Fluid flow models in performance analysis

O. Boxma, B. Zwart

Fluid flow models in performance analysis

Onno Boxma   Bert Zwart

March 28, 2018

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 $i \in \{1, 2, \ldots, N\}$ alternates between activity (on) periods $A_{i,1}, A_{i,2}, \ldots$ during which it generates fluid at a constant rate r_i, and silence (off) periods $S_{i1}, S_{i2}, \ldots$ 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 [59] 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. Such 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 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 [48] and Dai [20], and to [16] for a textbook treatment. (ii) Efficient computational and matrix-analytic [2, 3, 4, 51, 52] 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., [43]; 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 [33, 34, 35]; 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 [36].

Model and results: (i) Feedback fluid queues. Anick, Mitra and Sondhi [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 [39], 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 [50], $Q(y)$ 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 [50]; in [14] the buffer size is allowed to be infinite, and the background process has only two states.

Model and results: (ii) Linear stochastic fluid networks. In a series of pa-

pers [27, 28, 29], 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 [29]. 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. 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 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 [29] prove that the Laplace-Stieltjes transform of the steady-state buffer content W is given by

$$E[\mathrm{e}^{-\alpha W}] = \exp(-\int_0^\infty \eta(\mathrm{e}^{-rs}\alpha)\mathrm{d}s). \tag{1}$$

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 [26]. 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].

3 Flow level 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

4

[30], is that packet-level dynamics can be interpreted as implementation of decentralized solutions of optimization problems. This 'reversed engineering' point of view led to the field of network utility maximization [17, 57]. In the context of performance analysis under stochastic dynamics, a key paper is [41].

Model. The above-mentioned insights led to an important class of stochastic systems which in its simplest form can be described as follows. Consider a network with J links, and suppose that link j, $j = 1, \ldots, J$, 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 $A\Lambda \leq C$ and possibly additional individual rate constraints $\Lambda_r/n_r \leq d_r$. $U_r(x)$ 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 [32] 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 $r \geq 1$ 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

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

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 [32] 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 [40]. 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 [25, 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

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

The expected queue length at route r can now be approximated by solving $z^* = \lambda_r E[\min\{D_r, B_r z_r^*/\Lambda_r^*\}]$. This procedure is developed rigorously using fluid limits in [45, 44].

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. [25] 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 [38] 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 [32].

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 [39] for large deviation asymptotics in the case of a large number of users, and to [49] 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 [37] and Miyazawa and Rolski [42]. 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 [31] and Electric Vehicle Charging [6]. Finally, we also feel that time-varying fluid models deserve much more attention [56, 58].

References

[1] I.J.B.F. Adan, E.A. van Doorn, J.A.C. Resing and W.R.W. Scheinhardt (1998). Analysis of a single-server queue interacting with a fluid reservoir. *Queueing Systems* **29**, 313-336.

[2] S. Ahn and V. Ramaswami (2003). Fluid flow models and queues – a connection by stochastic coupling. *Stochastic Models* **19**, 325-348.

[3] S. Ahn and V. Ramaswami (2005). Efficient algorithms for transient analysis of stochastic fluid flow models. *J. Appl. Probab.* **42**, 531-549.

[4] N. Akar and K. Sohraby (2004). Infinite- and finite-buffer Markov fluid queues: A unified analysis. *J. Appl. Probab.* **41**, 557-569.

[5] D. Anick, D. Mitra and M.M. Sondhi (1982). Stochastic theory of a data-handling system with multiple sources. *Bell Syst. Techn. J.* **61**, 1871-1894.

[6] A. Aveklouris, M. Vlasiou and B. Zwart. (2017). A stochastic resource sharing model for electric vehicle charging. https://arxiv.org/abs/1711.05561

[7] F. Baskett, K.M. Chandy, R.R. Muntz and F.G. Palacios (1975). Open, closed and mixed networks of queues with different classes of customers. *J. ACM* **22**, 248-260.

[8] T. Bonald and A. Proutière (2002). Insensitivity in processor-sharing networks. *Performance Evaluation* **49**, 193-209.

[9] T. Bonald and A. Proutière (2004). On performance bounds for balanced fairness. *Performance Evaluation* **55**, 25-50.

[10] S. Borst, R. Egorova, B. Zwart (2014). Fluid limits for bandwidth-sharing networks in overload. *Mathematics of Operations Research* **39**, 533-560.

