BUSY PERIODS IN FLUID QUEUES WITH MULTIPLE EMPTYING INPUT STATES

A. J. FIELD,[*][**] AND

P. G. HARRISON,[*] *Imperial College London*

Abstract

A semi-numerical method is derived to compute the Laplace transform of the equilibrium busy period probability density function in a fluid queue with constant output rate when the buffer is nonempty. The input process is controlled by a continuous-time semi-Markov chain (CTSMC) with n states such that in each state the input rate is constant. The holding time in states with net positive output rate—so-called *emptying states*—is assumed to be an exponentially distributed random variable, whereas in states with net positive input rate—so-called *filling states*—it may have an arbitrary probability distribution. The result is demonstrated by applying it to various systems, including fluid queues with two on–off input sources. The latter exercise in part shows consistency with prior results but also solves the problem in the case where there are two emptying states. Numerical results are presented for selected examples which expose discontinuities in the busy period distribution when the number of emptying states changes, e.g. as a result of increasing the fluid arrival rate in one or more states of the controlling CTSMC.

Keywords: Fluid queue; busy period; CTSMC; discontinuity

2010 Mathematics Subject Classification: Primary 68U99
Secondary 90C99

1. Introduction

Fluid models have long been used to approximate traffic flows in systems comprising very large numbers of discrete entities flowing around a network. The traditional 'building block' node in such models is a fluid queue fed by one or more sources, each of which is governed by some modulated arrival process. The states, also termed *phases*, of the modulating process have an associated fluid arrival rate and state holding time probability distribution that can be used, for example, to distinguish different classes of traffic, including 'off', or 'silent', periods where the traffic arrival rate is 0. Fluid queueing models have seen widespread application, particularly in the analysis of packet flows in communication networks and in the analysis of congestion control regimes; see, for example, [9] and [13].

Fluid queues have been analysed by various methods, and for a variety of models of their inputs, to yield equilibrium performance measures such as the distribution and moments of the fluid level [6], [10]–[12], [14] and of the busy period. The busy period has been analysed by Boxma and Dumas [4] for the case where the queue is fed by an arbitrary number of on–off arrival streams. The off-periods are assumed to be exponentially distributed, but the on-periods may have an arbitrary distribution. The fluid arrival rate of each input stream that is in the

on-state is constant and assumed to be larger than the service rate of the queue. Thus, the queue can become empty only when all arrival streams are in the off-state. The analysis of [3] relaxes this restriction, but requires that the on-periods have exponentially distributed state holding times. Their analysis exploits this property to allow the continuous fluid flow process to be modelled as a discrete 'counting' process, by uniformization.

As well as being useful in their own right, modulated fluid queues are also important as a *building block queue* in a compositional approximation for networks of fluid queues, as described in [5]. This leads to exact results for the equilibrium mean fluid level in each buffer in a tandem network of queues at equilibrium, using intermediate Markov-modulated arrival processes (MMAPs) with three states; exact results for the variance can be achieved by using MMAPs with four states [7]. However, the current approach requires that the fluid arrival rate in each input phase is either 0 or greater than the constant output rate of the fluid server. This is because the MMAP parameterisation is based on matching moments against those of the busy period of the upstream queue, using the results of [4]. In more general fluid systems, however, there will be input states in which fluid arrives at a positive rate less than the node's maximum output rate, and on-periods will be generally distributed (e.g. nonexponential) random variables. We remark that a similar compositional approach has been developed in [2] for linear tandem networks fed by an M/M/1 source queue. In that work, the fluid flow between each pair of queues in the chain is represented by an on–off process whose on-period is distributed as the busy period of an intermediate M/M/1 queue.

In this paper, we consider a queue whose fluid arrival rate is piecewise constant in time, its value being determined by the current state of an independent, continuous-time semi-Markov chain (CTSMC)—the modulating chain. The holding time of a CTSMC state in which the net *output* rate is positive (i.e. the server's output rate exceeds the input rate in this state) is assumed to be an exponentially distributed random variable; we refer to these states as *emptying states*. A state with net positive *input* rate (where the server's output rate is less than the input rate in this state) may have an arbitrary holding time probability distribution; we refer to these states as *filling states*.

A *busy period* is defined to be the contiguous length of time that the buffer's fluid level is greater than 0, i.e. the time elapsed between the end of one idle period, when the buffer is continually empty, to the beginning of the next idle period. The system is assumed to be at equilibrium so that it has almost surely finite busy periods. Idle periods can occur only in input states with net positive output rate. Importantly, this includes situations where the buffer is empty and the fluid arrival rate is positive but less than the specified constant output rate. In such a case, the actual output rate is nonzero, equal to the arrival rate.

Note that if the fluid arrival rate is constrained to be either 0 or greater than the output rate, as in previous treatments, then the busy period is synonymous with a period during which the server is continually emitting fluid at a positive rate. In the present contribution, however, it is possible for the server to be emitting fluid even though the buffer is empty. This is a significantly more general queue than one with a simple on–off input process, the solution to which is well known even when on-periods have an arbitrary probability distribution [4], [5]. It is also more general than a fluid queue with multiple on–off sources, each of which must have input rate greater than the service rate in its on-state, as in the model of [4] referred to above. Moreover, it extends the model of [3] in that the filling states can have an arbitrarily distributed state holding time, although the assumed input models are quite different.

One of the main contributions of the paper is an expression for the Laplace transform of the busy period's probability density function in the above queue (Section 3.3). This leads

to a numerical solution method that can be applied to systems with an arbitrary number of emptying and filling states. In general, this requires the numerical solution of a complex system of fixpoint equations for each point in the Laplace domain and, in turn, the solution of a polynomial equation of degree n_e, the number of emptying states. Although complex, the method is shown to be tractable for systems of modest size using MATHEMATICA® (see http://www.wolfram.com/products/mathematica/index.html) (Section 5.1). We demonstrate the application of the result to some specific scenarios that are simple enough to be treated symbolically in Section 4. Furthermore, in cases for which there is a known analytical solution, we verify that our results are consistent with the literature.

The mth moment of the busy period can, in principle, be computed numerically via the mth derivative of the above Laplace transform at the origin. The direct numerical approach turns out to be unstable for derivatives higher than the first. The alternative, symbolic approach requires the calculation of the nth derivative of each root of the corresponding polynomial equation, for $1 \le n \le m$. Again, the numerical computation of these root derivatives is unstable, except for the mean value (Section 5). A key result, presented in Proposition 3 of Section 5.1.1, defines the mth derivative of a root of an arbitrary polynomial in terms of the derivatives of the polynomial's *coefficients*. This is a direct recursive formulation that avoids the need for numerical differentiation, requiring only the numerical value of the root of the given polynomial equation at the point concerned (the origin for the present problem), which is usually quickly computed by standard mathematical software. This approach to calculating moments is not only efficient, but is also numerically stable for all the test cases considered. The robust computation of moments is of particular significance to the compositional approach outlined above, which is based on moment matching between the busy period of a fluid queue and the state holding times in an MMAP.

Some numerical results for both the moments and density function of the busy period are presented in Section 5. These results show that the busy period distribution exhibits discontinuities at the instants where the number of emptying states changes, e.g. as a result of a continuous scaling of one or more load parameters. This behaviour is counterintuitive, but is shown to be a natural consequence of the instantaneous transition of one or more states in the underlying CTMC from being emptying states to filling states. We further show that the discontinuities can lead to either a positive or negative jump in the moments of the busy period distribution, depending on the parameterisation of the model.

2. The model

We first formally describe the model in terms of its semi-Markov-modulated input process and its net output rate in each state of that process. The solution for the busy period is then derived as the Laplace–Stieltjes transform of its probability distribution function, using conditional expectation arguments.

2.1. Model definition

We consider a single fluid queue, comprising a server that outputs fluid at a constant rate when it has a positive quantity of fluid stored in its reservoir, or buffer, and a time-homogeneous semi-Markov-modulated input (or arrival) stream. We use the following notation:

- there are n states, or phases, in the state space $\mathcal{S}$ of the CTSMC, which has probability transition matrix $\boldsymbol{P} = (p_{ij} \mid 1 \le i, \ j \le n)$, defined at state transition instants;

- the state holding time in state i, given that the next state is j, is H_{ij};

- the fluid arrival rate in phase i is the constant λ_i volume units of fluid per unit time;

- the rate at which the server outputs fluid when its buffer is nonempty is μ volume units of fluid per unit time;

- the diagonal (net input) *rate matrix* $\boldsymbol{R} = \mathrm{diag}(r_1, \ldots, r_n)$, where $r_i = \lambda_i - \mu$ for $1 \le i \le n$, and the rate vector $\vec{r} = (r_i, \ldots, r_n)$);

- the subset of *filling states* with positive net input rate is $\mathcal{F} = \{i \in \mathcal{S} \mid r_i > 0\} \subset \mathcal{S}$;

- the subset of *emptying states* with negative net input rate is $\mathcal{E} = \{i \in \mathcal{S} \mid r_i < 0\} = \mathcal{S} \setminus \mathcal{F}$;

- $n_{\mathrm{f}} = |\mathcal{F}|$ and $n_{\mathrm{e}} = |\mathcal{E}|$ so that $n_{\mathrm{f}} + n_{\mathrm{e}} = n$;

- each state i is assumed to be associated with a nonzero net arrival rate of fluid ($r_i \ne 0$);

- the vector $\boldsymbol{e} = (1, 1, \ldots, 1)$.

