Proceedings of the 8th International Conference on Applied Informatics Eger, Hungary, January 27–30, 2010. Vol. 2. pp. 275–285.

Finite-source Retrial Queues with Applications^{*}

Patrick Wüchner^a, János Sztrik^b, Hermann de Meer^a

^aFaculty of Informatics and Mathematics, University of Passau, Germany e-mail: {patrick.wuechner,hermann.demeer}@uni-passau.de

> ^bFaculty of Informatics, University of Debrecen, Hungary e-mail: jsztrik@inf.unideb.hu

Abstract

We consider a single-server retrial queueing system with a finite number of homogeneous sources generating requests after a random time. The service of an arriving customer finding the server idle starts immediately; otherwise the request enters into an orbit from where the calls are repeated after a random time. Upon completion of a service, with a certain probability the server searches for an orbiting customer. We assume the search time to be negligible.

While modeling generation, retrial, and service times by their first two moments, we are interested in steady-state performance measures. The novelty of the investigation is the discussion of the second moment's effect on the performance of finite-source retrial queues with orbital search.

The MOSEL-2 tool is used to formulate and solve the problem resulting in various illustrative numerical examples showing the effect of the the moments and the search probability on the mean response time of the requests.

Keywords: retrial queueing systems, finite number of sources, orbital search, performance tool, performance measures, hypoexponential distribution, generalized exponential distribution

MSC: 60K25, 68M20

^{*}This research is partially supported by the German-Hungarian Intergovernmental Scientific Cooperation, HAS-DFG, 436 UNG 113/197/0-1, by the Hungarian Scientific Research Fund, OTKA K60698/2006, by the SOCIONICAL project (IP, FP7 Call 3, ICT-2007-3-231288), by the ResumeNet project (IP, FP7 Call 3, ICT-2007-3-231288), and by the Network of Excellence and EuroNF (IST, FP7, ICT-2007-1-216366). The work is supported by the TAMOP 4.2.1./B-09/1/KONV-2010-0007 project. The project is implemented through the New Hungary Development Plan, co-financed by the European Social Fund and the European Regional Development Fund.

1. Introduction

Frequently when modeling practical applications, random system behavior cannot be described exactly enough by assuming exponentially distributed random variables. Then, in addition to the random variables' means, at least their second moments, often represented by the variance or squared coefficient of variation, should also be taken into account. The mean and the variance can easily be calculated from measurement samples.

In this paper, we discuss the influence of the variance of the random generation, retrial, and service times on the performance of finite-source retrial queues with orbital search. Hence, the present paper is a natural generalization of the authors' work [13] that quantitatively investigates the server's search for orbiting customers in finite-source retrial queues within an exponentially distributed setting.

While there is a considerable amount of work on *infinite-source* retrial queues with non-exponentially distributed generation, service, and/or retrial times (see [3] and references therein), *finite-source* retrial queues are discussed relatively rarely in the literature in this regard (see, e.g., [8, 2]). Additionally, we are not aware of any previous results on non-exponentially distributed generation, retrial, and/or service times in the context of finite-source retrial queues with orbital search.

In this paper, to preserve mathematical tractability of the model, we follow the method of phases (see Sec. 2.2) to approximate general distributions. This allows us to apply the Markovian methodology (see, e.g., [6, Ch. 3]) for model analysis. Moreover, we restrict ourself to the single-server case.

The remainder of this paper is organized as follows. In Sec. 2, we present a detailed model description, the phase-type approximation leading to the underlying Markov chain, and the performance measures of interest. In Sec. 3, the model is validated against related work and the influence of the generation, retrial, and service times' second moments on the mean response time of the finite-source retrial queue with orbital search is discussed. Finally, we summarize the paper and give directions to future work in Sec. 4.

2. Investigated Retrial Queueing Model

We first define more formally the model under investigation together with its parameters and performance measures of interest.

2.1. Model Description

The retrial queue with orbital search is illustrated in Fig. 1. The model parameters are summarized in Tab. 1. Each of the K sources generates a maximum of 1 job at a time after a generally distributed random generation time with mean value \bar{t}_{λ} and squared coefficient of variation (SCV) c_{λ}^2 . We call $\lambda = 1/\bar{t}_{\lambda}$ the generation rate. A job leaving the sources enters the server unless the latter is already busy.



