Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

A fluid system with coupled input and output, and its application to bottlenecks in ad hoc networks

M.R.H. Mandjes, F. Roijers

REPORT PNA-R0603 MARCH 2006

Centrum voor Wiskunde en Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

Copyright © 2006, Stichting Centrum voor Wiskunde en Informatica
P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

A fluid system with coupled input and output, and its application to bottlenecks in ad hoc networks

ABSTRACT
This paper studies a fluid queue with coupled input and output. Flows arrive according to a Poisson process, and when n flows are present, each of them transmits traffic into the queue at a rate c/(n + 1), where the remaining c/(n + 1) is used to serve the queue. We assume exponentially distributed flow sizes, so that the queue under consideration can be regarded as a system with Markov fluid input. The rationale behind studying this queue lies in ad hoc networks: bottleneck links have roughly this type of sharing policy. We consider four performance metrics: (i) the stationary workload of the queue, (ii) the queueing delay, i.e., the delay of a 'packet' (a fluid particle) that arrives at the queue at an arbitrary point in time, (iii) the flow transfer delay, i.e., the time elapsed between arrival of a flow and the epoch that all its traffic has been put into the queue, and (iv) the sojourn time, i.e., the flow transfer time increased by the time it takes before the last fluid particle of the flow is served. For each of these random variables we compute the Laplace transform. The corresponding tail probabilities decay exponentially, as is shown by a large-deviations analysis.

2000 Mathematics Subject Classification: 60K25
Keywords and Phrases: fluid queues, flow-level analysis, sojourn times, ad hoc networks
Note: This work has been carried out partly in the Dutch BSIK/BRICKS and Easy Wireless projects.

A fluid system with coupled input and output, and its application to bottlenecks in ad hoc networks

Michel Mandjes* and Frank Roijers†

February 28, 2006

Abstract

This paper studies a fluid queue with coupled input and output. Flows arrive according to a Poisson process, and when n flows are present, each of them transmits traffic into the queue at a rate $\mathrm{C}/(n+1)$, where the remaining $\mathrm{C}/(n+1)$ is used to serve the queue. We assume exponentially distributed flow sizes, so that the queue under consideration can be regarded as a system with Markov fluid input. The rationale behind studying this queue lies in ad hoc networks: bottleneck links have roughly this type of sharing policy. We consider four performance metrics: (i) the stationary workload of the queue, (ii) the queueing delay, i.e., the delay of a 'packet' (a fluid particle) that arrives at the queue at an arbitrary point in time, (iii) the flow transfer delay, i.e., the time elapsed between arrival of a flow and the epoch that all its traffic has been put into the queue, and (iv) the sojourn time, i.e., the flow transfer time increased by the time it takes before the last fluid particle of the flow is served. For each of these random variables we compute the Laplace transform. The corresponding tail probabilities decay exponentially, as is shown by a large-deviations analysis.

*M. Mandjes (email: michel@cwi.nl) is with CWI, P.O. Box 94079, 1090 GB Amsterdam, the Netherlands. He is also affiliated to Korteweg-de Vries Institute, University of Amsterdam, Amsterdam, the Netherlands, and EURANDOM, Eindhoven, the Netherlands. This work has been carried out partly in the Dutch BSIK/BRICKS project.

†F. Roijers (email: f.roijers@telecom.tno.nl) is with TNO ICT, Brasserplein 2, P.O. Box 5050, 2600 GB Delft, The Netherlands. He is also affiliated to CWI, Amsterdam, the Netherlands. This work has been carried out partly in the SENTER-NOVEM funded project Easy Wireless.

1 Introduction

Standard Markov fluid queues consist of *traffic sources* feeding into a *queue* that is emptied at a constant rate, say C. The sources are for instance of the exponential on-off type: they alternate between activity periods (with a duration that is exponentially distributed with mean μ^{-1} during which traffic is generated at some fixed rate, say p) and silences (which have an exponential distribution with mean λ^{-1}). If there are N of such sources (i.i.d.), and if $\mathrm{N}p > \mathrm{C}$, every now and then the buffer of the queue fills. Under the stability condition $\mathrm{N}pf < \mathrm{C}$, with $f := \lambda/(\lambda+\mu)$ the fraction of time each source is on, the queue's workload has a steady-state distribution, say $W^\star$. A detailed performance analysis of this workload is known, see e.g. [1].
In practical situations, however, often the role played by the sources and the queue is rather different. In this paper we consider a stylized model of a so-called bottleneck in an ad hoc wireless communication network; here it suffices to understand the working of ad hoc networks at an abstract level, but for more details, particularly on Quality-of-Service aspects, we refer to the excellent survey [3]. Flows, for instance arriving according to a Poisson process, wish to send their information through the bottleneck node. The complicating issue, however, is that the total transmission capacity is fixed (say C), and this capacity should be used both to feed the traffic from the flows into the bottleneck node, *and* to serve the queue of the bottleneck node. A common situation is that, when n flows are present, each of these uses $\mathrm{C}/(n+1)$ to transmit their traffic into the queue, while the remaining capacity $\mathrm{C}/(n+1)$ is used to drain the queue. The question arises whether the analysis techniques for standard Markov fluid models carry over to these fluid systems with coupled input and output. From a practical perspective, one is interested in characterizing the steady-state workload, the queueing delay, etc., in terms of expectations, variances, and higher moments, but also their tail behavior.

Standard Markov fluid queues have been studied extensively. In the seminal studies [1, 15] a system of differential equations (known as Kolmogorov forward equations) is derived for $\mathbb{P}(W^\star \leq x, N^\star = n)$, where $N^\star$ is the number of sources in the on-state in steady-state. Later these results have been extended in many dimensions. To mention a few: one has considered heterogeneous sources, sources with a more general structure than exponential on-off, see e.g. [20], there have been rather explicit results for the case that the sources have a so-called birth-death structure [9] or have a countably infinite state-space, see e.g. [26], and also models have been studied in which the source behavior depends on the current workload [17, 24]. In addition there has been considerable interest in so-called large-buffer asymptotics, i.e., expansions of $\mathbb{P}(W^\star > x)$ for large x; these relate nicely to a notion of effective bandwidths [11, 14].

The goal of the present paper is to extend the results for standard Markov fluid queues to our model of a bottleneck in an ad hoc network. Interestingly, not even the stability criterion is completely trivial, as essentially all traffic has to be 'served' twice (it has to be transmitted into the queue, and subsequently it has to be served by the queue); as a result the common stability condition that the mean input rate, say m, be smaller than C does not apply. In *Section 2* we present, besides a detailed model description, the correct stability requirement.
The second aim is to characterize the steady-state workload distribution. It is not hard to see that this can be analyzed by setting up a system of Kolmogorov forward equations, but the special structure

allows more explicit results. The crucial helpful property of our model with coupled input and output is that the queue drains only when there are no flows present. This property entails that our model strongly resembles the classical M/G/1 queueing model, and hence the Laplace transform (LT) of the steady-state workload distribution can be given explicitly. These results are presented in *Section 3.*
In standard Markov fluid queues there is a one-to-one mapping between the buffer content that a 'fluid particle' sees upon arrival, and the delay it has: if it sees x units traffic in the queue, it leaves the queue after x/C units of time. As a consequence, for standard Markov fluid queues, the queueing delay distribution follows immediately from the steady-state workload distribution. This is not the case for our model with coupled input and output; more specifically, when considering a tagged fluid particle that arrived at time 0, flows arriving in the future have impact on the service capacity available to the queue, and hence also on the delay of the fluid particle. This makes the analysis of the queueing delay non-standard. A full characterization of its Laplace transform, also relying on the results of Section 3, is given in *Section 4.*
In *Section 5* we study the flow transfer delay, i.e., the time it takes before the flow has transmitted all its traffic into the queue. This delay is essentially the absorption time of a certain continuous-time Markov chain. Again, the solution is given in terms of Laplace transforms.
The sojourn time of a flow is defined as the flow transfer time of an arbitrary flow increased by the time it takes before the last fluid particle of the flow is served. As these two components are correlated the Laplace transform of the sojourn time does not immediately follow from the results of Sections 4 and 5. The derivation of the transform of the sojourn time explicitly uses the fact that the buffer content cannot decrease during any flow transfer time. These issues are dealt with in *Section 6.*
Having the Laplace transforms of the workload, queueing delay, and flow transfer delay at our disposal, a next question is how the tails of these distributions behave. In *Section 7* it is shown that they decay exponentially, and, relying on large-deviations tools, the decay rates are derived.
Section 8 concludes and identifies a few challenging subjects for future research. In particular, it is discussed to what class of sharing policies (between the flows and the queue) our results can be extended.

2 Model and background

In this section, we first give a detailed description of our model. Then we derive the steady-state distribution of the number of flows simultaneously present in the system, allowing us to give a precise stability condition.