For a generic continuous random variable X, we denote its probability distribution function by $X(t) = \mathrm{P}(X \le t)$ and the Laplace–Stieltjes transform (LST) of this distribution by $X^*(\theta) = \mathrm{E}[e^{-\theta X}]$, where $\mathrm{E}[\cdot]$ and $\mathrm{E}[\cdot \mid \cdot]$ denote the expectation and conditional expectation operators, respectively. We denote the density function by $x(t) = X'(t)$, the derivative of the distribution function, with Laplace transform $X^*(\theta)$.

The modulating CTSMC is defined with more rigour in terms of the Markov renewal process $\{(X_n, T_n) \colon n \ge 0\}$, where T_n is the time of the nth transition ($T_0 = 0$) and X_n is the state at (just after) the nth transition. The required CTSMC is therefore $\{Z(t) = X_{N(t)}, \ t \ge 0\}$, where $N(t) = \max\{n \colon T_n < t\}$ and the (time-homogeneous) kernel K of (X_n, T_n) is (for all n)

$$K(i, j, t) = \mathrm{P}(X_{n+1} = j, \ T_{n+1} - T_n \le t \mid X_n = i) = p_{ij} H_{ij}(t),$$

where $p_{ij} = \mathrm{P}(X_{n+1} = j \mid X_n = i)$.

We further define the following random variables.

T_{ij} The time elapsed between entry to state $i \in \mathcal{F}$ and the first subsequent transit from some state in $\mathcal{F}$ to state $j \in \mathcal{E}$—representing the sum of the lengths of the on-periods entered and the event of exit out of $\mathcal{F}$ to state $j \in \mathcal{E}$.

Y_{ij} The net volume of fluid collected in the buffer during the period T_{ij}, i.e. during successive on-periods that start on entry to state $i \in \mathcal{F}$ and end on transit from some state in $\mathcal{F}$ to state $j \in \mathcal{E}$.

V_{ij} The busy period starting in filling state $i \in \mathcal{F}$ (immediately after a transition from an emptying state) and ending in emptying state $j \in \mathcal{E}$.

$R_{ij}(x)$ The time to clear the buffer when it has x units of fluid in it, the current state is $i \in \mathcal{E}$, and the buffer becomes empty in state $j \in \mathcal{E}$ (just after that time has elapsed).

E_i Exponentially distributed holding time for state $i \in \mathcal{E}$, with mean value $1/\gamma_i$. Thus, for $i \in \mathcal{E}$, H_{ij} is equal in distribution to E_i for all $j \in \mathcal{S}$.

We define $T_{ij} = Y_{ij} = \infty$ if the first subsequent transition after leaving state $i \in \mathcal{F}$ to a state in $\mathcal{E}$ is to $l \ne j$.

Finally, let $L_{ij}(\beta, \alpha)$ be the Laplace transform of $R_{ij}^*(x, \alpha)$—itself the LST of $R_{ij}(x)(t)$—with respect to x, i.e.

$$L_{ij}(\beta, \alpha) = \int_0^\infty e^{-\beta x} R_{ij}^*(x, \alpha) \, \mathrm{d}x.$$

2.1.1. *Server vacations.* We remark that the above formulation in terms of the *net* input rate (r_i in state i) allows for the service rate to also depend on the phase of the underlying CTSMC. It is thus possible, for example, to model server vacations, provided the vacation policy is independent of the state of the buffer.

2.2. The busy period

The analysis of a busy period is based on the observation that it comprises an initial filling phase, followed by a sequence of emptying phases interspersed with nested, or 'pseudo', busy periods. A pseudo-busy period ends when the fluid level (which is rising at the commencement of the pseudo-busy period) falls back to its original level.

For a simple two-state on–off process ($n = 2$, $\lambda_1 = 0$ (off), $\lambda_2 = \lambda$ (on)), the Laplace transform of the busy period ($V^*(\theta)$) is easily obtained, e.g. by conditional expectation arguments, in terms of the LST of the on-period's probability distribution, $W^*(\theta)$:

$$V^*(\theta) = W^*(\theta\rho + b(\rho - 1)(1 - V^*(\theta))) \quad \text{(see, e.g. [5]).}$$

For the more general case, where the input process is governed by the semi-Markov process defined earlier, the situation is complicated by the fact that there can now be phase transitions during both a *filling period*, where the input process follows a succession of filling states, and an *emptying period*, where the input process follows a succession of emptying states.

Following the line of reasoning typically used for the simple on–off case above, we observe that the time required to clear a units of fluid from an instant of entering a state $f \in \mathcal{F}$ has the same probability distribution as the sum of a busy period V_{fe}, that starts in state f and ends in state $e \in \mathcal{E}$, and the time to remove a units of fluid starting in state e (see Figure 1). The argument is that the original a units are served *last*, the order of service being immaterial as far as busy periods are concerned.

Using the nomenclature introduced earlier, a busy period begins at a transition instant from some state $l \in \mathcal{E}$ to a state $i \in \mathcal{F}$ when the buffer is empty. When the system next returns to a state $k \in \mathcal{E}$, after a time T_{ik}, the buffer will hold Y_{ik} units of fluid, which will eventually be cleared in state $j \in \mathcal{E}$ after a time $R_{kj}(Y_{ik})$. The length of the busy period that started in state
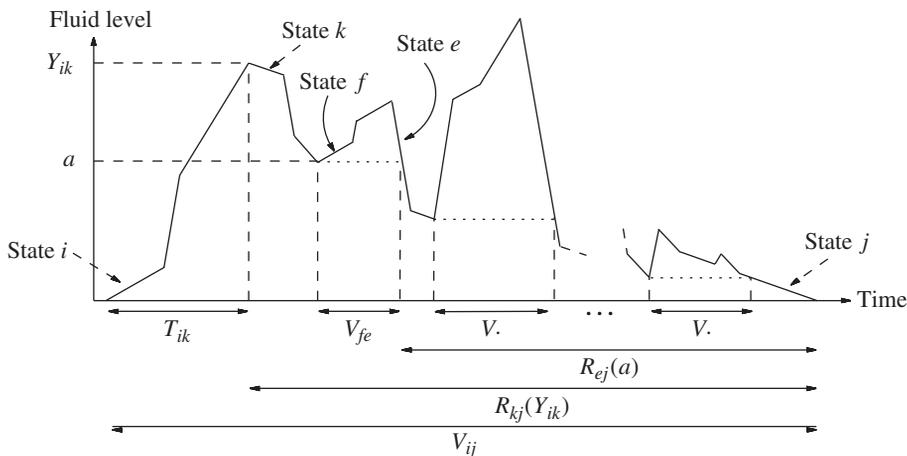

FIGURE 1: The busy period for an n-phase arrival process.

$i \in \mathcal{F}$ and ends in state $j \in \mathcal{E}$ is then

$$V_{ij} = T_{ik} + R_{kj}(Y_{ik}).$$

This is illustrated in Figure 1. Note that the problem differs from that considered in [4] in that it generalises the notion of a single off-state, where the fluid arrival rate is 0, to any number of emptying states in which the fluid arrival rate is less than the output rate of the server. In particular, the duration of an emptying period is no longer, in general, exponentially distributed; cf. the simple on–off process outlined above and other processes with a single emptying state that has a negative exponential holding time.

3. Model solution

The primary objective of this paper is to determine the Laplace transform, $V_{ij}^*(\alpha)$, of the probability density function of the busy period in the above queue, and (in Appendix A) its moments. The busy period has two components: the initial filling period, which accumulates x (say) units of fluid, and the period required to clear those x units. Therefore, $V_{ij}^*(\alpha)$ is defined in terms of the (joint) Laplace transform of the joint probability density of the dependent random variables T_{ik} and Y_{ik}, for states $i \in \mathcal{F}$ and $k \in \mathcal{E}$, and the Laplace transform of the clearing time's probability density, i.e. $R_{kj}^*(x, \alpha)$ for states $k, j \in \mathcal{E}$. These are derived in Sections 3.1 and 3.2, respectively. The derivation of $V_{ij}^*(\alpha)$, in terms of these two transforms, appears in Section 3.3.

3.1. Time to clear a buffer

We first examine the time to clear a given amount of fluid x in the buffer when in one of the emptying states. Since the holding time in such a state is exponentially distributed, it makes no difference how long the modulating chain has already spent in it up to the present time instant. We first prove the following result.

