Retrial Queueing Models: A Survey on Theory and Applications

Tuan Phung-Duc
Faculty of Engineering, Information and Systems
University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan
Email: tuan@sk.tsukuba.ac.jp

Abstract

Retrial phenomenon naturally arises in various systems such as call centers, cellular networks and random access protocols in local area networks. This paper gives a comprehensive survey on theory and applications of retrial queues in these systems. We investigate the state of the art of the theoretical researches including exact solutions, stability, asymptotic analyses and multidimensional models. We present an overview on retrial models arising from real world applications. Some open problems and promising research directions are also discussed.

1 Introduction

The loss models (including Erlang loss model) assume that an arriving customer that sees the service area being fully occupied is blocked and is lost forever. On the other hand, in models with an infinite waiting capacity, a customer waits until being served. However, there are various situations in our everyday life and in various systems where blocked customers are not willing to wait and they temporarily leave the service facility for a while but try again after some random time. A blocked customer is said to be in a virtual waiting room called *orbit* before retrying to occupy a server again. These situations are modeled as retrial queues.

For example, in a call center, if a customer makes a phone call when all the agents are busy, the customer will try to make a phone call again after some random time. In computer networks, if a packet is lost, the packet may be retransmitted at a later time by a retransmission mechanism such as the TCP (Transmission Control Protocol) [24, 25]. In these applications, the orbit is virtual and cannot be observed. Figure 1 represents a general multiserver retrial queue.

In many applications, the customers in the orbit act independently of each other, thus the retrial rate depends on the number of customers in the orbit. As a result, the underlying Markov chains of retrial queues have a spatially non-homogeneous structure. Due to the spatial non-homogeneity of the underlying Markov chains, the analysis of retrial queues is more complex and challenging than that of standard queues. For an extensive comparison of standard and retrial queueing systems, the readers are referred to the paper of Artalejo and Falin [15]. Even for the M/M/c/c retrial queue, where retrial interval of customers follows an exponential distribution, analytical solutions are obtained in only a few special cases [15, 57].

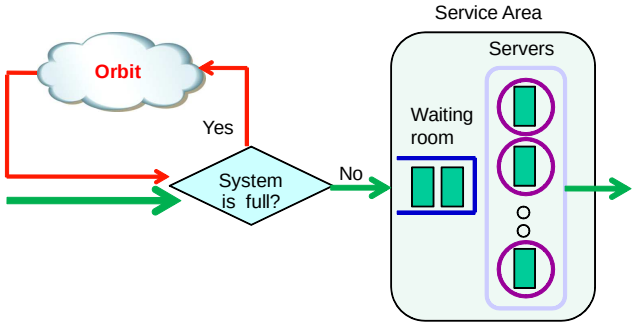

Figure 1: A general multiserver retrial queue.

Classical retrial policy

In a retrial queue with the classical retrial policy, each blocked customer stays in the orbit for an exponentially distributed time independently of other customers. As a result, the retrial rate is proportional to the number of customers in the orbit. The classical retrial policy naturally arises from applications such as call center and telephone exchange system, where customers in the orbit act independently because a retrial customer cannot observe the behavior of the others. Retrial queues with the classical retrial policy have attracted many researchers since they naturally arise from various applications with random access. Under some Markovian settings of the service time distribution and the arrival process, the underlying stochastic process is a level-dependent quasi-birth-and-death (LDQBD) process where the level is the number of customers in the orbit and the phase represents the states of the servers. For general LDQBD processes, some general numerical algorithms are avail-

able [31, 99]. The sparsity of the block matrices of the LDQBD processes of retrial queues could be used to develop efficient computational algorithm [101].

Constant retrial policy

There are several applications in communication networks, where the retrial of customers is controlled. It means that the retrial rate may not depend on the number of customers in the orbit. For example, Choi et al. [35] study the stability of the CSMA/CD (Carrier Sense Multiple Access with Collision Detection) protocol, by a retrial queue with a constant retrial rate. Avrachenkov and Yechiali [24, 25] use a retrial queueing network with constant retrial rate to model TCP traffic. Constant retrial rate could be interpreted by the so-called "calling for blocked customers". When the server is idle, it calls blocked customers one by one. The time for the server to pick up a blocked customer could be interpreted as the retrial time.

Artalejo et al. [12] formulate a multiserver retrial queue (M/M/c/c) with constant retrial rate by a level-independent QBD process which could be analyzed efficiently using matrix analytic methods invented by Neuts [87, 93]. As in the classical retrial rate, the block matrices of the level-independent QBD process are also sparse leading to efficient algorithms for the stationary distribution. These algorithms are discussed by Artalejo et al. [12] using a matrix analytic method and by Do et al. [] using the spectral expansion method [44, 45].

The structure and aim of this chapter are as follows. First in Section 2, we provide a comprehensive review of retrial queues in real world applications. Second, in Section 3, we present the main results on the analysis of retrial queueing models. The aim of this survey is to provide a guide for researches who want to enter and deepen the understanding of the field of retrial queues. To this end, we point out some open problems and promising research topics.

2 Retrial Queues in Applications

In this section, we present several retrial queueing models arising from real world applications.

2.1 Call Centers

A call center is important for a company because it provides a channel for customers to contact the company. In a call center, agents are the people who answer the calls from the customers. When a customer makes a phone call, if there is an idle call agent, the customer is immediately answered by the call agent. If all the agents are busy, the customer may hear some massage such as "the system is busy at the moment, please wait for a moment". At this moment, the customer either hangs up the phone immediately or continues to hear the message. In the former case, the customer may try again after some time. Customers who decide to wait for a free call agent may renege if the waiting time is too long. These customers may also make a phone call later.

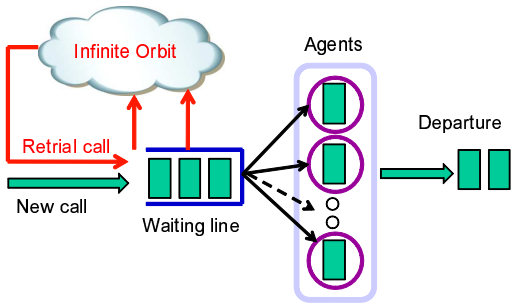

Figure 2: A retrial queueing model for call centers.

One of the most important performance measures for a call center is the blocking probability that a customer cannot find a free call agent upon arrival. From the customer's point of view, a low blocking probability is desirable. In order to keep the blocking probability to be low, a simple solution is to increase the number of agents. However, from a management point of view we need to minimize the number of agents, due to the fact that the cost of a call center is mainly the human cost [121]. In order to achieve the customer satisfaction under some constraint on the cost, we need some mathematical model to express the trade-off between the customer satisfaction and the human cost. A queueing model is one of the most appropriate mathematical models for the design of call centers. In addition, in order to capture the retrial phenomenon as presented above, a retrial queueing model is expected to be more appropriate than the corresponding standard queueing model [62, 81, 121, 1]. See Figure 2 for a simple retrial queueing model for call centers. For a detailed explanation on call centers, we refer to the book by Stolletz [121]. Numerical results by Phung-Duc et

al. [101] show that approximating retrial flow by a Poisson process leads to a large error under the fast retrial regime. This is an evidence for modelling call centers by retrial queues.

