Journal of Systems Scalability
http://dx.doi.org/10.62762/JSS.2025.500072

PERSPECTIVE

Retrial Queues: Scaling Limits

Tuan Phung-Duc[1,2,*]

[1] Department of Policy and Planning Sciences, Institute of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan
[2] Center for Artificial Intelligence Research (C-AIR), Tsukuba Institute for Advanced Research (TIAR), University of Tsukuba, Tsukuba, Ibaraki 305-8577, Japan

Abstract

Retrial queues arise in various applications such as call centers, services, and computer networks. The study of retrial queues is an important research branch of Queueing Theory. Retrial queues are characterized by the feature that customers who cannot receive service upon arrival do not queue but retry to enter the server after some random time. This makes the analysis of retrial queues more difficult than that of corresponding models without retrials. While the latter can be considered the limit of the former as the retrial time tends to infinity, some scaling limits are needed to obtain a scaled version of the number of retrial customers as the retrial time tends to zero, because the number of retrial customers explodes under this limiting regime. In this paper, a brief survey is provided on the major scaling limits of these models, and open problems are discussed. These limits can be used as approximations of complex systems with retrials.

Keywords: scaling limits, fluid limit, diffusion limit, heavy traffic, retrial queues, networks.

Submitted: 22 December 2025
Accepted: 28 April 2026
Published: 30 April 2026

Vol. 1, **No.** 2, 2026.
10.62762/JSS.2025.500072

***Corresponding author:**
✉ Tuan Phung-Duc
tuan@sk.tsukuba.ac.jp

1 Introduction

Retrial queues naturally arise in our daily life as well as in various systems such as call centers and computer communication systems. In traditional queues, customers who cannot receive service upon arrival wait at the queue and eventually are served. In retrial queues, customers who find the server busy upon arrival instead of joining the queue leave the system and retry after a random time. During this time, the customers stay in a virtual waiting line called the orbit. Those customers repeatedly try until they successfully find an empty server. Customers in the orbit retry independently of other customers, so the arrival rate from the orbit depends on the number of customers there. This dependence makes retrial queues analysis more challenging than counterpart models with a buffer in front of the server(s). The M/G/1 retrial queue and its variant models are solved, and their joint queue length distribution is given in terms of integrals [1]. For the simplest M/M/c/c models with retrials, it is challenging to obtain analytical solutions for the joint distributions of the number of customers in the orbit and the number of busy servers. Solutions for the case $c = 2$ are given in terms of a hypergeometric function, while those for $c = 3, 4$ are given in terms of continued fractions. For queueing networks with retrials, the results are scarce. It is known that these networks do not possess

Citation
Phung-Duc, T. (2026). Retrial Queues: Scaling Limits. *Journal of Systems Scalability*, 1(2), 39–42.

a product-form solution, and thus, analytical results are even more challenging.

The purpose of this paper is to clarify some classes of multidimensional retrial systems for which simple and elegant scaling limits are (or might be) available. We focus on the fluid/diffusion limits of retrial queues in the heavy-traffic regime, where the retrial rate is extremely low. In this case, the number of customers in the orbit explodes. However, if we scale this number appropriately, we can obtain the normal distribution in the stationary regime and the diffusion process in the non-stationary case. These results are then used for approximations of these models when the retrial rate is small.

For some related scaling regimes, we refer to [3] for the diffusion limit of a GI/G/1 retrial queue when the arrival rate, buffer size, and retrial rates are scaled. Furthermore, in [4] a diffusion limit is derived for a multiserver retrial queue with non-stationary input when the arrival rate is scaled.

2 Problem Statement and Existing Results

We review the results for M/M/c/c retrial queues under slow retrial regime. We assume that customers arrive at the system according to a Poisson process with rate λ. Service times of customers follow the exponential distribution with mean $1/\mu$. Blocked customers join the orbit and retry independently with rate σ. Let $C(t)$ and $N(t)$ denote the number of busy servers and that of customers in the orbit at time t, respectively. It is easy to see that $\{(C(t), N(t)); t \geq 0\}$ forms a Markov chain on the state space $\{0, 1, 2, \ldots, c\} \times \{0, 1, 2, \ldots\}$. It is shown in [2] for a more general model that, $\lim_{\sigma \to 0} \sigma N(t/\sigma) = x(t)$, where $x(t)$ is the solution of (1).

$$x'(t) = a(x), \tag{1}$$

where $a(x) = (\lambda + x)R_c(x) - x(1 - R_c(x))$ and $R_c(x)$ is the blocking probability of the M/M/c/c loss system with arrival rate $\lambda + x$ and x was used instead of $x(t)$, according to the context. It should be noted that the first term and the second term of $a(x)$ represent the flow of blocked customers entering the orbit and the retrial flow of customers from the orbit that successfully capture an idle server, respectively. Furthermore, we obtain that $\sqrt{\sigma}(N(t/\sigma) - x(t)/\sigma)$ weakly converges to a diffusion process $y(t)$ whose coefficients are expressed in terms of $a(x)$ and other known quantities [2]. In this paper, all the convergences mean weak convergence.