2.1 Model

Consider a queueing system at which flows arrive according to a Poisson process, transmit traffic into a queue, and leave when ready. When there are n flows active, any flow can transmit its traffic into the queue at rate $\mathrm{C}/(n+1)$, while a rate $\mathrm{C}/(n+1)$ is used to serve the queue; as a consequence, the queue only drains when there are no flows present, while it stays at the same level if exactly one flow is active. Suppose that we impose the admission control policy that the system accommodates maximally $\mathrm{N} \in \mathbb{N}$ flows simultaneously; in this way each active flow (as well as the queue) is guaranteed

at least a transmission rate $\mathrm{C}/(\mathrm{N}+1)$.
We let N_t denote the number of flows present (i.e., feeding traffic into the queue) at time t. It is not hard to see that, under the assumption of exponentially distributed flow sizes (with mean μ^{-1}) and interarrival times with mean λ^{-1}, the process N_t constitutes a Markov chain on $\{0,\ldots,\mathrm{N}\}$, with generator matrix

$$Q := \begin{pmatrix} -\lambda & \lambda & & & & \\ \mu_1\mathrm{C} & -\mu_1\mathrm{C}-\lambda & \lambda & & & \\ & \mu_2\mathrm{C} & -\mu_2\mathrm{C}-\lambda & \lambda & & \\ & & \ddots & \ddots & \ddots & \\ & & & & \mu_{\mathrm{N}}\mathrm{C} & -\mu_{\mathrm{N}}\mathrm{C} \end{pmatrix}, \tag{1}$$

where $\mu_n := \mu n/(n+1)$. When $N_t = n$, the aggregate traffic rate generated by the flows is $r_{\mathrm{I},n} := \mathrm{C}n/(n+1)$, while the queue's output rate is $r_{\mathrm{O},n} := \mathrm{C}/(n+1)$, such that the net rate of change of the queue is 0 when $Q_t = N_t = 0$, and otherwise, for $n \in \{0,\ldots,\mathrm{N}\}$,

$$r_{\mathrm{A},n} := r_{\mathrm{I},n} - r_{\mathrm{O},n} = \mathrm{C}\,\frac{n-1}{n+1}.$$

Define $R_{\mathrm{I}} := \mathrm{diag}\{r_{\mathrm{I}}\}$, $R_{\mathrm{O}} := \mathrm{diag}\{r_{\mathrm{O}}\}$, and $R_{\mathrm{A}} := R_{\mathrm{I}} - R_{\mathrm{O}}$.

Two variants of this model. In a first variant, one lets $\mathrm{N} \to \infty$, thus getting a countably infinite state space. This means that there is no admission control imposed on the number of flows.
In a second variant, there are N sources that can be potentially active, and each source has a silence time that is exponentially distributed with mean λ^{-1}. The $q_{n,n+1}$ should be $(\mathrm{N}-n)\lambda$ rather than λ (for $n = 0,\ldots,\mathrm{N}-1$).

2.2 Stability condition

Due to the sharing of the service capacity between the flows and the queue, the stability condition of this model is not standard. Also, a fraction of the flows is rejected because they enter when already N flows are present. In this subsection we find the stability condition and the blocking probability.
To find a condition on λ, μ and C under which the queue is stable, we first determine the equilibrium distribution π of $(N_t)_{t\in\mathbb{R}}$. Trivially, the balance equations are

$$\pi_n \mu_n \mathrm{C} = \pi_{n-1}\lambda, \quad n = 1,\ldots,\mathrm{N}.$$

Recursively solving these equations, it is not hard to derive, with ϱ defined as $\lambda/(\mu\mathrm{C})$, that

$$\pi_n = \frac{\varrho^n(n+1)}{\sum_{k=0}^{\mathrm{N}} \varrho^k(k+1)}.$$

Standard calculus on the geometric series yields

$$\begin{aligned} \sum_{k=0}^{\mathrm{N}} \varrho^k(k+1) &= \frac{\mathrm{d}}{\mathrm{d}\varrho}\left(\sum_{k=0}^{\mathrm{N}} \varrho^{k+1}\right) = \frac{\mathrm{d}}{\mathrm{d}\varrho}\left(\varrho\,\frac{1-\varrho^{\mathrm{N}+1}}{1-\varrho}\right) \\ &= \frac{1-\varrho^{\mathrm{N}+1}(\mathrm{N}+2)+\varrho^{\mathrm{N}+2}(\mathrm{N}+1)}{(1-\varrho)^2}. \end{aligned}$$

The equilibrium condition of the fluid model is $\sum_{n=0}^{\mathrm{N}} \pi_n r_{\mathrm{A},n} < 0$, after considerable algebra translating into

$$\frac{-1 + 2\varrho - \varrho^{\mathrm{N}+1}\mathrm{N} + \varrho^{\mathrm{N}+2}(\mathrm{N}-1)}{1 - \varrho^{\mathrm{N}+1}(\mathrm{N}+2) + \varrho^{\mathrm{N}+2}(\mathrm{N}+1)} \cdot \mathrm{C} < 0.$$

Due to the PASTA-property, the probability of an arbitrary arriving flow being blocked is

$$\pi_{\mathrm{N}} = \frac{\varrho^{\mathrm{N}}(\mathrm{N}+1)(1-\varrho)^2}{1 - \varrho^{\mathrm{N}+1}(\mathrm{N}+2) + \varrho^{\mathrm{N}+2}(\mathrm{N}+1)}.$$

Special case of $\mathrm{N} \to \infty$. Interestingly, for $\mathrm{N} \to \infty$, the equilibrium probabilities π_n have the form $(1-\varrho)^2(n+1)\varrho^n$, and the equilibrium condition $(-1+2\varrho)\mathrm{C} < 0$, or, equivalently, $2\lambda/\mu < \mathrm{C}$. The latter condition has an appealing interpretation. In the model with $\mathrm{N} \to \infty$, the input process is essentially a Poisson stream (arriving at rate λ) of flows that have mean size μ^{-1}. Every flow has to be processed twice: first it has to be put into the queue, and then it has to be served by the queue. This immediately leads to the stability condition $2\lambda/\mu < \mathrm{C}$.

3 Steady-state workload distribution

In this section we study the steady-state workload of the queue. As mentioned in the introduction, one could set up a system of Kolmogorov forward equations as in [1], which, in conjunction with the proper boundary condition, characterize the distribution function (in terms of eigenvalues and eigenvectors of some matrix). Due to the specific structure of our model, however, rather explicit results for the Laplace transform of the steady-state workload can be given. In particular we exploit the property that the buffer content only decreases when no flows are present, and the fact that these periods have an exponential duration, cf. for instance [6, 26]. As a consequence, our model is closely related to the family of M/G/1 systems.

3.1 Busy periods

In our analysis of the steady-state workload distribution, we need the notion of busy periods. A busy period B is, in this context, defined as a period that starts at an epoch at which $(N_t)_{t\in\mathbb{R}}$ jumps from 0 to 1, and ends at a moment that it jumps from 1 to 0. We introduce the auxiliary quantity B_n, for $n = 1, \ldots, \mathrm{N}$:

$$B_n := \inf\{t \geq 0 : N_t = n-1 \mid N_0 = n\};$$

evidently $B \stackrel{\mathrm{d}}{=} B_1$. In our analysis we also need the distribution of T, the *net* amount of traffic entering the queue (i.e., the increase of the buffer content) during B. Define $A(s,t) := \int_s^t r_{\mathrm{A},N_u}\mathrm{d}u$. Then $T \stackrel{\mathrm{d}}{=} T_1$, with

$$T_n \stackrel{\mathrm{d}}{=} A(0, B_n).$$

Analysis of the Laplace transform. Using standard arguments, cf. [13, 21, 23], we find the recursion, for $n = 1, \ldots, \mathrm{N}-1$,

$$\mathbb{E}e^{-sT_n} = \frac{\lambda}{\lambda + \mu_n\mathrm{C} + r_{\mathrm{A},n}s}\,\mathbb{E}e^{-sT_{n+1}}\mathbb{E}e^{-sT_n} + \frac{\mu_n\mathrm{C}}{\lambda + \mu_n\mathrm{C} + r_{\mathrm{A},n}s}, \tag{2}$$

while for $n = \mathrm{N}$ the random variable T_n is exponentially distributed with mean $r_{\mathrm{N}}/(\mu_{\mathrm{N}}\mathrm{C})$:

$$\mathbb{E}e^{-sT_{\mathrm{N}}} = \frac{\mu_{\mathrm{N}}\mathrm{C}}{\mu_{\mathrm{N}}\mathrm{C} + r_{\mathrm{A,N}}s}. \tag{3}$$

The above implies that $\mathbb{E}e^{-sT}$ is the solution of a finite recursion, of which the starting condition is known (namely $\mathbb{E}e^{-sT_{\mathrm{N}}}$). The nature of the formula for $\mathbb{E}e^{-sT}$ is an N-fold iterated fraction.

Mean and second moment. Similarly to the above, we can find a recursion for the mean. It reads

$$\mathbb{E}T_n = \frac{r_{\mathrm{A},n}}{\mu_n\mathrm{C}} + \frac{\lambda}{\mu_n\mathrm{C}}\mathbb{E}T_{n+1} = \ldots = \sum_{i=n}^{\mathrm{N}} \frac{\lambda^{i-n} r_{\mathrm{A},i}}{\mu_n \cdots \mu_i \mathrm{C}^{i-n+1}} = \frac{1}{n\mu}\sum_{i=n}^{\mathrm{N}} \varrho^{i-n}(i-1).$$