Phung-Duc and Kawanishi [104] consider two-way communication retrial queueing systems as the models for blended call centers. Queueing models without retrials for blended call centers are proposed and analyzed in [32]. Two-way communication retrial queues are studied by Artalejo and Phung-Duc in [21, 22] where some analytical results such as the stability condition and generating functions are obtained. The main contribution in Phung-Duc and Kawanishi [104] is to propose an efficient computational algorithm for multiserver retrial queues with distinct distributions of incoming and outgoing calls. The algorithm [104] could be considered as a matrix version of the one by Phung-Duc et al. [101]. In two-way communication queueing systems, a server not only receives incoming calls but also makes outgoing calls to outside in its idle time. This is the situation in blended call centers where the operators may make outgoing calls to the customers for some marketing purposes etc.

Furthermore, Phung-Duc and Kawanishi [105] analyze a fairly general and practical retrial model for inbound call centers with after-call work and abandonment. After-call work is a typical task in call centers where after a conversation with a customer, the operator should do some after-call work for that customer after the customer departs from the system. It means that a call line is released for a newly arrived customer. In [105], the effects of retrials by blocked and abandoned customers on the waiting time distribution are investigated.

2.2 Cellular Communication Networks

In a cellular network, the service area is divided into cells. Users (mobile stations) in each cell are served by a base station with a limited number of channels. Therefore, only a limited number of users can communicate at the same time. A base station serves two types of calls: fresh calls and handover calls. A fresh call is made by a user that stays inside the current cell and a handover call is made by a user that has been traveling from an adjacent cell into the current cell. Because a handover call has been communicating by a channel in an adjacent cell, the call should be assigned a channel upon its arrival into the current cell as soon as possible for a continuous communication. Therefore, a handover call

should be given a priority over a fresh call.

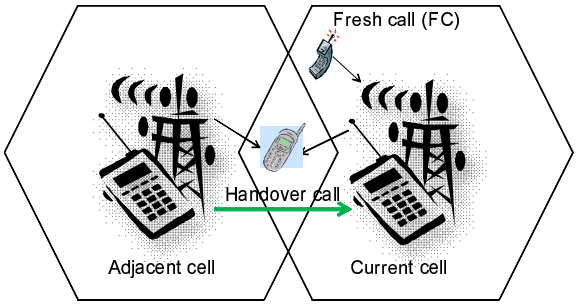

Figure 3: Two adjacent cells in cellular networks.

There are several channel assignment policies that provide some priority for handover calls over fresh calls [45, 123]. In a guard channel policy, a number of channels are reserved for handover calls. The rest of channels are equally shared by both fresh and handover calls. In a fractional guard channel policy, fresh calls are accepted with a probability depending on the number of currently occupied channels, while handover calls are accepted as long as an idle channel is available [114]. Both fresh calls and handover calls may be blocked due to the limit on the number of channels. In modern cellular communication systems (e.g. mobile phone system), if a call is blocked, a redial can be made easily, for example just by pushing only one button. In some applications, blocked calls are automatically redialed. Thus, the consideration of retrial calls is very important while designing these systems [123].

Figure 3 represents two adjacent cells in a cellular networks. In Figure 4, a buffer for handover calls represents overlap areas of the current cell with the adjacent cells. Handover calls in the overlap areas can receive signals from both the adjacent cell and the current cell. Thus, if a channel in the current cell is not yet available, the handover call can continue to communicate using the occupying channel in the adjacent cell. However, the handover call is terminated if it exceeds an overlap area but no idle channel in the targeting cell is available. In this case, the handover call may attempt again after some random time as a fresh call.

In addition to the pure Markovian models presented in [45, 123], more general models with correlated arrival processes such as Markovian Arrival Processes (MAP) and Batch Markovian Arrival Processes (BMAP) and phase type service time distributions are investi-

gated in [2, 37, 74, 75]. A model under random environment is presented in [20]. In all these models, the performance measures are directly calculated from the stationary probabilities. Economou and Lopez-Herrero [52] provide some more sophisticated performance measures such as waiting time distribution, idle time of guard channels etc.

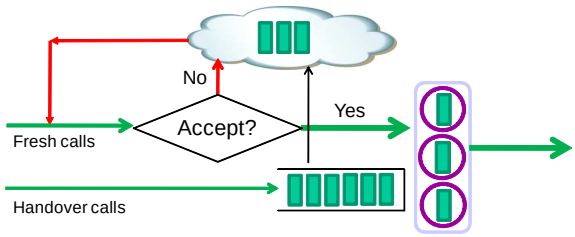

Figure 4: A model of cellular networks with overlapping cells.

2.3 Local Area Networks

In a local area network (LAN), multiple nodes share a physical link (channel) in order to transmit their data (packets). Assuming that multiple nodes send their packets at the same time, a collision may occur and all the packets will be destroyed.

Because nodes in a LAN are located closely to each other, propagation delays are much shorter than data transmission time. Therefore nodes can obtain useful information about the channel by sensing the presence of signals on the channel. Taking this property into account, CSMA (Carrier Sense Multiple Access) protocols have been developed. The fundamental idea of a CSMA protocol is that it senses the channel before transmitting data. In a 1-persistent CSMA protocol, when a node is ready to send its data, the node checks if the channel is busy. If the channel is busy, then the node continues to sense the link until the channel is idle and then immediately sends a frame. In case of collision, the node waits for a random period of time and tries to transmit again. The problem of the 1-persistent CSMA protocol is that in a highly loaded condition, there may be several nodes waiting for the availability of the channel and they send data at the same time when the channel becomes idle. Therefore, collisions occur with a high probability.

Another protocol is non-persistent CSMA, in which if a node is ready to send data, it senses the channel and transmits data immediately if the link is available, otherwise the node waits for a random time and tries to retransmit again. A p-persistent CSMA protocol

is considered to be a hybrid of 1-persistent CSMA and non-persistent CSMA protocols. In a p-persistent CSMA, if the channel is idle, the node transmits a frame with probability p and delays its transmission for one unit time (channel propagation delay) with probability $1 - p$. If the channel is busy, the node continues to sense until the channel becomes idle and repeat the same procedure to send its frame.

Because the collisions in CSMA protocols cannot be completely avoided, the retrial phenomenon occurs in local area networks, e.g. wireless networks. Therefore, retrial queueing models are considered to be more appropriate than standard queueing models in modelling and performance analysis of these protocols [35, 67]. The readers are referred to the book by Kurose and Ross for a detailed explanation on protocols for local area networks [83]. Because the behaviors of nodes in radio networks depend on each other, analysis of an exact model representing the states of all nodes is very challenging. Assuming that the behaviors of nodes are independent, Bianchi [28] demonstrates a simple analysis of this complex system. Recently, Fiems and Phung-Duc [60] propose a light traffic analysis by a series expansion method for a retrial model which represents all the states of nodes concurrently. The accuracy of the model is validated by simulations. The series expansion method [60] might be useful for other retrial models for local area networks with random access protocols.

2.4 Retrial Queues for Cognitive Networks

Recently, cognitive radio networks are extensively studied. The spectrum for wireless networks is physically limited. On the other hand, the amount of traffic by smart phones and other devices vastly increases day by day. However, most of bandwidth is granted to licensed users (primary users). The bandwidth is not always used by primary users. The idea of cognitive networks is to provide the opportunity for secondary users to use this bandwidth when it is not used by the primary users. Secondary users are interrupted upon the arrivals of primary users. Queueing analysis of cognitive networks has been presented by Konishi et al. [80] using a multiserver priority model without buffer where interrupted secondary users are lost. In practice, interrupted users may retry again. From this point of view, appropriate retrial queueing models might be useful for cognitive systems. Wang et al. [127, 128] use some simple M/M/1/1 retrial queueing models to study strategic behav-

ior of secondary users in cognitive networks. Dudin et al. [50] propose a multiclass retrial queueing model for cognitive systems. Salameh et al. [124]study retrial queueing models taking into account the sensing time of interrupted secondary users. Because cognitive radio is promising solution for the insufficiency of wireless spectrum [80], new retrial models for various scenarios and technologies in cognitive systems are promising topics for future researches.

