Queueing Systems (2022) 100:481–483
https://doi.org/10.1007/s11134-022-09759-6

Optimization of time-dependent queueing systems

Raik Stolletz[1]

Received: 4 February 2022 / Accepted: 28 February 2022 / Published online: 27 March 2022

1 Introduction

A variety of decisions in service operations, manufacturing, and logistics are supported using stationary queueing models. However, parameters of those models can be time-dependent. For example, the arrival rate and the staffing level significantly vary over the day in call centres. In manufacturing, additional production lines are put into operation or workers gain experience over time. Another application is the adjustment of buffer capacities in between machines, for example via dynamic KANBAN control.

Such systems can be modelled as time-dependent queues where one or several parameters change over time. Exact analytical approaches for performance analysis are known under certain assumptions. However, approximation methods are often applied to analyse performance measures over time, see the survey [7]. Even though time-dependent performance analysis is challenging by itself, an additional fruitful direction of research is the optimization of parameters of time-dependent queueing systems.

2 Problem statement

A steady state cannot be reached in queues with (i) time-dependent demand, (ii) time-dependent capacities, or (iii) changing queueing configurations. For example, customer demand or order releases may change over time (varying arrival rates). Capacity of a station may vary based on fatigue of personnel, learning behaviour, changing machine status, or a varying number of parallel servers (varying service rates). The queue configuration may change because of dynamic routing policies or changes in the number of buffer spaces (waiting space). Those sources of dynamics lead to one or several time-dependent parameters describing the queueing system. Performance measures of interest are also time-dependent or are aggregated over a longer planning horizon.

Several methods for time-dependent performance analysis are developed for systems with different assumptions, see the classification in [7]. Decisions in time-dependent queues may be related to any parameter of the original queueing system and can be classified according to

✉ Raik Stolletz
stolletz@bwl.uni-mannheim.de

[1] Business School, University of Mannheim, Mannheim, Germany

(i) Demand and arrival management: Decisions on the arrival rate are about releases into the queue, for example, in appointment scheduling, in order releases planning for manufacturing, or in acceptance of demand within revenue management.
(ii) Queue configurations: Decisions on routing policies are affected by the time-dependent evolvement of parameters. The size of the waiting space may also be a time-dependent decision.
(iii) Server capacities: The optimization of server capacities includes decisions on the number of servers and/or decisions on the server characteristics, for example, the skills or the processing rate.

Both static decisions over the entire planning horizon and time-dependent decisions are possible.

3 Discussion

Methods for the optimization of time-dependent queueing systems are manifold and often based on heuristics. The methods can be classified as follows:

Exact approaches: Optimization problems can be formulated as Markov decision problems under certain assumptions and can be used to prove insights on the structure of the solution, see, for example, [6]. However, they are often solved heuristically because of the state-space explosion. For example, approximate dynamic programming techniques or new methods like reinforcement learning are applied, see [4].

Sequential approaches: Sequential approaches separate the queueing from the optimization part. For example, resource requirements per period are derived from a queueing model. They then serve as constraints in deterministic shift scheduling, see, for example, [2].

Iterative approaches: In contrast to sequential approaches, such methods iterate between an evaluation part and an optimization part to get a new candidate solution, see, for example, [1]. Queueing approaches or simulation evaluates the candidate solution. Low-fidelity high-fidelity approaches may refine the evaluation part over the iterations to speed up the procedure. The optimization part uses the evaluated former solution(s) and often applies heuristic procedures.

Integrated approaches: Integrated approaches transform analytical evaluations of queueing systems into optimization models. However, the resulting optimization problems can be linear or nonlinear mathematical models which are difficult to solve with exact methods, see, for example, [8].

Stochastic programming: Stochastic programming replaces probability distributions by samples or scenarios and can handle time-dependent parameters or decisions, see, for example, [3]. However, it results in deterministic optimization problems which can be huge based on the number of considered samples.

A statistical or machine learning model can also be applied, as a surrogate model, for performance evaluation within the sequential, iterative, or integrated approaches, see, for example, [5].

When sketching a road map for further research on the optimization of time-dependent queueing systems, one has to consider the queueing models, the solutions methods, and new managerial insights.

Certain applications already got attention with respect to time-dependent queueing models, for example staffing and shift scheduling or appointment scheduling. For new and interesting models, traditional optimization models in stationary systems can be extended to relevant time-dependent parameters or decisions. In particular, complex systems with multiple stages, heterogeneous job characteristics, or general distributions deserve larger research efforts.

From a methodological point of view, further analytical results and additional numerical methods are of high interest. In addition, a structured comparison of methods can give guidelines to which method is suitable for which type of system, for example dependent on the load, the size of the system, or the type of decision variables.

Managerial insights about effects of time-dependent parameters or decisions together with stochastic assumptions should be investigated. For example, [8] describe and explain the inverse capacity boost when deciding on the time-dependent processing rate. The optimized processing rate may increase above the stationary optimum shortly before a demand decrease. Future research should analytically discuss those effects in addition to numerical investigations.

To summarize, further research should show the impact of solving optimization problems in time-dependent queues. Exact methods or reliable heuristics should be developed or extended such that they can be applied to larger problems or complex systems. Analytical methods are necessary to derive general insights in the structure of the solution. New effects in the structure of the solutions and managerial insights can be shown numerically and founded analytically.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Atlason, J., Epelman, M.A., Henderson, S.G.: Optimizing call center staffing using simulation and analytic center cutting-plane methods. Management Science **54**(2), 295–309 (2008)
2. Defraeye, M., Van Nieuwenhuyse, I.: Staffing and scheduling under nonstationary demand for service: A literature review. Omega **58**, 4–25 (2016)
3. Grass, E., Fischer, K.: Two-stage stochastic programming in disaster management: A literature survey. Surveys in Operations Research and Management Science **21**(2), 85–100 (2016)
4. Koole, G.: An Introduction to Business Analytics. Lulu. com, (2019)
5. Li, S., Wang, Q., Koole, G.: Predicting call center performance with machine learning. In INFORMS International Conference on Service Science, pages 193–199. Springer, (2018)
6. Roubos, A., Bhulai, S., Koole, G.: Flexible staffing for call centers with non-stationary arrival rates. In: Markov decision processes in practice, pages 487–503. Springer, (2017)
7. Schwarz, J.A., Selinka, G., Stolletz, R.: Performance analysis of time-dependent queueing systems: Survey and classification. Omega **63**, 170–189 (2016)
8. Vogel, J., Stolletz, R.: Anticipation of future demand changes in capacity planning: The inverse capacity boost. Available at SSRN, (2021)