Queues with Time-Varying Arrival Rates: A Bibliography

Ward Whitt

Industrial Engineering and Operations Research
Columbia University, New York, NY, 10027; mc3983@columbia.edu

March 14, 2016

Abstract

This is a bibliography of research on the performance analysis of queueing systems with time-varying arrival rates.

Keywords: time-varying arrival rates, nonstationary queues, transient analysis, staffing to stabilize performance, nonhomogeneous Poisson processes

1 Introduction

This is an annotated bibliography of research on queues with time-varying arrival rates, largely focusing own my own research. Most of this research has focused on many-server queues, which can be approached via the more tractable infinite-server queues, as can be seen by the surveys in Green et al. (2007) and Whitt (2007a, 2013). There are important applications to call centers and healthcare, as can be seen from the survey by Aksin et al. (2007) and the recent papers by Armony et al. (2015), Kim et al. (2015a,b), Kim and Whitt (2014b) and Yom-Tov and Mandelbaum (2014).

This account is divided into three sections: §2 focuses on infinite-server queues and §3 focuses on many-server queues, while §4 focuses on single-server queues, which has been the focus of some of our recent research, e.g., Ma and Whitt (2015a,b, 2016) and Whitt (2014b, 2015b).

2 Infinite-Server Queues

As discussed in Whitt (2016b), 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.

There is a long history of research on the IS model. A major focus for me began in joint work with Bill Massey, starting with Eick et al. (1993a,b) and Massey and Whitt (1993), and then turning to its implications for staffing finite-capacity systems in Massey and Whitt (1994a,b, 1996, 1997a,b) and Jennings et al. (1996).

Heavy-traffic limit theorems are especially tractable, as perhaps best shown in Glynn and Whitt (1989b). More on heavy-traffic limits is in Whitt (1982) and Pang and Whitt (2010a,b, 2012a,b, 2013).

Other papers on IS models include Duffield and Whitt (1997), Duffield et al. (2001), Goldberg and Whitt (2008), McCalla and Whitt (2002), Nelson and Taaffe (2004a,b), Whitt (2014a) and Yom-Tov and Mandelbaum (2014)

2.1 Little's Law and The Time-Varying Version

In general, the fundamental Little's law ($L = \lambda W$), as in Little (1961, 2011), Stidham (1974) and El-Taha and Stidham (1999), is intimately connected to the infinite-server (IS) queueing model, as emphasized in the early review paper by Whitt (1991a). Little's law can be important for estimation, as discussed in Glynn and Whitt (1989a), Little and Graves (2008), Lovejoy and Desmond (2011), Kim and Whitt (2013a) and Mandelbaum (2010).

The IS model with a time-varying (TV) arrival rate is in turn intimately connected to the time-varying Little's law (TVLL) in Bertsimas and Mourtzinou (1997), Fralix and Riano (2010), Kim and Whitt (2013b).

3 Many-Server Queues with Time-Varying Arrival Rates

As reviewed in Green et al. (2007) and Whitt (2007a, 2013), many-server queues with time-varying arrival rates can often be analyzed by exploiting related IS models.

3.1 Performance Analysis

There is a substantial literature on the performance analysis of many-server queues with time-varying arrival rates. We have emphasized the modified offered load (MOL) approach originally due to Jagerman (1975); see Massey and Whitt (1994a,b, 1996, 1997a,b), Davis et al. (1995), Crescenzo and Nobile (1995) and Grier et al. (1997).

Closure approximations were first proposed by Rothkopf and Oren (1979). New versions have been developed by Massey and Pender (2013); see also Pender (2015) and Pender and Massey (2014).

Also see Hampshire and Massey (2010), Heyman and Whitt (1984), Massey (2002),Leung et al. (1994) and Liu and Whitt (2011c).

3.2 Staffing to Stabilize Performance

We first applied results for IS models in Eick et al. (1993a,b) to develop staffing algorithms for many-server queues in Jennings et al. (1996). The initial offered-load (OL) approach was refined to a modified OL (MOL) method in Massey and Whitt (1994a), which followed earlier work by Jagerman (1975) for loss models. The OL method is an alternative to the pontwise-stationary approximation (PSA) discussed in Green and Kolesar (1991), Green et al. (2007) and Whitt (1991b, 2007a, 2013).