Lemma 1. *For $i, j \in \mathcal{E}$,*

$$\left[\beta - \frac{\alpha + \gamma_i}{r_i} \right] L_{ij}(\beta, \alpha) = \delta_{ij} - \sum_{k \in \mathcal{E}} \frac{\gamma_i p_{ik} L_{kj}(\beta, \alpha)}{r_i} - \sum_{k \in \mathcal{F}} \sum_{l \in \mathcal{E}} \frac{\gamma_i p_{ik} V_{kl}^*(\alpha) L_{lj}(\beta, \alpha)}{r_i}.$$

Proof. Following the line of reasoning described in Section 2.2, commonly applied to busy period analysis, let the current (emptying) state be $i \in \mathcal{E}$. We then have

$$R_{ij}(x) = \begin{cases} \infty & \text{if } E_i > -x/r_i \wedge i \neq j, \\ -x/r_i & \text{if } E_i > -x/r_i \wedge i = j, \\ E_i + R_{kj}(x + r_i E_i) & \text{if } E_i \leq -x/r_i \wedge i \to k \in \mathcal{E}, \\ E_i + V_{kl} + R_{lj}(x + r_i E_i) & \text{if } E_i \leq -x/r_i \wedge i \to k \in \mathcal{F} \wedge l \in \mathcal{E}, \end{cases}$$

where the right arrow symbol denotes a single-step transition in the modulating CTSMC. Hence,

$$\begin{aligned} R_{ij}^*(x, \alpha) &= \mathrm{E}[e^{-\alpha R_{ij}(x)}] \\ &= \mathrm{E}[\mathrm{E}[e^{-\alpha R_{ij}(x)} \mid E_i]] \\ &= \delta_{ij} e^{\gamma_i x/r_i} e^{\alpha x/r_i} + \sum_{k \in \mathcal{E}} p_{ik} \int_0^{-x/r_i} \gamma_i e^{-\gamma_i t} \mathrm{E}[e^{-\alpha(t + R_{kj}(x + r_i t))}] \, dt \\ &\quad + \sum_{k \in \mathcal{F}} \sum_{l \in \mathcal{E}} p_{ik} \int_0^{-x/r_i} \gamma_i e^{-\gamma_i t} \mathrm{E}[e^{-\alpha(t + V_{kl} + R_{lj}(x + r_i t))}] \, dt \end{aligned}$$

$$= \delta_{ij} e^{(\alpha+\gamma_i)x/r_i} + \sum_{k \in \mathcal{E}} p_{ik} \int_0^{-x/r_i} \gamma_i e^{-(\alpha+\gamma_i)t} R_{kj}^*(x+r_i t, \alpha)\, dt$$

$$+ \sum_{k \in \mathcal{F}} \sum_{l \in \mathcal{E}} p_{ik} V_{kl}^*(\alpha) \int_0^{-x/r_i} \gamma_i e^{-(\alpha+\gamma_i)t} R_{lj}^*(x+r_i t, \alpha)\, dt.$$

Therefore,

$$L_{ij}(\beta, \alpha) = \delta_{ij} \int_0^\infty e^{-(\beta-(\alpha+\gamma_i)/r_i)x}\, dx$$

$$+ \sum_{k \in \mathcal{E}} \gamma_i p_{ik} \int_{x=0}^\infty \int_{t=0}^{-x/r_i} e^{-\beta x - (\alpha+\gamma_i)t} R_{kj}^*(x+r_i t, \alpha)\, dx\, dt$$

$$+ \sum_{k \in \mathcal{F}} \sum_{l \in \mathcal{E}} \gamma_i p_{ik} V_{kl}^*(\alpha) \int_{x=0}^\infty \int_{t=0}^{-x/r_i} e^{-\beta x - (\alpha+\gamma_i)t} R_{lj}^*(x+r_i t, \alpha)\, dx\, dt$$

$$= \delta_{ij}(\beta - (\alpha+\gamma_i)/r_i)^{-1}$$

$$+ \sum_{k \in \mathcal{E}} \gamma_i p_{ik} \int_{t=0}^\infty \int_{y=0}^\infty e^{-\beta y} R_{kj}^*(y, \alpha) e^{-(\alpha+\gamma_i-\beta r_i)t}\, dt\, dy$$

$$+ \sum_{k \in \mathcal{F}} \sum_{l \in \mathcal{E}} \gamma_i p_{ik} V_{kl}^*(\alpha) \int_{t=0}^\infty \int_{y=0}^\infty e^{-\beta y} R_{lj}^*(y, \alpha) e^{-(\alpha+\gamma_i-\beta r_i)t}\, dt\, dy,$$

by changing the domains of integration to $0 \le t < \infty$ and $-r_i t \le x < \infty$, and the integration variable from x to $y = x + r_i t$. The result now follows.

In order to solve for $R_{i,j}^*(x, \alpha)$, we make the observation that the Laplace transform of R^* itself with respect to x can be expressed in partial fractions, each of which is in a form that can be inverted analytically. This enables us to write $R_{i,j}^*(x, \alpha)$ in closed form as a sum of weighted exponential terms. To this end, we define the following matrices:

- the $n_e \times n_e$ matrix $L = (L_{ij}(\beta, \alpha))$ for $i, j \in \mathcal{E}$;

- the $n_e \times n_e$ identity matrix I_e;

- the $n_e \times n$ matrix $U = (u_{ij})$ by $u_{ij} = -p_{ij}\gamma_i/r_i$ for $i \in \mathcal{E}$ and $j \in \mathcal{S}$;

- the $n_e \times n_e$ diagonal matrix $D = (d_{ij})$ by $d_{ij} = [\beta - (\alpha+\gamma_i)/r_i]\delta_{ij}$ for $i, j \in \mathcal{E}$;

- and the $n \times n_e$ matrix $W = (w_{ij})$ by

$$w_{ij} = \begin{cases} V_{ij}^*(\alpha) & \text{if } i \in \mathcal{F}, \\ \delta_{ij} & \text{if } i \in \mathcal{E}, \end{cases}$$

for $j \in \mathcal{E}$.

Then the lemma above may be written in the form $DL = I_e + UWL$ so that $L = (D - UW)^{-1}$.

Note that of the matrices other than L (the dependent variable), only matrix D depends on β, and in it every matrix element is a linear function of β. The same therefore also applies to the $n_e \times n_e$ matrix $M = D - UW$ so that its determinant, $\Delta := |M|$, is a polynomial of degree n_e in β and all its cofactors are polynomials of degree $n_e - 1$. If $\{b_s(\alpha) \mid 1 \le s \le r\}$

are the distinct roots of the equation $\Delta = 0$ (in β) and m_s is the multiplicity of root $b_s(\alpha)$ (so that $\sum m_s = n_e$), then the determinant of $\boldsymbol{M}$ can be written as $\Delta = \prod_{s=1}^{r}(\beta - b_s)^{m_s}$.

Each element in the inverse matrix $\boldsymbol{L} = (\boldsymbol{D} - \boldsymbol{U}\boldsymbol{W})^{-1}$ (assuming this exists, i.e. that Δ is nonzero) is a rational function of β and can therefore be written in partial fractions as

$$L_{ij}(\beta, \alpha) = (\boldsymbol{D} - \boldsymbol{U}\boldsymbol{W})_{ij}^{-1} = \sum_{s=1}^{r} \sum_{k=1}^{m_s} \frac{a_{sk;ij}(\alpha)}{(\beta - b_s(\alpha))^k}$$

for $i, j \in \mathcal{E}$, where the $a_{sk;ij}(\alpha)$ are independent of β, defined by (dropping the parameter α when the meaning is clear)

$$a_{sk;ij} = \frac{1}{(m_s - k)!} \left[\frac{\mathrm{d}^{m_s - k}}{\mathrm{d}\beta^{m_s - k}} \left(\frac{c_{ji}}{\Delta_s} \right) \right]_{\beta = b_s},$$

where $\boldsymbol{C} = (c_{ij} \mid i, j \in \mathcal{E})$ is the cofactor matrix of $\boldsymbol{M}$, whence $\boldsymbol{L} = \boldsymbol{C}^{\top}/\Delta$, and $\Delta_s = \Delta/(\beta - b_s)^{m_s}$. We define Δ_s to emphasize the fact that the term $(\beta - b_s)^{m_s}$ must symbolically cancel the corresponding term in Δ, when computing numerically, in order to avoid infinities; indeed, $\Delta_s \equiv \prod_{1 \leq i \neq s \leq n_e}(\beta - b_i)^{m_i}$.

Since n_e will usually be small—no more than about 10—the constants $\{a_{sk;ij}, b_s, m_s\}$ can easily be computed, for each α, by mathematical software such as MATHEMATICA, the choice of the authors. Note that the coefficients in the equation $|\boldsymbol{M}| = 0$ are functions of the sought after $V_{ij}^*(\alpha)$. Hence, fixpoint equations must be posed, which are solved iteratively. Except in very simple cases, the values of the $a_{sk;ij}(\alpha)$ must be calculated numerically on each iteration. Note also that, unless there is some structural symmetry in the construction of the modulating CTSMC (such as may happen in a network with identical queues and external arrival processes), it is very rare to find any $m_s \neq 1$. In this case we have the much simpler expression $a_{sk;ij} = [c_{ji}/\Delta_s]_{\beta = b_s}$.

Symbolic solutions are not usually feasible for $n_e > 2$—impossible for $n_e > 5$—and we will be relying on numerical methods, for example in calculating moments.

We note that the following properties must hold in order for equilibrium to exist:

- $|\boldsymbol{M}| \neq 0$ so that $\boldsymbol{M}$ has an inverse;

- the roots b_s all have negative real part, unless the corresponding terms $a_{sk;ij}$ are 0.

Certainly these apply to all the examples we have considered. The Laplace transform can now be inverted to give the following result.

Proposition 1. *For $i, j \in \mathcal{E}$ and real $x > 0$,*

$$R_{ij}^*(x, \alpha) = \sum_{s=1}^{r} \sum_{k=1}^{m_s} \frac{a_{sk;ij}}{(m_s - k)!} x^{m_s - k} \mathrm{e}^{b_s x}.$$

Proof. The proof is immediate, using the standard result for the Laplace transform of a gamma density function, or by noting that

$$\frac{\mathrm{d}^m}{\mathrm{d}\theta^m} \int_0^\infty \mathrm{e}^{-\theta x} f(x)\,\mathrm{d}x = (-1)^m \int_0^\infty \mathrm{e}^{-\theta x} x^m f(x)\,\mathrm{d}x$$

for all functions f when the integrals and derivatives exist.