In particular,

$$\mathbb{E}T = \frac{1}{\mu}\cdot\frac{\varrho}{(1-\varrho)^2}\left(1 - \varrho^{\mathrm{N}-1}\mathrm{N} + \varrho^{\mathrm{N}}(\mathrm{N}-1)\right);$$

for the case $\mathrm{N} \to \infty$, this converges to the clean expression $\varrho/(\mu(1-\varrho)^2)$. For the second moment we can develop a recursion in the same way, again by distinguishing between the period where the number of flows is n, and the first jump afterwards. We obtain

$$\begin{aligned}\mathbb{E}T_n^2 &= \frac{2r_{\mathrm{A},n}^2}{(\lambda+\mu_n\mathrm{C})^2} + \frac{2r_{\mathrm{A},n}}{\lambda+\mu_n\mathrm{C}}\frac{\lambda}{\lambda+\mu_n\mathrm{C}}(\mathbb{E}T_n + \mathbb{E}T_{n+1}) \\ &\quad + \frac{\lambda}{\lambda+\mu_n\mathrm{C}}(\mathbb{E}T_n^2 + 2\mathbb{E}T_n\mathbb{E}T_{n+1} + \mathbb{E}T_{n+1}^2),\end{aligned}$$

with $\mathbb{E}T_{\mathrm{N}}^2 = 2r_{\mathrm{A,N}}^2/(\mu_{\mathrm{N}}\mathrm{C})^2$. The recursion can be restated as $\mathbb{E}T_n^2 = \alpha_n\mathbb{E}T_{n+1}^2 + \beta_n$, with $\alpha_n := \lambda/(\mu_n\mathrm{C})$, and

$$\beta_n := 2\frac{r_{\mathrm{A},n}^2}{\mu_n\mathrm{C}(\lambda+\mu_n\mathrm{C})} + 2\frac{r_{\mathrm{A},n}\lambda}{\mu_n\mathrm{C}(\lambda+\mu_n\mathrm{C})}(\mathbb{E}T_n + \mathbb{E}T_{n+1}) + 2\frac{\lambda}{\mu_n\mathrm{C}}\mathbb{E}T_n\mathbb{E}T_{n+1};$$

notice that β_n, $n = 1, \ldots, \mathrm{N}$, are known numbers, in view of the formulae for $\mathbb{E}T_n$ above. The solution of the recursion is, with the 'empty product' defined as 1,

$$\mathbb{E}T_n^2 = \left(\sum_{i=n}^{\mathrm{N}-1}\left(\prod_{j=n}^{i-1}\alpha_j\right)\beta_i\right) + \left(\prod_{j=n}^{\mathrm{N}-1}\alpha_j\right)\mathbb{E}T_{\mathrm{N}}^2.$$

In particular, by inserting $n = 1$ we derive the second moment of T:

$$\mathbb{E}T^2 = \left(\sum_{i=1}^{\mathrm{N}-1}\varrho^{i-1}i\beta_i\right) + 2\varrho^{\mathrm{N}-1}\mathrm{N}\frac{r_{\mathrm{A,N}}^2}{(\mu_{\mathrm{N}}\mathrm{C})^2}.$$

3.2 Steady-state workload

The steady-state workload, say $W^\star$, is, according to Reich's formula [22], distributed as

$$W^\star \stackrel{\mathrm{d}}{=} M := \sup_{t\geq 0} A(-t,0) \stackrel{\mathrm{d}}{=} \sup_{t\geq 0} A(0,t),$$

where the second equality in distribution is due to the reversibility of $(N_t)_{t\in\mathbb{R}}$. In this subsection, we derive an explicit expression for the LT of M. Define

$$M_i := \sup_{t\geq 0}\{A(0,t) \mid N_0 = i\};$$

clearly $\mathbb{E}e^{-sM} = \sum_{n=0}^{\mathrm{N}} \pi_n \mathbb{E}e^{-sM_n}$, hence we have to find expressions for $\mathbb{E}e^{-sM_n}$, for $n = 0, \ldots, \mathrm{N}$.
As for $n = 1, \ldots, \mathrm{N}$, during periods B_n the queue does not decrease, the random variables T_n are nonnegative almost surely. In fact, $A(0,t)$ attains its maximum either at time 0, or at an epoch at which N_t jumps from 1 to 0. These observations lead to the following equality in distribution:

$$M_n \stackrel{\mathrm{d}}{=} T_n + T_{n-1} + \cdots T_1 + M_0,$$

with $B_n, B_{n-1}, \ldots, B_1, M_0$ independent. This entails that

$$\mathbb{E}e^{-sM_n} = \mathbb{E}e^{-sM_0} \cdot \prod_{i=1}^{n} \mathbb{E}e^{-sT_i},$$

for $n = 0, \ldots, \mathrm{N}$ (again defining the empty product to be 1). With a recipe to compute $\mathbb{E}e^{-sT_i}$ given in the previous section, we are left to compute $\mathbb{E}e^{-sM_0}$.
We now introduce an embedding that facilitates easy computation of the LT of M_0. Starting in 0, the maximum of $A(0,t)$ over $t \geq 0$ equals the maximum of $\sum_{j=0}^{i}(X_j - Y_j)$ over $i = 0, 1, \ldots$, with the X_j i.i.d. samples, distributed as T, and the Y_j i.i.d. samples from an exponential distribution with mean C/λ (where also the sequences X_j and Y_j are independent). The LT of the latter maximum is given by the celebrated Pollaczek-Khinchine formula, see for instance [2], so that we arrive at

$$\mathbb{E}e^{-sM_0} = \left(1 - \frac{\lambda \mathbb{E}T}{\mathrm{C}}\right) \frac{s}{s - (\lambda/\mathrm{C})(1 - \mathbb{E}e^{-sT})}.$$

Our final result is stated in the following theorem.

Theorem 3.1 *The LT of the steady-state workload is given by, $s \geq 0$,*

$$\mathbb{E}e^{-sW^\star} = \mathbb{E}e^{-sM} = \sum_{n=0}^{\mathrm{N}} \pi_n \left(1 - \frac{\lambda \mathbb{E}T}{\mathrm{C}}\right) \frac{s}{s - (\lambda/\mathrm{C})(1 - \mathbb{E}e^{-sT})} \left(\prod_{i=1}^{n} \mathbb{E}e^{-sT_i}\right),$$

where the $\mathbb{E}e^{-sT_i}$ follow from (2) *and* (3).

Moreover, we can also consider the joint distribution of the steady-state workload $W^\star$ and number of flows $N^\star$. It turns out that

$$\mathbb{E}(e^{-sW^\star} 1\{N^\star = n\}) = \pi_n \left(1 - \frac{\lambda \mathbb{E}T}{\mathrm{C}}\right) \frac{s}{s - (\lambda/\mathrm{C})(1 - \mathbb{E}e^{-sT})} \left(\prod_{i=1}^{n} \mathbb{E}e^{-sT_i}\right). \tag{4}$$

The above results also enable calculation of the mean steady-state workload:

$$\mathbb{E}W^\star = \left(\frac{1}{2} \frac{\lambda \mathbb{E}T^2}{\mathrm{C} - \lambda \mathbb{E}T}\right) + \left(\sum_{n=0}^{\mathrm{N}} \left(\pi_n \sum_{i=1}^{n} \mathbb{E}T_i\right)\right),$$

following the convention that the empty sum is defined as 0.

4 Queueing delay distribution

As argued in the introduction, it is a nontrivial step to translate the steady-state workload distribution into the queueing delay distribution: for standard Markov fluid queues the buffer content seen by a fluid particle arriving, say at time 0, fully determines the epoch at which it will leave the queue, whereas in our system with coupled input and output the arrivals and departures of flow after 0 has impact. In the first subsection we analyze the so-called *virtual queueing delay*, i.e., the delay experienced by a fluid particle arriving at a random point in time (i.e., a 'time average'), whereas the second subsection characterizes the queueing delay of an arbitrary fluid particle (i.e., a 'traffic average').

4.1 Virtual queueing delay

Let $D^\star$ denote the delay experienced by a fluid particle arriving at the queue in steady state, say for ease at time 0; this type of delay is sometimes referred to as virtual queueing delay. Let $O(0,t)$ denote the amount of output capacity available in the interval $[0,t)$. Then, cf. [16, Section III],

$$\begin{aligned}\mathbb{E}e^{-sD^\star} &= \int_0^\infty e^{-st}\mathbb{P}(D^\star = t)\mathrm{d}t = \int_0^\infty e^{-st}\mathbb{P}(W^\star = O(0,t))\mathrm{d}t \\ &= \sum_{n=0}^{\mathrm{N}}\int_0^\infty e^{-st}\mathbb{P}(W^\star = O(0,t), N^\star = n)\mathrm{d}t.\end{aligned}$$

Now define, for $z \geq 0$, the random variable τ_z as the time until z units of service have become available:

$$\tau_z := \inf\{t \geq 0 : O(0,t) = z\} = \inf\left\{t \geq 0 : \int_0^t r_{\mathrm{O},N_s}\mathrm{d}s = z\right\};$$

notice that $O(0,t)$ is increasing in t. Using this notion, we get, with some abuse of notation,

$$\mathbb{E}e^{-sD^\star} = \sum_{n=0}^{\mathrm{N}}\int_0^\infty e^{-st}\mathbb{P}(\tau_{W^\star} = t, N^\star = n)\mathrm{d}t,$$

which equals, remarking that $O(0,t)$ depends on $(W^\star, N^\star)$ just through $N^\star$,

$$\sum_{n=0}^{\mathrm{N}}\int_0^\infty\int_0^\infty e^{-st}\mathbb{P}(W^\star = z, N^\star = n)\mathbb{P}(\tau_z = t \mid N^\star = n)\mathrm{d}z\mathrm{d}t.$$

Now we interchange the order of integration, to get

$$\sum_{n=0}^{\mathrm{N}}\int_0^\infty \mathbb{E}(e^{-s\tau_z} \mid N^\star = n)\mathbb{P}(W^\star = z, N^\star = n)\mathrm{d}z.$$

Hence, to further compute this expression, we need to evaluate $\mathbb{E}(e^{-s\tau_z} \mid N^\star = n)$. Fortunately, we have the following proposition at our disposal, cf. [5] and the appendix of [14].

Proposition 4.1 *Consider an irreducible, finite-state (with states $0,\ldots,\mathrm{N}$), continuous-time Markov chain $(X_t)_{t\in\mathbb{R}}$ with generator Q. Let r be a componentwise positive vector of dimension N, and $R := \mathrm{diag}\{r\}$. Define*

$$\tau_z := \inf\left\{t \geq 0 : \int_0^t r_{X_s}\mathrm{d}s = z\right\},$$

and $\xi_n(s,z) := \mathbb{E}(e^{-s\tau_z} \mid X_0 = n)$. *Then, with* $\xi(s,z) = (\xi_1(s,z),\ldots,\xi_{\rm N}(s,z))^{\rm T}$, *and* $\mathbf{1}$ *an* $({\rm N}+1)$*-dimensional vector with 1's,*

$$\xi(s,z) = \exp((R^{-1}Q - sR^{-1})z)\mathbf{1}. \tag{5}$$

In addition, the eigenvalues $\delta_0(s),\ldots,\delta_{\rm N}(s)$ *of* $R^{-1}Q - sR^{-1}$ *are real, negative, and unique* $(s>0)$.