There is a large body of recent literature, including Jennings et al. (1996), Green et al. (2007), Feldman et al. (2008), He et al. (2014), Defraeye and Niewenhuyse (2013) and Liu and Whitt (2012c, 2014c, 2016).

3.3 Fluid Models and MSHT Limits

Many-server heavy-traffic (MSHT) limits were established for a large class of Markovian models by Mandelbaum et al. (1998); also see Puhalskii (2013). The FLLN's lead to deterministic fluid models, while the refined FCLT's lead to diffusion approximations. A survey of martingale methods for proving the FLLN's and FCLT's is in Pang et al. (2007) and Whitt (2007b).

A general fluid model for the non-Markovian $G_t/GI/s + GI$ model was developed in Whitt (2006) and Liu and Whitt (2012a). Additional research is in Liu and Whitt (2011a,b, 2012a,b, 2014a,b) and Whitt (2015a, 2016c,a).

3.4 Arrival Process Models

The natural arrival process model when there is a time-varying arrival rate is a nonhomogoeneous Poisson process (NHPP), but that model has been questioned, Avramidis et al. (2004), Besbes et al. (2010), Brown et al. (2005), Ibrahim et al. (2012), Jongbloed and Koole (2001), Kim and Whitt (2014a,b, 2013c), Kim et al. (2015a,b). A new test of an NHPP is developed in Kim and Whitt (2014a,b) based on Durbin (1961), Lewis (1965).

Ways to test and model non-Poisson nonstationary arrival processes are studied in Massey and Whitt (1994b), Gebhardt and Nelson (2009), Nelson and Gerhardt (2011), He et al. (2014) and Zhang et al. (2014). Ways to fit or approximate the arrival rate function were studied in Massey et al. (1996).

The recent work in He et al. (2014) exploits indices of dispersion or scaled variance-time curves, as in Cox and Lewis (1966), Sriram and Whitt (1986),Fendick et al. (1989, 1991), Fendick and

Whitt (1989).

3.5 Loss Models

Some research has focused on the time-varying or transient performance of loss models with TV arrival rate functions; see Abate and Whitt (1998) and Li and Whitt (2014), Li et al. (2015).

3.6 Controls

Some of our research has been devoted to controls that are flexible enough to apply to TV queueing systems. First, analysis of overload controls is contained in Duffield and Whitt (1997) and Perry and Whitt (2013, 2014a,b, 2016). Related queue-and-idleness-ratio (QIR) controls in Gurvich and Whitt (2009a,b,c) apply in normal operation to TV systems,

Methods for delay prediction and delay announcements have been devised for TV systems; see Armony et al. (2009), Ibrahim and Whitt (2012).

3.7 Healthcare

For healthcare applications, see Armony et al. (2015), Jacobson et al. (2006), Kim et al. (2015a,b), Kim and Whitt (2014b), Shi et al. (2015) and Yom-Tov and Mandelbaum (2014).

3.8 Call Centers

See Aksin et al. (2007) and Brown et al. (2005)

3.9 Periodic Queues

For more on periodic many-server queues, see Dong and Whitt (2015) and Whitt (2015a)

4 Single-Server Queues with Time-Varying Arrival Rates

Single-server queues with TV arrival rates are more difficult; there is less work on them. Early work was done by Newell (1968a,b,c) and followed up by Massey (1985) and Mandelbaum and Massey (1995). We have started to work in this area too, as can be seen from Ma and Whitt (2015a,b) and Whitt (2014b, 2015b)

For general background, see Heyman and Whitt (1984) and Rolski (1981, 1989).

For interesting asymptotics involving queues that are sometimes overloaded, see Choudhury et al. (1997).

4.1 Periodic Queues

Early analysis of the periodic $M_t/GI/1$ model was done by Harrison and Lemoine (1977), Lemoine (1981). The structure of more general periodic queues was studied by Ross (1978) and Rolski (1981, 1989).

