Editorial introduction: special issue on product forms, stochastic matching, and redundancy

Kristen Gardner, Pascal Moyal

Noname manuscript No.
(will be inserted by the editor)

Editorial Introduction: Special Issue on Product Forms, Stochastic Matching, and Redundancy

Kristen Gardner · Pascal Moyal

Received: date / Accepted: date

In the analysis of queueing systems, one's goal typically is to derive closed-form expressions for performance metrics of interest, e.g., the distribution of the number of jobs in the system or of response time. Often, one assumes that the system under consideration has Markovian arrival and service processes, thus enabling the use of Markov chain analysis. The vast majority of the time, however, the associated Markov chain is complicated and does not admit a straightforward analysis. Consequently, a wealth of techniques have been developed to solve intricate Markov chains, either exactly or in approximation. Of particular interest to us in this special collection of papers in *Queueing Systems* are processes that yield a so-called "product-form" stationary distribution.

Historically, the notion of a product-form stationary distribution dates back to the 1960s, referring to results on networks of queues where it is shown that the state of each individual queue within the network is independent of the states of all other queues. Here, "product form" refers to the fact that the stationary distribution of the network state as a whole can be expressed as a product of the marginal stationary distributions for the individual queues. Such networks include those developed by Jackson [31], Kelly [33,34], and Whittle [47,48]; BCMP networks generalize these earlier results, while still preserving the product-form stationary distribution [9]. The above results all were derived using the "local balance" technique [19,47], wherein the global balance equation for each state is decomposed into several independent "local" equations, and it is shown separately that the desired form for the stationary distribution satisfies each local equation. The concept of local balance is closely tied to that of quasi-reversibility: any system that is quasi-reversible will have a stationary distribution that satisfies (some form of) local balance. Interestingly, however, not all product-form systems are quasi-reversible. For example, G-networks, which allow for "signals" and "negative customers" when jobs are routed within the network, admit a product-form stationary distribution but are not quasi-reversible [27,28].

More recently, the term "product-form" has been used to refer to the situation where the stationary distribution of an *individual* queue can be expressed as a product of terms, with one term corresponding to each job in the system. This type of product-form result came to prominence with the seminal work of Krzesinski on Order Independent (OI) queues [11,12,35,36]. The OI queue is described by states of the form $\mathbf{c} = (c_1, \ldots, c_n)$, where there are n jobs in the system, listed in order of arrival, so that c_1 is the class of the oldest job and c_n the most recent arrival. The service process is such that (1) the *total* service rate allocated to the first i jobs, $\mu(c_1, \ldots, c_i)$, can depend only on the classes of those jobs, and not on their order, (2) the service rate provided to the ith job in the queue, $\Delta_i\mu(c_1, \ldots, c_i) := \mu(c_1, \ldots, c_i) - \mu(c_1, \ldots, c_{i-1})$, can depend only on jobs 1 through i (and not on any jobs that arrived later than job i), and (3) a strictly positive service rate is allocated to the first job in the queue, i.e., $\mu(c) > 0$ for any class c. The key result of Krzesinski is that the stationary distribution of the OI queue is given by

$$\pi(\mathbf{c}) = C \cdot \prod_{i=1}^{n} \frac{\lambda_i}{\mu(c_1, \ldots, c_i)}, \tag{1}$$

K. Gardner
Department of Computer Science, Amherst College
E-mail: kgardner@amherst.edu

P. Moyal
Université de Lorraine / Inria PASTA
E-mail: pascal.moyal@univ-lorraine.fr

where C is a normalizing constant.