For (1), the existence of a unique equilibrium point κ such that $a(\kappa) = 0$ is guaranteed under the stability condition $\lambda < c\mu$ (Theorem 2.6 in [1]). An interesting observation is that the equation $a(x) = 0$ is equivalent to finding the unique positive root of a c-degree polynomial for which a simple explicit solution is available for the case $c = 1, 2$ and an explicit but complicated solution is available for $c = 3, 4$ and only a numerical solution for $c > 4$. This is similar to the M/M/c/c retrial queue with an arbitrary σ for which a simple solution is available for $c = 1, 2$, a continued fraction solution for $c = 3, 4$, and a generalized continued fraction solution for $c > 4$. From the uniqueness of κ, under $\sigma \to 0$, $C(t)$ and $N(t)$ are independent. Furthermore, $C(t)$ is asymptotically the number of busy servers in the M/M/c/c system with arrival rate $\lambda + \kappa$ and $N(t)$ asymptotically follows the Gaussian distribution with mean κ/σ and variance D/σ, where D is expressed in terms of κ and other known parameters [1]. Furthermore, $(\sigma N(t/\sigma) - \kappa)/\sqrt{\sigma}$ weakly converges to an Ornstein-Uhlenbeck (OU) process, as $\sigma \to 0$ [1].

3 Open problems and discussion

3.1 Models with one infinite dimension and several finite dimensions

In recent research, some results in Section 2 have been extended to the M/M/c/c retrial queue with setup times and related models [2]. These models are common in the sense that they are expressed in terms of a three dimensional Markov chain $\{(C_1(t), C_2(t), N(t)); t \geq 0\}$, where for example, $C_1(t)$ and $C_2(t)$ are the number of servers in setup and that in service, respectively. For these models, we have $\lim_{\sigma \to 0} \sigma N(t/\sigma) = x(t)$, where $x(t)$ satisfies

$$x'(t) = \lambda \mathrm{E}_1 R(x) - x \mathrm{E}_2 R(x), \tag{2}$$

where the matrix $R(x) = (R_{i,j}(x))$ ($R_{i,j}(x) = P(C_1(t) = i, C_2(t) = j)$) collects the stationary probabilities of Markov chain $(C_1(t), C_2(t))$ for the corresponding loss model without retrial with arrival rate $\lambda + x$. In (2), E_1 is the summing operator over the states of the servers such that arriving customers join the orbit, while E_2 is the summing operator over the states of the servers such that retrial customers successfully enter an idle server. See Figure 1 for the image of the interaction between servers and orbit.

Open problem 1: Can we obtain similar results (OU process as a scaling limit of the number of customers in the orbit) as for the M/M/c/c retrial queue in the stationary regime as presented in Section 2?

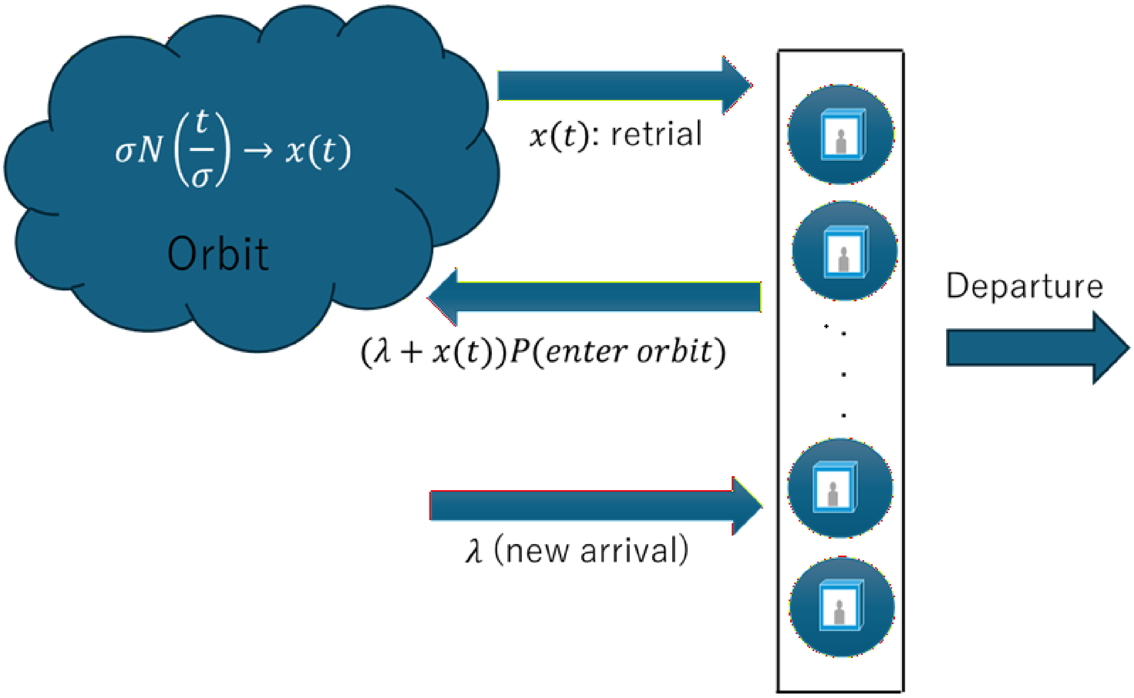

Figure 1. Interaction between servers and orbit.