Figure 1: Visualization of retrial queue with orbital search.

Unlike classical first-come first-served (FCFS) queueing systems, an arriving job finding the server busy does not enter a queue but enters the orbit instead. Since jobs can only enter the orbit if a job is located at the server, the maximum number of jobs at the orbit is K - 1. Orbiting jobs are re-trying to get access to the server after a generally distributed random retrial time with mean value \bar{t}_{ν} and SCV c_{ν}^2 . We call $\nu = 1/\bar{t}_{\nu}$ the retrial rate.

The server processes jobs with a generally distributed random service time with mean value \overline{t}_{μ} and SCV c_{μ}^2 . We call $\mu = 1/\overline{t}_{\mu}$ the service rate. After service, the processed job returns to the sources, starting a new job generation process. At the same time, with a probability of p, the server searches in the orbit for a job to be processed next immediately (depicted by the *link* symbol in Fig. 1). We call p the orbital search probability. With probability 1 - p the server does not conduct orbital search.

Depertor	Symbol	Bango	Default Values				
ranneter	Symbol	nange	(a)	(b)	(c)		
Number of sources (jobs)	K	\mathbb{N}	3	4	3		
Mean generation time	$\overline{t}_{\lambda} = 1/\lambda$	\mathbb{R}^+	$1 \dots 10^4$	4/3	8		
SCV of generation time	c_{λ}^2	\mathbb{R}^+	1	0.5, 1, 5	1		
Mean service time	$\overline{t}_{\mu} = 1/\mu$	\mathbb{R}^+	1	1	1		
SCV of service time	c_{μ}^2	\mathbb{R}^+	1	0.5, 1, 5	1		
Mean retrial time	$\overline{t}_{\nu} = 1/\nu$	\mathbb{R}^+	$200, 10^3$	≈ 0	4		
SCV of retrial time	c_{ν}^2	\mathbb{R}^+	1	1	1		
Search probability	p	(0, 1)	$10^{-8} \dots 0.99$	≈ 1	0.01, 0.5, 0.99		

Table 1: Main model parameters with default values for validation and evaluation scenarios: (a) Section 3.1.1, (b) Section 3.1.2, (c) Section 3.2.

2.2. Phase-Type Approximation

Remember that the generation, service, and retrial time distributions of the model described in Sec. 2.1 are given by (at least) their mean \overline{t} and SCV c^2 . To map this model to a stochastic model that is mathematically more tractable, we approximate these distributions by phase-type constructs depending on the SCV c^2 of the respective distribution as follows.

- $c^2 = 1$: In this case, the general distribution is replaced by an exponential distribution with rate parameter $\mu' = \frac{1}{t}$ which comprises a single phase only.
- $c^2 > 1$: In this case, the generalized exponential (GE) distribution (see, e.g, [9, 7]) is chosen. The GE distribution has several advantages (when compared to the hyperexponential or Cox distributions): (1) it involves the least assumptions and arbitrary information in terms of higher moments (see [10]), (2) it has been shown that the GE distribution less likely underestimates mean queue lengths (see [10]), and (3), most important, its application does not increase the state space of the underlying stochastic model since there is only a single phase. The parameters of the GE distributions are given by $\mu' = \frac{p'}{t}$ and $p' = \frac{2}{1+c^2}$ (see [7]).
- $c^2 < 1$: In this case, the general distribution is replaced by a hypoexponential distribution. The implementable recursive algorithm provided in [1] can be used here to determine the rates of each of the $\lceil \frac{1}{c^2} \rceil$ phases needed for the approximation.

In the following, we refer to the resulting number of exponentially distributed phases of the approximated generation, service, and retrial time distributions with r_{λ} , r_{μ} , and r_{ν} , respectively. Due to space constraints, we refrain from presenting more details on the phase-type approximation here. Moreover, the MOSEL-2 tool takes care of generating the suitable phase-type constructs automatically.