Product forms have been of particular interest in recent years because they emerge in the analysis of two important application domains: redundancy systems and stochastic matching models. In a *redundancy system*, when a job arrives it joins the queue at all servers with which it is compatible, where a bipartite graph specifies the job-server compatibilities. One can thus think of the job as having "copies" present in multiple servers' queues; the extra copies are immediately removed from the system when the first copy either enters service (cancel-on-start) or completes service (cancel-on-completion). The redundancy system with cancel-on-completion was shown to have a product-form stationary distribution in [26]; it was later shown that this result for redundancy systems follows directly from the product-form for OI queues [13]. In *skill-based service systems*, each arriving customer is assigned to some server at which it is processed; the matchings between customers and servers are again subject to compatibility constraints given by a bipartite graph. Under the so-called FCFS-ALIS (First-Come, First Served - Assign the Longest Idling Server) matching rule, such a system has a product-form stationary distribution [3,17,46]. By *stochastic matching system* or *matching queue*, we refer to an extension of skill-based service systems in which items leave the system as soon as they are matched. The compatibility structure between classes of items may again be a bipartite graph, to account for settings in which items represent customers/servers (or demands/supplies, donors/receivers, and so on). This is the case e.g. in the seminal papers [1,16], in which it is assumed that customers and servers enter and depart the system in pairs. In [1], quasi-reversibility arguments are used to show that the stationary distribution has a product form in the particular case where matching occurs according to FCFM (First Come, First Matched). In [16], policies other than FCFM are addressed. To account for applications such as kidney exchange programs, dating websites or assemble-to-order systems, this class of matching systems has naturally been extended to systems in which arrivals are single (not pairwise), and the compatibility structure is not necessarily a bipartite graph; such extensions include a simple graph [20,21,37–39], a graph with self-loops [10,15], or an hypergraph (thereby gathering groups of elements of cardinality more than two) [29,40,44]. These settings have also been extended to systems with reneging, as in [6,32], for instance. Many of these works address matching systems beyond the FCFM matching policy; it is significant that the stationary distribution then fails to have a product form, and is in general not even known explicitly. Several works seek to identify connections between the redundancy and stochastic matching models and their associated results, see e.g., [2,7,8,25].

Despite the wealth of literature that focuses on deriving product-form stationary distributions in matching and redundancy systems, the work of understanding performance in these systems does not end with the product-form results. In particular, the state space needed to model these systems is complicated; obtaining performance metrics of interest such as the distribution of the number of jobs in the system requires aggregating a combinatorially explosive number of states. In a few special cases in which the compatibility graph is highly structured, exact closed-form analysis of the per-class number of jobs in the system is feasible [14,22,24]. However, much of the ongoing work in the space of redundancy and stochastic matching systems focuses on developing approaches that leverage—or go beyond—the product form stationary distribution to gain insight into the system performance. One approach is to use the product-form stationary distribution as a starting point for light-traffic or heavy-traffic analysis [18, 45]. In the space of redundancy systems, several works aim to characterize the stability region and understand how it changes as a function of the scheduling policy and correlation among the copies' service requirements [4,42,41], see also [5] and the references therein. Other work explores the implications of relaxing the assumption that service times are exponentially distributed [23,30,43].

This special collection of papers in *Queueing Systems* aims to bring into conversation the streams of literature focusing on stochastic matching systems and redundancy systems, with the goal of highlighting the connections between these models, including (but not limited to) product form results. The first part of the special collection contains four papers. In "A fluid approximation for a matching model with general reneging distributions" (by A. Avecklouris, A.L. Puha and A.R. Ward), a fluid approximation is obtained for a two-sided matching queue with reneging, in which the reneging time distribution is general, i.e. not necessarily exponential, thereby leading to a natural measure-valued representation. In "Heavy traffic analysis of multi-class bipartite queueing systems under FCFS" (by L.A. Hillas, R. Caldentey and V. Gupta), a heavy-traffic analysis of a bipartite queueing system under the FCFS-ALIS discipline is undertaken to characterize the steady state distribution of the waiting times of the customer classes. Furthermore, it is shown that the asymptotic matching probabilities of the various classes is insensitive to the direction along which the system approaches heavy traffic. In "Multi-component Matching Queues in Heavy Traffic" (by B. Xie), another heavy traffic analysis is proposed, this time for a matching queue of K types of perishable components, with applications to assemble-to-order systems. The heavy-traffic

limit is characterized by coupled stochastic integral equations, allowing, among other interesting features, for the derivation of an "asymptotic Little's Law" for each queue. In "Efficient scheduling in redundancy systems with general service times" (by E. Anton, R. Righter, and I.M. Verloop), two-level scheduling policies for redundancy systems are considered wherein the first-level policy determines which job class is served, and the second-level policy selects a job within the chosen class. It is shown that prioritizing job classes with a lower degree of redundancy is preferable when the job size distribution is New-Better-Than-Used, while the opposite is true for New-Worse-Than-Used distributions.