3.2. Volume of fluid collected

To find the busy period, all we now have to do is find out how much fluid is collected in the buffer during a sequence of filling states (e.g. Y_{ik} in Figure 1). If the buffer was empty just prior to starting this sequence then the busy period is the sum of the times taken to pass through the sequence and then to clear the buffer as per the preceding analysis. To this end, we compute the following joint Laplace transforms for $i \in \mathcal{F}$ and $j \in \mathcal{E}$:

$$G_{ij}(\theta, \phi) = \mathrm{E}[e^{-\theta T_{ij} - \phi Y_{ij}}].$$

These are simply determined as follows.

Proposition 2. *For $i \in \mathcal{F}$ and $j \in \mathcal{E}$,*

$$G_{ij}(\theta, \phi) = p_{ij} H_{ij}^*(\theta + \phi r_i) + \sum_{k \in \mathcal{F}} p_{ik} H_{ik}^*(\theta + \phi r_i) G_{kj}(\theta, \phi). \tag{1}$$

Proof. We have

$$T_{ij} = \begin{cases} H_{ij} & \text{if } i \to j \in \mathcal{E}, \\ H_{ik} + T_{kj} & \text{if } i \to k \in \mathcal{F}, \end{cases}$$

and, similarly,

$$Y_{ij} = \begin{cases} r_i H_{ij} & \text{if } i \to j \in \mathcal{E}, \\ r_i H_{ik} + Y_{kj} & \text{if } i \to k \in \mathcal{F}. \end{cases}$$

Consequently,

$$\begin{aligned}
G_{ij}(\theta, \phi) &= p_{ij} \mathrm{E}[e^{-\theta H_{ij} - \phi r_i H_{ij}}] + \sum_{k \in \mathcal{F}} p_{ik} \mathrm{E}[e^{-\theta H_{ik} - \phi r_i H_{ik}} e^{-\theta T_{kj} - \phi Y_{kj}}] \\
&= p_{ij} H_{ij}^*(\theta + \phi r_i) + \sum_{k \in \mathcal{F}} p_{ik} H_{ik}^*(\theta + \phi r_i) G_{kj}(\theta, \phi),
\end{aligned}$$

by the Markov property that holds at the instant when $i \to k$.

3.3. The busy period

We are now in a position to determine the busy period itself. Such a time interval begins at a transition instant from a state $l \in \mathcal{E}$ to a state $i \in \mathcal{F}$ when the buffer is empty. When the system next returns to a state k, say, in $\mathcal{E}$, after a time T_{ik}, the buffer will hold Y_{ik} units of fluid, which will be cleared in state $j \in \mathcal{E}$ after a time $R_{kj}(Y_{ik})$. The length of the busy period that started in state $i \in \mathcal{F}$ and ends in state $j \in \mathcal{E}$ is then $T_{ik} + R_{kj}(Y_{ik})$.

We therefore have the main result of this paper.

Theorem 1. *For $i \in \mathcal{F}$ and $j \in \mathcal{E}$,*

$$V_{ij}^*(\alpha) = \sum_{k \in \mathcal{E}} \sum_{s=1}^{r} \sum_{l=1}^{m_s} \frac{(-1)^{m_s - l} a_{sl;kj} G_{ik}^{(0, m_s - l)}(\alpha, -b_s)}{(m_s - 1)!},$$

where

$$G_{ij}^{(0,m)}(x, y) = \left. \frac{\partial^m G_{ij}(\theta, \phi)}{\partial \phi^m} \right|_{\theta = x, \, \phi = y}.$$

Proof. Let the random variable K denote the first state in $\mathcal{E}$ entered by the modulating CTSMC after its initial state $i \in \mathcal{F}$. Then

$$V_{ij} = T_{iK} + R_{Kj}(Y_{iK}),$$

and so

$$
\begin{aligned}
V_{ij}^* &= \mathrm{E}[\mathrm{E}[e^{-\alpha T_{iK}} e^{-\alpha R_{Kj}(Y_{iK})} \mid K, T_{iK}]] \\
&= \mathrm{E}[e^{-\alpha T_{iK}} R_{Kj}^*(Y_{iK}, \alpha)] \\
&= \sum_{s=1}^{r} \sum_{l=1}^{m_s} \mathrm{E}\left[\frac{a_{sl;Kj}}{(m_s - l)!} \mathrm{E}[Y_{iK}^{m_s - l} e^{-\alpha T_{iK}} e^{b_s Y_{iK}} \mid K] \right] \\
&= \sum_{s=1}^{r} \sum_{l=1}^{m_s} \frac{(-1)^{m_s - l}}{(m_s - l)!} \mathrm{E}[a_{sl;Kj} G_{iK}^{(0, m_s - l)}(\alpha, -b_s)] \\
&= \sum_{k \in \mathcal{E}} \sum_{s=1}^{r} \sum_{l=1}^{m_s} \frac{(-1)^{m_s - l} a_{sl;kj} G_{ik}^{(0, m_s - l)}(\alpha, -b_s)}{(m_s - l)!}.
\end{aligned}
$$

4. Some examples

4.1. Single off-state

Consider a fluid queue with a single source defined by a CTSMC with a single emptying state, 0 say, typically called the 'off-state' and $n - 1$ filling states (or 'on-states') $1, \ldots, n - 1$. Then Proposition 2 gives, for $1 \le i \le n - 1$,

$$G_{i0}(\theta, \phi) = p_{i0} H_{i0}^*(\theta + \phi r_i) + \sum_{k \in \mathcal{F}} p_{ik} H_{ik}^*(\theta + \phi r_i) G_{k0}(\theta, \phi).$$

The matrix $\boldsymbol{M}$ is 1×1 with single element $\beta - (\alpha + \gamma_0)/r_0 + \sum_{j=1}^{n-1} p_{0j} \gamma_0 V_{j0}^*(\alpha)/r_0$, so that $r = 1, m_1 = 1, a_{11;00} = 1$, and $b_1 = [\alpha + \gamma_0(1 - \sum_{j=1}^{n-1} p_{0j} V_{j0}^*(\alpha))]/r_0$.

In the special case that there is also a single on-state, labelled 1, with $p_{10} = p_{01} = 1$ and $G_{10}(\theta, \phi) = H_{10}^*(\theta + \phi r_1)$, Theorem 1 then yields

$$V_{10}^*(\alpha) = a_{11;11} G_{10}(\alpha, -b_1) = H_{10}^*\left(\alpha - \frac{[\alpha + \gamma_0(1 - V_{10}^*(\alpha))] r_1}{r_0} \right).$$

This is the well-known result for the single on–off source with general on-time and exponential off-time [4], [5]. Of course, in this very simple case, we can use Lemma 1 directly in Theorem 1, by-passing Proposition 2.

4.1.1. *Multiple sources.* Consider first the case of two on–off sources, with exponential off-periods, to a single fluid queue with output rate less than either of the input rates of the sources in their on-states. This is the situation in [4] when there are two sources, each with exponential on-period. The modulating CTSMC is now a CTMC with state transition diagram as in Figure 2. The single off-state is labelled 0, the state with only source 1 on is labelled 1, the state with only source 2 on is labelled 2, and the state with both sources on is labelled 3. The off-to-on and on-to-off rates at source $i = 1, 2$ are respectively denoted by γ_i and κ_i. The fluid arrival rates for sources 1 and 2 are λ_1 and λ_2, respectively. The arrival rate in each state of the modulating
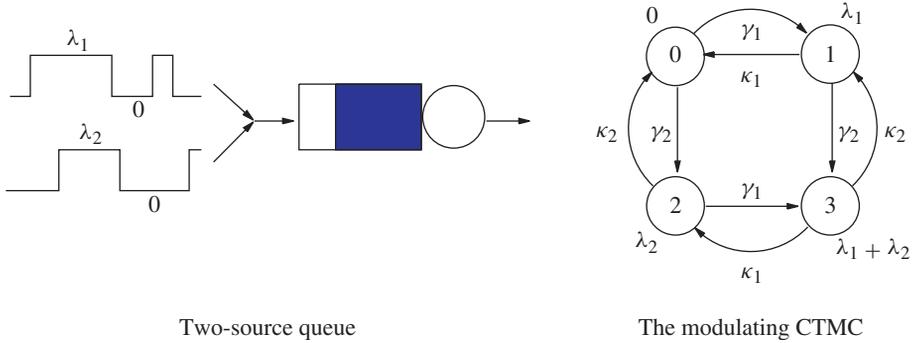
FIGURE 2: A fluid queue with two on–off sources.