[11] S. Borst, R. Núñez-Queija, B. Zwart (2006). Sojourn time asymptotics in processor-sharing queues. *Queueing Systems* **53**, 31-51.

[12] S. Borst, B. Zwart. (2005). Fluid queues with heavy-tailed $M/G/\infty$ input. *Mathematics of Operations Research* **30**, 852-879.

[13] O.J. Boxma and V. Dumas (1998). Fluid queues with long-tailed activity period distributions. *Computer-Communications* **21**, 1509-1529.

[14] O.J. Boxma, H. Kaspi, O. Kella and D. Perry (2005). On/off storage systems with state-dependent input, output and switching rates. *Prob. Eng. Inform. Sci.* **19**, 1-14.

[15] O.J. Boxma and D. Perry (2009). On the cycle maximum of mountains, dams and queues. *Commun. in Statistics - Part A. Theory and Methods* **38**, 2706-2720.

[16] Hong Chen and David D. Yao (2001). *Fundamentals of Queueing Networks. Performance, Asymptotics, and Optimization.* Applications of Mathematics 46. Stochastic Modelling and Applied Probability. Springer-Verlag, New York.

[17] M. Chiang, S. Low, A. Calderbank and J.C. Doyle (2007). Layering as optimization decomposition: a mathematical theory of network architectures. *Proceedings of the IEEE* **95**, 255-312.

[18] J.W. Cohen (1979). The multiple phase service network with generalized processor sharing. *Acta Informatica* **12**, 245-284.

[19] J.W. Roberts (ed.) (1992). Information Technologies and Sciences. COST 224. Performance Evaluation and Design of Multiservice Networks. Final report of the COST project.

[20] J.G. Dai (1995). On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Ann. Appl. Probab.* **5**, 49-77.

[21] K. Debicki, A.B. Dieker and T. Rolski (2007). Quasi-product forms for Lévy-driven fluid networks. *Mathematics of Operations Research* **32**, 629-647.

[22] Krzysztof Debicki and Michel Mandjes (2015). *Queues and Lévy Fluctuation Theory*. Springer-Verlag, New York.

[23] R. Gibbens, S. Sargood, C. Van Eijl, F. Kelly, H. Azmoodeh, R. Macfadyen and N. Macfadyen (2000). Fixed-point models for the end-to-end performance analysis of IP networks, in: 13th ITC Special Seminar: IP Traffic Management, Modeling and Management.

[24] P. Jelenkovic, P. Momcilovic (2003). Asymptotic loss probability in a finite buffer queue with heterogeneous heavy-tailed fluid on-off processes. *Annals of Applied Probability* **13**, 576-603.

[25] W. Kang, F.P. Kelly, N.H. Lee, and R.J. Williams (2009). State space collapse and diffusion approximation for a network operating under a fair bandwidth sharing policy. *Ann. Appl. Probab.* **19**, 1719-1780.

[26] O. Kella and W. Stadje (2002). Markov-modulated linear fluid networks with Markov additive input. *J. Appl. Probab.* **39**, 413-420.

[27] O. Kella and W. Whitt (1992). A tandem fluid network with Lévy input. In: *Queueing and Related Models*, U.N. Bhat and I.V. Basawa (eds.), Clarendon Press, Oxford, pages 112-128.

[28] O. Kella and W. Whitt (1996). Stability and structural properties of stochastic fluid networks. *J. Appl. Probab.* **33**, 1169-1180.

[29] O. Kella and W. Whitt (1999). Linear stochastic fluid networks. *J. Appl. Probab.* **36**, 244-260.

[30] F. Kelly (1997). Charging and rate control for elastic traffic. *European Transactions on Telecommunications* **8**, 33–37.

[31] F. Kelly and R. Williams (2010). Heavy traffic on a controlled motorway. In: N.H. Binghem and C.M. Goldie (eds.), *Probability and Mathematical Genetics: Papers in Honour of Sir John Kingman*, Cambridge University Press, Cambridge, pp. 416-445.

[32] F. Kelly and E. Yudinova (2014). *Stochastic Networks*. Cambridge University Press.

[33] L. Kosten (1974). Stochastic theory of a multi-entry buffer (I). *Delft Progress Report* **1**, 10-18.

[34] L. Kosten (1974). Stochastic theory of a multi-entry buffer (II). *Delft Progress Report* **1**, 44-50.

[35] L. Kosten and O.J. Vrieze (1975). Stochastic theory of a multi-entry buffer (III). *Delft Progress Report* **1**, 103-115.