Heavy traffic limits were established by Falin (1989) and Whitt (2014b).

4.2 Staffing to Stabilize Performance

We have recently studied service-rate controls for stabilizing the performance of TV single-server queues in Whitt (2015b).

4.3 Simulation

New simulation methods for periodic queues are contained in Ma and Whitt (2015a,b, 2016) and Xiong et al. (2015)

References

d

J. Abate and W. Whitt. Calculating transient characteristics of the Erlang loss model by numerical transform inversion. *Stochastic Models*, 14(3):663–680, 1998.

O. Z. Aksin, M. Armony, and V. Mehrotra. The modern call center: a multi-disciplinary perspective on operations management research. *Production Oper. Management*, 16:665–688, 2007.

M. Armony, N. Shimkin, and W. Whitt. The impact of delay announcements in many-server queues with abandonment. *Operations Research,*, 57(1):66–81, 2009.

M. Armony, S. Israelit, A. Mandelbaum, Y. Marmor, Y. Tseytlin, and G. Yom-Tov. Patient flow in hospitals: a data-based queueing-science perspective. *Stochastic Systems*, 5(1):146–194, 2015.

A. N. Avramidis, A. Deslauriers, and P. L'Ecuyer. Modeling daily arrivals to a telephone call center. *Management Sci.*, 50:896–908, 2004.

D. Bertsimas and G. Mourtzinou. Transient laws of nonstationary queueing systems and their applications. *Queueing Systems*, 25:315–359, 1997.

O. Besbes, R. Philliips, and A. Zeevi. Testing the validity of a demand model: an operations perspective. *Manufacturing and Service Oper. Management*, 12(1):162–183, 2010.

L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: a queueing-science perspective. *J. Amer. Stat. Assoc.*, 100:36–50, 2005.

G. L. Choudhury, A. Mandelbaum, M. I. Reiman, and W. Whitt. Fluid and diffusion limits for queues in slowly changing random environments. *Stochastic Models*, 13(1):121–146, 1997.

D. R. Cox and P. A. W. Lewis. *The Statistical Analysis of Series of Events*. Methuen, London, 1966.

A. Di Crescenzo and A. G. Nobile. Diffusion approximation to a queueing system with time-dependent arrival and service rates. *Queueing Systems*, 19:41–62, 1995.

J. L. Davis, W. A. Massey, and W. Whitt. Sensitivity to the service-time distribution in the nonstationary Erlang loss model. *Management Sci.*, 41(6):1107–1116, 1995.

M. Defraeye and I. Van Niewenhuyse. Controlling excessive waiting times in small service systems with time-varying demand: an extension of the ISA algorithm. *Decision Support Systems*, 54(4):1558–1567, 2013.

J. Dong and W. Whitt. Using a birth-and-death process to estimate the steady-state distribution of a periodic queue. *Naval Research Logistics*, 2015. published on line, December 30, 2015, DOI 10.1002nav.21672.

N. G. Duffield and W. Whitt. Control and recovery from rare congestion events in a large multi-server system. *Queueing Systems*, 26:69–104, 1997.

N. G. Duffield, W. A. Massey, and W. Whitt. A nonstationary offered-load model for packet networks. *Telecommunication Systems*, 13(3-4):271–296, 2001.

J. Durbin. Some methods for constructing exact tests. *Biometrika*, 48(1):41–55, 1961.

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

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

M. El-Taha and S. Stidham. *Sample-Path Analysis of Queueing Systems*. Kluwer, Boston, 1999.

G. I. Falin. Periodic queues in heavy traffic. *Advances in Applied Probability*, 21:485–487, 1989.

Z. Feldman, A. Mandelbaum, W. A. Massey, and W. Whitt. Staffing of time-varying queues to achieve time-stable performance. *Management Sci.*, 54(2):324–338, 2008.

K. W. Fendick and W. Whitt. Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue. *Proceedings of the IEEE*, 71(1):171–194, 1989.

