Computer Communications 131 (2018) 22–25

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Review

Fluid flow models in performance analysis

Onno Boxma[*,a], Bert Zwart[a,b]

[a] Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
[b] Centrum Wiskunde & Informatica (CWI), P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

ABSTRACT

We review several developments in fluid flow models: feedback fluid models, linear stochastic fluid networks and bandwidth sharing networks. We also mention some promising new research directions.

1. Introduction

A key concept in the performance analysis of computer and communication networks is time-scale separation. A separation of time scales enables to analyse different dynamics in a tractable way, distinguishing between fast dynamics and slow dynamics. In the context of computer and communication networks, fast dynamics are often associated with the behavior of data packets, while slow dynamics (sometimes called burst scale) are associated with files or users, often called flows. When the number of flows is fixed, it is often possible to analyze the long-term behavior of protocols that are designed to deal with packets. The steady-state behavior of these packet-level dynamics is then assumed to be achieved instantaneously and deterministically when looking at the system at a slow time scale, in which the main source of randomness is not associated with packet-level dynamics, but with arrivals and departures of flows. This gives rise to fluid flow models.

The survey paper *Fluid queues with long-tailed activity period distributions* [13], which appeared in this journal 20 years ago, was devoted to a particular class of fluid flow models, viz., fluid queues fed by a number of on/off sources. A fluid queue is a buffer that receives fluid input from a number, say N, of sources, and from which fluid drains at a constant rate. Source $i \in \{1, 2, \cdots, N\}$ alternates between activity (on) periods $A_{i,1}, A_{i,2}, \cdots$ during which it generates fluid at a constant rate r_i, and silence (off) periods $S_{i1}, S_{i2}, \cdots$ during which it generates no fluid. The successive activity periods of a source are assumed to be i.i.d. (independent, identically distributed) random variables, and the same holds for the silence periods. Furthermore, independence is assumed between all activity and all silence periods of all sources.

Such a fluid queue model fed by on/off sources has been used to capture the behaviour of a wide range of computer and communication networks at the burst scale. Initially, all on- and off-period distributions were assumed to have exponential tails. Triggered by a host of measurements of actual communication traffic, the focus in [13] was on the situation in which at least one source has activity periods which do *not* have an exponential tail, but instead are long-tailed. Such activity periods, which behave fundamentally different from exponentially distributed activity periods, may have a severe impact on the tail behaviour of the buffer content and the busy period (non emptiness) distribution of the buffer. That impact was qualitatively described in [13].

After the appearance of [13], much research has been devoted to a further study of the impact of heavy-tailed – in particular, regularly varying – on-periods on the buffer content distribution of which we mention the most general results that exist to date. In [60] it is shown that the buffer content distribution is asymptotically equivalent to that in a reduced system. The reduced system consists of a 'dominant' subset of the flows, with the original service rate subtracted by the mean rate of the other flows. The dominant set is shown to be determined via a knapsack formulation. The dominant set consists of a 'minimally critical' set of on-off flows with regularly varying on-periods. A related result for a model where flows are initiated by a Poisson process is [12], while analogous results for finite buffers have been considered in [24].

In the present note, the focus is on two related topics regarding flow level models. The first topic concerns the exact analysis of two families of fluid queueing systems, viz., feedback fluid queues and linear stochastic fluid networks. The second topic is bandwidth sharing networks. Bandwidth networks are related to fluid queues, but explicitly take into account the feedback loop that exists in congestion-aware packet level protocols. In each of these two topics, the idea of time-scale separation is directly used in stochastic modeling. We would like to remark that it is also very well possible to start with a more comprehensive model, and establish a time-scale separation in a more formal, endogenous way. In the context of applied probability and stochastic networks, one typically begins with a detailed 'intractable' model, and then applies probabilistic scaling techniques, which lead to the

* Corresponding author.
E-mail addresses: o.j.boxma@tue.nl (O. Boxma), bert.zwart@cwi.nl (B. Zwart).

https://doi.org/10.1016/j.comcom.2018.07.009
Received 3 April 2018; Received in revised form 15 June 2018; Accepted 1 July 2018
Available online 18 July 2018

identification of relevant time scales, and the associated reduction in model complexity in an endogenous way [55].