2.5 Retrial Queues in other Applications

Beside concrete applications mentioned in previous subsections, retrial queues are also ubiquitous in other applications. Retrial queues for optical networks are presented in [96, 61, 29]. Furthermore, some emerging technology such as cloud computing, one can also find various situations where retrial queues are applicable. For example, single server retrial queues with setup time are proposed for power-saving servers [107, 110]. Phung-Duc and Kawanishi [113] investigate the impact of retrial phenomenon on power saving data centers by an M/M/c/c retrial queue with setup time. In cloud systems, the computing unit and the storage unit may be separated leading to the transmission time between them. A retrial model for such a situation is initiated in [109]. Furthermore, in cloud systems, the capacity of the server can be scaled according to the workload in the system. Taking this into account, Phung-Duc et al. [108, 111] propose single server retrial queueing models with speed scaling and setup time where the speed of the server is proportional to the number of jobs in the system.

3 Models and Methodologies

In the analysis of retrial queueing models, we need to keep track not only of the state of the servers but also the number of customers in the orbit. Under the assumption that retrial customers behave independently, the retrial flow by repeated customers makes the underlying stochastic process non-homogeneous. As a result, the analysis of retrial queues is more difficult than the corresponding models with infinite waiting room. Indeed, the model with infinite waiting room could be obtained from the corresponding retrial models by taking the limit as the retrial time tends to zero. In this section, we show methods and analytical results for some major retrial queueing models.

3.1 M/G/1/1 Retrial Queues and Their Variants

In this model, there is only one server and there is not a waiting room before the server. Customers arrive at the system according to a Poisson process and the service time for a customer is arbitrarily and identically distributed. An arriving customer immediately occupies the server if it is idle. Otherwise, the customer will join the orbit from which he will retry again in an exponentially distributed time with positive mean. This model could be analyzed by either embedded Markov chain or by complementary variable method as for the original M/G/1 model without retrials.

Let $\pi_{i,n}$ denote the stationary probability that the server is at state i ($i = 0$ if the server is idle and $i = 1$ if the server is busy) and there are n customers in the orbit. Furthermore, let $\Pi_i(z) = \sum_{n=0}^{\infty} \pi_{i,n} z^n$ $(i = 0, 1)$. For this model, the generating functions, i.e., $\Pi_i(z)$ $(i = 0, 1)$ of the number of customers in the orbit can be obtained in an integral form [57]. Furthermore, $\pi_{0,0}$ is given in an integral form and other probabilities $\pi_{i,j}$ are recursively computed. If the service time is exponentially distributed, the integral forms of $\Pi_i(z)$ $(i = 0, 1)$ become explicit. Furthermore, under the light-tailed assumption of the service time, asymptotic formula for the joint queue length distribution is known. In particular, $\pi_{0,n} \asymp C_0 n^{a-1} \sigma^{-n}$ and $\pi_{1,n} \asymp C_1 n^a \sigma^{-n}$ as $n \to \infty$ [78] for some positive constants C_0, C_1, a and $\sigma > 1$.

Artalejo and Phung-Duc [22] study M/G/1/1 retrial queue with two-way communication where the server makes an outgoing call in an exponentially distributed idle time with mean $1/\alpha$. In [22], incoming calls and outgoing calls follow two distinct arbitrary distributions. Under light-tailed assumption of the service time distributions of incoming and outgoing calls, asymptotic analysis of the joint queue length is also presented in [22] using a simpler method in comparison with that of Kim et al. [78] for the M/G/1/1 retrial queue without outgoing calls. Under some heavy tailed assumptions of the service times, the queue length asymptotics is presented in Shang et al. [116] for the M/G/1/1 model and by Yamamuro [129] for M^X/G/1/1 retrial queue.

Heavy traffic asymptotics is presented in Falin [57]. In particular, when the traffic intensity is close to one, the scaled queue length distribution tends to Gamma distribution. Furthermore, when the retrial rate is extremely low, the distribution of the scaled number of customers in the orbit tends to Gaussian distribution. Sakurai and Phung-Duc [115]

study heavy traffic analysis for M/G/1/1 retrial queue with two-way communication. In addition to the two heavy traffic regimes: i) traffic intensity is close to 1 and ii) extremely slow retrial rate, the authors also study the regime iii) where length of an outgoing call is extremely long. Sakurai and Phung-Duc prove that in regime iii), the distribution for the scaled number of customers in the orbit tends to Gaussian distribution. We refer to recent work of Nazarov et al. [91] and Fedorova [59] for recent development on heavy traffic analysis of retrial queues without and with random environment.

An extension of the M/G/1/1 retrial queues with nonpersistent customers is proposed by Yang et al. [131]. In this model, a blocked fresh customer joins the orbit with probability p while a blocked retrial customer joins the orbit with probability q, respectively. With the complementary probability, they abandon joining the orbit. In the case $q = 1$, the analysis is almost the same as that of the basic model with $p = q = 1$. However, the analysis for the case $q < 1$ is essentially different from that of the case $q = 1$. In particular, although the equations for the partial generating functions of the joint queueing length distribution are obtained, the solution for these equations is not obtained. It is shown that all the quantities interest such as the joint stationary joint queue length distribution and the factorial moments are expressed in terms of the utilization of the server which is unknown. A numerical algorithm is then developed to determine the utilization of the server. However, analytical expression for the utilization remains an open problem.

3.2 Multiple Server Models

While the single server model is relatively tractable in comparison with that without retrials, multiserver retrial queues are much more difficult than those without retrials. The basic M/M/c/c retrial queues with classical retrial rate (i.e. exponential retrial intervals) has been studied by Cohen [38]. In this system customers arrive at the system according to a Poisson process with rate λ, the service time of a customer is exponentially distributed with mean $1/\nu$ and the retrial intervals are exponentially distributed with mean $1/\mu$.

In this model, let $\pi_{i,j}$ denote the joint stationary probability that there are i ($i = 0, 1, \ldots, c-1, c$) busy servers and j customers ($j \in \mathbb{Z}_+$) in the orbit. Furthermore, let $\Pi_i(z) = \sum_{j=0}^{\infty} \pi_{i,j} z^j$ ($i = 0, 1, \ldots, c$). Explicit expressions for $\pi_{i,j}$ and $\Pi_i(z)$ are obtained for the case $c = 1$. For $c = 2$, $\pi_{i,j}$ and $\Pi_i(z)$ are expressed in terms of hypergeometric

functions [71]. For the case $c \geq 3$, it is challenging to obtain analytical solutions for $\pi_{i,j}$ and $\Pi_i(z)$. Pearce [95] constructs a solution to the balance equations in terms of generalized continued fraction. Phung-Duc et al. [97] prove for the case $c = 3, 4$ that $\pi_{i,j}$ is expressed in terms of the minimal solution of a three-term recurrence relation and thus in terms of continued fractions.

Phung-Duc et al. [98] extend the analysis to M/M/c/c ($c \leq 4$) retrial queues with nonpersistent customers where a blocked fresh customer and a blocked retrial customer joins the orbit with probability p and q, respectively. As shown in [98], the solution for the model $q = 1$ is almost the same as that of the basic model with $p = q = 1$, while the solution for the case $q < 1$ exhibits a different structure. Furthermore, an accurate algorithm is developed to calculate these continued fractions with pre-specified accuracy [98].