K. W. Fendick, V. Saksena, and W. Whitt. Dependence in packet queues. *IEEE Trans Commun.*, 37: 1173–1183, 1989.

K. W. Fendick, V. Saksena, and W. Whitt. Investigating dependence in packet queues with the index of dispersion for work. *IEEE Trans Commun.*, 39(8):1231–1244, 1991.

B. H. Fralix and G. Riano. A new look at transient versions of Little's law. *J. Appl. Prob.*, 47:459–473, 2010.

I. Gebhardt and B. L. Nelson. Transforming renewal processes for simulation of non-stationary arrival procsses. *INFORMS Journal on Computing*, 21:630–640, 2009.

P. W. Glynn and W. Whitt. Indirect estimation via $L = \lambda W$. *Oper. Res.*, 37:82–103, 1989a.

P. W. Glynn and W. Whitt. A new view of the heavy-traffic limit for infinite-server queues. *Advances in Applied Probability*, 23(1):188–209, 1989b.

D. Goldberg and W. Whitt. The last departure time from an $M_t/G/\infty$ queue with a terminating arrival process. *Queueing Systems*, 58:77–104, 2008.

L. V. Green and P. J. Kolesar. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Sci.*, 37:84–97, 1991.

L. V. Green, P. J. Kolesar, and W. Whitt. Coping with time-varying demand when setting staffing requirements for a service system. *Production Oper. Management*, 16:13–29, 2007.

N. Grier, W. A. Massey, T. McKoy, and W. Whitt. The time-dependent Erlang loss model with retrials. *Telecommunications Systems*, 7:253–265, 1997.

I. Gurvich and W. Whitt. Queue-and-idleness-ratio controls in many-server service systems. *Mathematics of Operations Research*, 34(2):363–396, 2009a.

I. Gurvich and W. Whitt. Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing and Service Oper. Management*, 11:237–253, 2009b.

I. Gurvich and W. Whitt. Service-level differentiation in many-server service systems via queue-ratio routing. *Operations Research*, 58(2):316–328, 2009c.

R. C. Hampshire and W. A. Massey. Dynamic optimization with applications to dynamic rate queues. *Tutorials in Operations Research*, 27:208–247, 2010.

J. M. Harrison and A. J. Lemoine. Limit theoorems for periodic queues. *Journal of Applied Probability*, 14: 566–576, 1977.

B. He, Y. Liu, and W. Whitt. Stabilizing performance in nonstationary queues with non-Poisson arrivals. Columbia University, working paper, 2014.

D. P. Heyman and W. Whitt. The asymptoic behavior of queues with time-varying arrival. *Journal of Applied Probability*, 21(1):143–156, 1984.

R. Ibrahim and W. Whitt. Real-time delay estimation based on delay history in many-server queues with time-varying arrivals. *Production and Operations Management*, 20(5):654–667, 2012.

R. Ibrahim, P. L'Ecuyer, N. Regnard, and H. Shen. On the modeling and forecasting of call center arrivals. *Proceedings of the 2012 Winter Simulation Conference*, 2012:256–267, 2012.

S. H. Jacobson, S. N. Hall, and J. R. Swisher. Discrete-event simulation of health care systems. In R. W. Hall, editor, *Patient Flow: Reducing Delay in Healthcase Delivery*, chapter 8, pages 211–252. Springer, 2006.

D. L. Jagerman. Nonstationary blocking in telephone traffic. *Bell System Tech. J.*, 54:625–661, 1975.

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

G. Jongbloed and G. Koole. Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17:307–318, 2001.

S. Kim and W. Whitt. Statistical analysis with Little's law. *Operations Research*, 61(4):1030–1045, 2013a.

S. Kim and W. Whitt. Estimating waiting times with the time-varying Little's law. *Probability in the Engineering and Informational Sciences*, 27:471–506, 2013b.

S. Kim and W. Whitt. Using simulation to study statistical tests for arrival processes and service time models for service systems. In R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, editors, *Proceedings of the 2013 Winter Simulation Conference*, pages 1–10, 2013c.

S. Kim and W. Whitt. Choosing arrival process models for service systems: Tests of a nonhomogeneous Poisson process. *Naval Research Logistics*, 17:307–318, 2014a.