In this brief note, lack of space forces us to largely ignore several interesting areas of research on fluid queues. To mention a few: (i) The idea of replacing a stochastic model with its associated deterministic dynamics (called fluid model) is quite useful to determine whether or not a stochastic model is positive recurrent. These ideas are not treated here; we refer to the groundbreaking papers of Rybko and Stolyar [47] and Dai [20], and to [16] for a textbook treatment. (ii) Efficient computational and matrix-analytic [2–4,50,51] methods for the steady-state and transient analysis of fluid flow models, mostly based on a connection between fluid queues and quasi birth-death processes. (iii) Studies on stochastic storage and dam processes, which may be viewed as (predecessors of) fluid flow models; see, e.g., [44]; see also the mountain processes as discussed, a.o., in [15].

The remainder of this note is organized as follows. Section 2 is devoted to exact results for feedback fluid queues and linear stochastic fluid networks. Flow level models for bandwidth sharing are discussed in Section 3. Section 4 mentions interesting new developments and contains some suggestions for further research.

2. Exact analysis of feedback fluid queues and linear stochastic fluid networks

Origin. Pioneering work concerning the probabilistic analysis of fluid queues fed by on/off sources was done by Kosten [34–36]; in those three papers he considered an infinite number of sources, successively taking exponential, Erlang and hyperexponential on-period distributions. Several years later it was followed by the breakthrough paper [5], which triggered intensive research on all kinds of extensions; some of these are summarized in the COST report [19] and the survey [37].

Model and results: (i) Feedback fluid queues. Anick, Mitra and Sondhi (AMS) [5] consider a buffer that is fed by N on/off sources. It is assumed that all N on- and off periods are exponentially distributed. The steady-state buffer content distribution is described in terms of a set of differential equations, using spectral analysis. This AMS model may be viewed as a Markov modulated fluid model (MMFM): If an underlying Markov process is in state, say, j, then the input rate into the buffer is r_j. In several studies, Scheinhardt et al. (see, e.g., [1]) consider extensions of such MMFM by allowing a form of feedback: Not only is the behaviour of the buffer content determined by some background process, but also does the behaviour of the background process depend on the current buffer level. The rates r_j become rates $r_j(y)$ when the buffer content equals y, and the matrix Q that determines the transitions between different background states becomes $Q(y)$. Due to this feedback, the background process no longer is a Markov process.

Such feedback fluid queues may represent the behaviour of certain production- and communication systems in which the network and the sources interact. In [40], feedback fluid queues are used to study feedback schemes in access networks; the functions $Q(y)$ and $r_j(y)$ are here taken piecewise constant. In [49], $Q(y)$ and the $r_j(y)$ are allowed to depend continuously on y. The stationary distribution for the two-dimensional process consisting of background state and buffer content is described by a set of ordinary differential and algebraic equations; this set of equations is solved explicitly for the case of only two background states. The buffer size is assumed to be finite in [49]; in [14] the buffer size is allowed to be infinite, and the background process has only two states.

Feedback fluid queues with piecewise constant functions $Q(y)$ and $r_j(y)$ have recently also been used for performance modelling in various other areas than access networks, such as optical buffering using fiber delay lines [57] and battery life times under stochastic charging/discharging periods [25].