Proof. A straightforward conditioning argument yields, with $q_j := -q_{jj}$,

$$\xi_n(s,z) = \sum_{m\neq n} \xi_m(s, z - r_n\,\Delta t)q_{nm}\Delta t \;+\; \xi_n(s, z - r_n\,\Delta t)e^{-s\Delta t}(1 - q_n\,\Delta t) \;+\; o(\Delta t).$$

Now writing $e^{-s\Delta t} = 1 - s\Delta t + O((\Delta t)^2)$, subtracting $\xi_n(s, z - r_n\,\Delta t)$ from both sides, dividing the equation by $r_n\Delta t$, and letting $\Delta t \downarrow 0$, we arrive at

$$\frac{\partial}{\partial z}\xi_n(s,z) = \sum_{m=1}^{\rm N} \frac{q_{nm}}{r_n}\xi_n(s,z) - \xi_n(s,z)\frac{s}{r_n}.$$

In matrix-notation, we have that

$$\frac{\partial}{\partial z}\xi(s,z) = (R^{-1}Q - sR^{-1})\xi(s,z),$$

which yields (5).
Geršgorin's circle theorem [19] implies that each eigenvalue of $M(s) = (m_{ij})_{i,j=0}^{\rm N} := R^{-1}Q - sR^{-1}$ is in at least one of the disks

$$\left\{ z\in\mathbb{C} : \left| z - \frac{q_{ii}-s}{r_i}\right| < \sum_{j\neq i}\frac{q_{ij}}{r_i}\right\},$$

and hence all eigenvalues are in the left half plane. Furthermore, the matrix $M(s)$ is real and tridiagonal with $m_{i,i+1}m_{i+1,i} > 0$ for $i = 0,\ldots,{\rm N}-1$, and hence all its eigenvalues are real and unique, see again [19]. □

Apply Proposition 4.1, with continuous-time Markov chain N_t governed by Q as defined by (1), and $R := R_{\rm O}$ (which is indeed componentwise positive). Recalling that all eigenvalues $\delta_0(s),\ldots,\delta_{\rm N}(s)$ of $M(s) := R_{\rm O}^{-1}Q - sR_{\rm O}^{-1}$ are different, so that we can write, for constants γ_{mn} with $m,n = 0,\ldots,{\rm N}$,

$$\mathbb{E}(e^{-s\tau_z} \mid N^\star = n) = \sum_{m=0}^{\rm N} \gamma_{mn}e^{\delta_m(s)\,z}. \tag{6}$$

Then we have found an explicit expression of the LT of the virtual queueing delay.

Theorem 4.2 *For* $s>0$,

$$\mathbb{E}e^{-sD^\star} = \sum_{n=0}^{\rm N}\sum_{m=0}^{\rm N}\gamma_{mn}\mathbb{E}(e^{\delta_m(s)W^\star}1\{N^\star = n\}),$$

where the γ_{mn} *are as in* (6). *The* $\delta_n(s)$, *for* $n = 0,\ldots,{\rm N}$, *are the eigenvalues of* $R_{\rm O}^{-1}Q - sR_{\rm O}^{-1}$ *(which are negative). An expression for* $\mathbb{E}(e^{-sW^\star}1\{N^\star = n\})$ *is available from Theorem 3.1.*

4.2 'Packet-average' queueing delay

Informally, the previous section gave the LT of the queueing delay 'at an arbitrary point in time'. Clearly, there is a bias between the delay $D^\star$ 'at an arbitrary point in time' and delay $\bar{D}^\star$ 'seen by an arbitrary fluid molecule'. The correction to be made is rather straightforward:

$$\mathbb{E}e^{-s\bar{D}^\star} = \sum_{n=0}^{\mathrm{N}} \left(\frac{r_{\mathrm{I},n}}{\sum_{k=0}^{\mathrm{N}} \pi_k r_{\mathrm{I},k}} \right) \sum_{m=0}^{\mathrm{N}} \gamma_{mn} \mathbb{E}(e^{\delta_m(s)W^\star} 1\{N^\star = n\}),$$

cf. Asmussen [2, Prop. 7.2].

5 Flow transfer delay distribution

Now we focus on the time F it takes for an arbitrary arriving flow to transmit its traffic. We define the transfer time as the time between arrival and the epoch that its last fluid particle has been transmitted into the queue.

5.1 Flow transfer delay

Let the process $(Z_i)_{i\in\mathbb{N}}$ correspond to the number of flows present at (i.e., *just after*) arrival epochs. This process is a Markov chain, with, say, transition matrix $P = (p_{mn})_{m,n=1}^{\mathrm{N}}$. It is clear that Z_i can jump only one level up, or in other words, $p_{mn} = 0$ for all $n > m+1$. It can be verified easily that, for $m = 1, \ldots, \mathrm{N}$ and $n = 1, \ldots, m+1$,

$$p_{mn} = \left(\prod_{k=n}^{m} \frac{\mu_k \mathrm{C}}{\lambda 1\{k \neq n\} + \mu_k \mathrm{C}} \right) \frac{\lambda}{\lambda + \mu_{n-1}\mathrm{C}}.$$

From this the equilibrium distribution π^Z can be computed efficiently due to the fact that the chain can jump just one level upwards. More directly, however, one can argue that we can use the PASTA-property here, such that

$$\pi_n^Z := \frac{\pi_{n-1}}{\sum_{m=0}^{\mathrm{N}-1} \pi_m}. \tag{7}$$

We can now compute the LT of the flow transfer delay. Define F as the transfer delay of a tagged flow, that arrives at, say, time 0, when there are $n-1$ flows present (i.e., there are n flows immediately after the arrival of the tagged flow), $n = 1, \ldots, \mathrm{N}$. We compute, for $n = 1, \ldots, \mathrm{N}$ and $m = 0, \ldots, \mathrm{N}-1$,

$$\phi_{nm}(s) := \mathbb{E}(e^{-sF} 1\{N_{F+} = m\} \mid N_0 = n).$$

A standard linear system can be written down, for $n = 1, \ldots, \mathrm{N}-1$, cf. the analysis for the finite-capacity processor-sharing queue in [4, Section II]:

$$\phi_{nm}(s) = \frac{1}{\lambda + \mu_n \mathrm{C} + s} \left(\lambda \phi_{n+1,m}(s) + \frac{n-1}{n} \mu_n \mathrm{C}\, \phi_{n-1,m}(s) + \frac{1}{n} \mu_n \mathrm{C}\, 1\{n-1 = m\} \right);$$

here the fraction $1/n$ is the probability that at a departure epoch it is the tagged flow that leaves. We also have

$$\phi_{\mathrm{N}m}(s) = \frac{1}{\mu_{\mathrm{N}}\mathrm{C} + s} \left(\frac{\mathrm{N}-1}{\mathrm{N}} \mu_{\mathrm{N}}\mathrm{C}\, \phi_{\mathrm{N}-1,m}(s) + \frac{1}{\mathrm{N}} \mu_{\mathrm{N}}\mathrm{C}\, 1\{\mathrm{N}-1 = m\} \right).$$

We have thus derived, for fixed $m = 0, \ldots, \mathrm{N}-1$ and s, N linear equations in N unknowns; as in [4] it can be shown that the corresponding matrix is, for any $s > 0$, diagonally dominant and thus non-singular, and hence there is a unique solution. The transform of the flow transfer delay of an arbitrary customer now reads

$$\mathbb{E}e^{-sF} = \sum_{n=1}^{\mathrm{N}} \sum_{m=0}^{\mathrm{N}-1} \pi_n^Z \, \phi_{nm}(s). \tag{8}$$

5.2 Representation of flow transfer delay with a phase-type distribution

Alternatively, the flow transfer delay distribution can also be found through a system of Kolmogorov equations. Defining

$$f_{nm}(t) := \mathbb{P}(F > t, N_{F+} = m \mid N_0 = n),$$

it is standard to derive through the usual Δt-argumentation, for $n = 1, \ldots, \mathrm{N}$ and $m = 0, \ldots, \mathrm{N}-1$,

$$\begin{aligned} f_{nm}(t+\Delta t) &= f_{n+1,m}(t)\, \lambda \Delta t \, 1\{n < \mathrm{N}\} \, + \, f_{n-1,m}(t)\, \mu_n \mathrm{C} \frac{n-1}{n} \Delta t \, 1\{n > 1\} \\ &\quad + f_{nm}(t) \left(1 - (\lambda \, 1\{n < \mathrm{N}\} + \mu_n \mathrm{C} \, 1\{n > 1\}) \Delta t\right), \end{aligned}$$

immediately leading to

$$\begin{aligned} f'_{nm}(t) &= \lambda \, 1\{n < \mathrm{N}\} f_{n+1,m}(t) \, + \, \mu_n \mathrm{C} \frac{n-1}{n} \, 1\{n > 1\} f_{n-1,m}(t) \\ &\quad - (\lambda \, 1\{n < \mathrm{N}\} + \mu_n \mathrm{C} \, 1\{n > 1\}) f_{nm}(t). \end{aligned}$$