S. Kim and W. Whitt. Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing and Service Oper. Management*, 16(3):464–480, 2014b.

S Kim, P. Vel, W. Whitt, and W. C. Cha. Poisson and non-Poisson properties of appointment-generated arrival processes: The case of an endocrinology clinic. *Operations Research Letters*, 43:247–253, 2015a.

S. Kim, W. Whitt, and W. C. Cha. A data-driven model of an appointment-generated arrival processes at an outpatient clinic. Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html, presented at INFORMS 2015, 2015b.

A. J. Lemoine. On queues with periodic poisson input. *Journal of Applied Probability*, 18:889–900, 1981.

K. K. Leung, W. A. Massey, and W. Whitt. Traffic models for wireless communication networks. *IEEE J. Selected Areas in Communication*, 12(8):1353–1364, 1994.

P. A. W. Lewis. Some results on tests for Poisson processes. *Biometrika*, 52(1):67–77, 1965.

A. Li and W. Whitt. Approximate blocking probabilities for loss models with independence and distribution assumptions relaxed. *Performance Evaluation*, 80:82–101, 2014.

A. Li, W. Whitt, and J. Zhao. Staffing to stabilize blocking in loss models with time-varying arrival rates. *Probability in the Engineering and Informational Sciences*, 2015.

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

J. D. C. Little. Little's law as viewed on its 50^{th} anniversary. *Oper. Res.*, 59:536–539, 2011.

J. D. C. Little and S. C. Graves. Little's law. In D. Chhajed and T. J. Lowe, editors, *Building Intuition: Insights from Basic Opertations Management Models and Principles*, chapter 5, pages 81–100. New York, 2008.

Y. Liu and W. Whitt. A network of time-varying many-server fluid queues with customer abandonment. *Oper. Res.*, 59:835–846, 2011a.

Y. Liu and W. Whitt. Large-time asymptotics for the $G_t/M_t/s_t + GI_t$ many-server fluid queue with abandonment. *Queueing Systems*, 67:145–182, 2011b.

Y. Liu and W. Whitt. Nearly periodic behavior in the the overloaded $G/D/S + GI$ queue. *Stochastic Systems*, 1(2):340–410, 2011c.

Y. Liu and W. Whitt. The $G_t/GI/s_t + GI$ many-server fluid queue. *Queueing Systems*, 71:405–444, 2012a.

Y. Liu and W. Whitt. A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading. *Oper. Res. Letters*, 40:307–312, 2012b.

Y. Liu and W. Whitt. Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper. Res.*, 60(6):1551–1564, 2012c.

Y. Liu and W. Whitt. Many-server heavy-traffic limits for queues with time-varying parameters. *Annals of Applied Probability*, 24(1):378–421, 2014a.

Y. Liu and W. Whitt. Algorithms for time-varying networks of many-server fluid queues. *Informs Journal on Computing*, 26(1):59–73, 2014b.

Y. Liu and W. Whitt. Stabilizing performance in networks of queues with time-varying arrival rates. *Probability in the Engineering and Informational Sciences*, 28(4):414–449, 2014c.

Y. Liu and W. Whitt. Stabilizing performance in a service system with time-varying arrivals and customer feedback. Columbia University, Available at: http://www.columbia.edu/∼ww2040/allpapers.html, 2016.

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

N. Ma and W. Whitt. Efficient simulation of non-Poisson non-stationary point processes to study queueing approximations. *Statistics and Probability Letters*, 102:202–207, 2015a.

N. Ma and W. Whitt. Using simulation to study service-rate controls to stabilize performance in a single-server queue with time-varying arrival rate. In L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, editors, *Proceedings of the 2015 Winter Simulation Conference, Huntington Beach, CA, December 6-9, 2015*, pages 1–12. ACM, 2015b.

N. Ma and W. Whitt. A simulation algorithm for periodic queues. Columbia University, in preparation, 2016.

A. Mandelbaum. Lecture notes on Little's law, course on service engineering. The Technion, Israel, http://iew3.technion.ac.il/serveng/Lectures/lectures.html, 2010.