[36] V.G. Kulkarni (1997). Fluid models for single buffer systems. In: J.H. Dshalalow (ed.), *Frontiers in Queueing*, CRC Press, Boca Raton (Fl.), pp. 321-338.

[37] P. Lieshout and M. Mandjes (2008). Asymptotic analysis of Lévy-driven tandem queues. *Queueing Systems* **60**, 203-226.

[38] M. Mandjes and J. Kim (2001). Analysis of a phase transition phenomenon in packet networks. *Advances in Applied Probability* **31**, 360-380.

[39] M.R.H. Mandjes, D. Mitra and W.R.W. Scheinhardt (2003). Models of network access using feedback fluid queues. *Queueing Systems* **44**, 365-398.

[40] Laurent Massoulié (2007). Structural properties of proportional fairness: stability and insensitivity. *Ann. Appl. Probab.* **17**, 809-839.

[41] L. Massoulié and J. Roberts (1999). Bandwidth sharing: objectives & algorithms. *Proc. of IEEE Infocom*, 1395-1403.

[42] M. Miyazawa and T. Rolski (2009). Tail asymptotics for a Lévy-driven tandem queue with an intermediate input. *Queueing Systems* **63**, 323-353.

[43] N.U. Prabhu (1998). *Stochastic Storage Processes: Queues, Insurance Risk, and Dams*. Springer, New York.

[44] J.E. Reed and B. Zwart (2014). Limit theorems for bandwidth sharing networks with rate constraints. *Oper. Res.* **62**, 1453-1466.

[45] M. Remerova, J. Reed and B. Zwart (2014). Fluid limits for bandwidth-sharing networks with impatience. *Mathematics of Operations Research* **39**, 746-774.

[46] S. Resnick and G. Samorodnitsky (2001). Steady state distribution of the buffer content for $M/G/\infty$ input fluid queues. *Bernoulli* **7**, 191-210.

[47] J.W. Roberts (2004). A survey on statistical bandwidth sharing. *Computer Networks* **45**, 319-332.

[48] A.N. Rybko and A.L. Stolyar (1992). On the ergodicity of random processes that describe the functioning of open queueing networks. *Problemy Peredachi Informatsii* **28**, 3-26. (translation in *Problems Inform. Transmission* **28**, 199-220).

[49] Y. Sakuma and M. Miyazawa (2004). Asymptotic behavior of the loss probability for a feedback finite fluid queue with downward jumps. In: W. Yue, Y. Takahashi and H. Takagi (eds.), *Advances in Queueing Theory and Network Applications*, Springer, New York, pp. 195-211.

[50] W. Scheinhardt, N. van Foreest and M. Mandjes (2005). Continuous feedback fluid queues. *Oper. Res. Letters* **33**, 551-559.

[51] A. da Silva Soares and G. Latouche (2002). Further results on the similarity between fluid queues and QBDs. In: G. Latouche and P. Taylor (eds.), *Matrix-Analytic Methods: Theory and Applications*, World Scientific, Singapore, pp. 89-106.

[52] A. da Silva Soares and G. Latouche (2006). Matrix-analytic methods for fluid queues with finite buffers. *Performance Evaluation* **63**, 295-314.

[53] M. Vlasiou, J. Zhang and Zwart (2014). Insensitivity of Proportional Fairness in Critically Loaded Bandwidth Sharing Networks. Under review.

[54] N. Walton (2009). Proportional fairness and its relationship with multiclass queueing networks. *Ann. Appl. Probab.* **19**, 2301-2333.

[55] W. Whitt (2002). *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and their Application to Queues*. Springer, New York.

[56] W. Whitt (2017). Time-varying queues. *Queueing Models and Service Management*, to appear.

[57] Y. Yi and M. Chiang (2008). Stochastic network utility maximization. *European Transactions on Telecommunications* **19**, 421-442.

[58] B. Zhang and B. Zwart (2012). Fluid models for many-server Markovian queues in a changing environment. *Oper. Res. Letters* **40**, 573-577.

[59] Bert Zwart, Sem Borst and Michel Mandjes (2004). Exact asymptotics for fluid queues fed by multiple heavy-tailed on-off flows. *Ann. Appl. Probab.* **14**, 903-957.

Addresses:

Onno Boxma: Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands; o.j.boxma@tue.nl

Bert Zwart: Centrum Wiskunde & Informatica (CWI), P.O. Box 94079, 1090 GB Amsterdam, The Netherlands, and Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands; bert.zwart@cwi.nl