Queueing Systems Vol. 100
Special Issue: 100 views on queues

Metastability in Queues

Jing Dong

February 10, 2022

1 Introduction

We refer to metastability as the phenomenon where the many-server fluid limit of the queueing model has multiple asymptotically stable equilibria. For the fluid dynamical system, it will converge to one of the equilibria depending on the initialization. For the stochastic system, it will fluctuate in one equilibrium region for a long time before transitioning to another equilibrium region. This phenomenon arises in many communication, service, and healthcare applications [5,6]. To demonstrate the prevalence of the phenomenon, we start by giving two examples.

Example 1 (Service slowdown) Slowdown of service rate when the system is congested is a widely observed phenomenon. This can be caused by staff fatigue, deterioration of health condition due to delayed treatment in healthcare, or other physiological or psychological reasons [7,1]. We note that if the customers require longer service times when the system is congested, this can cause a snowball effect, leading to more severe congestions. Following [3], we consider a fluid model with arrival rate λ, abandonment rate θ, and s servers. The service rate is state-dependent, i.e., $\mu(q) = \mu_0 + \delta \exp(-b(q-s)^+/s)$. In particular, the fluid dynamics takes the form

$$dq(t)/dt = f(q(t)) := \lambda - (q(t)-s)^+\theta - (q(t)\wedge s)\mu(q(t)).$$

In certain parameter regimes, the system has two stable equilibria: one associated with good performance (zero queue); the other with bad performance (large queue and and a high level of abandonment). Figure 1 demonstrates the flow rate of the fluid model and a sample path of a stochastic system with bi-stability. Note that the fluid model has two asymptotically stable equilibria: 100 and 181 (The equilibrium 129 is unstable). For the stochastic system, it evolves in two equilibrium regions: one around 100×5 and the other around 181×5. The transition from one region to the other is random.

Jing Dong
Decision, Risk, and Operations Division, Graduate School of Business, Columbia University, USA. E-mail: jing.dong@gsb.columbia.edu

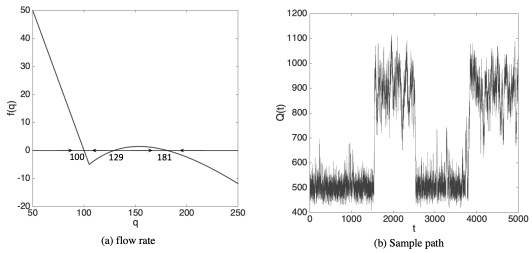

Fig. 1 Flow rate and sample path of a bi-stable multi-server queue with load-dependent service rate ($\lambda = 100$, $s = 105$, $\mu_0 = 0.6$, $\delta = 0.4$, $b = 1.5$. For the stochastic system, we scale λ and s up by 5.)

Example 2 (Service speedup with returns) In many service systems, faster service rate can be applied when the system is congested (e.g., early discharge in a congested hospital ward). Although speeding up during periods of congestion may address a present congestion issue, it may exacerbate the problem by increasing the need for rework in the future [2]. In [2], a modified Erlang-R model in which both the service rates and the return rates are state-dependent are analyzed. They show that the system can exhibit bi-stability in certain parameter regimes.

For a stochastic system that exhibits metastability, its stationary distribution (if exists) can have multiple modes. Estimating metamodal distributions can be challenging and the underlying stochastic process can have a slow convergence rate to stationarity. For example, consider a simple birth-and-death process on $\{0,1,2\}$. The state-dependent birth and death rates are $\lambda_0 = \varepsilon$, $\lambda_1 = 1$, $\lambda_2 = 0$, and $\mu_0 = 0$, $\mu_1 = 1$, $\mu_2 = \varepsilon$, respectively. For small ε, the stationarity distribution is highly concentrated on 0 and 2. The asymptotic variance (also known as the time-averaged variance constant) is $2/(\varepsilon(2+\varepsilon))$ [9], which can be arbitrarily large when ε is small.

2 Problem statement

Theoretical performance analysis Metastability has given rise to a growing literature in statistical physics and theoretical probability [8]. The large deviation theory has been one of the main tools for theoretical analysis of these systems [4]. It can be used to characterize the exit probability and exit path from an equilibrium regime, which is a rare event when the random perturbation is small.

For performance analysis of queues that exhibit metastability, we have a relatively good understanding of its fluid dynamics, especially when the process dimension is low. However, *how to conduct meaningful and rigorous diffusion-scale analysis* remains quite open. One possible direction is to develop a diffusion approximation around each equilibrium and use the large deviation theory to quantify transitions between equilibria.

System design and control In queueing applications, different equilibria can correspond to vastly different system performance. Even though in the fluid model, we can ensure that the system converges to a good equilibrium by proper initialization, in the stochastic

system, the system can move to any of the equilibrium regions due to random fluctuations. Two approaches can be taken to ensure good system performance: 1) Design the system in an appropriate way to eliminate metastability; 2) Introduce effective controls to prevent the system moving to the "bad" equilibria.

To demonstrate the two approaches, we revisit the example in Figure 1. For approach 1), increasing the service capacity can completely eliminate the second equilibrium. However, the system will have a lower server utilization around the remaining equilibrium. For approach 2), we can implement an admission control policy, where we block the incoming customers when the queue reaches a certain threshold. As demonstrated in [3], with a properly chosen threshold, we only need to block a small proportion of customers while maintaining the system around the good equilibrium. Intuitively, this is because once $q < 129$ in Figure 1(a), the flow rate will drive the system to the good equilibrium. If blocking is not feasible, we can also consider to temporarily increase the capacity when the queue reaches a certain threshold. If we have such flexibility, this can lead to a lower staffing cost than the proposed approach 1. More generally, *how to design the system or control to avoid equilibria with bad performance* is of both theoretical interests and practical relevance.

3 Discussion

We only give two simple examples where metastability arises. When taking customer and/or agent's strategic behavior into account, there can be more complicated systems with metastability, in which even characterizing the mean-field equilibria can be challenging. In addition, many service/communication systems also experience highly time-varying demand. Defining a proper notion of metastability (e.g., multiple periodic equilibria) in time-varying systems is a challenge on its own. In summary, *identifying metastability, conducting meaningful performance analysis, and designing effective control policies for these systems* are all worth exploring.

References

1. C. W. Chan, V. F. Farias, and G. J. Escobar. The impact of delays on service times in the intensive care unit. *Management Science*, 63(7):2049–2072, 2017.
2. C. W. Chan, G. Yom-Tov, and G. Escobar. When to use speedup: An examination of service systems with returns. *Operations Research*, 62(2):462–482, 2014.
3. J. Dong, P. Feldman, and G. B. Yom-Tov. Service systems with slowdowns: Potential failures and proposed solutions. *Operations Research*, 63(2):305–324, 2015.
4. M. I. Freidlin and A. D. Wentzell. *Random Perturbations of Dynamical Systems*, volume 260. Springer Science & Business Media, 2012.
5. R. Gibbens, P. Hunt, and F. Kelly. Bistability in communication networks. *Disorder in Physical Systems*, pages 113–128, 1990.
6. Y. Hu, C. W. Chan, and J. Dong. Optimal scheduling of proactive service with customer deterioration and improvement. *Management Science*, 2021.
7. D. S. Kc and C. Terwiesch. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science*, 55(9):1486–1498, 2009.
8. E. Olivieri and M. E. Vares. *Large deviations and metastability*. Cambridge University Press, 2005.
9. W. Whitt. Asymptotic formulas for Markov processes with applications to simulation. *Operations Research*, 40(2):279–291, 1992.