A. Mandelbaum and W. A. Massey. Strong approximations for time-dependent queues. *Mathematics of Operations Research*, 20(1):33–64, 1995.

A. Mandelbaum, W. A. Massey, and M. I. Reiman. Strong approximations for Markovian service networks. *Queueing Systems*, 30:149–201, 1998.

W. A. Massey. Asymptotic analysis of the time-varying $M/M/1$ queue. *Mathematics of Operations Research*, 10(2):305–327, 1985.

W. A. Massey. The analysis of queues with time-varying rates for telecommunication models. *Telecommunication Systems*, 21:173–204, 2002.

W. A. Massey and J. Pender. Gasussian skewness approximation for dynamic rate multi-server queues with abandonment. *Queueing Systems*, 75:243–277, 2013.

W. A. Massey and W. Whitt. Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems*, 13(1):183–250, 1993.

W. A. Massey and W. Whitt. An analysis of the modified offered load approximation for the nonstationary Erlang loss model. *Ann. Appl. Probab.*, 4:1145–1160, 1994a.

W. A. Massey and W. Whitt. Unstable asymptotics for nonstationary queues. *Math. Oper. Res.*, 19(2): 267–291, 1994b.

W. A. Massey and W. Whitt. Stationary-process approximations for the nonstationary Erlang loss model. *Oper. Res.*, 44(6):976–983, 1996.

W. A. Massey and W. Whitt. Peak congestion in multi-server service systems with slowly varying arrival rates. *Queueing Systems*, 25:157–172, 1997a.

W. A. Massey and W. Whitt. Uniform acceleration expansions for Markov chains with time-varying rates. *Annals of Applied Probability*, 9(4):1130–1155, 1997b.

W. A. Massey, G. A. Parker, and W. Whitt. Estimating the parameters of a nonhomogeneous Poisson process with linear rate. *Telecommunication Systems*, 5:361–388, 1996.

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

B. Nelson and M. Taaffe. The $Ph_t/Ph_t/\infty$ queueing system: Part I the single node. *INFORMS J. Comput.*, 16:266–274, 2004a.

B. Nelson and M. Taaffe. The $[Ph_t/Ph_t/\infty]^K$ queueing system: Part II the multiclass network. *INFORMS J. Comput.*, 16:275–283, 2004b.

B. L. Nelson and I. Gerhardt. Modeling and simulating renewal nonstationary arrival processes to facilitate analysis. *Journal of Simulation*, 5:3–8, 2011.

G. F. Newell. Queues with time dependent arrival rates, I: the transition through saturation. *Journal of Applied Probability*, 5:436–451, 1968a.

G. F. Newell. Queues with time dependent arrival rates, II: the maximum queue and the return to equilibrium. *Journal of Applied Probability*, 5:579–590, 1968b.

G. F. Newell. Queues with time dependent arrival rates, III: a mild rush hour. *Journal of Applied Probability*, 5:591–606, 1968c.

G. Pang and W. Whitt. Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems*, 65:325–364, 2010a.

G. Pang and W. Whitt. Continuity of a queueing integral representation in the m1 topology. *Annals of Applied Probability*, 20(1):214–237, 2010b.

G. Pang and W. Whitt. The impact of dependent service times on large-scale service systems. *Manufacturing and Service Oper. Management*, 14(2):262–278, 2012a.

G. Pang and W. Whitt. Infinite-server queues with batch arrivals and dependent service times. *Probability in the Engineering and Information Sciences*, 26:197–220, 2012b.

G. Pang and W. Whitt. Two-parameter heavy-traffic limits for infinite-server queues with dependent service times. *Queueing Systems*, 73(2):119–146, 2013.

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

J. Pender. Nonstationary loss queues via cumulant moment approximations. *Probability in the Engineering and Information Sciences*, 29:27–49, 2015.

J. Pender and W. A. Massey. Approximating and stabilizing dynamic rate Jackson networks with abandonment. Cornell University, http://people.orie.cornell.edu/jpender/, 2014.

O. Perry and W. Whitt. A fluid limit for an overloaded x model via a stochastic averaging principle. *Mathematics of Operations Research*, 38(2):294–349, 2013.