CTSMC is written next to the state in the figure, and we have $\lambda_1 < \mu$, $\lambda_2 < \mu$, and $\lambda_1 + \lambda_2 > \mu$. As before, the output rate of the server, when busy, is μ. From Theorem 1 we have, for $i = 1, 2$,

$$V_{i0}^*(\alpha) = G_{i0}\left(\alpha, -\frac{\alpha + \gamma_0(1 - \sum_{j=1}^{n-1} p_{0j} V_{j0}^*(\alpha))}{r_0}\right),$$

where $\gamma_0 = \gamma_1 + \gamma_2$ and the generating functions G_{j0}, $j = 1, 2, 3$, are the solutions for g_j in the equations given by Proposition 2:

$$g_1 = e_1 p_{10} + p_{13} e_1 g_3, \qquad g_2 = e_2 p_{20} + p_{23} e_2 g_3, \qquad g_3 = e_3 p_{31} g_1 + e_3 p_{32} g_2, \qquad (2)$$

where $e_j = E_j^*(s_j) = \eta_j/(\eta_j + s_j)$, $s_j = \alpha - b_1 r_j = \alpha - [\alpha + \gamma_0(1 - \sum_{j=1}^{n-1} p_{0j} V_{j0}^*(\alpha))] r_j / r_0$, and $1/\eta_j$ is the mean holding time in state j, so that $\eta_1 = \kappa_1 + \gamma_2$, $\eta_2 = \kappa_2 + \gamma_1$, $\eta_3 = \kappa_1 + \kappa_2$. The state transition probability matrix of the modulating CTSMC, $\boldsymbol{P} = (p_{ij} \mid 0 \le i, j \le 3)$, is

$$\boldsymbol{P} = \begin{pmatrix} 0 & \dfrac{\gamma_1}{\gamma_1 + \gamma_2} & \dfrac{\gamma_2}{\gamma_1 + \gamma_2} & 0 \\[2mm] \dfrac{\kappa_1}{\kappa_1 + \gamma_2} & 0 & 0 & \dfrac{\gamma_2}{\kappa_1 + \gamma_2} \\[2mm] \dfrac{\kappa_2}{\kappa_2 + \gamma_1} & 0 & 0 & \dfrac{\gamma_1}{\kappa_2 + \gamma_1} \\[2mm] 0 & \dfrac{\kappa_2}{\kappa_1 + \kappa_2} & \dfrac{\kappa_1}{\kappa_1 + \kappa_2} & 0 \end{pmatrix}.$$

Eliminating g_3 from (2) and simplifying, we find that

$$(\kappa_1 + \kappa_2 + \omega_1 + \omega_2)[(\kappa_1 + \omega_1)g_1 - \kappa_1]$$
$$+ \gamma_2[(\kappa_1 + \kappa_2 + \omega_1 + \omega_2)g_1 g_2 - \kappa_2 g_1 - \kappa_1 g_2] = 0,$$
$$(\kappa_1 + \kappa_2 + \omega_1 + \omega_2)[(\kappa_2 + \omega_2)g_2 - \kappa_2]$$
$$+ \gamma_1[(\kappa_1 + \kappa_2 + \omega_1 + \omega_2)g_1 g_2 - \kappa_2 g_1 - \kappa_1 g_2] = 0,$$

where $\omega_1 = s_1 + \gamma_2(1 - g_2)$ and $\omega_2 = s_2 + \gamma_1(1 - g_1)$ so that $s_3 = \omega_1 + \omega_2$. These equations have the solution $g_i = \kappa_i/(\kappa_i + \omega_i)$ for $i = 1, 2$, which is that obtained by Boxma and Dumas [4], as expected.

In fact, their result allows generally distributed on-times, provided the input rate of every source exceeds the output rate of the server, giving only one off-state. This was proved using properties specific to the busy period random variable: by observing that if a source i turns on at time T_1 when the fluid level in the buffer is $X_1 > 0$ (i.e. during a busy period), the time that elapses up to the instant T_2 when the fluid level next returns to X_1 is just a busy period that starts (during an off-period) with source i turns on. Our method cannot be applied as it stands because the corresponding input process is not a CTSMP. This is because when a new source turns on during the on-period of some other source, the Markov property does not hold since the residual on-period is not independent of the partially completed on-period, i.e. of the time in the past at which the source turned on. However, using the same busy period 'trick', our method could be adapted similarly, if required. Instead of repeating this work, we consider input processes with more than one off-period ($n_e > 1$) so that, for example, we can deal with a fluid queue with multiple input sources, not all of which have input rates greater than the output rate of the server.

4.2. Two off-states

In general, when $n_e > 1$, we first need to compute the determinant $\Delta(\alpha, \beta)$ to get the inverse matrix (L) of M, expanded in partial fractions. This is a polynomial equation of degree n_e, which in general requires numerical methods to solve. However, in the special case that $n_e = 2$, representing two input states with net positive fluid output rate, we can continue symbolically. This situation occurs when exactly one of several sources to a fluid queue has input rate less than the server's output rate, μ say.

In this model, state 0 represents both sources being off, state 1 represents source 1 only being on, emitting fluid at rate $\lambda_1 < \mu$, state 2 represents source 2 only being on, emitting fluid at rate $\lambda_2 > \mu$, and state 3 represents both sources being on. The state transition system is the same as that in Figure 2, except that states 0 and 1 are now both emptying states. We assume the same transition rates as above ($\kappa_i, \gamma_i, i = 1, 2$) and state holding time distributions (exponential in all cases).

4.2.1. The functions G_{ij}. The net input rates in this model are $r_0 = -\mu, r_1 = \lambda_1 - \mu, r_2 = \lambda_2 - \mu$, and $r_3 = \lambda_1 + \lambda_2 - \mu$, and (1) then becomes

$$G_{20}(\theta, \phi) = E_2^*(\theta + \phi r_2) p_{20} + E_2^*(\theta + \phi r_2) p_{23} G_{30}(\theta, \phi),$$
$$G_{21}(\theta, \phi) = E_2^*(\theta + \phi r_2) p_{23} G_{31}(\theta, \phi),$$
$$G_{30}(\theta, \phi) = E_3^*(\theta + \phi r_3) p_{32} G_{20}(\theta, \phi),$$
$$G_{31}(\theta, \phi) = E_3^*(\theta + \phi r_3) p_{31} + E_3^*(\theta + \phi r_3) p_{32} G_{21}(\theta, \phi).$$

Using the above parameters, $E_2^*(s_2) = (\kappa_2 + \gamma_1)/(\kappa_2 + \gamma_1 + s_2)$, $E_3^*(s_3) = (\kappa_1 + \kappa_2)/(\kappa_1 + \kappa_2 + s_3)$, $p_{20} = \kappa_2/(\kappa_2 + \gamma_1)$, $p_{23} = \gamma_1/(\kappa_2 + \gamma_1)$, $p_{31} = \kappa_2/(\kappa_1 + \kappa_2)$, and $p_{32} = \kappa_1/(\kappa_1 + \kappa_2)$. The matrix $G = (G_{ij} \mid i = 2, 3, \ j = 0, 1)$ is therefore defined by

$$G = \frac{\kappa_2}{(\kappa_2 + \gamma_1 + s_2)(\kappa_1 + \kappa_2 + s_3) - \gamma_1 \kappa_1} \begin{pmatrix} \kappa_1 + \kappa_2 + s_3 & \gamma_1 \\ \kappa_1 & \kappa_2 + \gamma_1 + s_2 \end{pmatrix},$$

where $s_2 = \theta + \phi r_2$ and $s_3 = \theta + \phi r_3$.

4.2.2. Determination of b_s and $a_{s1;ij}$. To use Theorem 1, we now need the matrix M to calculate the terms b_s and $a_{s1;kj}$ for $j, k, s = 1, 2$; we restrict ourselves to the case where the roots b_s are distinct. Using the notation of Section 3.1, the matrix $D = \mathrm{diag}(\beta - (\alpha + \gamma_i)/r_i)$ $(i = 0, 1)$,

$u_{ij} = -p_{ij}\gamma_i/r_i$ ($i = 0, 1$, $j = 0, 1, 2, 3$), and $w_{ij} = V_{ij}^*(\alpha)$ if $i = 2, 3$, $j = 0, 1$, and $w_{ij} = \delta_{ij}$ if $i, j = 0, 1$. Hence,

$$M = \begin{pmatrix} \beta - \dfrac{\alpha + \gamma_1 + \gamma_2(1 - V_{20}^*(\alpha))}{r_0} & \dfrac{\gamma_1 + \gamma_2 V_{21}^*(\alpha)}{r_0} \\ \dfrac{\kappa_1 + \gamma_2 V_{30}^*(\alpha)}{r_1} & \beta - \dfrac{\alpha + \kappa_1 + \gamma_2(1 - V_{31}^*(\alpha))}{r_1} \end{pmatrix}$$

and $b_1(\alpha)$, $b_2(\alpha)$ are the roots (assumed distinct) of the equation

$$\beta^2 - \left(\frac{\alpha + \gamma_1 + \gamma_2(1 - V_{20}^*(\alpha))}{r_0} + \frac{\alpha + \kappa_1 + \gamma_2(1 - V_{31}^*(\alpha))}{r_1} \right)\beta$$

$$+ \frac{(\alpha + \gamma_1 + \gamma_2(1 - V_{20}^*(\alpha)))(\alpha + \kappa_1 + \gamma_2(1 - V_{31}^*(\alpha))) - (\gamma_1 + \gamma_2 V_{21}^*(\alpha))(\kappa_1 + \gamma_2 V_{30}^*(\alpha))}{r_0 r_1} = 0.$$

The terms $a_{11;ij}$ and $a_{21;ij}$ are the (i, j)th elements of the matrices

$$A_{11} = \frac{1}{b_1(\alpha) - b_2(\alpha)} \begin{pmatrix} b_1(\alpha) - \dfrac{\alpha + \kappa_1 + \gamma_2(1 - V_{31}^*(\alpha))}{r_1} & -\dfrac{\gamma_1 + \gamma_2 V_{21}^*(\alpha)}{r_0} \\ -\dfrac{\kappa_1 + \gamma_2 V_{30}^*(\alpha)}{r_1} & b_1(\alpha) - \dfrac{\alpha + \gamma_1 + \gamma_2(1 - V_{20}^*(\alpha))}{r_0} \end{pmatrix}$$

and

$$A_{21} = -\frac{1}{b_1(\alpha) - b_2(\alpha)} \begin{pmatrix} b_2(\alpha) - \dfrac{\alpha + \kappa_1 + \gamma_2(1 - V_{31}^*(\alpha))}{r_1} & -\dfrac{\gamma_1 + \gamma_2 V_{21}^*(\alpha)}{r_0} \\ -\dfrac{\kappa_1 + \gamma_2 V_{30}^*(\alpha)}{r_1} & b_2(\alpha) - \dfrac{\alpha + \gamma_1 + \gamma_2(1 - V_{20}^*(\alpha))}{r_0} \end{pmatrix}.$$

Defining the matrix $V = (V_{ij}^*(\alpha) \mid i = 2, 3, \ j = 0, 1)$, we finally have

$$V_{ij}^*(\alpha) = \sum_{k=0}^{1} \sum_{s=1}^{2} a_{s1;kj} G_{ik}(\alpha, -b_s(\alpha)),$$

$$V = G(\alpha, -b_1(\alpha)) \cdot A_{11} + G(\alpha, -b_2(\alpha)) \cdot A_{21}.$$

5. Numerical experiments

We present below numerical results for some simple single-server systems. For each case considered, we are able to produce identical behaviour, modulo confidence intervals, from an equivalent fluid queue simulation. (We used a fluid extension of the JINQS library, see http://www.doc.ic.ac.uk/~ajf/Research/manual.pdf, for this purpose.)

