The Infinite-Server Queueing Model: The Center of the Many-Server Queueing Universe (i.e., More Relevant Than It Might Seem)

Ward Whitt

Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699, USA

Abstract

The infinite-server (IS) queueing model is more widely applicable than it might seem. Given that queueing science is primarily concerned with congestion (e.g., waiting and blocking) associated with limited resources, on the surface the IS model may seem useless and uninteresting. However, it is remarkable what a central role the IS model plays. First, both the classic and transient versions of Little's law can be expressed in terms of the IS model. Second, IS models often serve as surprisingly good and useful approximations for multi-server queueing systems. The idealized IS model with time-varying arrival rate is useful for understanding the physics of corresponding many-server queues. Third, asymptotic results for IS models can be useful tools for proving corresponding asymptotic results for systems with only finitely many servers. Finally, and arguably of greatest importance, for multi-server systems with time-varying arrivals, the IS model serves as the basis for important offered-load analysis, which characterizes the total load faced by the system, and serves as the basis for much useful engineering analysis.

Keywords: infinite-server queueing models, Little's law, $L = \lambda W$, transient version of $L = \lambda W$, offered load analysis, offered load, modified offered load
2000 MSC: 60K25, 90B22, 90B22

1. Introduction

If people in the field of operations research and management science were asked to identify the *most applicable* and *least applicable* models of queueing theory, surely many would fairly select $L = \lambda W$ (Little's law, [14, 16]) as one of the most applicable and *unfairly* select the infinite-server (IS) queueing model as one of the least applicable.

First, the relation $L = \lambda W$ deserves its celebrity. The relation $L = \lambda W$ states that the average number of customers (items) waiting in line (in a system), L, is equal to the arrival rate λ multiplied by the average waiting time (time spent in system) per customer, W. Under very general conditions, the relation is valid for both long-run averages of individual sample paths and expected values of stationary random variables in stochastic models; see [1, 5, 37, 38, 39, 41, 42]. As emphasized by [16], the relation $L = \lambda W$ has been widely applied; also see [15, 21, 23].

In contrast, performance results for the infinite-server (IS) queueing model would surely be on few people's short list of general widely applicable results. Indeed, given that queueing science is primarily concerned with congestion (e.g., waiting and blocking) associated with limited resources, on the surface the IS model may seem useless and uninteresting. However, it is remarkable what a central role the IS model plays.

We mention the IS model together with $L = \lambda W$ to make a point: Actually, the relation $L = \lambda W$ can be viewed as a *consequence* of the IS model theory. Both Little's law and its time-varying generalization can be expressed in the setting of the IS model. They can be viewed as *applications* of IS theory. The essential point here is that there are important *connections* between Little's law and the IS model. Establishing connections between existing theories may be just as important as establishing the theories themselves. Certainly, **connections are a key part of understanding**.

In this paper we make the case that the IS model is more useful than it might appear. First, in §2 we indicate that both the classic and transient versions of the relation $L = \lambda W$ can be expressed in the setting of the IS model. Second, in §3 we indicate that the IS model often serves as a useful approximation for the more obviously appropriate queueing models with only finitely many servers. Third, we point out that IS models provide important tools for proving theorems for many-server queues, such as heavy-traffic limit theorems. Finally, and arguably of greatest importance, for multi-server systems with time-varying arrivals, the IS model serves as the basis for important offered-load analysis, which characterizes the total load faced by the system, which serves as the basis for much useful engineering analysis.

2. The Connections to Little's Law

Theorem 1. *The results for Little's law can be recast in terms of the IS model.*

For recent papers, focusing on the applied value of Little's law, see [16, 21]. For a review of Little's law, see [39]; for the connection mentioned above, see the statement in italics on p. 238. For more, see [1, 5, 37, 38, 39, 41, 42].

Theorem 2. *The results for the transient Little's law can be recast in terms of the IS model with an arrival process having time-varying rate.*

For papers on time-varying Little's law, see [2, 7, 28, 29]. None of this mentions the IS model. However, to see the key connection, see Remark 2.3 of [24] and §4 of [17].

3. Approximations for Multi-Server Queues

3.1. Basic Theory for IS Models

1. Exact results with nonhomogeneous Poisson arrival processes [3, 4, 24]
2. Heavy-traffic limits for $G_t/GI/\infty$ [12, 31] and references therein.
3. Heavy-traffic limits for $G_t/G/\infty$ and with dependent service times [32, 33, 34]

We will concentrate on the paper [3]. This paper discusses the applied significance of the basic IS theory, emphasizing the case of time-varying arrival rates. It considers the basic $M_t/GI/\infty$ model with nonhomogeneous Poisson arrival process and i.i.d. service times that are independent of the arrival process. The theory for this IS model was developed long before [3]; the paper [3] reviews a classic proof and discusses the engineering implications of the key results.