O. Perry and W. Whitt. Diffusion approximation for an overloaded x model via a stochastic averaging principle. *Queueing Systems*, 76:347–401, 2014a.

O. Perry and W. Whitt. Achieving rapid recovery in an overload control for a large-scale service system. *INFORMS Journal on Computing*, 27(3):491–506, 2014b.

O. Perry and W. Whitt. Chattering and congestion collapse in an overload switching control.k. Columbia University, Available at: http://www.columbia.edu/∼ww2040/allpapers.html, 2016.

A. A. Puhalskii. On the $M_t/M_t/K_t + M_t$ queue in heavy traffic. *Math. Methods Oper. Res.*, 78:119–148, 2013.

T. Rolski. Queues with nonstationary input stream: Ross's conjecture. *Advances in Applied Probability*, 13: 603–618, 1981.

T. Rolski. Queues with nonstationary inputs. *Queueing Systems*, 5:113–130, 1989.

S. M. Ross. Average delay in queues with nonstationary poisson arrivals. *Journal of Applied Probability*, 15: 602–609, 1978.

M. H. Rothkopf and S. S. Oren. A closure approximation for the nonstationary $M/M/s$ queue. *Management Science*, 25(6):522–534, 1979.

P. Shi, M. C. Chou, J. G. Dai, D. Ding, and J. Sim. Models and insights for hospital inpatient operations: time-dependent ED boarding time. *Management Sccience*, 62(1):1–28, 2015.

K. Sriram and W. Whitt. Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE Journal on Selected Areas in Communications*, SAC-4(6):833–846, 1986.

S. Stidham. A last word on $L = \lambda W$. *Oper. Res.*, 22:417–421, 1974.

W. Whitt. On the heavy-traffic limit theorem for $GI/G/\infty$ queue. *Advances in Applied Probability*, 14(1): 171–190, 1982.

W. Whitt. A review of $L = \lambda W$. *Queueing Systems*, 9:235–268, 1991a.

W. Whitt. The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Management Science*, 37(3):307–314, 1991b.

W. Whitt. Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management*, 15(1):88–102, 2006.

W. Whitt. What you should know about queueing models to set staffing requirements in service systems. *Naval Research Logistics*, 54(5):476–484, 2007a.

W. Whitt. Proofs of the martingale fclt. *Probability Surveys*, 4:268–302, 2007b.

W. Whitt. Offered load analysis for staffing. *Manufacturing and Service Operations Management*, 15(2): 166–169, 2013.

W. Whitt. The steady-state distribution of the $M_t/M/\infty$ queue with a sinusoidal arrival rate function. *Operations Research Letters*, 42:311–318, 2014a.

W. Whitt. Heavy-traffic limits for queues with periodic arrival processes. *Operations Research Letters*, 42: 458–461, 2014b.

W. Whitt. Many-server limits for periodic infinite-server queues. Columbia University, Availabel at: http://www.columbia.edu/∼ww2040/allpapers.html, 2015a.

W. Whitt. Stabilizing performance in a single-server queue with time-varying arrival rate. *Queueing Systems*, 81:341–378, 2015b.

W. Whitt. Many-server limits for periodic infinite-server queues. Columbia University, Available at: http://www.columbia.edu/∼ww2040/allpapers.html, 2016a.

W. Whitt. The infinite-server queueing model: more relevant than it might seem. in preparation, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html, 2016b.

W. Whitt. A Poisson limit for the departure process from a queue with many slow server. Columbia University, Available at: http://www.columbia.edu/∼ww2040/allpapers.html, 2016c.

Y. Xiong, D. J. Murdoch, and D. A. Stanford. Perfect sampling of a single server queue with periodic Poisson arrivals. *Queueing Systems*, 80:15–33, 2015.

G. Yom-Tov and A. Mandelbaum. Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing and Service Oper. Management*, 16(2):283–299, 2014.

X. Zhang, L.J. Hong, and P. W. Glynn. Timescales in modeling call center arrivals. Working paper, Stanford University, 2014.