Define the matrix $Q^\star = (q^\star_{mn})_{m,n=1}^{\mathrm{N}}$ through $q^\star_{n,n-1} := q_{n,n-1}\,(n-1)/n$, and $q^\star_{mn} := q_{mn}$ otherwise. Then we have that the vector $f_m(t) := (f_{m1}(t), \ldots, f_{m\mathrm{N}}(t))^{\mathrm{T}}$ satisfies $f'_m(t) = Q^\star f_m(t)$. Now also observe that the starting condition $f_{mn}(0)$ (again, fix m) follows from

$$\begin{aligned} (\lambda \, 1\{n < \mathrm{N}\} + \mu_n \mathrm{C} \, 1\{n > 1\}) f_{nm}(0) &= \lambda \, 1\{n < \mathrm{N}\} f_{n+1,m}(0) + \frac{n-1}{n} \, \mu_n \mathrm{C} f_{n-1,m}(0) + \\ &\quad \frac{1}{n} \, \mu_n \mathrm{C} 1\{n-1 = m\}; \end{aligned}$$

we call the solution $\bar{f}_m := (\bar{f}_{m1}, \ldots, \bar{f}_{m\mathrm{N}})^{\mathrm{T}}$. We thus have obtained that

$$f_m(t) = \exp(Q^\star t) \bar{f}_m.$$

As $Q^\star$ is strictly diagonally dominant, it is non-singular. Using Geršgorin's theorem, one can prove that the eigenvalues $\bar{\delta}_1, \ldots, \bar{\delta}_{\mathrm{N}}$ have a negative real part. In addition, as $q^\star_{m,m+1} q^\star_{m+1,m} > 0$ and $Q^\star$ is a real and tridiagonal matrix, all eigenvalues are real and unique [19]. These observations imply that we can find constants $\bar{\gamma}_{nm}$ such that

$$\mathbb{P}(F > t \mid N_0 = n) = \sum_{m=1}^{\mathrm{N}} \bar{\gamma}_{nm} e^{\bar{\delta}_m t}. \tag{9}$$

Now we can rewrite LT (8) as follows. Observe that

$$\mathbb{E}e^{-sU} = 1 - \int_0^\infty \mathbb{P}(U > u) s e^{-su} \mathrm{d}u,$$

for any random variable U on $[0,\infty)$ for which these expectations exist. Hence, we obtain that, using that $\sum_{n=1}^{\mathrm{N}} \bar{\gamma}_{mn} = 1$ for all m, and $\sum_{m=1}^{\mathrm{N}} \pi_m^Z = 1$,

$$\begin{aligned}\mathbb{E}e^{-sF} &= 1 - \sum_{m=1}^{\mathrm{N}} \pi_m^Z \left(\sum_{n=1}^{\mathrm{N}} \bar{\gamma}_{nm} \frac{s}{-\bar{\delta}_n + s} \right) = \sum_{m=1}^{\mathrm{N}} \pi_m^Z \left(\sum_{n=1}^{\mathrm{N}} \bar{\gamma}_{nm} \frac{-\bar{\delta}_n}{-\bar{\delta}_n + s} \right) \\ &= \sum_{n=1}^{\mathrm{N}} \bar{\gamma}_n \frac{-\bar{\delta}_n}{-\bar{\delta}_n + s}, \quad \text{with} \quad \bar{\gamma}_n := \sum_{m=1}^{\mathrm{N}} \pi_m^Z \bar{\gamma}_{nm}.\end{aligned}$$

We conclude that F has a phase-type distribution, with shape parameters $-\bar{\delta}_1, \ldots, -\bar{\delta}_{\mathrm{N}}$ and weights $\bar{\gamma}_1, \ldots, \bar{\gamma}_{\mathrm{N}}$ (where the latter vector sums to 1).

5.3 Mean transfer delay

Consider the mean transfer delay of a flow that finds $n-1$ flows upon arrival ($n = 1, \ldots, \mathrm{N}$), i.e.,

$$\mathbb{E}(F \mid N_0 = n) =: \eta_n;$$

at time 0 there are n flows present, including the tagged flow. Clearly, η_n is characterized through the N linear equations

$$(\lambda\, 1\{n < \mathrm{N}\} + \mu_n \mathrm{C}\, 1\{n > 1\})\, \eta_n = 1 + \lambda\, 1\{n < \mathrm{N}\} \eta_{n+1} + \frac{n-1}{n}\, \mu_n \mathrm{C}\, 1\{n > 1\} \eta_{n-1}.$$

Interestingly, these equations can be solved iteratively, as follows. The first equation gives η_2 in terms of η_1. Then consider the second equation; this gives η_3 in terms of η_1 and η_2, and hence also η_3 in terms of η_1 alone. Continuing in this way, we derive from the j th equation η_{j+1} in terms of η_1. After the $(\mathrm{N}-1)$-st equation we have η_1 up to η_{N} expressed in terms of η_1. Plug these into the N-th equation, and solve η_1, and implicitly also $\eta_2, \ldots, \eta_{\mathrm{N}}$. This procedure, however, does not lead to attractive explicit expressions.

Mean flow transfer delay $\mathbb{E}F$. First consider the limiting case of $\mathrm{N} \to \infty$. Then it turns out that the above equations *do* allow a nice explicit solution. Inspired by the results for the processor-sharing queue [25], we try the 'linear solution' $\eta_n = \vartheta_{\mathrm{I}} + \vartheta_{\mathrm{II}}\, n$. Plugging these into our recursion yields the remarkably simple expressions

$$\vartheta_{\mathrm{I}} = \frac{1}{\mu \mathrm{C}} \frac{1}{2-\varrho}, \quad \vartheta_{\mathrm{II}} = \frac{1}{\mu \mathrm{C}} \frac{3}{2-\varrho},$$

so that

$$\mathbb{E}(F \mid N_0 = n) = \frac{1}{\mu \mathrm{C}} \frac{n+3}{2-\varrho}.$$

The unconditioned mean file transfer delay (of an accepted flow) now reads (use PASTA)

$$\mathbb{E}F = \sum_{n=0}^{\infty} \pi_n \mathbb{E}(F \mid N_0 = n+1) = \sum_{n=0}^{\infty} \varrho^n (n+1)(1-\varrho)^2 \frac{1}{\mu \mathrm{C}} \frac{n+4}{2-\varrho} = \frac{2}{\mu \mathrm{C} - \lambda} = \frac{2}{\mu \mathrm{C}} \frac{1}{1-\varrho}.$$

We remark that the latter quantity can be computed also in a direct way, as follows. The mean number of flows in the system is $\sum_{n=0}^{\infty} n \varrho^n (n+1)(1-\varrho)^2 = 2\varrho/(1-\varrho)$, and with 'Little' we get the desired.

'Little' can of course also be used when $\mathrm{N} < \infty$; the advanatge is that then we do not need explicit expressions for $\mathbb{E}(F \mid N_0 = n)$ to compute $\mathbb{E}F$. It yields

$$\mathbb{E}F = \frac{\sum_{n=0}^{\mathrm{N}} n\pi_n}{\lambda(1-\pi_{\mathrm{N}})} = \frac{\sum_{n=0}^{\mathrm{N}} n\varrho^n(n+1)(1-\varrho)^2}{\lambda(1+\varrho^{\mathrm{N}+1}\mathrm{N}-\varrho^{\mathrm{N}}(\mathrm{N}+1))} = \frac{1}{\mu\mathrm{C}}\frac{\sum_{n=0}^{\mathrm{N}-1}\varrho^n(n+1)(n+2)(1-\varrho)^2}{1+\varrho^{\mathrm{N}+1}\mathrm{N}-\varrho^{\mathrm{N}}(\mathrm{N}+1)};$$

an explicit (though unattractive) expression for the numerator can be derived by differentiating the finite geometric series $\sum_{n=0}^{\mathrm{N}} \varrho^n = (1-\varrho^{\mathrm{N}+1})/(1-\varrho)$ twice.

Mean flow transfer delay $\mathbb{E}F(x)$ of a flow of size x. We can also compute the expected flow transfer delay (of an accepted flow) *given* that the flow has size x. It is given by [6]

$$\mathbb{E}F(x) = \frac{x}{\mathrm{C}}\frac{1}{1-\pi_{\mathrm{N}}}\left(\sum_{n=0}^{\mathrm{N}-1}\varrho^n\frac{c_{n+1}}{n!}\right)\Bigg/\left(\sum_{n=0}^{\mathrm{N}}\varrho^n\frac{c_n}{n!}\right),$$

where c_n is the fraction of the service rate C that is dedicated to a single flow, when there are n flows present, i.e., $1/(n+1)$. This formula, which is remarkably enough linear in x, can be simplified to

$$\mathbb{E}F(x) = \frac{x}{\mathrm{C}}\frac{\sum_{n=0}^{\mathrm{N}-1}\varrho^n(n+1)(n+2)}{\sum_{n=0}^{\mathrm{N}-1}\varrho^n(n+1)} = \frac{fx}{\mathrm{C}}, \text{ with } f := \frac{\sum_{n=0}^{\mathrm{N}-1}\varrho^n(n+1)(n+2)(1-\varrho)^2}{1+\varrho^{\mathrm{N}+1}\mathrm{N}-\varrho^{\mathrm{N}}(\mathrm{N}+1)};$$

by integrating x out, the above expression for $\mathbb{E}F$ is recovered.

6 Sojourn time distribution

In this section we analyze the sojourn time of flows in the system, which is in fact the flow transfer time, increased by the time it takes to serve the last fluid particle of the flow. Notice that these two components are *not* independent, and as a consequence the LT of the sojourn time does not follow immediately from our earlier results.

We first describe the state of the system just after an arrival of an accepted flow. Then we study the transform of the flow transfer time *jointly with* the increase of the buffer during this period. Finally we use these ingredients to find the LT of the sojourn time.