Model and results: (ii) Linear stochastic fluid networks. In a series of papers [28–30], Kella and Whitt have developed the elegant theory of *linear stochastic fluid networks*. Such fluid networks have random external inputs, but all internal flows are deterministic and continuous like fluid. We focus in particular on the linear stochastic fluid networks introduced in [30]. Linear stochastic fluid networks arise as the limit of normalized networks of infinite server queues with batch arrivals, in which the batch sizes grow. In those fluid networks, just like in networks of infinite server queues, the movement of separate particles can be thought of as being mutually independent, conditional on the time and place of entering the network. The resulting tractability makes linear stochastic fluid networks a strong candidate for approximations of discrete queueing networks. Kella and Whitt do not consider stochastic behaviour of individual particles, but specify what happens to deterministic proportions of the arriving fluid. Instead of independent, identically distributed service times with distribution $G(\cdot)$, the following service mechanism is employed. Let $G = \{G(x, t)|(x, t) \in \mathcal{R}^2\}$ be a stochastic process with $0 \leq G(\cdot, \cdot) \leq 1$ and $G(x, t)$ non-decreasing in t. The meaning of the stochastic process is the following: A proportion $G(x, t)$ of any input arriving at time x leaves at time t. $A(s, t)$ denotes the external input in a fluid queue during $(s, t]$. Very detailed buffer content results can be obtained when there is a proportional release rate r (which, in a discrete setting, can be interpreted as departure rate which is proportional to the number of jobs) and the input process has stationary and independent increments, i.e., it is a Lévy process. Let $\eta(\cdot)$ denote the Laplace exponent of the Lévy process. Kella and Whitt [30] prove that the Laplace-Stieltjes transform of the steady-state buffer content W is given by

$$E[e^{-\alpha W}] = \exp\left(-\int_0^\infty \eta(e^{-rs}\alpha)ds\right). \tag{1}$$

They also provide a matrix form of this result, for the case of a network of m fluid queues, m-dimensional Lévy input process, proportional release rate vector, and proportional routing matrix.

A generalization to Markov-modulated linear fluid networks is presented by Kella and Stadje [27]. Under the assumption that the external input is a multivariate Markov additive process, they provide stability conditions and show how to compute transient and stationary characteristics of the networks under consideration.

For the case of stochastic fluid networks with a tree structure, driven by a multidimensional Lévy process, Debicki et al. [21] obtain elegant results for the joint distribution of the buffer contents and the ages of the busy periods (uninterrupted period of time a buffer has been non-empty) and of the idle periods. These and other results for Lévy-driven fluid queues are also discussed in [22].

While linear stochastic fluid models can be regarded as continuous analogs or fluid limits of open networks of infinite-server queues, they also appear in different applications; for an application to power systems we refer to [52].

3. Bandwidth sharing network models for internet congestion

Origin. The TCP/IP protocol has been an important building block of the modern internet. Within the context of performance analysis, both packet-level models and flow-level models for TCP exist. The key idea, proposed by Kelly [31], is that packet-level mechanisms can be interpreted as implementation of decentralized solutions of optimization problems, i.e., the TCP protocol can be seen as a distributed way of solving a particular optimization problem. This 'reversed engineering' point of view gave rise to the field of network utility maximization [17,58]. These developments led to a class of stochastic processes on networks called bandwidth sharing networks. In a bandwidth sharing network, flows with random duration arrive according to a random process and need to be processed along different routes. The bandwidth available at the bottlenecks of these routes (each bottleneck may correspond to multiple routes) is shared in a way that extends processor sharing in a single-node case [42].

Model. In its simplest form, a bandwidth sharing network can be

described as follows. Consider a network with J links, and suppose that link j, $j = 1, ...,J$, can process work at rate C_j. There are R classes of users called routes. A flow on route r uses a subset of the links, encoded by a $0-1$ matrix A with 0-1 elements A_{jr}. Flows on route r arrive according to a Poisson arrival process of rate λ_r and have service requirements that are exponential with rate μ_r. Consider now a static situation in which the number of flows along each route is given by a vector n. The connection with packet-level dynamics is now modeled by assuming that flows along route r are served with rate $\Lambda_r(n)$, where $\Lambda(n)$ can be characterized as the solution of a concave programming problem of the form

$$\Lambda(n) = \arg\max \sum_r n_r U_r(\Lambda_r/n_r), \tag{2}$$

subject to the network capacity constraints $A\Lambda \le C$ and possibly additional individual rate constraints $\Lambda_r/n_r \le d_r$. $U_r(x)$ is the utility for a user of route r if its service rate equals x. When the system is single node and single class, this reduces to a processor sharing queue. In the network set-up, the choice $U_r(x) = w_r \log x$, called weighted proportional fairness, is the most popular, though different choices of utility function correspond to different packet-level dynamics - we refer to [33] for a textbook treatment.