Tail asymptotic analysis for the queue length distribution of M/M/c/c retrial queues has been extensively investigated. Liu and Zhao [86] show that $\pi_{c-i,j}$ is of the order of $j^{a-i}\rho^j$ as $j \to \infty$ for some constant a where $\rho = \lambda/(c\nu) < 1$, using a matrix analytic method. The analysis of Liu and Zhao [86] is based on the series expansion (up to second order) of the rate matrix R_n of the underlying level-dependent QBD process in terms of $1/n$. The series expansion is extended to any order by Phung-Duc [106] for M/M/c/c models with two types of non-persistent customers. It should be noted only the last row of R_n is non-zero in these models. Kajiwara and Phung-Duc [72] further extend the analysis of Phung-Duc [106] to an M/M/c/c retrial model with one guard channel for priority and retrial customers where the last two rows of the rate matrix R_n are non-zero. Kim et al. [76, 77] refine the results of [86] by a generating function approach.

A simple and accurate fixed point approximation is proposed by Cohen [38] for the case where the retrial rate is relatively small in comparison with the service rate. In such a situation, the total arrival flow by both fresh customers and retrial ones is approximated by a Poisson process whose rate is the solution of a fixed point equation (See e.g. Falin [57] for details). Let $\lambda + r$ denote the arrival rate of the approximated Poisson process, where r is the additional arrival rate due to retrial customers. Let $B(\lambda, c)$ denote the blocking probability of the corresponding Erlang-loss system without retrials where $\nu = 1$. The additional arrival rate r is the solution of

$$r = (\lambda + r)B(\lambda + r, c). \tag{1}$$

The main reason for this approximation is that under low retrial regime, the retrial flow is likely to form a Poisson process. Recently, the heavy traffic analysis (also known as Halfin-Whitt regime) for equation (1) is presented by Avram et al. [27].

For a numerical computation of the joint stationary distribution, we need somehow to truncate the orbit at some truncation point. The simplest truncation method is limiting the number of customers in the orbit to N. Thus, a blocked customer that sees N customers in the orbit is lost. Another truncation method is to modify the structure of the original Markov chain after the truncation point. The modification makes Markov chain analytically tractable. This is referred to as generalized truncation in the literature [57, 92, 8, 14]. The idea in [57, 14] is to disregard the states (after the truncation level (orbit size)) having small probability mass.

More general models with MAP or BMAP arrivals and phase-type service time are also investigated [30, 79, 74, 73] by means of the so-called quasi-Toeplitz Markov chains [79].

3.3 Stability Conditions

The stability condition of the basic M/M/c/c retrial queue is simply given by $\lambda < c\nu$ or the offered load (λ/ν) is less than the number of servers. The proof of this result is based on an appropriate Lyapunov function of a linear form $f(i, j) = ai+j$, where i is the number of busy servers and j is the number of customers in the orbit [57]. Artalejo and Phung-Duc [21] derived the necessary and sufficient condition for M/M/c/c retrial queues with two-way communication where an idle server may make an outgoing call after some exponentially distributed idle time with mean $1/\alpha$. In [21], incoming calls and outgoing calls follow two distinct exponential distributions. It turns out that the stability condition of the M/M/c/c retrial queues with two-way communication coincides with that of the corresponding model without outgoing calls, i.e., $\lambda < c\nu$. Phung-Duc and Dragieva [112] obtain the stability condition for a multiserver retrial queue with interaction between servers and orbit where not only customers retry but also servers call for customers from the orbit. The Lyapunov function for the models in [21, 112] is of the form $f(i, j, k) = ai + bj + k$ where i and j are variables representing the states of the servers and k is the number of customers in the orbit.

For retrial models whose LDQBD process has complex phase structure, it is efficient

to use the Lyapunov function proposed by Diamond and Alfa [41]. This approach has been used to show the stability condition for MAP arrival models, model with after call work [103, 105] and model with setup time [113].

In a recent paper, Shin [118] proves that the stability condition for the multiclass M/M/c/c retrial queue is that the total offered load of all classes is less than the number of servers. The proof of Shin [118] is based on a Lyapunov function that is a linear combination of the numbers of customers in the orbits of all classes and the states of the servers. The stability condition concerns when there are a very large number of customers in the orbit. Thus, the time that a retrial customer reaches an idle server is almost zero in such a situation. This is an intuitive observation that the stability condition of retrial queues coincides with that of corresponding models with infinite buffer. Recently, Dayar and Can Orhan [39] prove the stability condition for multiclass MAP/PH/c retrial queues with cyclic PH retrial times. The proof in [39] is based on a Lyapunov function which is also a linear function of the numbers of customers in the orbit. It should be noted that the Lyapunov function approach is applied for Markovian models only.

In a fairly general class of retrial queues (both single and multiple class models) with classical retrial rate, the coincidence in the stability conditions between retrials models and non-retrial models is confirmed by Morozov et al. [89, 90] for more general non-Markovian models (i.e., arbitrary renewal arrivals and arbitrary service time) using the regenerative approach. The regenerative approach of Morozov et al. is also used to prove other models with constant retrial rate [3, 4]. The result in [90] generalizes that of Shin [118].

3.4 Multiclass Models

3.4.1 Classical retrial policy

We consider the multiclass M/G/1/1 retrial model with classical retrial rate where m classes of customers arrive at the server according to m distinct Poisson processes with rate $\lambda_1, \lambda_2, \ldots, \lambda_m$. The service times of m classes of customers follow m distinct arbitrary distributions. A blocked customer of class k, joins the k-th orbit and retries to enter the server after some exponentially distributed time with mean $1/\mu_k$ ($k = 1, 2, \ldots, m$). For this model and its variants the stability conditions are available and the means number of customers (also mean waiting time) in the orbit for each class are obtained [57, 55, 82].

Grishechkin [70] studies single server retrial queue with structured batch arrivals and investigates some heavy traffic limits for the queue length process. Many open questions for these models such as heavy traffic for slow retrial case, waiting time distribution are still open for further investigations. Under multiserver settings, only the stability conditions are known [89, 90, 118, 39].

3.4.2 Constant retrial policy

Multiclass M/G/1/1 retrial queues with constant retrial rates have been paid much attention in recent years. For this model, under Markovian assumptions, i.e., the service times of each class follow a distinct exponential distribution; the underlying Markov chain is multi-dimensional. In the case of two classes, boundary problems are formulated and some information on the means number of customers of two classes in the orbit may be derived. Furthermore, information on the joint generating functions of the numbers of customers in the orbit are obtained [26] using Riemann-Hilbert boundary value problems. Song et al. [120] study the same model using a kernel approach and obtain tail asymptotic for the queue length of each class. Dimitriou [42, 43] studies some extended models for network coding in relay nodes in wireless networks.

3.5 Priority and Related Models

3.5.1 Classical retrial policy

In retrial queues with priority, blocked priority customer can wait in front of the server while blocked normal customer joins the orbit. Under the classical retrial rate setting, bivariate generating functions for the joint queue length distribution of the number of customers in normal queue and that in the orbit is obtained in an integral form [36, 57, 58]. For this model, heavy traffic analysis (the total traffic intensity tends to 1) is carried out by Falin et al. [58]. Recently, Walraevens et al. [125] derive tail asymptotic formulas for the stationary distribution of the number of customers in the orbit. A detailed survey on related models is presented in [36].