2.3. Underlying Markov Chain

The system state of the presented retrial queue at time t (including the phase-type approximations) can be mapped to a continuous-time Markov chain (CTMC, see [6]) $\tilde{X}(t) = (\tilde{C}(t), \tilde{N}(t), \tilde{\Lambda}(t))$, where $\tilde{C}(t), \tilde{N}(t)$, and $\tilde{\Lambda}(t)$ describe the state of the phase-type service, retrial, and generation process at time t, respectively:

- $\tilde{C}(t) = (\tilde{C}_1(t), \tilde{C}_2(t), \dots, \tilde{C}_{r_{\mu}}(t))$, where $\tilde{C}_j(t) = 0, 1$ $(1 \le j \le r_{\mu})$ is the state of the *j*-th phase of the service process at time *t*.
- $\tilde{N}(t) = (\tilde{N}_{1,1}(t), \tilde{N}_{1,2}(t), \dots, \tilde{N}_{K-1,r_{\nu}}(t))$, where $\tilde{N}_{i,j}(t) = 0, 1$ $(1 \leq i \leq K-1, 1 \leq j \leq r_{\nu})$ is the state of the *j*-th phase of the *i*-th job following the retrial process at time *t*.

• $\tilde{\Lambda}(t) = (\tilde{\Lambda}_{1,1}(t), \tilde{\Lambda}_{1,2}(t), \dots, \tilde{\Lambda}_{K-1,r_{\lambda}}(t)), \text{ where } \tilde{\Lambda}_{i,i}(t) = 0, 1 \ (1 \le i \le K, 1 \le K))$ $i < r_{\lambda}$ is the state of the *i*-th phase of the *i*-th job following the generation process at time t.

We refer the interested reader to [13, Fig. 3] for a graphical representation of the CTMC for K = 4 and $c_{\lambda}^2 = c_{\mu}^2 = c_{\nu}^2 = 1$.

The state space $\mathbf{S}_{\tilde{X}}$ of the CTMC $\tilde{X}(t)$ is finite and irreducible, i.e., all states are pairwise reachable from each other within finite time. Therefore, the CTMC is ergodic for all values of the mean generation time \bar{t}_{λ} . From now on, we assume the system to be in the steady state, i.e., $t \to \infty$. Then, the stationary state probabilities of the CTMC are given by:

$$P(\tilde{c}, \tilde{n}, \tilde{l}) = \lim_{t \to \infty} P(\tilde{C}(t) = \tilde{c}, \, \tilde{N}(t) = \tilde{n}, \, \tilde{\Lambda}(t) = \tilde{l}) \,, \tag{2.1}$$

where \tilde{c} , \tilde{n} , and \tilde{l} are the state vectors describing the distribution of jobs across the phases of the service, retrial, and generation process, respectively.

The stationary state probabilities of the CTMC can be derived, e.g., by applying standard Markovian analysis (see [6, Chap. 3]). In the following, we assume that the stationary state probabilities are known.

Let us now define by X(t) = (C(t), N(t)) the stochastic process of the workload distribution at time t, i.e., the distribution of jobs across the sources, server, and orbit, with the number of jobs at the server C(t), the number of jobs in the orbit N(t). Note that the number of request-generating sources $\Lambda(t)$ results from $\Lambda(t) =$ K - (C(t) + N(t)).

The stationary state probabilities of X(t) can then be derived from the stationary state probabilities of the Markov chain X(t) (recall Eq. 2.1) and are given by

$$P(c,n) = \lim_{t \to \infty} P(C(t) = c, N(t) = n) = \sum_{\substack{(\tilde{c}, \tilde{n}, \tilde{l}) \in \mathbf{S}_{\tilde{X}} \\ \text{where } \sum_{h=1}^{r_{\mu}} \tilde{c}_{h} = c \\ \text{and } \sum_{i=1}^{K-1} \sum_{i=1}^{r_{\nu}} \tilde{n}_{ij} = n} P(\tilde{c}, \tilde{n}, \tilde{l})$$
(2.2)

with c = 0, 1 (server state) and $n = 0, \ldots, K - 1$ (number of orbiting jobs).

2.4. Main Performance Measures

Assuming the stationary state probabilities of X(t) (Eq. (2.2)) are known, we are now able to derive the most important performance measures as follows:

- Utilization of the server:
- $\rho_{\mu} = \sum_{n=0}^{K-1} P(1,n).$ $\overline{N} = \sum_{c=0}^{1} \sum_{n=0}^{K-1} nP(c,n).$ • Mean number of jobs in the orbit:

3. Numerical Results

For the model evaluation, we use the MOSEL-2 (cf. [12]) tool. The tool allows to conveniently specify the model described in Sec. 2.1 by the help of a textual modeling language. Due to space constraints, we refrain from reproducing the MOSEL-2 models in this paper. The models can be obtained from the first author on request. Alternatively, a related model discussing the purely exponential case $(c_{\lambda}^2 = c_{\mu}^2 = c_{\nu}^2 = 1)$ is presented in [13].