Results. Like in the case of fluid queues, much of the initial work focused on exact analysis, and on determining whether results for exponential processing times can be extended to more general distributions. This can be related to the analysis of insensitivity in classical product-form queueing networks in the following illustrative way. Consider a network topology with J links and $R = J + 1$ routes, where one route (number 0) is using all links, and route $r \ge 1$ is using link r. When also $C_j = 1$, and $U_r(x) = \log x$, the optimization problem (2) can be solved explicitly, leading to the expression

$$\Lambda_0(n) = \frac{n_0}{\sum_{r=0}^J n_r}, \qquad \Lambda_r(n) = 1 - \Lambda_0(n), \quad r \ge 1. \tag{3}$$

One can then apply existing results on classical product-form queueing networks dating back to [7,18] to conclude that the invariant distribution $\pi(n)$ is given by

$$\pi(n) = \frac{\prod_{r=1}^J (1-\rho_0-\rho_r)}{(1-\rho_0)^{J-1}} \binom{\sum_{r=0}^J n_r}{n_0} \prod_{r=0}^J \rho_r^{n_r}, \tag{4}$$

with $\rho_r = \lambda_r/\mu_r$ - see Chapter 8.5 of [33] for a textbook treatment. An appealing property of these results is that the assumption of exponentially distributed flow sizes can be relaxed towards a dense class of distributions.

This type of result can be generalized further to so-called hypercube topologies. In [8] it is shown that the only instances yielding product-form results are such topologies. Moreover, these networks need to operate under the unweighted proportional fairness allocation mechanism $U_r(x) = \log x$ and have identical link capacities.

This has led to follow-up research focusing on bounds [9] as well as modifications of proportional fairness [41]. Another line of more recent research is aimed at simplifications using probabilistic model reduction techniques such as fluid and diffusion (heavy traffic) approximations. For the case of networks operating under proportional fairness, a number of results has been obtained that lead to heavy-traffic approximations that have a tractable limiting distribution. These results suggest that proportional fairness is approximately insensitive if the network operates in heavy traffic [26,53]. Bandwidth sharing networks in overload have been studied in [10].

All of the above results hold under the assumption that there is no upper bound on how fast a flow can be transmitted, i.e. $d_r = \infty$. In practice, it would take many flows to saturate a link, so that C_i can be orders of magnitude bigger than d_r. This leads to different scalings, and nontrivial behavior of fluid approximations, especially when also abandonments exist.

Nevertheless, it is still possible to come up with tractable approximations of the performance in this case, even when service time and deadline have a general joint distribution. To give a taste of the results that can be obtained, assume that flows on route r have service requirement B_r and generic deadline D_r, with (B_r, D_r) a general two-dimensional random vector. In this case, it can be shown that the following procedure leads to an accurate approximation of the expected queue length vector in steady state. Define $g_r(x) = \lambda_r E[\min\{xD_r, B_r\}]$ and let G_r be a function with derivative $G_r'(x) = U_r'(g_r^{-1}(x))$. Define now Λ^* as the solution of the concave continuous optimization problem

$$\min_{\Lambda: A\Lambda \le C, \Lambda_r \le g_r(d_r)} \sum_j G_j(\Lambda_j). \tag{5}$$

The expected queue length at route r can now be approximated by solving

$$z^* = \lambda_r E[\min\{D_r, B_r z_r^*/\Lambda_r^*\}]. \tag{6}$$

This procedure is developed rigorously using fluid limits in [45,46]. To get some intuition, note that (5) emerges by combining the Karush-Kuhn Tucker equations for $\Lambda(z)$ with equation (6) which can be interpreted as Little's law, where the RHS is an approximation of the expected sojourn time of a job, assuming its service rate is constant and equal to Λ_r^*/z_r^*.