6.1 Situation at flow arrival epochs

Here the PASTA-property applies. In other words: the joint distribution of the workload and the number of flows just after an arrival of an accepted flow is given by (4). Therefore, associating time 0 with the accepted flow arrival, we write, for $n = 1, \ldots, \mathrm{N}$,

$$\chi_n(s) := \mathbb{E}(e^{-sW_0}1\{N_0 = n\}) = \frac{\pi_{n-1}}{\sum_{m=0}^{\mathrm{N}-1}\pi_m}\left(1-\frac{\lambda\mathbb{E}T}{\mathrm{C}}\right)\frac{s}{s-(\lambda/\mathrm{C})(1-\mathbb{E}e^{sT})}\left(\prod_{i=1}^{n}\mathbb{E}e^{sT_i}\right), \quad (10)$$

cf. also (7).

6.2 Joint transform of flow transfer delay and workload increment

The goal of this subsection is to compute the transform of the transfer delay F of a job that finds $n-1$ jobs upon arrival ($n = 1, \ldots, N$), jointly with the increment of the workload in this period, say ΔW, and the number of flows present at the end of the transfer (not counting the flow that just left) N_{F+}:

$$\psi_{nm}(\vec{s}) := \mathbb{E}(e^{-s_1 F - s_2 \Delta W} 1\{N_{F+} = m\} \mid N_0 = n),$$

with $\vec{s} \equiv (s_1, s_2)$. Notice that the workload cannot decrease during the flow transfer, and, as a consequence, the distribution of ΔW depends on the past only through N_0 (importantly, the value of W_0 does not play a role).
The $\psi_{nm}(\vec{s})$ satisfy, for $n = 1, \ldots, \mathrm{N}-1$, the following system of equations:

$$\psi_{nm}(\vec{s}) = \frac{1}{\lambda + \mu_n \mathrm{C} + s_1 + r_{\mathrm{A},n} s_2} \left(\lambda \psi_{n+1,m}(s) + \frac{n-1}{n} \mu_n \mathrm{C}\, \psi_{n-1,m}(s) + \frac{1}{n} \mu_n \mathrm{C}\, 1\{n-1 = m\} \right). \quad (11)$$

We also have

$$\psi_{\mathrm{N}m}(\vec{s}) = \frac{1}{\mu_{\mathrm{N}} \mathrm{C} + s_1 + r_{\mathrm{A},\mathrm{N}} s_2} \left(\frac{\mathrm{N}-1}{\mathrm{N}}\, \mu_{\mathrm{N}} \mathrm{C}\, \psi_{\mathrm{N}-1,m}(s) + \frac{1}{\mathrm{N}}\, \mu_{\mathrm{N}} \mathrm{C}\, 1\{\mathrm{N}-1 = m\} \right). \quad (12)$$

For fixed m and $\vec{s}$, these form a system of linear equations, which is (as earlier) non-singular.

6.3 Sojourn time

In our analysis, we use the following decomposition of the sojourn time S: S can be written as the sum of

- the flow transfer delay,
- and the time required to process the last particle of the flow. The buffer content at the end of the flow transfer time can be decomposed into
  - (i) the amount of traffic in the buffer at the epoch the flow arrived,
  - (ii) the net amount of fluid that entered the buffer during the flow transfer delay.

Above we have seen that the workload at flow arrival (intersected with the event that n flows are present) is characterized through the LT $\chi_n(s)$. On the other hand, the net amount of fluid entering the queue, jointly with the flow transfer delay and intersected with the event that when the tagged flow leaves there are m flows present, given that at the start of the flow transfer n flows were transmitting, is characterized through LT $\psi_{nm}(s)$. Combining these gives, with some abuse of notation, and with τ_z as defined before, the following expression for the LT of S:

$$\begin{aligned} \mathbb{E}e^{-sS} &= \mathbb{E} \exp(-sF - s\tau_{W_0 + \Delta W}) \\ &= \int_0^\infty \int_0^\infty \sum_{n=1}^{\mathrm{N}} \sum_{m=0}^{\mathrm{N}-1} \mathbb{P}(W_0 = x, N_0 = n) \\ &\qquad \mathbb{E}(e^{-sF} 1\{\Delta W = y, N_{F+} = m\} \mid N_0 = n) \mathbb{E}(e^{-s\tau_{x+y}} \mid N_0 = m) \mathrm{d}x \mathrm{d}y. \end{aligned}$$

Now using Proposition 4.1, this expression equals

$$\int_0^\infty \int_0^\infty \sum_{n=1}^{\mathrm{N}} \sum_{m=0}^{\mathrm{N}-1} \mathbb{P}(W_0 = x, N_0 = n)$$
$$\mathbb{E}(e^{-sF} 1\{\Delta W = y, N_{F+} = m\} \mid N_0 = n) \sum_{k=0}^{\mathrm{N}} \gamma_{km} e^{\delta_k(s)\,(x+y)} \mathrm{d}x\mathrm{d}y.$$

We have proven the following result.

Theorem 6.1 *For $s > 0$,*

$$\mathbb{E}e^{-sS} = \sum_{n=1}^{\mathrm{N}} \sum_{m=0}^{\mathrm{N}-1} \sum_{k=0}^{\mathrm{N}} \gamma_{km} \chi_n(-\delta_k(s)) \psi_{nm}(s, -\delta_k(s)),$$

where the γ_{mn} are as in (6), $\chi_n(\cdot)$ as in (10), and $\psi(\cdot)$ defined through (11) and (12).

Remark 6.2 The above procedure also yields the joint LT of the flow transfer time F, and the time $\tau_{W_0+\Delta W}$ it takes to serve the last fluid particle of the flow:

$$\mathbb{E}\exp(-s_1 F - s_2 \tau_{W_0+\Delta W}) = \sum_{n=1}^{\mathrm{N}} \sum_{m=0}^{\mathrm{N}-1} \sum_{k=0}^{\mathrm{N}} \gamma_{km} \chi_n(-\delta_k(s_2)) \psi_{nm}(s_1, -\delta_k(s_2)).$$

This formula (implicitly) describes the correlation between F and $\tau_{W_0+\Delta W}$. ♢

7 Tail probabilities

In this section, we study the tail behavior of $W^\star$, $D^\star$, and F, and S. More specifically, we show that these three random variables decay exponentially, and, in addition, we identify the associated decay rate. We first recall the following collection of results, which were proven in, e.g., [14], relying on the Gärtner-Ellis theorem [8, Thm. 2.3.6]. A key role is played by the asymptotic logarithmic moment generating function (mgf), or cumulant function, and its properties.

Proposition 7.1 *Consider an irreducible, finite-state (with states $0, \ldots, \mathrm{N}$), continuous-time Markov chain $(X_t)_{t\in\mathbb{R}}$ with generator Q and equilibrium distribution π. Let r be a vector of dimension N such that $m_{\mathrm{A}} := \sum_{n=0}^{N} \pi_n r_n < 0$, and $R := \mathrm{diag}\{r\}$. Define $A(s,t) := \int_s^t r_{X_u} \mathrm{d}u$.*

1. *The asymptotic logarithmic mgf of $A(0,t)$, i.e.,*

$$\Lambda_{\mathrm{A}}(\theta) := \lim_{t\to\infty} \frac{1}{t} \log \mathbb{E}\exp(\theta A(0,t)),$$

is a convex function, and equals the largest eigenvalue of $Q + \theta R$, irrespective of the value of X_0. With $q_i := \sum_{j\neq i} q_{ij}$, we have that $\Lambda_{\mathrm{A}}(\theta)$ exists for all θ smaller than

$$\min\left\{\frac{q_i}{r_i} : r_i > 0\right\}.$$

2. *For any* $x > m_{\rm A}$,

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}\left(\frac{A(0,t)}{t} > x\right) = -I_{\rm A}(x),$$

with $I_{\rm A}(x) := \sup_\theta(\theta x - \Lambda_{\rm A}(\theta))$; $I_{\rm A}(\cdot)$ *is convex,* $I_{\rm A}(m_{\rm A}) = 0$. *Similarly, for* $x < m_{\rm A}$,

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}\left(\frac{A(0,t)}{t} < x\right) = -I_{\rm A}(x).$$

3. *For the steady-state workload* $W^\star$, *which is distributed as* $\sup_{t\geq 0} A(-t,0)$, *it holds that*

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbb{P}(W^\star > x) = -\theta^\star.$$

Here $\theta^\star$ *is the smallest positive eigenvalue solving the eigensystem* $-\theta Rx = Qx$. *Alternatively,* $\theta^\star$ *is characterized as the unique positive solution of* $\Lambda_{\rm A}(\theta) = 0$. *Yet a third way of computing the decay rate is*

$$\theta^\star = \inf_{m>0} I_{\rm A}(m)/m. \tag{13}$$

Remark 7.2 An intuitive explanation of the relation (13) is the following. $I_{\rm A}(m)$ can be interpreted as the cost incurred for the process $A(0,t)$ to generate traffic at rate m; evidently there is no cost involved when sending at the average rate $m_{\rm A}$ (reflected by $I_{\rm A}(m_{\rm A}) = 0$), but there is a positive cost for sending at a higher (or lower) rate. Suppose the process generates traffic at rate $m > 0$. Then it takes about x/m to reach buffer level x, and the cost made is $I_{\rm A}(m)/m$. There is an evident trade-off between the numerator and the denominator: when choosing m small but positive, the cost per unit of time are relatively low, but it takes long to reach x, whereas the opposite applies when choosing m large. We conclude that the 'most likely speed' $m^\star$ is the minimizing argument in

$$x \left(\inf_{m>0} I_{\rm A}(m)/m\right), \tag{14}$$

where (14) roughly equals $-\log \mathbb{P}(W^\star > x)$, for x large. ♢

7.1 Decay rate of steady-state workload

The decay rate $\theta^\star$ of $W^\star$ follows immediately from Proposition 7.1, with continuous-time Markov chain N_t governed by Q as defined by (1), and $R := R_{\rm A}$: $\theta^\star$ is the smallest positive eigenvalue of the system $-\theta R_{\rm A} x = Qx$. In fact, one can prove the stronger statement that $\mathbb{P}(W^\star > x)\exp(\theta^\star x)$ converges to some constant $\kappa > 0$ for $x \to \infty$, and even, for $n = 0, \ldots, {\rm N}$,

$$\lim_{x\to\infty} \mathbb{P}(W^\star > x, N^\star = n)\exp(\theta^\star x) = \kappa_n, \tag{15}$$

for $\kappa_n > 0$, see for instance [15].