First we outline the main issues involved in the numerical computations of the density function and moments of the busy period.

5.1. Computational algorithms

The LST of the busy period probability distribution is given by the equation in Theorem 1. Since the $a_{sl;kj}(\alpha)$ and $b_s(\alpha)$ terms all depend on $\{V_{ij}^*(\alpha) \mid i \in \mathcal{F}, \ j \in \mathcal{E}\}$, which is what we seek, this is a complex fixpoint equation for the matrix

$$V^*(\alpha) = (V_{ij}^*(\alpha) \mid i \in \mathcal{F}, \ j \in \mathcal{E}).$$

The natural algorithm for computing the value of $V^*(\alpha)$ at any given value of α is to make an initial guess and successively update the current guess by substituting it into the right-hand side of the equation. The existence of a fixpoint solution follows from Brouwer's theorem, V^* being bounded, but convergence of this particular algorithm has not been proved. Our implementation converged quickly for all the examples we tried, in which values were required at very many points α in calls from a Laplace transform numerical inverter (see below).

5.1.1. *Moments.* The moments of the busy period can be obtained by differentiating V^* at the origin, avoiding the need to invert V^*. However, the derivatives of $a_{sl;kj}$ and b_s are not simple to obtain, as the defining expressions are functions of V^* themselves, as noted above. Their nth derivatives can be expressed in terms of all the derivatives of V^* up to the nth, leading to an incremental sequence of fixpoint problems for the ith derivative, $i = 1, 2, \ldots, n$. Numerical solutions can then be computed iteratively in the same way as described above for the LST. In particular, the derivatives of $b_1, \ldots, b_{n_e}$ require care, these being the roots of a polynomial equation of degree n_e. For $n_e = 1$ or 2, the roots can be computed as relatively simple symbolic expressions. This is more problematic for higher n_e and impossible for $n_e > 4$.

The problem can be overcome using the following proposition, which defines the mth derivative of a root, b, of an arbitrary polynomial at a specified point in terms of the derivatives of the polynomial's coefficients and the value of b at that point.

Proposition 3. *For the polynomial equation*

$$\sum_{i=1}^{n} c_i(\alpha) x^i = 0, \tag{3}$$

the mth derivative of a root $x = b(\alpha)$ with respect to α is

$$b^{(m)} = -\frac{1}{\sum_{i=0}^{n} i c_i b^{i-1}}$$
$$\times \left(\sum_{i=0}^{n} \sum_{k=0}^{m-1} \binom{m}{k} c_i^{(m-k)} D_{k-1,0}(i) + \sum_{i=0}^{n} c_i \sum_{j=0}^{m-2} \binom{m-1}{j} D_{m-1-j,1}(i) b^{(j+1)} \right),$$

where, for $0 \le l \le k$,

$$D_{kl}(i) = I(l \le i - 1) \frac{i!}{(i-l-1)!} b^{i-l-1} b^{(k-l+1)} + \sum_{j=0}^{k-l-1} \binom{k-l}{j} D_{k-j,l+1}(i) b^{(j+1)},$$

wherein $D_{-1,0}(i) = b^i$ and $I(\cdot)$ is the indicator function.

The proof is given in Section A.2 of Appendix A. This result can be used to compute the required mth derivatives of the roots of the polynomial equation of Section 3.1 at $\alpha = 0$ (the roots themselves being obtained numerically), whilst avoiding the numerical problems outlined above. The details of functions that compute the derivatives of the $a_{sl;kj}$ and $G_{i,j}$, and, hence, the moments of V^*, are also given in Appendix A.

In principle, the derivatives of the b_i could be computed numerically. In practice, however, the computation of derivatives higher than the first turns out to be unstable for step sizes less than about 10^{-8}, using the default precision of MATHEMATICA. For step sizes greater than this, accuracy is inadequate (more than 10% error) even for the second derivative, although it is

satisfactory for the first derivative. Working with the highest level of precision available might make some more derivatives numerically viable, but the computational cost would be greater and almost certainly instability would arise in higher derivatives.

5.2. Results

In the examples we now consider, both the moments and probability density function of the busy period were computed using MATHEMATICA. The moments were calculated by directly coding the functions listed in Section A.2, using the above proposition in particular. The density functions were computed by using the fixed Talbot method to numerically invert the Laplace transform V^*, itself computed iteratively as described above [1].

We assume throughout that there is a single server with a fixed service rate of $\mu = 10$. We consider first a system with two independent input processes, each defined by a Markov-modulated on–off process with the following characteristics.

- Exponentially distributed state holding times in each state, with transition rates 1 (off $\rightarrow$ on) and 2 (on $\rightarrow$ off) for both inputs.

- A fluid arrival rate of λ_1 or λ_2 for the first or, respectively, second input process when that process is in the on-state.

The composite arrival process therefore has a modulating CTMC with four states, as per Figure 2. Its probability transition matrix was given in Section 4.2. The fluid arrival rates for states 0–3 are respectively 0 (both off), λ_1 (input 1 on), λ_2 (input 2 on), and $\lambda_1 + \lambda_2$ (both on).

We compute the mean fluid level of the busy period, using the result in [5], as λ_2 is increased, whilst fixing λ_1 at 12; the experiment is then repeated for $\lambda_1 = 16$. Note that the stability condition is $\lambda_1 + \lambda_2 < 30$, since the proportion of time spent by each (independent) source in its on-state is $\frac{1}{3}$ and so the sources contribute net traffic at rate $(\lambda_1 + \lambda_2)/3$. For $\lambda_2 < 10$, there are two emptying states (0 and 1) and, for $\lambda_2 > 10$, there is only one emptying state (0). We are particularly interested in the behaviour of the system across the transition point where $\lambda_2 = \mu = 10$.

Figure 3 shows the mean fluid level in the buffer which, as expected, increases continuously with λ_2. Figure 4 shows the mean busy period. The immediately striking artefact is the discontinuity at $\lambda_2 = \mu$. With this parameterisation, the mean busy period drops at $\lambda_2 = \mu$ and then increases continuously in accordance with the results reported in [4] (see Section 4.1.1). The behaviour of the busy period for $\lambda_2 < \mu$ (two off-states) illustrates the additional capabilities of our model. Note that a similar discontinuity is exhibited in the higher moments, as illustrated in Figure 5, which shows the standard deviation of the busy period for the same parameterisation.

The discontinuity occurs when there is a change in the number of emptying states (or filling states), here as a result of modifying the load on the system in terms of the fluid arrival rate in one of the states of the controlling CTSMC. (Note that if an emptying state becomes a filling state as a result of a change in load, the state holding time is required to be distributed exponentially.) We might expect, therefore, that if two similar load parameters are modified together, so that the number of emptying states varies between 1 and 3, then a discontinuity will emerge along each of the two dimensions of change. This is verified in Figure 6, which shows the behaviour of the mean busy period as both λ_1 and λ_2 are varied between 6 and 12; note that four continuous subregions are formed.
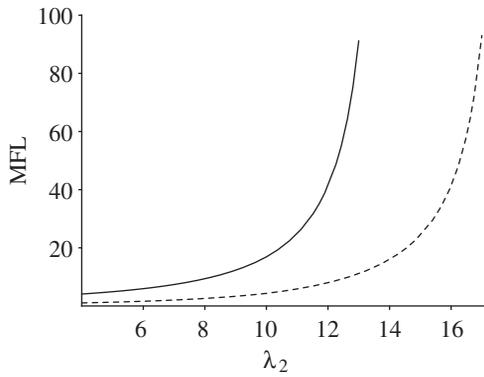
FIGURE 3: Mean fluid level (MFL) in a queue with two inputs, $\lambda_1 = 16$ (*solid line*) and $\lambda_1 = 12$ (*dashed line*).
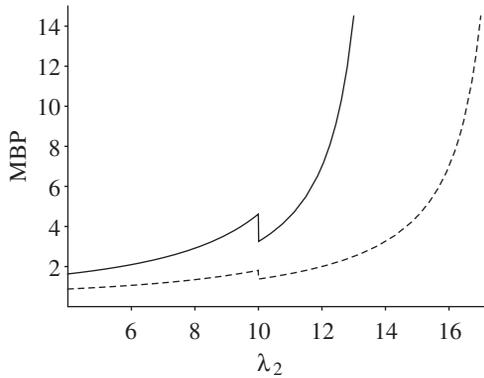