While the four papers in this first part of the special collection each focus on a particular system—the first three on a stochastic matching system, and the fourth on a redundancy system—we draw attention to the fact that the close connections between these models mean that insights obtained in one model are likely applicable to the other. We hope, therefore, that grouping these papers together in a special collection will inspire new directions for future work. A second part of the special collection will follow.

References

1. I. Adan, A. Bušić, J. Mairesse, and G. Weiss. Reversibility and further properties of fcfs infinite bipartite matching. *Mathematics of Operations Research*, 43(2):598–621, 2018.
2. I. Adan, I. Kleiner, R. Righter, and G. Weiss. FCFS parallel service systems and matching models. *Performance Evaluation*, 127:253–272, 2018.
3. I. Adan and G. Weiss. A skill based parallel service system under FCFS-ALIS—steady state, overloads, and abandonments. *Stochastic Systems*, 4(1):250–299, 2014.
4. E. Anton, U. Ayesta, M. Jonckheere, and I. M. Verloop. On the stability of redundancy models. *Operations Research*, 69(5):1540–1565, 2021.
5. E. Anton, U. Ayesta, M. Jonckheere, and I. M. Verloop. A survey of stability results for redundancy systems. In *Modern Trends in Controlled Stochastic Processes: Theory and Applications, V. III*, pages 266–283. Springer, 2021.
6. A. Aveklouris, L. DeValve, M. Stock, and A. R. Ward. Matching impatient and heterogeneous demand and supply. *arXiv preprint arXiv:2102.02710*, 2021.
7. U. Ayesta, T. Bodas, J. Dorsman, and M. Verloop. A token-based central queue with order-independent service rates. *Operations Research*, 70(1):545–561, 2022.
8. U. Ayesta, T. Bodas, and M. Verloop. On a unifying product form framework for redundancy models. *Performance Evaluation*, 127:93–119, 2018.
9. F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *Journal of the ACM (JACM)*, 22(2):248–260, 1975.
10. J. Begeot, I. Marcovici, P. Moyal, and Y. Rahme. A general stochastic matching model on multigraphs. *ALEA: Latin American Journal of Probability and Mathematical Statistics*, 18(2):1325–1351, 2021.
11. S. Berezner, C. Kriel, and A. Krzesinski. Quasi-reversible multiclass queues with order independent departure rates. *Queueing Systems*, 19:345–359, 1995.
12. S. Berezner and A. Krzesinski. Order independent loss queues. *Queueing Systems*, 23(1-4):331–335, 1996.
13. T. Bonald and C. Comte. Balanced fair resource sharing in computer clusters. *Performance Evaluation*, 116:70–83, 2017.
14. T. Bonald, C. Comte, and F. Mathieu. Performance of balanced fairness in resource pools: A recursive approach. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):1–25, 2017.
15. A. Busic, A. Cadas, J. Doncel, and J.-M. Fourneau. Product form solution for the steady-state distribution of a markov chain associated with a general matching model with self-loops. In *European Workshop on Performance Engineering*, pages 71–85. Springer, 2022.
16. A. Bušić, V. Gupta, and J. Mairesse. Stability of the bipartite matching model. *Advances in Applied Probability*, 45(2):351–378, 2013.
17. R. Caldentey, E. H. Kaplan, and G. Weiss. Fcfs infinite bipartite matching of servers and customers. *Advances in Applied Probability*, 41(3):695–730, 2009.
18. E. Cardinaels, S. Borst, and J. S. van Leeuwaarden. Heavy-traffic universality of redundancy systems with assignment constraints. *Operations Research*, 2022.
19. K. Chandy. The analysis and solutions for general queueing networks. In *Proceedings of the Sixth Anual Princeton Conference on Information Sciences and Systems*, pages 224–228, 1972.
20. C. Comte. Stochastic non-bipartite matching models and order-independent loss queues. *Stochastic Models*, 38(1):1–36, 2022.
21. C. Comte, F. Mathieu, and A. Bušić. Stochastic dynamic matching: A mixed graph-theory and linear-algebra approach. *arXiv preprint arXiv:2112.14457*, 2021.
22. K. Gardner, M. Harchol-Balter, E. Hyytiä, and R. Righter. Scheduling for efficiency and fairness in systems with redundancy. *Performance Evaluation*, 116:1–25, 2017.
23. K. Gardner, M. Harchol-Balter, A. Scheller-Wolf, and B. Van Houdt. A better model for job redundancy: Decoupling server slowdown and job size. *IEEE/ACM transactions on networking*, 25(6):3353–3367, 2017.
24. K. Gardner, M. Harchol-Balter, A. Scheller-Wolf, M. Velednitsky, and S. Zbarsky. Redundancy-d: The power of d choices for redundancy. *Operations Research*, 65(4):1078–1094, 2017.
25. K. Gardner and R. Righter. Product forms for FCFS queueing models with arbitrary server-job compatibilities: an overview. *Queueing Systems*, 96(1-2):3–51, 2020.
26. K. Gardner, S. Zbarsky, S. Doroudi, M. Harchol-Balter, E. Hyytiä, and A. Scheller-Wolf. Queueing with redundant requests: exact analysis. *Queueing Systems*, 83:227–259, 2016.
27. E. Gelenbe. Product-form queueing networks with negative and positive customers. *Journal of applied probability*, 28(3):656–663, 1991.