Eleven things to note in Theorem 1:

1. $Q(t)$ has a Poisson distribution for each t, and so has one parameter, its mean $m(t)$.
2. The stochastic process $\{Q(t) : t \geq 0\}$ is *not* a Poisson process.
3. The mean $m(t)$ has a simple expression as an integral (or a double integral); see the first lines display on p. 733.

4. From (3), the mean can be expressed as the integral of the arrival rate over a random service time, S, prior to time t.
5. From (3), the mean can be expressed as the PSA OL $m_{PSA}(t) \equiv \lambda(t)E[S]$ plus a random time lag by S_e (S_e in (1) instead of S).
6. The departure process is also a nonhomogeneous Poisson process.
7. From (4), the departure rate is the arrival rate with a random time lag, equal to a random service time S (S, not S_e).
8. Remarkably, $Q(t)$ is independent of $\{D(s) : s \leq t\}$.
9. The proof of Theorem 1 follows from Figure 1.
10. The Poisson arrival process and i.i.d. service times leads to a Poisson random measure representation.
11. The formula for the mean does not actually depend on the Poisson property, yielding the so-called time-varying Little's law. (This is observed in Remark 2.3 of [24].)

Other things to note:

1. Theorem 2 (Fig. 3): joint distribution of $Q(t)$ and $Q(t+u)$
2. Corollary 4: The simple relation between $m(t)$ and PSA $m_{PSA}(t) \equiv \lambda(t)E[S]$ with M service
3. Theorem 9: The quadratic approximation (e.g., Taylor): (14) showing the time lag and space shift.
4. (20): The approach to steady state in a stationary model starting empty: $m(t)/m(\infty) = P(S_e \leq t)$

3.2. Where the Model Directly Captures the Main Issues

1. movement through space, network structure [24, 43]
2. The Poisson Arrival Location Model (PALM) [26]
3. produce life cycle dynamics [22]
4. wireless communication [13] (conversations by people moving through space)

4. A Tool for Proofs for Multi-Server Models

1. Reed's approach to $G/G/N$ [36]
2. The new FWLLN and FCLT for the $G_t/GI/s_t + GI$ model in [19, 20].

5. Offered Load Analysis to Cope with Time-varying Arrivals

See [9] for survey of ways to cope with time-varying arrival rates in service systems.

1. Modified Offered Load Approximations §4.3 of [9] and [25, 27, 43]
2. Stabilizing Performance [6, 10, 18, 43]

Acknowledgement. This research was supported by NSF grant CMMI 1066372.

References

[1] F. Baccelli, P. Bremaud. *Elements of Queueing Theory: Palm-Martingale Calculus and Stochastic Recurrences*, second ed., Springer, New York, 2003.

[2] D. Bertsimas, G. Mourtzinou. Transient laws of nonstationary queueing systems and their applications. *Queueing Systems* **25** 115–155.

[3] Eick, S. G., W. A. Massey, W. Whitt. 1993a. The physics of the $M_t/G/\infty$ queue. *Oper. Res.* **41** 731–742.

[4] Eick, S. G., W. A. Massey, W. Whitt. 1993b. $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Sci.* **39** 241–252.

[5] M. El-Taha, S. Stidham, Jr. *Sample-Path Analysis of Queueing Systems*, Kluwer, Boston, 1999.

[6] Feldman, Z., A. Mandelbaum, W. A. Massey, W. Whitt. 2008. Staffing of time-varying queues to achieve time-stable performance. *Management Sci.* **54** 324–338.

[7] B. H. Fralix, G. Riano. 2010. A new look at transient versions of Little's law and $M/G/1$ preemptive last-come first-served queues. *J. Appl. Prob,* **47** 459–473.

[8] P. W. Glynn, W. Whitt. Indirect estimation via $L = \lambda W$. *Oper. Res,* **37** (1989) 82–103.

[9] L. V. Green, P. J. Kolesar, W. Whitt, Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management* **16** (2007) 13–39.

[10] Jennings, O. B., A. Mandelbaum, W. A. Massey, W. Whitt. 1996. Server staffing to meet time-varying demand. *Management Sci.* **42** 1383–1394.

[11] S. Kim, W. Whitt. Little's law: from measurements to prediction. In preparation. Columbia University, 2012.

[12] E. V. Krichagina, A. A. Puhalskii, A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Systems* **25** (1997) 235–280.

[13] K. Leung, W. A. Massey, W. Whitt. Traffic Models for Wireless Communication Networks. *IEEE Journal on Selected Areas in Communication*, vol. 12, No. 8, 1994.

[14] J. D. C. Little. A proof of the queueing formula: $L = \lambda W$. *Oper. Res.* **9** (1961) 383–387.