FIGURE 4: Mean busy period (MBP) in a queue with two inputs, $\lambda_1 = 16$ (*solid line*) and $\lambda_1 = 12$ (*dashed line*).
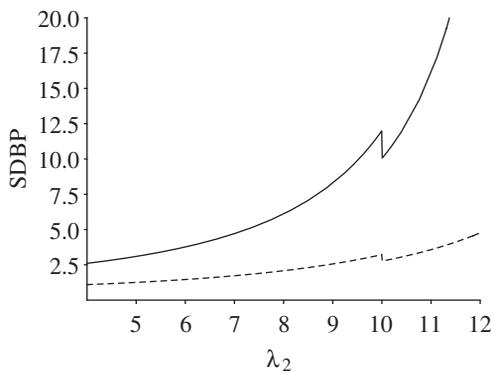

FIGURE 5: Standard deviation of busy period (SDBP) in a queue with two inputs, $\lambda_1 = 16$ (*solid line*) and $\lambda_1 = 12$ (*dashed line*).
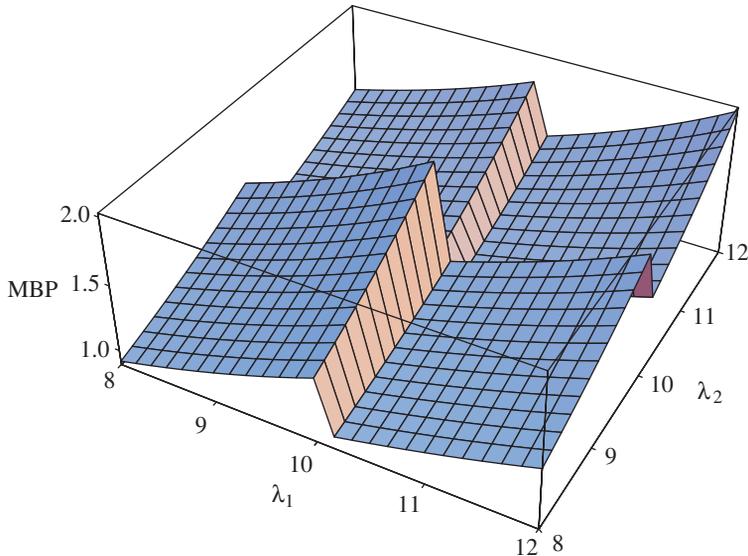
FIGURE 6: Mean busy period for a four-state model, as a function of λ_1 and λ_2.

5.3. Discontinuity

The observed discontinuity in the distribution of the busy period is counterintuitive and warrants further explanation. The behaviour is a consequence of the instantaneous transition of a state from being an emptying state to a filling state (or vice versa) as the load is increased (or decreased), and can be understood by considering how busy periods are formed. To illustrate the effect, consider a slightly simpler model in which there are just three states (0,1,2) whose fluid arrival rates are respectively $\lambda_0 = 0$, $\lambda_1 = \mu - \varepsilon$ for some $\varepsilon > 0$, and $\lambda_2 > \mu$. This has two emptying states (0 and 1) and one filling state (2).

A possible sample path of such a process is illustrated in Figure 7. The first time series (top) shows the fluid arrival rate as a function of time and the second time series (bottom) shows the volume of fluid in the buffer. For the time period shown, the input process first oscillates between states 0 and 1, both of which are emptying states, at a time when the buffer is empty. The buffer remains empty throughout this time since the fluid arrival rate in both states is less than μ. On the next transition to state 2, the buffer begins to fill, since $\lambda_2 > \mu$. At the end of the ensuing busy period, the buffer empties whilst the input process is in state 1.

Now consider what would happen for the identical sequence of transitions in the case where $\lambda_1 = \mu + \varepsilon$, so that state 1 now becomes a filling state. This is illustrated in Figure 8. Busy periods that existed previously will be extended in length by two mechanisms. Firstly, by the fact that state 1 is now a filling state, so that the buffer volume is increasing in state 1, rather than decreasing, this extra fluid requires additional time to be removed. Secondly, the busy period may start earlier than before, as the result of a transition from state 0 to state 1, when the buffer used to be empty prior to the start of the busy period considered above. If this additional fluid is not drained in time, during a sojourn in state 0, the effect is to add a prefix to that busy period. Both of these effects are illustrated in Figure 8.

Counter to this lengthening of the busy period is the possibility that new busy periods formed by the above mechanism may 'die young' as a result of the system later transiting to the (only remaining) empty state, and the buffer draining completely before the next busy period forms.
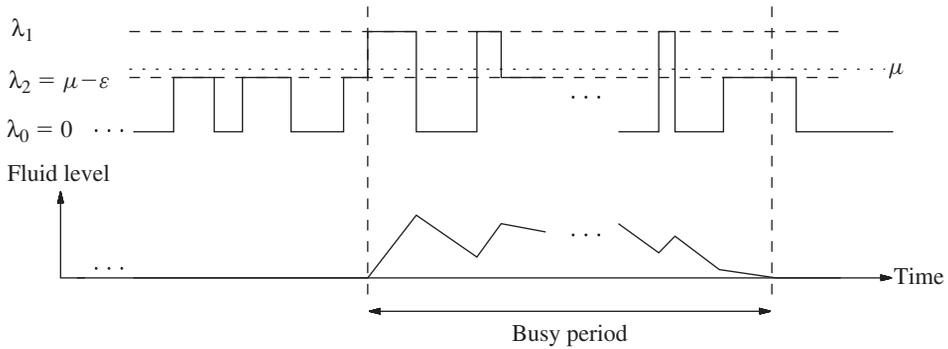
FIGURE 7: Sample path for a three-phase input process with two emptying states.
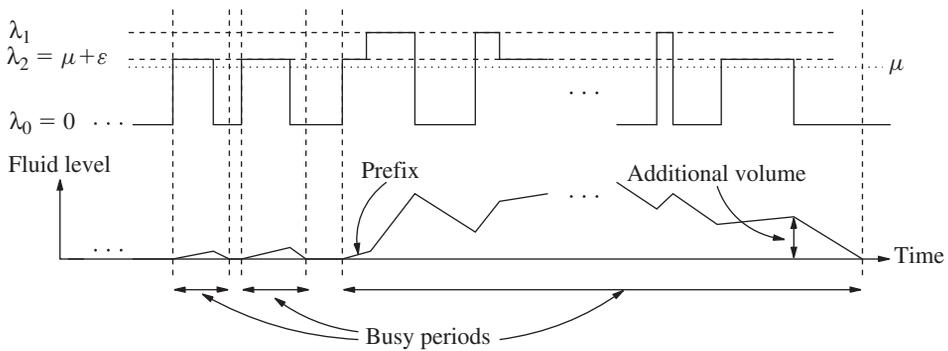

FIGURE 8: Sample path for a three-phase input process with one emptying state.

The effect of this is to introduce additional busy periods of relatively short duration. This is illustrated in Figure 8, where the first two transitions from state 0 to state 1 introduce short-lived busy periods that terminate following a subsequent transition back to state 0.

For this example, the intuition is thus as follows. If there is a relatively high probability of state 2 being entered on a transition from state 1, then the first effect dominates, and the busy periods will, on average, be extended in length. If the converse is true, however, then the second effect dominates and sufficiently many short-lived busy periods will be formed to reduce their average duration overall. In short, the effect of crossing the transition point is a step change in the shape of the busy period distribution.

Note that a reduction in the average length of a busy period does not imply that the server does less work overall. Clearly, if more fluid is introduced per unit time on average, then the server must spend more time removing that additional fluid. Note also that the change in behaviour is instantaneous at the point where λ_1 exceeds μ, hence the discontinuity.

In order to illustrate the effect quantitatively, we set up the three-state system as follows.

- Constant service rate 10.

- Exponentially distributed state holding times in each state, with a transition rate parameter of 0.5 in each case.

- Fluid arrival rate 0 in state 0 and 12 in state 1, with the rate in state 2 varying from 9.7 to 10.3, i.e. crossing the transition point where the state-2 rate matches the service rate.
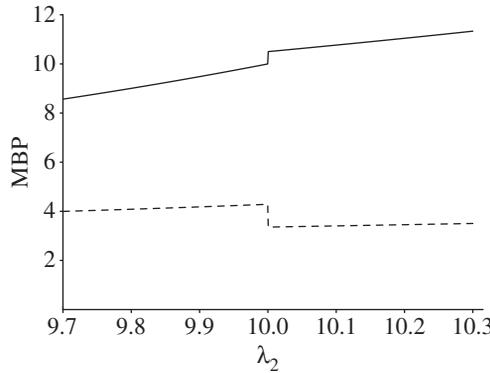
FIGURE 9: Mean busy period for a three-state input process, $p_{2,0} = 0.1$ (*solid line*) and $p_{2,0} = 0.9$ (*dashed line*).
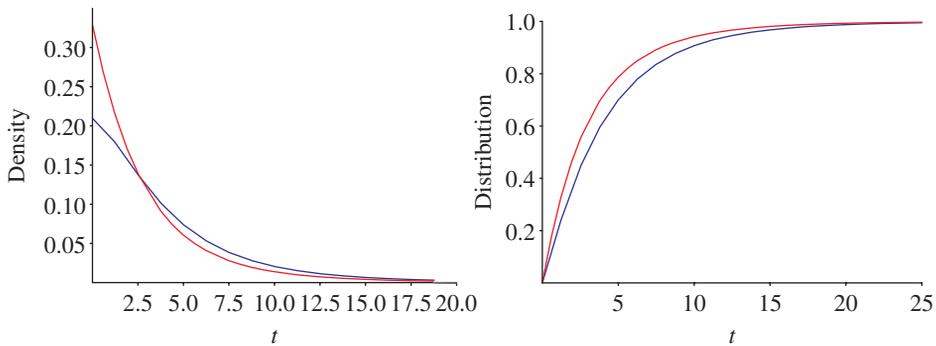

FIGURE 10: Busy period density and cumulative distribution function (CDF) for the three-state model. The flatter density corresponds to $\lambda_1 = 9.99$ (lower curve on the CDF). When λ_1 is increased to 10.01 (upper curve on the CDF), proportionally more short-lived busy periods are formed and the density shifts upwards at $t = 0$.