After specifying the model, the MOSEL-2 evaluation environment takes care of approximating the general distributions by phase-type constructs based on the given mean value and variance (according to Sec. 2.2) as well as generating and solving the underlying CTMC. The results can be accessed in form of steady-state probabilities of the stochastic process X(t) or directly, i.e., more conveniently, in form of, for example, node utilizations $(\rho_{\lambda}, \rho_{\nu}, \rho_{\mu})$ and mean node workloads $(\overline{\Lambda}, \overline{N})$. Together with the equations given in Sec. 2.4, all main performance measures of interest can then easily be obtained.

3.1. Model Validation

3.1.1. Validation for $c_{\lambda}^2 = c_{\mu}^2 = c_{\nu}^2 = 1$ with Retrials

First, we validate our model using results available in [13, Fig. 4]. Since [13] considers finite-source retrial queues with exponentially distributed generation, service, and retrial times only, we set $c_{\lambda}^2 = c_{\mu}^2 = c_{\nu}^2 = 1$ in our current model. All other model parameters are adopted from [13, Tab. 3] and replicated in column (a) of Tab. 1 for the reader's convenience.

The resulting trend of the mean response time \overline{T} (y-axis) depending on the generation rate λ (x-axis) is shown in Fig. 2 for different values of the retrial rate ν and orbital search probability p. Fig. 2 exactly matches [13, Fig. 4]. For more detailed discussions of the exponential case resulting in the maximum mean response time illustrated in Fig. 2, we refer the interested reader to [13] and [14].



response time for parameters given in column (a) of Tab. 1 and $c_{\lambda}^2 = c_{\mu}^2 = c_{\nu}^2 = 1$.

Figure 3: Influence of generation time's SCV c_{λ}^2 on the mean response time \overline{T} for different search probabilities p.

3.1.2. Validation for $c_{\lambda}^2 \neq 1$ and/or $c_{\mu}^2 \neq 1$ without Retrials

In this section, we compare the proposed model with G/G/1/K - FCFS closed tandem queueing network models without retrials. For this, we set $p \approx 1$ and $\nu \approx \infty$ to achieve a non-retrial, FCFS-equivalent service discipline. Additionally, the source model is adapted to account for state-independent arrival rates.

Results for the node utilizations in a G/G/1/K-FCFS closed tandem queueing network model are derived by Kouvatsos in [9, Tab. 3] by using the maximum entropy method (MEM; see [9] for details). We produced additional results for an equivalent tandem network using WinPEPSY (see [5], [4], or [6, Sec. 12.1.5]) by applying discrete-event simulation (cf. [6, Chap. 11]) and Marie's method (cf. [6, Sec. 10.1.4.2]).

For comparison, we adopt the model parameters used in [9] as summarized in column (b) of Tab. 1. The variable values of c_{λ}^2 and c_{μ}^2 are given in Tab. ?? together with the obtained results. It can be seen that our results are very close to the ones obtained by the other methods in general. The best fits across all methods are achieved for the exponential case $(c_{\lambda}^2 = c_{\mu}^2 = 1)$. For $c_{\lambda}^2 > 1$ and/or $c_{\mu}^2 > 1$, our results (column \mathcal{P}) are very close to the ones derived using the MEM (column \mathcal{K} , cf. [9]) since both approaches use the GE distribution. For $c_{\lambda}^2 = c_{\mu}^2 = 0.5$, our results are closer to the simulation results (column \mathcal{S}) than the ones derived with Marie's method (column \mathcal{M}) or the MEM.

3.2. Influence of Variance on Mean Response Time

Unless stated otherwise, in this section, we use the default parameters defined in column (c) of Tab. 1.