28. E. Gelenbe. G-networks: a unifying model for neural and queueing networks. *Annals of Operations Research*, 48(5):433–461, 1994.
29. I. Gurvich and A. Ward. On the dynamic control of matching queues. *Stochastic Systems*, 4(2):479–523, 2015.
30. T. Hellemans and B. Vanhoudt. Analysis of redundancy (d) with identical replicas. *ACM SIGMETRICS Performance Evaluation Review*, 46(3):74–79, 2019.
31. J. Jackson. Jobshop-like queueing systems. *Management Science*, 10(1):131–142, 1963.
32. M. Jonckheere, P. Moyal, C. Ramírez, and N. Soprano-Loto. Generalized max-weight policies in stochastic matching. *Stochastic Systems*, 13(1):40–58, 2023.
33. F. Kelly. Networks of queues with customers of different types. *Journal of applied probability*, 12(3):542–554, 1975.
34. F. Kelly. Networks of queues. *Advances in Applied Probability*, 8(2):416–432, 1976.
35. A. Krzesinski and R. Schassberger. Product form solutions for multiserver centers with hierarchical concurrency constraints. *Probability in the Engineering and Informational Sciences*, 6(2):147–156, 1992.
36. A. E. Krzesinski. Order independent queues. In *Queueing Networks: A Fundamental Approach*, pages 85–120. Springer, 2010.
37. J. Mairesse and P. Moyal. Stability of the stochastic matching model. *Journal of Applied Probability*, 53(4):1064–1077, 2016.
38. P. Moyal, A. Bušić, and J. Mairesse. A product form for the general stochastic matching model. *Journal of Applied Probability*, 58(2):449–468, 2021.
39. P. Moyal and O. Perry. On the instability of matching queues. *Annals of Applied Probability*, 27(6), 2017.
40. M. Nazari and A. L. Stolyar. Reward maximization in general dynamic matching systems. *Queueing Systems*, 91:143–170, 2019.
41. Y. Raaijmakers and S. Borst. Achievable stability in redundancy systems. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 4(3):1–21, 2020.
42. Y. Raaijmakers, S. Borst, and O. Boxma. Redundancy scheduling with scaled bernoulli service requirements. *Queueing Systems*, 93:67–82, 2019.
43. Y. Raaijmakers, S. C. Borst, and O. J. Boxma. Delta probing policies for redundancy. *ACM SIGMETRICS Performance Evaluation Review*, 46(3):72–73, 2019.
44. Y. Rahme and P. Moyal. A stochastic matching model on hypergraphs. *Advances in Applied Probability*, 53(4):951–980, 2021.
45. M. van Der Boor and C. Comte. Load balancing in heterogeneous server clusters: Insights from a product-form queueing model. In *2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS)*, pages 1–10. IEEE, 2021.
46. J. Visschers, I. Adan, and G. Weiss. A product form solution to a system with multi-type jobs and multi-type servers. *Queueing Systems*, 70(3):269–298, 2012.
47. P. Whittle. Nonlinear migration processes. *Bull. Inst. Int. Statist*, 42:642–647, 1967.
48. P. Whittle. Equilibrium distributions for an open migration process. *Journal of Applied Probability*, 5(3):567–571, 1968.