4. Outlook

While the literature on exact analysis can be seen as comprehensive, this is not the case for asymptotic methods, though the open issues mentioned in e.g. [26] constitute very tough open problems in probability theory. Less work has been done on making the connection between packet-level models and flow-level models rigorous. The only work we are aware of that rigorously connects such models is [39] for fluid queues and [54] for bandwidth sharing models; a non-rigorous approach can be found in [23]. More work would be welcome in this direction. For a textbook discussion on connecting the shadow prices appearing in flow level models (2) to packet-level models, see [33].

In view of the fact that only few fluid queue models succumb to an exact analysis (cf. Section 2), there is a considerable need for asymptotic results. For feedback fluid queues, we refer to [40] for large deviation asymptotics in the case of a large number of users, and to [48] for the asymptotic behavior of the loss probability when the buffer is finite. For tandem fluid queues that do not fall in the class of linear stochastic fluid networks as discussed in Section 2, interesting asymptotic studies are presented, a.o., by Lieshout and Mandjes [38] and Miyazawa and Rolski [43]. To obtain exact results for the delay of users in bandwidth sharing networks is quite hard (in Section 3 we focused on queue length), but it is possible to develop tail asymptotics, as is surveyed in [11].

Another appealing direction for future research is to consider the above-mentioned classes of models in different application areas. Examples so far are road traffic [32] and Electric Vehicle Charging [6]. Finally, we also feel that time-varying fluid models deserve much more attention [56,59].

Acknowledgment

The research of Onno Boxma is supported by the NWO Gravitation Project NETWORKS, Grant Number 024.002.003. The research of Bert Zwart is supported by the NWO VICI grant 639.033.413.

References

[1] I.J.B.F. Adan, E.A. van Doorn, J.A.C. Resing, W.R.W. Scheinhardt, Analysis of a single-server queue interacting with a fluid reservoir, Queueing Syst. 29 (1998) 313–336.
[2] S. Ahn, V. Ramaswami, Fluid flow models and queues – a connection by stochastic coupling, Stochastic Models 19 (2003) 325–348.