Another way to characterize $\theta^\star$ is as follows [18]. Let U_{mn} be the value of $A(0, V_n)$ conditional on $N_0 = m$, where V_n is the epoch of the first entrance of N_t for $t > 0$ to state n. Then $\theta^\star$ can be alternatively characterized as the unique positive solution of $\mathbb{E}e^{\theta U_{mm}} = 1$; remarkably, in [18] it is shown this solution is identical for any $m = 0, \ldots, {\rm N}$. Now consider $m = 0$. Then U_{00} is distributed

as $E+T$, with E exponentially distributed with mean λ^{-1}, T as defined in Section 3, and E and T independent. The equation $\mathbb{E}e^{\theta U_{00}} = 1$ then reduces to

$$\frac{\lambda}{\lambda+\theta \mathrm{C}}\,\mathbb{E}e^{\theta T},$$

or, equivalently, $\theta + (\lambda/\mathrm{C})(1-\mathbb{E}e^{\theta T}) = 0$. We conclude that the decay rate $\theta^\star$ coincides with (minus) the pole of $\mathbb{E}e^{-sW^\star}$, cf. Theorem 3.1.

7.2 Decay rate of queueing delay

We next characterize the exponential decay rate of the queueing delay. We here focus on the virtual queueing delay, but it can be verified easily that the same decay rate applies to the 'packet average'. We first define the cumulant function of the output process, as follows. For $\theta \in \mathbb{R}$,

$$\Lambda_{\mathrm{O}}(\theta) := \lim_{t\to\infty} \frac{1}{t} \log \mathbb{E} \exp(\theta O(0,t)).$$

This function equals the largest eigenvalue of $Q+\theta R_{\mathrm{O}}$, due to Proposition 7.1. We also define $I_{\mathrm{O}}(x) := \sup_\theta(\theta x - \Lambda_{\mathrm{O}}(\theta))$, and $m_{\mathrm{O}} := \sum_{n=0}^{\mathrm{N}} \pi_n r_{\mathrm{O},n}$. We first observe that, again due to Proposition 7.1, irrespective of the number of flows present at time 0,

$$\lim_{u\to\infty} \frac{1}{u} \log \mathbb{P}(O(o,u) \in [i\epsilon u, (i+1)\epsilon u)) = \zeta_i(\epsilon) := \begin{cases} -I_{\mathrm{O}}(i\epsilon) & \text{if } m_{\mathrm{O}} < i\epsilon; \\ -I_{\mathrm{O}}((i+1)\epsilon) & \text{if } m_{\mathrm{O}} > (i+1)\epsilon; \\ -I_{\mathrm{O}}(m_{\mathrm{O}}) = 0 & \text{if } i\epsilon < m_{\mathrm{O}} < (i+1)\epsilon, \end{cases}$$

explicitly using the convexity of $I_{\mathrm{O}}(\cdot)$. The following result is [8, Lemma 1.2.15].

Lemma 7.3 *For any finite index set $\mathcal{S}$, and $\omega_i(u) \geq 0$,*

$$\limsup_{u\to\infty} \frac{1}{u} \log \sum_{i\in\mathcal{S}} \omega_i(u) = \max_{i\in\mathcal{S}} \limsup_{u\to\infty} \frac{1}{u} \log \omega_i(u).$$

Now we have collected the prerequisites for the proof of the following result.

Theorem 7.4 *The decay rate of the virtual queueing delay equals*

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}(D^\star > t) = -\inf_m (I_{\mathrm{O}}(m) + \theta^\star m) = \Lambda_{\mathrm{O}}(-\theta^\star). \tag{16}$$

Proof. We first prove the first equality in (16). We start by establishing the *upper bound*. Conditioning on the value of $O(0,t)$,

$$\begin{aligned} \mathbb{P}(D^\star > t) &= \sum_{n=0}^{\mathrm{N}} \mathbb{P}(W^\star > O(0,t), N^\star = n) \\ &\leq \sum_{n=0}^{\mathrm{N}} \sum_{i=0}^{\infty} \mathbb{P}(W^\star \geq (i+1)\epsilon t, N^\star = n) \mathbb{P}(O(0,t) \in [i\epsilon t, (i+1)\epsilon t) \mid N^\star = n). \end{aligned} \tag{17}$$

It is clear that for some values of i there is no contribution, due to the fact that the rates in the vector r_{O} are between $\mathrm{C}/(\mathrm{N}+1)$ and C. Therefore, we can restrict ourselves to

$$i \in \mathcal{I}_\epsilon, \text{ where } \mathcal{I}_\epsilon := \left\{ i \in \mathbb{N} : \frac{\mathrm{C}}{\epsilon(\mathrm{N}+1))} - 1 \leq i \leq \frac{\mathrm{C}}{\epsilon} \right\}.$$

The decay rate of $\mathbb{P}(W^\star \geq (i+1)\epsilon t, N^\star = n)$ is $-\theta^\star(i+1)\epsilon$, independently of n, see (15). The decay rate of $\mathbb{P}(O(0,t) \in [i\epsilon t, (i+1)\epsilon t) \mid N^\star = n)$ is $\zeta_i(\epsilon)$, as given above, also independently of n. In view of Lemma 7.3, the decay rate of (17) is majorized by

$$\max_{i \in \mathcal{I}_\epsilon} \left(-\theta^\star(i+1)\epsilon + \zeta_i(\epsilon)\right).$$

Now let $\epsilon \downarrow 0$; using the continuity of $I_{\rm O}(\cdot)$, we arrive at

$$\sup_{m \in [{\rm C}/({\rm N}+1), {\rm C}]} (-\theta^\star m - I_{\rm O}(m)). \tag{18}$$

Now we present the *lower bound*, which is established in a similar fashion. Evidently, for any i,

$$\mathbb{P}(D^\star > t) \geq \mathbb{P}(W^\star \geq (i+1)\epsilon t, N^\star = n)\mathbb{P}(O(0,t) \in [i\epsilon t, (i+1)\epsilon t) \mid N^\star = n).$$

The decay rate of the right-hand side of the previous display is $-\theta^\star i\epsilon + \zeta_i(\epsilon)$; as this holds for any i, the supremum over i is still a lower bound. Taking $\epsilon \downarrow 0$, we obtain that the upper bound (18) is also lower bound.
We have now proven the first equality in (16); the second immediately follows from the duality relation $\Lambda_{\rm O}(\theta) = \sup_x(x\theta - I_{\rm O}(x))$, see for instance [10, Thm. VI.4.1]. □

Remark 7.5 There is an appealing alternative way to characterize this decay rate, cf. Remark 7.2. Consider the event that a fluid particle arriving at time 0 has (approximately) virtual delay t. Suppose that, after time 0, the queue drains at rate m, which costs $I_{\rm O}(m)$ per unit of time. In order to achieve delay t, the workload at time 0 should have been mt. Supposing that the queue built up at rate $m' > 0$ before time 0, with cost $I_{\rm A}(m')$ per unit of time, this took $(m/m')t$ time. In other words, we are to minimize

$$\inf_{m,m'>0} \left(I_{\rm A}(m') \, \frac{mt}{m'} + I_{\rm O}(m)\, t \right) = t \left(\inf_{m>0} (\theta^\star m + I_{\rm O}(m)) \right),$$

where the equality is due to (13). ♢

7.3 Decay rate of flow transfer delay

The decay rate of the flow transfer delay follows immediately from the phase-type distribution identified in Section 5. Directly from Equation (9), we see that

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbb{P}(F > x) = \bar{\delta} := \max_{n=1,\ldots,{\rm N}} \bar{\delta}_n,$$

i.e., the dominant eigenvalue of $Q^\star$.

7.4 Decay rate of sojourn time

We now turn our attention to the tail behavior of the sojourn time. This is a complicated issue, as long sojourn times are due to a combination of (i) a high workload when the flow enters, (ii) a large flow, (iii) a large amount work brought along by flows arriving during the flow transfer time of the tagged flow, (iv) a low service speed available to the queue after the flow transmission time (i.e., when the

complete flow has been put into the queue). We below sketch how the exponential decay rate can be computed; the arguments can be made precise as in Section 6.2.
Using the representation $S = F + \tau_{W_0+\Delta W}$, we condition on the values of $W_0, \Delta W$, and F. With some abuse of notation,

$$\begin{aligned}
&\mathbb{P}(F + \tau_{W_0+\Delta W} > t) \\
&\quad\approx \sum_{n=1}^{\mathrm{N}} \int_0^\infty \mathbb{P}(W_0 = zt \mid N_0 = n)\mathbb{P}(F + \tau_{z+\Delta W} > t, N_0 = n)\mathrm{d}z \\
&\quad\approx \sum_{n=1}^{\mathrm{N}} \sum_{m=0}^{\mathrm{N}-1} \int_0^\infty \int_0^\infty \int_0^\infty \mathbb{P}(W_0 = zt \mid N_0 = n)\mathbb{P}(F = ft, \Delta W = wt, N_0 = n, N_{F+} = m) \\
&\qquad\qquad \mathbb{P}(\tau_{zt+wt} > t - ft, N_0 = m)\mathrm{d}f\,\mathrm{d}w\,\mathrm{d}z,
\end{aligned}$$

with $f \in (0, 1)$. Now use the folk theorem that says that the decay rate of an integral equals the decay rate of the maximum of the integrand. We saw earlier that the exponential decay rate (x large) of $\mathbb{P}(W_0 = zt \mid N_0 = n)$ does not depend on n; likewise, the decay rates of the other two probabilities, $\mathbb{P}(F = ft, \Delta W = wt, N_0 = n, N_{F+} = m)$ and $\mathbb{P}(\tau_{zt+wt} > t - ft, N_0 = m)$, do not depend on m and n. They can be computed as follows:

- As before, for $z > 0$,

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}(W_0 = zt) = -z\theta^\star =: J_1(z).$$

- Similar to the decay rate of F being equal to $\max_{n=1,\ldots,\mathrm{N}} \bar{\delta}_n$, i.e., the infimum over all $s < 0$ for which $\mathbb{E}e^{-sF} < \infty$, we have that

$$\lim_{x\to\infty} \frac{1}{t} \log \mathbb{P}(F = ft, \Delta W = wt) = \inf\left\{s_1 f + s_2 w : \mathbb{E}e^{-s_1 F - s_2 \Delta W} < \infty\right\} =: J_2(f, w).$$

  Notice that this decay rate is larger than $-\infty$, as can be seen as follows. Suppose that T is the flow size of the tagged flow. Then, as each flow receives a rate of maximally $\mathrm{C}/2$, we have that $F \geq 2T/\mathrm{C}$. Hence, for $s_1 > -\mu\mathrm{C}/2$,

$$\mathbb{E}e^{-s_1 F - s_2 \Delta W} \geq \frac{\mu}{\mu + 2s_1/\mathrm{C}},$$

  and $\mathbb{E}e^{-s_1 F - s_2 \Delta W} = \infty$ for $s_1 \leq -\mu\mathrm{C}/2$.

- Also, as earlier,

$$\begin{aligned}
&\lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}(\tau_{zt+wt} > t - ft) \\
&\quad= \lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}(O(0, (1-f)t) < zt + wt) = -(1-f)I_{\mathrm{O}}\left(\frac{z+w}{1-f}\right) =: J_3(z, f, w),
\end{aligned}$$

  with $(z + w)/(1 - f) < m_{\mathrm{O}}$.

Collecting terms, we find that

$$\lim_{x\to\infty} \frac{1}{t} \log \mathbb{P}(S > t) = \sup_{z,f,w} \left(J_1(z) + J_2(f, w) + J_3(z, f, w)\right),$$

where the maximization is over all $z, w > 0$ and $f \in (0, 1)$, such that $(z + w)/(1 - f) < m_{\mathrm{O}}$.

8 Discussion and outlook

An important feature of our model is that there is just one state in which the queue drains. It has appeared that this is a key property in our analysis. Importantly, it entails that the dynamics of the number of flows in the system are not affected by the workload process. This enabled the compuation of the LT of the workload, as it brought us into the framework of M/G/1-type of models. Also, it implied that the workload cannot decrease during flow transfer; as a consequence ΔW (as used in Section 7) depends on N_0, and not on W_0. One could, however, think of other allocation policies (i.e., policies to distribute the capacity between the flows and the queue, as alternatives to the '$\mathrm{C}/(n+1)$-policy' used in this paper), which still have the desirable property that there is just one 'buffer drain state'. An example could be

$$r_{\mathrm{I},n} := \mathrm{C} \max\left\{\frac{n}{n+m}, \frac{1}{2}\right\},$$

for $m \in \mathbb{N}$, and $r_{\mathrm{O},n} := \mathrm{C} - r_{\mathrm{I},n}$. When m is chosen 0, each source gets a fraction $1/n$ of the capacity, and traffic is served by the queue (at rate C) only when no flows are present; compared to the model of the present paper, i.e., $m = 1$, the flow transfer delays will be smaller, while the queueing delay will be longer. In the other extreme, $m \to \infty$, each source gets $\mathrm{C}/(2n)$ and the queue $\mathrm{C}/2$, so that the sources suffer from long flow transfer delays, but the queue never fills. The choice $m = 1$ is in this sense a compromise.

Another interesting extension would relate to the situation *without* admission control. The complication is that the state-space of $(N_t)_{t\in\mathbb{R}}$ becomes (countably) infinite. The results of Section 3 carry over to this situation; still the LT of T can be computed by methods similar to those in [13, 21]. The results of the other sections will change; in any case all matrix-exponentials should be handled with care.

One could also study the situation of multiple bottleneck links that are sharing capacity. The complicating factor is that then the dynamics of the flows feeding into one queue will be affected by the workload process in other queues. As a result, this model has the flavor of coupled-processors systems as studied in, e.g., [12], which are notoriously hard to analyze. Other challenging extensions include: (i) non-exponential flow-size distribution (for instance regularly varying), (ii) heterogeneous flow types, (iii) allocation policies that do not depend only on the number of flows present, but also on the buffer content, cf. [24].

Acknowledgment

The authors are grateful to Hans van den Berg (TNO Information and Communication Technology and University of Twente) for bringing this model under their attention. Werner Scheinhardt (University of Twente) has provided useful remarks.

References

[1] D. ANICK, D. MITRA, and M. SONDHI. Stochastic theory of a data-handling system with multiple sources. *The Bell System Technical Journal*, 61: 1871 – 1894, 1982.

[2] S. ASMUSSEN. Ruin probabilities. Advanced Series on Statistical Science & Applied Probability, Vol. 2. World Scientific, London, 2000.

[3] T. BHEEMA REDDY, I. KARTHIGEYAN, B. S. MANOJ, and C. SIVA RAM MURTHY. Quality-of-Service provisioning in ad hoc wireless networks: a survey of issues and solutions. *Journal of Ad Hoc Networks.* To appear, 2005.

[4] S. BORST, O. BOXMA, and N. HEGDE. Sojourn times in finite-capacity processor-sharing queues, *Proceedings 1st EURO-NGI Conference,* Rome, Italy, 2005.

[5] A. BRANDT and M. BRANDT. On the distribution of the number of packets in the fluid flow approximation of packet arrival streams. *Queueing Systems*, 17: 275 – 315, 1994.

[6] J. COHEN. Superimposed renewal processes and storage with gradual input. *Stochastic Processes and Applications*, 2: 31– 58, 1974.

[7] J. COHEN. The multiple phase service network with generalized processor sharing. *Acta Informatica*, 12: 245 – 284.

[8] A. DEMBO and O. ZEITOUNI. Large deviations techniques and applications, 2nd edition. Springer Verlag, New York, 1998.

[9] E. VAN DOORN, A. JAGERS, and J. DE WIT. A fluid reservoir regulated by a birth death-process. *Stochastic Models,* 4: 457 – 472.

[10] R. ELLIS. Entropy, large deviations, and statistical mechanics. Springer Verlag, New York, 1985.

[11] A. ELWALID and D. MITRA. Effective bandwidth of general Markovian traffic sources and admission control of high-speed networks. *IEEE/ACM Transactions on Networking,* 1: 329 – 343, 1993.

[12] G. FAYOLLE and R. IASNOGORODSKI. Two coupled processors: the reduction to a Riemann-Hilbert problem. *Zur Wahrscheinlichkeitstheorie verwandte Gebiete,* 47: 325 – 351, 1979.

[13] F. GUILLEMIN and A. SIMONIAN. Transient characteristics of an M/M/∞ system. *Advances in Applied Probability,* 27: 862 – 888, 1995.

[14] G. KESIDIS, J. WALRAND, and C.S. CHANG. Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Transactions on Networking*, 1: 424 – 428, 1993.

[15] L. KOSTEN. Stochastic theory of data-handling systems with groups of multiple sources. In H. Rudin and W. Bux (eds.), *Performance of Computer Communication Systems*, 321 – 331, Elsevier Amsterdam, 1984.

[16] J.G. KIM and M. KRUNZ. Fluid analysis of delay and packet discard performance for QoS support in wireless networks. *IEEE Journal on Selected Areas in Communications*, 19: 384 – 395, 2001.

[17] M. MANDJES, D. MITRA, and W. SCHEINHARDT. Models of network access using feedback fluid queues. *Queueing Systems*, 44, 365 – 398, 2003.

[18] M. MANDJES and A. RIDDER. Finding the conjugate of Markov fluid processes. *Probability in the Engineering and Informational Sciences*, 9: 297 – 315, 1995.

[19] M. MARCUS and H. MINC. A survey of matrix theory and matrix inequalities. Allyn and Bacon, Rockleigh NJ, 1964.

[20] D. MITRA. Stochastic theory of a fluid model of producers and consumers coupled by a buffer. *Advances in Applied Probability,* 20: 646 – 676, 1988.

[21] J. PREATER. M/M/∞ transience revisited. *Journal of Applied Probability*, 34: 1061 – 1067, 1997.

[22] E. REICH. On the integro-differential equation of Takács. I. *Annals of Mathematical Statistics*, 29: 563 – 570, 1958.

[23] F. ROIJERS, M. MANDJES, and H. VAN DEN BERG. Analysis of congestion periods in an M/M/∞ queue. Submitted, 2005.

[24] W. SCHEINHARDT, N. VAN FOREEST, and M. MANDJES. Continuous feedback fluid queues. *Operations Research Letters,* 33, 551 – 559, 2005.

[25] B. SENGUPTA and D. JAGERMAN. A conditional response time of the M/M/1 processor-sharing queue. *AT&T Technical Journal*, 2, 409–421, 1985.

[26] J. VIRTAMO and I. NORROS. Fluid queue driven by an M/M/1 queue. *Queueing Systems*, 16: 373 – 386, 1994.