3.5.2 Constant retrial policy

Priority retrial queues with constant retrial rate are also paid much attention. Lower priority customer joins the orbit while high priority customer joins the priority queue. The

orbit operates as a single server queue where customers retry to occupy the server according to a FCFS manner and only the customer in the head of the orbit queue retries at a time. Under a pure Markovian setting, i.e., exponential retrial time and service time, Li and Zhao [85] obtain tail asymptotics of the priority queue given a fixed number of customers in the orbit queue. Gomez-Corral [64] considers the model with constant retrial rate where the retrial time of the customer in the head of the orbit and the service time are arbitrarily distributed. Atencia and Moreno [23] analyze the model under general retrial time and Bernoulli scheduling. In this model [23] an arriving customer that sees the server busy either joins the priority queue (normal queue) or the orbit queue. The authors [23] obtain the bivariate generating function of the number of customers in the orbit and that in the normal queue and the remaining service time or the remaining retrial time. The analysis of the models with both arbitrary service and retrial times is possible because at any time we need to keep track of either remain service time or remaining retrial time.

3.6 Queueing Networks with Retrials

In our everyday life, there are various service systems, in which customers are served by a number of servers in a certain order. For example, in a manufacturing system, a product is made and checked by a number of persons. In a computer system, a message is transmitted from a source to a destination through several devices such as computers, routers and switches. These systems can be modeled by queueing networks.

Queueing networks with retrials do not possess product form solution [16]. Thus, exact solution is obtained in a few special cases [102, 88]. Phung-Duc [102] obtains explicit solution for a simple two-node tandem network with classical retrials at the first node. Moutzoukis and Langaris [88] derive the explicit results for the tandem model with constant retrial rate and blocking at the first server. For complex retrial queueing networks, a practical approach is the fixed point approximation. For example, fixed point approximations are used to analyze some tandem models with retrials by Avrachenkov et al. [24, 25]. In fixed point approximation, the system is divided into multiple subsystems whose input parameters are unknown. Furthermore, these subsystems are assumed to be independent. The output of one subsystem is the input of another subsystem. After some iterative calculations, one will get a convergence determining all the unknown parameters. One drawback

of this methodology is that a rigorous proof of the convergence and the accuracy of the approximation are not always presented. Moreover, the approximation is basically valid only under the slow retrial regime. Numerical solutions for some simple tandem retrial models are presented in [74, 73, 122, 65].

Recently, Fiems and Phung-Duc [60] present a light-traffic analysis for finite-source retrial systems arising from CSMA protocols without collision. These systems could be formulated by multidimensional Markov chains which are also used to represent retrial queueing networks. The analysis by Fiems and Phung-Duc [60] is based on series expansion subject to the arrival rate around the origin and is validated by simulation. Thus, power series expansion may be useful for analyzing queueing networks with retrials.

3.7 Model with Orbital Search

Orbital search mechanism for customers in the orbit is introduced by Artalejo et al. [10], where upon service completion, with some probability the server can pick a job from the orbit with zero searching time. This mechanism is further extended in [49, 40, 84]. Chakravarthy et al. [33] investigate multiserver models where an idle server immediately picks up a customer from the orbit (zero searching time) with a probability or stays idle with the commentary probability. Recently, Dragieva and Phung-Duc [112, 48] propose a related model that the authors call the retrial queueing model with interaction between server and customers in the orbit. The main idea is that not only customers retry to capture an idle server (incoming calls) but the server also makes outgoing calls to retrial customers. The distributions of the durations of incoming calls and outgoing calls are different. Under M/M/1/1 settings, the authors in [48] obtain explicit solution for the generating functions of the joint queue length distribution and the stability condition is obtained for the M/M/c/c model.

3.8 Game Theoretic Analysis

Recently, game theoretic analysis of queues has been attracted much attention. Some authors study game theoretic analysis for retrial models. In particular, a series of works by Economou and Kanta [51, 53] provide detailed analysis for retrial models with constant retrial rate. Wang et al. [126, 127, 128] analyze also the models with classical retrial rate. To

be more precise, models in Wang et al. [127, 128] are devoted to strategic joining behavior of secondary users in cognitive networks. It should be noted that these references are devoted to M/M/1/1 type retrial queues only. Thus, the analyses of more general models might be promising future topics.

3.9 General Retrial Times

Most of researches assume the exponential retrial time. Only a few references are devoted to models with other retrial time distributions where each customer in the orbit acts independently of other customers [34, 117, 119]. In the current literature, there is not an exact analysis for this type of settings and only some approximations or simulations are presented. Thus, researches in this direction may be highly appreciated.

3.10 Other Performance Measures

In almost the work mentioned above, the main quantity of interest is the stationary queue length distribution. A few references are devoted to some new quantities of interest. In particular, channel idle period is analyzed by Artalejo and Gomez-Corral in [13]. Distributions of the successful and blocked events are studied in [5, 6, 7]. Maximum queue length in busy period is presented in [9] while that in a fixed time interval is analyzed by Gomez-Corral and Garcia [68]. Artalejo and Lopez-Herrero [17] analyze the distribution of the number of retrials of a tagged customer in M/G/1/1 and M/M/c/c retrial queues. Falin [54] obtain the distribution of the number of retrials for M/G/1/1 retrial queues. Dragieva [46, 47] studies the distribution of the number of retrials in model with finite source with arbitrary service time distribution. Gao et al. [63] obtain the distribution for number of retrials in an M/M/1/1 retrial queue with constant retrial rate and impatient customers.

4 Concluding Remarks

In this paper, we have surveyed the main theoretical results for retrial queueing models. We have also investigated retrial queueing models arising from real applications such as call centers, random access protocols, cellular networks etc. We hope that this paper can be served as a basic reference for researchers who want to enter and deepen this field. Because the retrial queue literature is rich, we also refer to some earlier survey papers [130, 56, 11,

66, 18], two books [57, 19] and the recent Special Issue [69]. Most of references in this paper is for continuous time retrial queues. We refer to Nobel [94] for a survey on results of discrete time retrial queues. Sections 1 and 2 are partially based on the dissertation of the author [100].

Acknowledgments

Tuan Phung-Duc was supported in part by JSPS KAKENHI Grant Number 26730011. The author would like to thank two anonymous reviewers whose comments greatly helped to improve the presentation of the paper. I would like to thank the Editors of the book, especially Professor Shoji Kasahara for giving me an opportunity to write this survey.

The author would like to devote this paper to the memory of Professor Jesus Artalejo who was a leading researcher in the fields of Queueing Theory and Mathematical Biology publishing more than one hundred papers and was a co-author and friend of the author of this paper.

References

[1] S. Aguir, F. Karaesmen, O. Z. Aksin and F. Chauvet. The impact of retrials on call center performance. *OR Spectrum*, **26** (2004), 353-376.

[2] A. S. Alfa and W. Li. PCS networks with correlated arrival process and retrial phenomenon. *IEEE Transactions on Wireless Communications*, **1** (2002), 630–637.

[3] K. Avrachenkov and E. Morozov. Stability analysis of GI/GI/c/K retrial queue with constant retrial rate. *Mathematical Methods of Operations Research*, **79** (2014), 273-291.

[4] K. Avrachenkov, E. Morozov and B. Steyaert. Sufficient stability conditions for multi-class constant retrial rate systems. *Queueing Systems*, **82** (2016), 149-171.

[5] J. Amador and J.R. Artalejo. On the distribution of the successful and blocked events in the M/M/c retrial queue: A computational approach. *Applied Mathematics and Computation*, **190** (2007), 1612-1626.

[6] J. Amador and J.R. Artalejo. The M/G/1 retrial queue: New descriptors of the customer's behavior. *Journal of Computational and Applied Mathematics*, **223** (2009), 15-26.