[3] S. Ahn, V. Ramaswami, Efficient algorithms for transient analysis of stochastic fluid flow models, J. Appl. Probab. 42 (2005) 531–549.
[4] N. Akar, K. Sohraby, Infinite- and finite-buffer Markov fluid queues: A unified analysis, J. Appl. Probab. 41 (2004) 557–569.
[5] D. Anick, D. Mitra, M.M. Sondhi, Stochastic theory of a data-handling system with multiple sources, Bell Syst. Techn. J. 61 (1982) 1871–1894.
[6] A. Aveklouris, M. Vlasiou, B. Zwart, A stochastic resource sharing model for electric vehicle charging, 2017. https://arxiv.org/abs/1711.05561.
[7] F. Baskett, K.M. Chandy, R.R. Muntz, F.G. Palacios, Open, closed and mixed networks of queues with different classes of customers, J. ACM 22 (1975) 248–260.
[8] T. Bonald, A. Proutière, Insensitivity in processor-sharing networks, Performance Evaluation 49 (2002) 193–209.
[9] T. Bonald, A. Proutière, On performance bounds for balanced fairness, Performance Evaluation 55 (2004) 25–50.
[10] S. Borst, R. Egorova, B. Zwart, Fluid limits for bandwidth-sharing networks in overload, Mathematics of Operations Research 39 (2014) 533–560.
[11] S. Borst, R. Núñez Queija, B. Zwart, Sojourn time asymptotics in processor-sharing queues, Queueing Syst. 53 (2006) 31–51.
[12] S. Borst, B. Zwart, Fluid queues with heavy-tailed $M/G/\infty$ input, Mathematics of Operations Research 30 (2005) 852–879.
[13] O.J. Boxma, V. Dumas, Fluid queues with long-tailed activity period distributions, Computer-Communications 21 (1998) 1509–1529.
[14] O.J. Boxma, H. Kaspi, O. Kella, D. Perry, On/off storage systems with state-dependent input, output and switching rates, Prob. Eng. Inform. Sci. 19 (2005) 1–14.
[15] O.J. Boxma, D. Perry, On the cycle maximum of mountains, dams and queues, Commun. in Statistics - Part A. Theory Methods 38 (2009) 2706–2720.
[16] H. Chen, D.D. Yao, Fundamentals of queueing networks. Performance, asymptotics, and optimization, Applications of Mathematics 46. Stochastic Modelling and Applied Probability, Springer-Verlag, New York, 2001.
[17] M. Chiang, S. Low, A. Calderbank, J.C. Doyle, Layering as optimization decomposition: a mathematical theory of network architectures, Proc. IEEE 95 (2007) 255–312.
[18] J.W. Cohen, The multiple phase service network with generalized processor sharing, Acta Informatica 12 (1979) 245–284.
[19] J.W. Roberts, Information technologies and sciences. COST 224. performance evaluation and design of multiservice networks, Final report of the COST project (1992).
[20] J.G. Dai, On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models, Ann. Appl. Probab. 5 (1995) 49–77.
[21] K. Debicki, A.B. Dieker, T. Rolski, Quasi-product forms for Lévy-driven fluid networks, Mathematics of Operations Research 32 (2007) 629–647.
[22] K. Debicki, M. Mandjes, Queues and Lévy Fluctuation Theory, Springer-Verlag, New York, 2015.
[23] R. Gibbens, S. Sargood, C.V. Eijl, F. Kelly, H. Azmoodeh, R. Macfadyen, N. Macfadyen, Fixed-point models for the end-to-end performance analysis of IP networks, 13th ITC Special Seminar: IP Traffic Management, Modeling and Management, (2000).
[24] P. Jelenkovic, P. Momcilovic, Asymptotic loss probability in a finite buffer queue with heterogeneous heavy-tailed fluid on-off processes, Annals of Applied Probability 13 (2003) 576–603.
[25] G.L. Jones, P.G. Harrison, U. Harder, T. Field, Fluid queue models of battery life, Proc. 2011 IEEE 19th Annual International Symposium on Modelling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), (2011), pp. 278–285. Singapore
[26] W. Kang, F.P. Kelly, N.H. Lee, R.J. Williams, State space collapse and diffusion approximation for a network operating under a fair bandwidth sharing policy, Ann. Appl. Probab. 19 (2009) 1719–1780.
[27] O. Kella, W. Stadje, Markov-modulated linear fluid networks with Markov additive input, J. Appl. Probab. 39 (2002) 413–420.
[28] O. Kella, W. Whitt, A tandem fluid network with Lévy input, in: U.N. Bhat, I.V. Basawa (Eds.), Queueing and Related Models, Clarendon Press, Oxford, 1992, pp. 112–128.
[29] O. Kella, W. Whitt, Stability and structural properties of stochastic fluid networks, J. Appl. Probab. 33 (1996) 1169–1180.
[30] O. Kella, W. Whitt, Linear stochastic fluid networks, J. Appl. Probab. 36 (1999) 244–260.
