Queueing Systems (2022) 100:169–171
https://doi.org/10.1007/s11134-022-09785-4

Queueing theory: past, present, and future

Peter W. Glynn[1]

Received: 6 February 2022 / Accepted: 28 February 2022 / Published online: 8 April 2022

As QUESTA celebrates its 100'th volume, now is an opportune time to reflect upon the past, present, and potential future directions of the field. Queueing theory, broadly speaking, concerns the design and analysis of resource-constrained systems in which customers need to potentially wait for access to a resource. This paradigm arises in many applied settings, ranging from service and health operations to communications and computer systems to ride-sharing and job-matching platforms to organ exchange marketplaces. In view of the pervasiveness of queueing-type phenomena in applied settings, one can reasonably expect that queueing issues will continue to be studied for a long time to come, as they continue to arise as a first-order consideration in both ongoing and new problem environments.

From a theory perspective, there are three different major modeling perspectives that can shape a given research contribution. The first is that of descriptive modeling, in which one may build a stylized model that is intended to shed qualitative insight into a queueing phenomenon of interest. A good example of this is the $M/M/1$ queue, which while too simple to expect good predictive power, carries valuable insights into the scaling behavior of congested systems when one manages the system so as to minimize idle time. As a second example of such theory, much of the literature on behavior of queues with light-tailed and heavy-tailed service times is fundamentally oriented towards providing qualitative understanding of the differences in how long customer delays develop in queueing-type systems, rather than as a vehicle for predicting delays in real systems.

The second perspective is that of prescriptive modeling, in which one focuses on optimal design and control of queueing systems. In various simplified settings, in which the incoming traffic to the system is time-stationary and Poisson or Brownian, one can identify priority (e.g. the $c - \mu$ rule) or threshold policies (as in admission control settings) that are optimal. This type of work can provide actionable insight into how real-world systems should be designed, because one can use these structured policies as a starting point for constructing well-behaved implementable policies having parameters that can then be tuned to the real-world environment.

✉ Peter W. Glynn
glynn@stanford.edu

[1] Management Science and Engineering, Stanford University, Stanford, CA, USA

The final major perspective concerns predictive modeling. In this setting, model fidelity and accuracy is a central issue, and the efficient integration of real-world data into the prediction methodology is typically important. The Queueing Network Analyzer [3] was likely developed in this spirit, at least in view of some of its applications. Similarly, simulation model-building is often guided by a desire to build and solve models that can exhibit enough model fidelity with the real world to quantitatively capture the key performance characteristics of the associated real-world environment.

Most of the historical literature on queues has focused on descriptive and prescriptive theory, in which simplified caricatures of real systems are studied. Furthermore, the insights have been largely generated through mathematical rather than computational analysis. Early algorithmic approaches to queueing theory included matrix-geometric methodology [2], numerical schemes for computing normalization constants for product-form closed networks (e.g., [1]), and use of simulation modeling as a vehicle to the analysis of systems.

Numerical methods for studying queues have attracted increasing academic interest over the last couple of decades. Nevertheless, compared to many other areas of applied mathematics, it is probably fair to say that computational methods are still relatively under-utilized by the queueing community. This likely has to do both with the dominant focus being on acquisition of descriptive and prescriptive insights, and with the fact that since queues describe man-made systems, the dominant behavior is not governed by physical laws. As a consequence, the intrinsic model error tends to be larger for today's queueing models than for many of the models that arise from physics-based systems. In such contexts, even high-accuracy numerics may not give strong agreement with the real system under consideration.

Today, machine learning is having a major impact across many scientific, technical, and business domains. A major question facing queueing theory, as with virtually every other academic discipline, concerns the degree to which the subject will be impacted in the future by machine learning. It seems fair to say that stylized models, in the service of acquiring descriptive and prescriptive insight, will continue to play a central role as queueing theory is extended to cover future technologies and systems. But machine learning may compete favorably in settings which demand strong predictive performance.

In building and managing systems, there can be a need both for predicting performance in planning and designing systems that will be deployed perhaps months or years in the future, and also for predicting performance over much shorter time scales of minutes or hours. One may have no data or very little data available from which to build reliable models in the planning context, so this may limit machine learning's applicability in such settings. However, in ride-sharing contexts, call centers, computing environments, and many others, large amounts of data can be collected. In such applications, there will be an opportunity to customize machine learning to the queueing context and to determine those settings in which such methodology can compete effectively with queueing-based predictive methods that fully incorporate the physical and economic principles that govern such systems.

To the extent that queueing-based predictive methodology will successfully compete, there will be a need for researchers to identify the statistical features in observed data that heavily impact queue performance, and to develop calibration procedures for models that can flexibly represent such features. Furthermore, one will need computational methods that can effectively analyze high fidelity queueing models fed by such input streams. In many applied settings, there will be an interest in using such models to optimize various operational choices (e.g., the number of vehicles to be made available in the next half hour on a ride-sharing platform), so the model used will then also need to be tractable from an optimization viewpoint. This will create a need for the queueing community to take a view of modeling that values statistical and computational tractability as being as desirable as has been mathematical tractability historically.

An example of a field that involves a man-made system in which statistical and computational questions play as large a role as does stylized modeling is that of quantitative finance. In the finance setting, building models that can be efficiently calibrated and priced has carried significant value both academically and in practice for many years, and as a consequence, the mix of papers published there looks quite different than is the case within the queueing community. It is also the case that the use of machine learning is rapidly developing within the finance sector. Whether this style of research becomes more dominant within the queueing community in the years to come will depend in large part on the extent to which predictive queueing applications become more important in practice, and on the degree to which the queueing community embraces this trend.

In conclusion, given the ever-broadening applicability of queues and the new possibilities that greater data collection and ubiquitous computing will create, queueing theory has an exciting and impactful future lying ahead of it.

References

1. Buzen, J.P.: Computational algorithms for closed queueing networks with exponential servers. Commun. ACM **16**(9), 527–531 (1973)
2. Neuts, M.F.: Matrix-geometric Solutions in Stochastic Models: An Algorithmic Approach. The Johns Hopkins University Press, Baltimore (1981)
3. Whitt, W.: The queueing network analyzer. Bell Syst. Tech. J. **62**(9), 2779–2815 (1983)