[15] Little, J. D. C., S. C. Graves. 2008. Little's law. Ch. 5 in *Building Intuition: Insights from Basic Operations Management Models and Principles*, D. Chhajed, T. J. Lowe (eds.), Springer, New York, 81–100.

[16] J. D. C. Little. Little's law as viewed on its 50th anniversary. *Oper. Res.* **59** (2011) 536–539.

[17] Y. Liu, W. Whitt, The $G_t/GI/s_t + GI$ many-server fluid queue. Columbia University, NY, NY (2011a) http://www.columbia.edu/∼ww2040/allpapers.html (submitted to *Queueing Systems*)

[18] Y. Liu, W. Whitt, Stabilizing customer abandonment in many-server queues with time-varying arrivals. Columbia University, NY, NY (2011b) http://www.columbia.edu/∼ww2040/allpapers.html

[19] Y. Liu, W. Whitt, Many-server heavy-traffic limit for queues with time-varying parameters. Columbia University, NY, NY (2011c) http://www.columbia.edu/∼ww2040/allpapers.html (submitted to *Annals of Applied Probability*)

[20] Y. Liu, W. Whitt, A many-server fluid limit for the $G_t/GI/s_t + GI_t$ queueing model experiencing periods of overloading. Columbia University, NY, NY (2011d) http://www.columbia.edu/∼ww2040/allpapers.html (submitted to *Operations Research Letters*)

[21] Lovejoy, W. S., J. S. Desmond. 2011. Little's law flow analysis of observation unit impact and sizing. *Acad. Emergency Medicine* **18** 183–189.

[22] C. McCalla, W. Whitt. 2002. A time-dependent queueing-network model to describe life-cycle dynamics of private-line telecommunication services. *Telecommunication Systems* **17** 9–38

[23] Mandelbaum, A. 2011. Flow basics: Little's law. Lecture 2, Lecture notes in a course on Service Engineering. Available at: http://iew3.technion.ac.il/serveng/Lectures/lectures.html (Accessed December 19,2011)

[24] W. A. Massey, W. Whitt. 1993. Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems* **13** 183–250.

[25] W. A. Massey, W. Whitt. 1994. An analysis of the modified offered load approximation for the nonstationary Erlang loss model. *Ann. Appl. Probabil.* **4** 1145–1160.

[26] W. A. Massey, W. Whitt. 1994. A Stochastic Model to Capture Space and Time Dynamics in Wireless Communication Systems. Probability in the Engineering and Informational Sciences, vol. 8, 1994, pp. 541-569.

[27] W. A. Massey, W. Whitt. 1997. Peak congestion in multi-server sewrvice systems with slowly varying arrival rates. *Queueing Systems* **25** 157–172.

[28] R. Mazumdar, R. Kannurpatti, C, Rosenberg. *J. Appl, Prob,* **28** (1991) 762–770.

[29] M. Miyazawa. Time-dependent rate conservation laws for a process defined with a stationary marked point process and their applications. *J. Appl, Prob,* **31** (1994) 114–129.

[30] G. Pang, R. Talreja, W. Whitt, Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys* **4** (2007) 193–267.

[31] G. Pang, W. Whitt, Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems* **65** (2010) 325–364.

[32] G. Pang, W. Whitt, Two-Parameter Heavy-Traffic Limits for Infinite-Server Queues with Dependent Service Times. *Queueing Systems*, to appear.

[33] G. Pang, W. Whitt, The Impact of Dependent Service Times on Large-Scale Service Systems. To appear in *Manufacturing and Service Operations Management.*

[34] G. Pang, W. Whitt, Infinite-Server Queues with Batch Arrivals and Dependent Service Times. To appear in *Probability in the Engineering and Information Sciences.*

[35] A. A. Puhalskii, J. Reed, On many-server queues in heavy traffic. *Ann. Appl. Prob.* **20** (2010) 129–195.

[36] J. Reed, The $G/GI/N$ queue in the Halfin-Whitt regime. *Ann. Appl. Prob.* **19** (2009) 2211–2269.

[37] R. Serfozo. *Introduction to Stochastic Networks*, Springer, New York, 1999.

[38] K. Sigman. *Stationary Marked Point Processes, An Intuitive Approach*, Chapman and Hall, New York, 1995.

[39] W. Whitt. A review of $L = \lambda W$ and extensions. *Queueing Systems* **9** (1991) 235–268.

[40] W. Whitt. *Stochastic-Process Limits*, Springer, New York, 2002.

[41] R. W. Wolff. *Stochastic Modeling and the Theory of Queues*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[42] R. W. Wolff. Little's law and related results. *Wiley Encyclopedia of Operations Research and Management Science*, J. J. Cochran (ed.), Wiley, New York, 2011.

[43] Yom-Tov, G., Mandelbaum, A.: The Erlang-R queue: time-varying QED queues with re-entrant customers in support of healthcare staffing. working paper, the Technion, Israel, 2010.