2	2	\overline{M}			ρ{λ}				$ ho_{\mu}$			
c_{λ}	c_{μ}	S	\mathcal{M}	\mathcal{P}	\mathcal{K}	S	\mathcal{M}	\mathcal{P}	\mathcal{K}	S	\mathcal{M}	\mathcal{P}
0.5	0.5	1.298	1.350	1.297	0.964	0.956	0.951	0.958	0.723	0.717	0.723	0.718
0.5	1.0	1.340	1.372	1.372	0.924	0.916	0.927	0.927	0.693	0.687	0.695	0.695
0.5	5.0	1.455	1.449	1.580	0.782	0.830	0.827	0.781	0.587	0.622	0.620	0.586
1.0	0.5	1.409	1.405	1.405	0.924	0.931	0.923	0.923	0.603	0.698	0.692	0.692
1.0	1.0	1.423	1.444	1.444	0.896	0.901	0.896	0.896	0.672	0.676	0.672	0.672
1.0	5.0	1.535	1.491	1.592	0.768	0.807	0.803	0.768	0.576	0.605	0.602	0.576
5.0	0.5	1.787	1.599	1.604	0.753	0.843	0.840	0.753	0.565	0.632	0.629	0.565
5.0	1.0	1.607	1.607	1.610	0.745	0.816	0.816	0.745	0.559	0.612	0.612	0.559
5.0	5.0	1.595	1.591	1.642	0.700	0.740	0.749	0.700	0.525	0.555	0.561	0.525

Table 2: Comparison to closed G/G/1/K - FCFS tandem queueing network model for $c_{\lambda}^2 \neq 1$ and/or $c_{\mu}^2 \neq 1$ (S: WinPEPSY with simulation; \mathcal{M} : WinPEPSY with Marie's method; \mathcal{P} : using phasetype approximation (Sec. 2.2); \mathcal{K} : results provided by [9]).

3.2.1. Influence of Generation Time Variance

In Fig. 3, the influence of the generation time's SCV c_{λ}^2 ($0.1 \le c_{\lambda}^2 \le 20$; x-axis) on the mean response time \overline{T} (y-axis) is shown for different values of the orbital search probability $p \in \{0.01, 0.5, 0.99\}$. Evidently, in contrast to the FCFS-equivalent case (p = 0.99), c_{λ}^2 has a significant effect on the mean response time \overline{T} for the cases of distinctive retrial behavior ($p \in \{0.01, 0.5\}$). Consequently, both the retrial behavior and the second moment of the generation time should be considered when discussing the performance of real systems.

For $p \in \{0.01, 0.5\}$, Fig. 3 shows a maximum of the mean response time \overline{T} for $c_{\lambda}^2 \approx 0.5$. Hence, the exponential case $c_{\lambda}^2 = 1$ can be neither considered as optimistic nor pessimistic performance bound.

Obviously, the FCFS-equivalent case (p = 0.99) gives an optimistic performance bound. Retrial systems will always perform worse with respect to the mean response time \overline{T} when compared to FCFS systems. It is well-known that the performance of finite-source G/M/1/N - FCFS queueing systems is independent of the generation time's SCV c_{λ}^2 (see, e.g., [11]). This is also reflected in Fig. 3, where \overline{T} is (basically) constant for p = 0.99. Moreover, for high values of c_{λ}^2 , \overline{T} apparently approaches the mean response time of G/M/1/N - FCFS queueing systems independently of p.

3.2.2. Influence of Retrial Time Variance

In Fig. 4, the influence of the retrial time's SCV c_{ν}^2 (0.1 $\leq c_{\nu}^2 \leq 20$; *x*-axis) on the mean response time \overline{T} (*y*-axis) is illustrated for different values of the orbital search probability $p \in \{0.01, 0.5, 0.99\}$.

Obviously, for orbital search probabilities close to 1, the model approaches the FCFS-equivalent case, and consequently, c_{ν}^2 has negligible effect on \overline{T} only. On the other hand, the influence is more significant for smaller values of p while (in

contrast to Fig. 3) a maximum of \overline{T} is not observable for $c_{\nu}^2 \neq 0$.

Similar to the discussion of c_{λ}^2 , it can be seen in Fig. 4 that assuming exponentially distributed retrial times $(c_{\nu}^2 = 1)$ provides neither an optimistic nor pessimistic performance bound. Likewise, for high values of c_{ν}^2 , \overline{T} approaches the mean response time of G/M/1/N - FCFS queueing systems independently of p.



Figure 4: Influence of retrial time's SCV c_{ν}^2 on the mean response time \overline{T} for different orbital probabilities p.