[7] J. Amador and J.R. Artalejo. Transient analysis of the successful and blocked events in retrial queues. *Telecommunications Systems*, **41** (2009), 255-265.

[8] V. V. Anisimov and J. R. Artalejo. Approximation of multiserver retrial queues by means of generalized truncated models. *Top*, **10** (2002), 51–66.

[9] J.R. Artalejo, A. Economou and M.J. Lopez-Herrero. Algorithmic analysis of the maximum queue length in a busy period for the M/M/c retrial queue. *INFORMS Journal on Computing*, **19** (2007), 121-126.

[10] J.R. Artalejo, V.C. Joshua and A. Krishnamoorthy An M/G/1 retrial queue with orbital search by the server. In J. R. Artalejo and A. Krishnamoorthy, Eds. *Advances in Stochastic Modeling*. Notable publications. Inc. New Jersey. 2002.

[11] J. R. Artalejo. A classified bibliography of research on retrial queues: Progress in 1990-1999. *Top*, **7**, (1999), 187-211.

[12] J. R. Artalejo, A. Gomez-Corral and M. F. Neuts. Analysis of multiserver queues with constant retrial rate. *European Journal of Operational Research*, **135** (2001), 569-581.

[13] J.R. Artalejo and A. Gomez-Corral. Channel idle periods in computer and telecommunication systems with customer retrials. *Telecommunication Systems*, **24** (2003), 29-46.

[14] J. R. Artalejo and M. Pozo. Numerical calculation of the stationary distribution of the main multiserver retrial queue. *Annals of Operations Research*, **116** (2002), 41-56.

[15] J. Artalejo and G. Falin. Standard and retrial queueing systems: a comparative analysis. *Revista matematica complutense*, **XV** (2002), 101-129.

[16] J.R. Artalejo and A. Economou. On the non-existence of product-form solutions for queueing networks with retrials. *Electronic Modeling*, **27** (2005), 13-19.

[17] J. R. Artalejo and M. J. Lopez-Herrero. On the distribution of the number of retrials. *Applied mathematical modelling*, **31** (2007), 478-489.

[18] J. R. Artalejo. Accessible bibliography on retrial queues: Progress in 2000-2009. *Mathematical and computer modelling*, **51** (2010), 1071-1081.

[19] J. R. Artalejo and A. Gomez-Corral. Retrial queueing systems: a computational approach. Berlin Heidelberg: Springer-Verlag, 2008.

[20] J. R. Artalejo and M. J. Lopez-Herrero. Cellular mobile networks with repeated calls operating in random environment. *Computers & Operations Research*, **37** (2010), 1158-1166.

[21] J. R. Artalejo and T. Phung-Duc. Markovian retrial queues with two way communication. *Journal of Industrial and Management Optimization*, **8** (2012), 781-806.

[22] J. R. Artalejo and T. Phung-Duc. Single server retrial queues with two way communication. *Applied Mathematical Modelling*, **37** (2013), 1811-1822.

[23] I. Atencia and P. Moreno. A single-server retrial queue with general retrial times and Bernoulli schedule. *Applied Mathematics and Computation*, **162** (2005), 855-880.

[24] K. Avrachenkov and U. Yechiali. Retrial networks with finite buffers and their application to internet data traffic. *Probability in the Engineering and Informational Sciences*, **22** (2008), 519-536.

[25] K. Avrachenkov and U. Yechiali. On tandem blocking queues with a common retrial queue. *Computers & Operations Research*, **37** (2010), 1174-1180.

[26] K. Avrachenkov, P. Nain and U. Yechiali. A retrial system with two input streams and two orbit queues. *Queueing Systems*, **77** (2014), 1-31.

[27] F. Avram, A. J. E. M. Janssen and J. S. H. Van Leeuwaarden. Loss systems with slow retrials in the Halfin-Whitt regime. *Advances in Applied Probability*, **45** (2013), 274-294.

[28] G. Bianchi. Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE Journal on selected areas in communications*, **18** (2000), 535-547.

[29] A. Murtuza Ali, O. J. Boxma J. Resing. Analysis and optimization of vacation and polling models with retrials. *Performance Evaluation*, **98** (2016), 52-69.

[30] L. Breuer, A. Dudin and V. Klimenok. A retrial BMAP/PH/N system. *Queueing Systems*, **40** (2002), 433-457.

[31] L. Bright and P.G. Taylor. Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. *Stochastic Models*, **11** (1995), 497-525.

[32] S. Bhulai and G. Koole. A queueing model for call blending in call centers. *IEEE Transactions on Automatic Control* , **48** (2003), 1434-1438.

[33] S.R. Chakravarthy, A. Krishnamoorthy, and V.C. Joshua, Analysis of a multi-server retrial queue with search of customers from the orbit, *Performance Evaluation*, **63** (2006), 776-798.

[34] S. R. Chakravarthy. Analysis of MAP/PH/c retrial queue with phase type retrials - simulation approach. In *Modern Probabilistic Methods for Analysis of Telecommunication Networks*, (2013), Springer Berlin Heidelberg, 37-49.

[35] B. D. Choi, Y. W. Shin and W. C. Ahn. Retrial queues with collision arising from unslotted CSMA/CD protocol. *Queueing Systems*, **11** (1992), 335-356.

[36] B. D. Choi and Y. Chang. Single server retrial queues with priority calls. *Mathematical and Computer Modelling*, **30** (1999), 7-32.

[37] B. D. Choi, Y. Chang and B. Kim. MAP1, MAP2/M/c retrial queue with guard channels and its application to cellular networks. *Top*, **7** (1999), 231-248.

[38] J.W. Cohen. Basic problems of telephone traffic theory and the influence of repeated calls, *Philips Telecommunication Review*, **18** (1957), 49-100.

[39] T. Dayar and M. Can Orhan. Steady-state analysis of a multiclass MAP/PH/c queue with acyclic PH retrials. *Journal of Applied Probability*. **53** (2016), 1098-1110.

[40] T. G. Deepak, A. N. Dudin, V. C. Joshua and A. Krishnamoorthy. On an M (X)/G/1 retrial system with two types of search of customers from the orbit. *Stochastic Analysis and Applications*, **31** (2003), 92-107.

[41] J. E. Diamond and A. S. Alfa. The MAP/PH/1 retrial queue. *Stochastic Models*, **14** (1998), 1151–1177.

[42] I. Dimitriou. A queueing model with two types of retrial customers and paired services. *Annals of Operations Research*, **238** (2016), 123-143.

[43] I. Dimitriou. A two class retrial system with coupled orbit queues. *Probability in the Engineering and Informational Sciences*.

[44] T. V. Do and R. Chakka. An efficient method to compute the rate matrix for retrial queues with large number of servers. *Applied Mathematics Letters*, **23** (2010), 638-643.

[45] T. V. Do. Solution for a retrial queueing problem in cellular networks with the fractional guard channel policy. *Mathematical and Computer Modelling*, **53** (2010), 2059-2066.

[46] V. I. Dragieva. A finite source retrial queue: number of retrials. *Communications in Statistics-Theory and Methods*, **42** (2013), 812-829.

[47] V. I. Dragieva. Number of retrials in a finite source retrial queue with unreliable server. *Asia-Pacific Journal of Operational Research*, **31** (2014), 1440005.

[48] V. Dragieva and T. Phung-Duc. Two-way communication M/M/1 retrial queue with server-orbit interaction. *Proceedings of The 11th International Conference on Queueing Theory and Network Applications (QTNA2016)*, (2016) 7 pages. ACM Digital Library.

