Research Society, 46*, 783–790.

Whitt, W. (1999). Partitioning customers into service groups. *Management Science, 45*, 1579–1592.