3.2 Multidimensional models

In Section 2, we study retrial queues with only one orbit. It means that the underlying Markov chain has only one infinite dimension. Our research question is whether we can extend the results in Section 2 to multiple-orbit problems? As a first step, we consider the M/M/1/1 retrial queue with multiple customer classes. In this model, K classes of customers arrive at the system according to K independent Poisson processes with rates λ_i $(i = 1, 2, \dots, K)$. Service times for class-i customers are exponentially distributed with mean $1/\mu_i$. Blocked class-i customers join the orbit and retry with rate σ_i independently of other customers in the same class and those of different classes. For this model, we consider the scaling regime $\sigma_i = \gamma_i \sigma$ as $\sigma \to 0$.

In particular, denoting $N_i(t)$ $(i = 1, 2, \dots, K)$ the number of customers in the i-th orbit at time t, we prove that in the stationary regime $\sigma N_i(t) \to \kappa_i$ (for some explicit κ_i) and the vector of $\sqrt{\sigma}(N_i(t) - \kappa_i/\sigma)$ follows a multivariate Gaussian distribution as $\sigma \to 0$ in Nazarov et al. [5]. However, the non-stationary regime has not yet been investigated.

Open problem 2: Can this result be extended to the multiclass M/G/1/1 retrial queue with general service time distributions? Can we obtain similar results as in Section 2? Some partial results were recently obtained for the M/G/1/1 setting, where the authors determined the asymptotic marginal distribution of the number of customers in each orbit [6]. Results on the joint distribution of the number of customers across all orbits remain open to further exploration.

Open problem 3: Can we extend the results for the multiclass M/M/c/c retrial queues? Can we prove the uniqueness of κ_i, which is the limit of $\sigma N_i(t/\sigma)$ as $\sigma \to 0$ under the stationary regime and non-stationary regime? Can we have a diffusion limit for the multiclass M/M/c/c model with $c \geq 1$. Can we obtain similar results as mentioned in Section 2? Can we have the limit for the vector of $(\sigma N_i(t/\sigma) - \kappa_i)/\sqrt{\sigma}$? Similar to the case of multiclass retrial queues, queueing networks with retrials warrant further research. These results provide a foundation for the performance analysis of large-scale systems with retrials.

Data Availability Statement

Not applicable.

Funding

This work was supported by the Kayamori Foundation of Informational Science Advancement.

Conflicts of Interest

The author declares no conflicts of interest.

AI Use Statement

The author declares that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Falin, G. I. and Templeton, J.G.C. (1997). *Retrial Queues* (First ed.). Chapman & Hall, London. [Link]

[2] Moiseev, A., Nazarov, A., & Paul, S. (2020). Asymptotic diffusion analysis of multi-server retrial queue with hyper-exponential service. *Mathematics, 8*(4), 531. [CrossRef]

[3] Atar, R., & Lev-Ari, A. (2018). Optimizing buffer size for the retrial queue: two state space collapse results in heavy traffic. *Queueing Systems, 90*(3), 225-255. [CrossRef]

[4] Mandelbaum, A., Massey, W. A., Reiman, M. I., Stolyar, A., & Rider, B. (2002). Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems, 21*(2), 149-171. [CrossRef]

[5] Nazarov, A., Phung-Duc, T., & Izmailova, Y. (2020, September). Multidimensional central limit theorem of the multiclass M/M/1/1 retrial queue. In *International Conference on Distributed Computer and Communication Networks* (pp. 298-310). Cham: Springer International Publishing. [CrossRef]

[6] Nazarov, A. A., & Fedorova, E. A. (2025). Method of the Marginal Asymptotic-Diffusion Analysis for Multiclass Retrial Queue Mn/GIn. *AUTOMATION AND REMOTE CONTROL, 86*(3), 235-248.

Tuan Phung-Duc is a Professor at Institute of Systems and Information Engineering, University of Tsukuba. He received a Ph.D. in Informatics from Kyoto University in 2011. He is currently serving as Co-Editor-in-Chief for Queueing Models and Service Management and associate editor for three journals: International Transactions in Operational Research; Mathematics; Journal of Systems Scalability. He served as Guest Editor of four special issues of Annals of Operations Research. He was the Chairman of the 10th International Workshop on Retrial Queues (WRQ2014) and the TPC co-chair of several international conferences: ASMTA2016, QTNA2018, QTNA2019, EAI VALUETOOLS2026, General co-chair of EAI VALUETOOLS2020, General chair of ASMTA2021/EPEW2021. Dr. Phung-Duc received the Research Encourage Award from The Operations Research Society of Japan in 2013. His research interests include Applied Probability, Stochastic Models and their Applications in Performance Analysis of Telecommunication and Service Systems.