[49] A. N. Dudin, A. Krishnamoorthy, V. C. Joshua and G. V. Tsarenkov (2004). Analysis of the BMAP/G/1 retrial system with search of customers from the orbit. *European Journal of Operational Research*, **157**(2004), 169-179.

[50] A. Dudin, M. Lee, O. Dudina and S. Lee. Analysis of priority retrial queue with many types of customers and servers reservation as a model of cognitive radio system. *IEEE Transactions on Communications*, **65**, (2017), 186-199.

[51] A. Economou and S. Kanta. Equilibrium balking strategies in the observable single-server queue with breakdowns and repairs. *Operations Research Letters*, **36** (2008), 696-699.

[52] A. Economou and M. J. Lopez- Herrero. Performance analysis of a cellular mobile network with retrials and guard channels using waiting and first passage time measures. *European Transactions on Telecommunications*, **20** (2009), 389-401.

[53] A. Economou and S. Kanta. Equilibrium customer strategies and social-profit maximization in the single- server constant retrial queue. *Naval Research Logistics*, **58** (2011), 107-122.

[54] G. I. Falin. On the waiting-time process in a single-line queue with repeated calls. *Journal of Applied Probability*, **23** (1986), 185-192.

[55] G. I. Falin. On a multiclass batch arrival retrial queue. *Advances in Applied Probability*, **20** (1988), 483-487.

[56] G. I. Falin. A survey of retrial queues. *Queueing Systems*, **7** (1990), 127-168.

[57] G. I. Falin and J. G. C. Templeton. Retrial Queues. Chapman & Hall, London, 1997.

[58] G. I. Falin, J. R. Artalejo and M. Martin. On the single server retrial queue with priority customers. *Queueing systems*, **14** (1993), 439-455.

[59] E. Fedorova. The second order asymptotic analysis under heavy load condition for retrial queueing system MMPP/M/1. *Communications in Computer and Information Science*, **564** (2015), 344-357.

[60] D. Fiems and T. Phung-Duc. Light-traffic analysis of queues with limited heterogenous retrials. In *Proceedings of 11th International Conference on Queueing Theory and Network Application (QTNA2016)*, 2016.

[61] D. Fiems, J. P. L. Dorsman and W. Rogiest. Analysing queueing behaviour in void-avoiding fibre-loop optical buffers. *Performance Evaluation*, **103** (2016), 23-40.

[62] N. Gans, G. Koole and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, **5** (2003), 79-141.

[63] S. Gao, X Niu and T. Li. Analysis of a constant retrial queue with joining strategy and impatient retrial customers. *Mathematical Problems in Engineering*, (2017), Article ID 9618215.

[64] A. Gomez-Corral. Stochastic analysis of a single server retrial queue with general retrial times. *Naval Research Logistics*, **46** (1999), 561-581.

[65] A. Gomez-Corral. A matrix-geometric approximation for tandem queues with blocking and repeated attempts. *Operations Research Letters*, **30** (2002), 360-374.

[66] A. Gomez-Corral. A bibliographical guide to the analysis of retrial queues through matrix analytic techniques. *Annals of Operations Research*, **141** (2006), 163-191.

[67] A. Gomez-Corral. On the applicability of the number of collisions in p-persistent CSMA/CD protocols. *Computers & Operations Research*, **37** (2010), 1199-1211.

[68] A. Gomez-Corral and M.L. Garcia. Maximum queue lengths during a fixed time interval in the M/M/c retrial queue. *Applied Mathematics and Computation*, **235** (2014), 124-136.

[69] A. Gomez-Corral and T. Phung-Duc (Eds.) Retrial queues and related models, *Annals of Operations Research*, **247** (2016).

[70] S. A.Grishechkin. Multiclass batch arrival retrial queues analyzed as branching processes with immigration *Queueing Systems*, **11** (1992), 395-418.

[71] T. Hanschke. Explicit formulas for the characteristics of the M/M/2/2 queue with repeated attempts. *Journal of Applied Probability*, **24** (1987), 486-494.

[72] K. Kajiwara and T. Phung-Duc. Multiserver queue with guard channel for priority and retrial customers. *International Journal of Stochastic Analysis*, 2016.

[73] C. S. Kim, S. H. Park, A. Dudin, V. Klimenok and G. Tsarenkov Investigation of the BMAP/G/1 $\rightarrow \cdot$ /PH/1/M tandem queue with retrials and losses. *Applied Mathematical Modelling* **34** (2010), 2926-2940

[74] C. S. Kim, V. Klimenok and O. Taramin. A tandem retrial queueing system with two Markovian flows and reservation of channels. *Computer & Operations Research*, **37** (2010), 1238-1246.

[75] C. Kim, V. I. Klimenok and A. N. Dudin. Analysis and optimzation of Guard Channel Policy in cellular mobile networks with account of retrials. *Computers & Operations Research*, **43** (2014), 181-190.

[76] J. Kim, J. Kim and B. Kim. Tail asymptotics of the queue size distribution in the M/M/m retrial queue. *Journal of Computational and Applied Mathematics*, **236** (2012), 3445-3460.

[77] J. Kim and B. Kim. Exact tail asymptotics for the M/M/m retrial queue with nonpersistent customers. *Operations Research Letters*, **40** (2012), 537-540.

[78] J. Kim, B. Kim and S. S. Ko. Tail asymptotics for the queue size distribution in an M/G/1 retrial queue. *Journal of Applied Probability*, **44** (2007), 1111-1118.

[79] V. Klimenok and A. Dudin. Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. *Queueing Systems*, **54** (2006), 245–259.

[80] Y. Konishi, H. Masuyama, S. Kasahara and Y. Takahashi. Performance analysis of dynamic spectrum handoff scheme with variable bandwidth demand of secondary users for cognitive radio networks. *Wireless Networks*, **19** (2013), 607-617.

[81] G. Koole and A. Mandelbaum. Queueing models of call centers: An introduction. *Annals of Operations Research*, **113** (2002), 41-59.

[82] V. G. Kulkarni Expected waiting times in a multiclass batch arrival retrial queue. *Journal of Applied Probability*, **23** (1986), 144-154.

[83] J. Kurose and K. Ross. Computer Networking: A Top-Down Approach. Pearson Education, Inc. 2010.

[84] A. Krishnamoorthy, T. G. Deepak and V. C. Joshua. An M/G/1 retrial queue with nonpersistent customers and orbital search. Stochastic Analysis and Applications, **23** (2005), 975-997.

[85] H. Li and Y.Q. Zhao. A retrial queue with a constant retrial rate, server downs and impatient customers. *Stochastic Models*, **21** (2005), 531-550.

[86] B. Liu and Y.Q. Zhao. Analyzing retrial queues by censoring. *Queueing Systems*, **64** (2010), 203-225.

[87] G. Latouche and V. Ramaswami. Introduction to matrix analytic methods in stochastic modelling. Philadelphia PA, 1999.

[88] E. Moutzoukis and C. Langaris. Two queues in a tandem with retrial customers. *Probability in the Engineering and Informational Sciences* **15** (2001), 311-325.

[89] E. Morozov. A multiserver retrial queue: regenerative stability analysis. *Queueing Systems*, **56** (2007), 157-168.

[90] E. Morozov and T. Phung-Duc. Stability analysis of a multiclass retrial system with classical retrial policy. *Performance Evaluation*, DOI: 10.1016/j.peva.2017.03.003 (2017).