Figure 5: Influence of service time's SCV c_{μ}^2 on the mean response time \overline{T} for different orbital probabilities p.

3.2.3. Influence of Service Time Variance

The influence of the service time's SCV c_{μ}^2 (0.1 $\leq c_{\mu}^2 \leq 20$; *x*-axis) on the mean response time \overline{T} (*y*-axis) is depicted for different values of the orbital search probability $p \in \{0.01, 0.5, 0.99\}$ in Fig. 5.

In contrast to Figs. 3 and 4, c_{μ}^2 has notable influence on \overline{T} for all values of p and hence should be considered for non-retrial and retrial queues alike. The shortest mean response time is still achieved in the FCFS-equivalent case where p = 0.99.

4. Conclusion

In this paper, we provide a model of a finite-source retrial queueing system with orbital search where the job generation, retrial, and service times are random variables modeled by their first two moments (represented by mean and squared coefficient of variation). Numerical results are obtained by employing the MOSEL-2 tool. The results show that the second moments have a fair effect on the mean response time of the investigated retrial queue.

Hence, both the retrial behavior and the second moment need to be taken into account when analyzing the performance of real systems. Assuming exponentially distributed job inter-arrival or service times and ignoring retrial effects might lead to severe overestimation of system performance with negative impact on the application.

In near future, we plan to exploit the proposed model for the analysis of wireless multi-hop communication.

References

- AL-BEGAIN, K. Using subclass of PH distributions for the modeling of empirical activities. In Proc. of 18th Int'l Conference on Statistics and Computer Sciences (ICSCS 93) (Cairo, 1993), pp. 141–152.
- [2] ALFA, A. S., AND ISOTUPA, K. S. An M/PH/k retrial queue with finite number of sources. Computers and Operations Research 31 (2004), 1455–1464.
- [3] ARTALEJO, J. R., AND GÓMEZ-CORRAL, A. Retrial Queueing Systems: A Computational Approach. Springer Verlag, 2008.
- [4] BAZAN, P., AND GERMAN, R. Approximate transient analysis of large stochastic models with WinPEPSY-QNS. Computer Networks 53, 8 (June 2009), 1289–1301.
- [5] BAZAN, P., GERMAN, R., AND BOLCH, G. WinPEPSY-QNS performance evaluation and prediction system for queueing networks. In *Proc. of MMB 2006* (2006).
- [6] BOLCH, G., GREINER, S., DE MEER, H., AND TRIVEDI, K. S. Queueing Networks and Markov Chains, 2 ed. John Wiley & Sons, New York, 2006.
- [7] DALLERY, Y. On modeling failure and repair times in stochastic models of manufacturing systems using generalized exponential distributions. *Queueing Systems 15*, 1 (March 1994), 199–209.
- [8] FALIN, G., AND ARTALEJO, J. A finite source retrial queue. Eur. J. Operational Research 108 (1998), 409–424.
- [9] KOUVATSOS, D. D. Maximum entropy and the G/G/1/N queue. Acta Informatica 23 (1986), 545–565.
- [10] KOUVATSOS, D. D. Entropy maximisation and queueing network models. Annals of Operations Research 48 (1994), 63–126.
- [11] SZTRIK, J. On the finite-source $\vec{G}/M/r$ queue. Eur. J. Operational Research 20 (1985), 261–268.
- [12] WÜCHNER, P., DE MEER, H., BARNER, J., AND BOLCH, G. A brief introduction to MOSEL-2. In *Proc. of MMB 2006* (2006).
- [13] WÜCHNER, P., SZTRIK, J., AND DE MEER, H. Homogeneous finite-source retrial queues with search of customers from the orbit. In *Proc. of MMB 2008* (2008).
- [14] WÜCHNER, P., SZTRIK, J., AND DE MEER, H. Investigating the mean response time in finite-source retrial queues using the algorithm by Gaver, Jacobs, and Latouche. *Annales Mathematicae et Informaticae 36* (2009), 143–160.

Patrick Wüchner, Hermann de Meer

Faculty of Informatics and Mathematics University of Passau Innstraße 43 94032 Passau Germany

János Sztrik

Faculty of Informatics University of Debrecen Egyetem tér 1 P.O. Box 12 4010 Debrecen Hungary