[31] F. Kelly, Charging and rate control for elastic traffic, Eur. Trans. Telecommun. 8 (1997) 33–37.
[32] F. Kelly, R. Williams, Heavy traffic on a controlled motorway, in: N.H. Bingham, C.M. Goldie (Eds.), Probability and Mathematical Genetics: Papers in Honour of Sir John Kingman, Cambridge University Press, Cambridge, 2010, pp. 416–445.
[33] F. Kelly, E. Yudinova, Stochastic Networks. Cambridge University Press, 2014.
[34] L. Kosten, Stochastic theory of a multi-entry buffer (I), Delft Prog. Rep. 1 (1974) 10–18.
[35] L. Kosten, Stochastic theory of a multi-entry buffer (II), Delft Prog. Rep. 1 (1974) 44–50.
[36] L. Kosten, O.J. Vrieze, Stochastic theory of a multi-entry buffer (III), Delft Prog. Rep. 1 (1975) 103–115.
[37] V.G. Kulkarni, Fluid models for single buffer systems, in: J.H. Dshalalow (Ed.), Frontiers in Queueing, CRC Press, Boca Raton (Fl.), 1997, pp. 321–338.
[38] P. Lieshout, M. Mandjes, Asymptotic analysis of Lévy-driven tandem queues, Queueing Syst. 60 (2008) 203–226.
[39] M. Mandjes, J. Kim, Analysis of a phase transition phenomenon in packet networks, Adv. Appl. Probab. 31 (2001) 360–380.
[40] M.R.H. Mandjes, D. Mitra, W.R.W. Scheinhardt, Models of network access using feedback fluid queues, Queueing Syst. 44 (2003) 365–398.
[41] L. Massoulié, Structural properties of proportional fairness: stability and insensitivity, Ann. Appl. Probab. 17 (2007) 809–839.
[42] L. Massoulié, J. Roberts, Bandwidth sharing: objectives & algorithms, Proc. of IEEE Infocom (1999) 1395–1403.
[43] M. Miyazawa, T. Rolski, Tail asymptotics for a Lévy-driven tandem queue with an intermediate input, Queueing Syst. 63 (2009) 323–353.
[44] N.U. Prabhu, Stochastic Storage Processes: Queues, Insurance Risk, and Dams, Springer, New York, 1998.
[45] J.E. Reed, B. Zwart, Limit theorems for bandwidth sharing networks with rate constraints, Oper. Res. 62 (2014) 1453–1466.
[46] M. Remerova, J. Reed, B. Zwart, Fluid limits for bandwidth-sharing networks with impatience, Mathematics of Operations Research 39 (2014) 746–774.
[47] A.N. Rybko, A.L. Stolyar, On the ergodicity of random processes that describe the functioning of open queueing networks, Problemy Peredachi Informatsii 28 (1992) 3–26. (translation in Problems Inform. Transmission 28, 199-220).
[48] Y. Sakuma, M. Miyazawa, Asymptotic behavior of the loss probability for a feedback finite fluid queue with downward jumps, in: W. Yue, Y. Takahashi, H. Takagi (Eds.), Advances in Queueing Theory and Network Applications, Springer, New York, 2004, pp. 195–211.
[49] W. Scheinhardt, N. van Foreest, M. Mandjes, Continuous feedback fluid queues, Oper. Res. Letters 33 (2005) 551–559.
[50] A.daSilva Soares, G. Latouche, Further results on the similarity between fluid queues and QBDs, in: G. Latouche, P. Taylor (Eds.), Matrix-Analytic Methods: Theory and Applications, World Scientific, 2002, pp. 89–106. Singapore
[51] A.daSilva Soares, G. Latouche, Matrix-analytic methods for fluid queues with finite buffers, Performance Evaluation 63 (2006) 295–314.
[52] K. Turitsyn, P. Vorobev, A. Zocca, B. Zwart, Frequency fluctuations in the presence of random generator breakdowns and wind power fluctuations, In preparation (2018).
[53] M. Vlasiou, J. Zhang, B. Zwart, Insensitivity of proportional fairness in critically loaded bandwidth sharing networks, Under review (2014).
[54] N. Walton, Proportional fairness and its relationship with multi-class queueing networks, Ann. Appl. Probab. 19 (2009) 2301–2333.
[55] W. Whitt, Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and their Application to Queues, Springer, New York, 2002.
[56] W. Whitt, Time-varying queues, Queueing Models and Service Management, to appear (2017).
[57] M.A. Yazici, N. Akar, Performance modeling of QoS differentiation in optical packet switching via FDL access limitation, Photonic Network Communications 34 (2017) 344–355.
[58] Y. Yi, M. Chiang, Stochastic network utility maximization, Eur. Trans. Telecommun. 19 (2008) 421–442.
[59] B. Zhang, B. Zwart, Fluid models for many-server Markovian queues in a changing environment, Oper. Res. Letters 40 (2012) 573–577.
[60] B. Zwart, S. Borst, M. Mandjes, Exact asymptotics for fluid queues fed by multiple heavy-tailed on-off flows, Ann. Appl. Probab. 14 (2004) 903–957.