[91] A. Nazarov and Y. Izmaylova. Asymptotic analysis retrial queueing system M/GI/1 with hyper exponential distribution of the delay time in the orbit and exclusion of alternative customers. *Communications in Computer and Information Science*. **638** (2016), 292-302.

[92] M. F. Neuts and B. M. Rao. Numerical investigation of a multiserver retrial model. *Queueing Systems*, **7** (1990), 169-190.

[93] M. F. Neuts. Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach. Johns Hopkins University Press, 1981.

[94] R. Nobel. Retrial queueing models in discrete time: a short survey of some late arrival models. *Annals of Operations Research*, **247** (2016), 37-63.

[95] C. E. M. Pearce. Extended continued fractions, recurrence relations and two-dimentional Markov processes. *Advances in Applied Probability*, **21** (1989), 357-375.

[96] T. Phung-Duc, H. Masuyama, S. Kasahara and Y. Takahashi. Performance analysis of optical burst switched networks with limited-range wavelength conversion, retransmission and burst segmentation. *Journal of the Operations Research Society of Japan*, **52** (2009), 58-74.

[97] T. Phung-Duc, H. Masuyama, S. Kasahara and Y. Takahashi. M/M/3/3 and M/M/4/4 retrial queues. *Journal of Industrial and Management Optimization*, **5** (2009), 431-451.

[98] T. Phung-Duc, H. Masuyama, S. Kasahara and Y. Takahashi. State-dependent M/M/c/$c+r$ retrial queues with Bernoulli abandonment. *Journal of Industrial and Management Optimization*, **6** (2010), 517-540.

[99] T. Phung-Duc, H. Masuyama, S. Kasahara and Y. Takahashi. A simple algorithm for the rate matrices of level-dependent QBD processes. *Proceedings of the 5th International Conference on Queueing Theory and Network Applications* (2010).

[100] T. Phung-Duc Retrial Queues and their Applications in Communication Systems. Doctor of Informatics, Department of Systems Science, Graduate School of Informatics, Kyoto University, Japan, March 2011.

[101] T. Phung-Duc, H. Masuyama, S. Kasahara and Y. Takahashi. A matrix continued fraction approach to multiserver retrial queues. *Annals of Operations Research*, **203** (2013), 161-183.

[102] T. Phung-Duc. An explicit solution for a tandem queue with retrials and losses. *Operational Research*, **12** (2012), pp. 189–207.

[103] T. Phung-Duc and K. Kawanishi. Multiserver retrial queues with after-call work. *Numerical Algebra, Control and Optimization*, **1** (2011), 639-656.

[104] T. Phung-Duc and K. Kawanishi. An efficient method for performance analysis of blended call centers with redial. *Asia-Pacific Journal of Operational Research*, **31** (2014), 1440008 [33 pages].

[105] T. Phung-Duc, K. Kawanishi. Performance analysis of call centers with abandonment, retrial and after-call work, *Performance Evaluation*, **80** (2014), pp. 43-62.

[106] T. Phung-Duc. Asymptotic analysis for Markovian queues with two types of non-persistent retrial customers. *Applied Mathematics and Computation*, **265** (2015), 768-784.

[107] T. Phung-Duc. M/M/1/1 Retrial Queues with Setup Time. *Advances in Intelligent Systems and Computing*, **383** (2015), 93-104.

[108] T. Phung-Duc. Analysis of an M/M/1 Retrial Queue with Speed Scaling. *Advances in Intelligent Systems and Computing*, **383** (2015), 113-124.

[109] T. Phung-Duc. Retrial Queue for Cloud Systems with Separated Processing and Storage Units. *Advances in Intelligent Systems and Computing*, **383** (2015), 143-151.

[110] T. Phung-Duc. Single server retrial queues with setup time. *Journal of Industrial and Management Optimization*, **13** (2017), 1329-1345.

[111] T. Phung-Duc, W. Rogiest and S. Wittevrongel. Single server retrial queues with speed scaling: Analysis and performance evaluation. *Journal of Industrial and Management Optimization*, (2017), DOI: 10.3934/jimo.2017025.

[112] T. Phung-Duc and V. Dragieva. Stability condition for a multiserver retrial queue with interaction between servers and orbit *to appear in AIP Conference Proceedings (2017).*

[113] T. Phung-Duc and K. Kawanishi. Impacts of retrials on power-saving policy in data centers. *Proceedings of The 11th International Conference on Queueing Theory and Network Applications (QTNA2016)* (2016), 4 pages, ACM Digital Library.

[114] T. Ramjee, D. Towsley and R. Nagarajan. On optimal call admission control in cellular networks. *Wireless Networks*, **3** (1997), 29-41.

[115] H. Sakurai and T. Phung-Duc. Scaling limits for single server retrial queues with two-way communication. *Annals of Operations Research*, **247** (2016), 229-256.

[116] W. Shang, L. Liu and Q. .L. Li. Tail asymptotics for the queue length in an M/G/1 retrial queue. *Queueing Systems*, **52** (2006), 193-198.

[117] Y. W. Shin and D. H. Moon. Approximation of M/M/c retrial queue with PH-retrial times. *European Journal of Operational Research*, **213** (2011), 205-209.

[118] Y. W. Shin and D. H. Moon. M/M/c Retrial queue with multiclass of customers. *Methodology and Computing in Applied Probability*, **16** (2014), 931-949.

[119] Y. W. Shin and D. H. Moon. Approximation of PH/PH/c retrial queue with PH-retrial time. *Asia-Pacific Journal of Operational Research*, **31** (2014), 1440010 [21 pages].

[120] Y. Song, Z. Liu and Y. Q. Zhao. Exact tail asymptotics: revisit of a retrial queue with two input streams and two orbits. *Annals of Operations Research*, **247** (2016), 97-120.

[121] R. Stolletz. Performance Analysis and Optimization of Inbound Call Centers. Springer-Verlag Berlin Heidelberg, 2003.

[122] O. Taramin A tandem queue with two Markovian inputs and retrial customers. *Computer Modelling and New Technologies*, **13** (2009), 38-47.

[123] P. Tran-Gia and M. Mandjes. Modeling of customer retrial phenomenon in cellular mobile networks. *IEEE Journal on Selected Areas in Communications*, **15** (1997), 1406-1414.

[124] O. Salameh, K. De Turck, H. Bruneel, C. Blondia and S. Wittevrongel. Analysis of secondary user performance in cognitive radio networks with reactive spectrum handoff. *Telecommunication Systems*, (2016). 1-12.

[125] J. Walraevens, D. Claeys and T. Phung-Duc. Asymptotics of queue length distributions in priority retrial queues. Preprint (2016).

[126] J. Wang and F. Zhang. Strategic joining in M/M/1 retrial queues. *European Journal of Operational Research*, **230** (2013), 76-87.

[127] F. Wang, J. Wang and W. Li. Game-theoretic analysis of opportunistic spectrum sharing with imperfect sensing. *EURASIP Journal on Wireless Communications and Networking*. (2016), 1-12.

[128] J. Wang and W. Li. Noncooperative and cooperative joining strategies in cognitive radio networks with random access. *IEEE Transactions on Vehicular Technology*, **65** (2016), 5624-5636.

[129] K. Yamamuro. The queue length in an M/G/1 batch arrival retrial queue. *Queueing Systems*, **70** (2012), 187-205.

[130] T. Yang and J. G. C. Templeton. A survey on retrial queues. *Queueing Systems*, **2** (1987), 201-233.

[131] T. Yang, M. Posner and J. G. C. Templeton. The M/G/1 retrial queue with nonpersistent customers. *Queueing Systems*, **7** (1990), 209-218.