- Probability transition matrix

$$\begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \\ 0.1 & 0.9 & 0 \end{pmatrix}.$$

In this scenario a transition from state 2 to state 1 is significantly more likely than a transition from state 2 to state 0, and the effect is to cause a shift upwards in the mean length of the busy period across the discontinuity. This is illustrated in Figure 9 ($p_{20} = 0.1$).

If the transition matrix is modified so as to reverse the bias in the transition probabilities (i.e. by setting $p_{20} = 0.9$ and $p_{21} = 0.1$) then, for $\lambda_1 > \mu$, more short-lived busy periods will be formed, as there is a relatively much higher probability of entering the emptying state than the other filling state. This is borne out in Figure 9 ($p_{20} = 0.9$). Note that similar behaviour can be effected by adjusting the state holding times, instead of the state transition probabilities, e.g. by decreasing (then increasing) the rate parameter of the exponential distribution for the state 2 holding time when the state transition probabilities are all set to 0.5.

The effect on the busy period distribution of crossing the discontinuity is shown in Figure 10 for $\lambda_2 = 9.99$ and $\lambda_2 = 10.01$ ($\varepsilon = 0.01$) for the second scenario above ($p_{20} = 0.9$, $p_{21} = 0.1$).

The increased abundance of relatively short busy periods when $\lambda > \mu$, compared with when $\lambda < \mu$, can be seen clearly by comparing the two density functions. In general, the effect on the busy period distribution of a change in the number of emptying (hence filling) states depends subtly on the relationship between the transition probabilities and state holding times in the modulating CTSMC. It is not immediately obvious how to predict the effect for a specific parameterisation.

6. Conclusion

We have derived an equation for the Laplace transform (V^*) of the busy period density function in a fluid queue fed by a CTSMC. The fluid arrival rate in each state can be an arbitrary constant, provided that the system is stable overall. Furthermore, for those states for which the arrival rate exceeds the service rate, the state holding time can have an arbitrary probability distribution, all other state holding times being exponentially distributed.

The busy period can be a useful measure in its own right (it measures the time that the server is 'pumping' at its full capacity, for example), but the ability to compute moments of the busy period is also significant in the compositional analysis of fluid queueing networks, where each queue is modelled by ring-fencing it from the rest of the network. The actual input process is then approximated by an intermediate MMAP, whose parameters are determined by matching moments with the busy period(s) of the upstream queue(s). The analysis we have developed enables a richer class of networks to be analysed, partly because of the relaxed assumption about state holding times, but also because of our ability to handle arbitrary numbers of emptying states. This is important in complex networks where the aggregate instantaneous fluid arrival rate at a queue may often be less than the queue's service rate. The ability to handle an arbitrary number of emptying states is also likely to be important when modelling systems with large numbers of inputs. For the system to be stable, it is inevitable that many of those inputs will be emitting 'work', i.e. fluid, at low rates.

Our approach leads to exact analytical results in some cases, but in general involves numerical methods to obtain the busy period density function and/or its moments. We have demonstrated that our approach is computationally tractable, in that the computation of $V^*(\alpha)$, which involves a complex fixpoint calculation, converges successfully for the (many) parameter values that were required to complete the numerical Laplace transform inversions providing the density functions reported. Where only the moments of the busy period are required, there is no need to work with V^* directly; numerically differentiating $V^*(\alpha)$ at $\alpha = 0$, using the equations provided in Appendix A, yields corresponding fixpoint equations for the moments. This direct calculation of the moments (as opposed to numerical integration using the density function obtained by inverting $V^*(\alpha)$) is more efficient as only one set of fixpoint equations is needed for each moment, corresponding to $\alpha = 0$; recall that one fixpoint equation is required for each α when dealing with $V^*(\alpha)$.

Numerical experiments with selected case study systems showed somewhat unexpected behaviour in respect of the busy period, as the load on the system is scaled. For example, a continuous increase in the load (fluid arrival rate) in an emptying state was shown to lead to a discontinuity in the busy period distribution as the load parameter exceeds the service rate. At this point the state switches from being an emptying state to a filling state. As expected, the fluid level distribution behaves continuously over the same parameter range. We found that, for each such transition, leading to either an increase or decrease in the number of emptying states, a discontinuity is found in the busy period distribution, although it proved hard to predict in advance the direction of the discontinuity, e.g. whether the moments shift up or down. The

effect observed depends subtly on the parameterisation of the model and it would be interesting to see whether it could be predicted in advance in a simple way. The parameterisation of the experiments used to explain the discontinuity effects in Section 5.3 were arrived at by a mixture of educated guesses and trial and error.

Finally, we note that our result for the busy period can be used to compute the distribution of the response time in a priority fluid queue that is partitioned into low- and high-priority fluid compartments. The modulating input process is a CTMC with three states: high priority fluid is assumed to arrive in state 1, low priority fluid is assumed to arrive in state 2, and state 0 represents the off-state (no input). Analogously with conventional priority queues, fluid in the low priority compartment is only drained when the high priority compartment is empty. The high priority fluid is processed as if the low priority fluid is completely absent from the system, which corresponds to making state 2 an off-state. In this case, the set of off-states is thus $\mathcal{E} = \{0, 2\}$ and the set of all states is $\mathcal{S} = \{0, 1, 2\}$, so that $n_e = 2$ and $n = 3$. We have previously used this approach to model response times in a flash memory device with correlated streams of read, write, and erase requests [8].

Appendix A. Moments of the busy period

Denoting the nth derivative of a function F by $F^{(n)}$, i.e. $F^{(n)}(\alpha) = \mathrm{d}^n F / \mathrm{d}\alpha^n$ for $n \geq 0$, we compute the mth moment of a busy period that starts in state $i \in \mathcal{F}$ and ends in state $j \in \mathcal{E}$ as $(-1)^m V_{ij}^{*(m)}(0)$. Applying Leibnitz's rule to Theorem 1 requires us to determine the nth derivatives $(1 \leq n \leq m)$ of $a_{sl;kj}$ and b_s at $\alpha = 0$, as well as the moments of the holding times H_{ij}. These are obtained in the subsections below, after which we give the result for the moments of V_{ij}. We will find the following results useful.

A.1. Higher derivative chain rules

Let f and g be $(i + 1)$th order differentiable functions in $\mathbb{R} \to \mathbb{R}$, and, for nonnegative integers $j \leq i$, define the functions $D_{ij} \colon ((\mathbb{R} \to \mathbb{R}) \times (\mathbb{R} \to \mathbb{R})) \to (\mathbb{R} \to \mathbb{R})$ by

$$D_{ij}(f, g)(x) = \frac{\mathrm{d}^{i-j}}{\mathrm{d}x^{i-j}}(f^{(j+1)}(g(x))g^{(1)}(x)) \tag{4}$$

so that, for $i \geq 0$, $D_{ii}(f, g, x) = f^{(i+1)}(g(x))g^{(1)}(x)$. For ease of reading, we omit the arguments f and g where the meaning is clear.

Lemma 2. *For $0 \leq j \leq i$,*

$$D_{ij}(x) = f^{(j+1)}(g(x))g^{(i-j+1)}(x) + \sum_{m=0}^{i-j-1} \binom{i-j}{m} D_{i-m,j+1}(x)g^{(m+1)}(x).$$

Proof. Using definition (4) for $0 \leq j \leq i$ and applying Leibnitz's rule, we obtain

$$D_{ij}(x) = \frac{\mathrm{d}^{i-j}}{\mathrm{d}x^{i-j}}[f^{(j+1)}(g(x))g^{(1)}(x)]$$

$$= \sum_{m=0}^{i-j} \binom{i-j}{m}\left[\frac{\mathrm{d}^{i-j-m}}{\mathrm{d}x^{i-j-m}}f^{(j+1)}(g(x))\right]g^{(m+1)}(x)$$

$$= \sum_{m=0}^{i-j-1} \binom{i-j}{m} \left[\frac{d^{i-j-1-m}}{dx^{i-j-1-m}} f^{(j+2)}(g(x))g^{(1)}(x) \right] g^{(m+1)}(x)$$
$$+ f^{(j+1)}(g(x))g^{(i-j+1)}(x).$$

Given the functions f and g, we assume that $f^{(i)}$ and $g^{(i)}$ are known for $0 \le i \le k+1$, given integer k. Thus, certainly each D_{ii} is also known. These values are used as base cases in the recursion to calculate all the D_{ij} for $0 \le j \le i \le k$, from which $D_{i0} = d^{i+1} f(g(x))/dx^{i+1}$, evaluated at $x = 0$, computes the numerical values of the $(i + 1)$th moments.

More generally, for a suitably differentiable function F of n variables, we denote the partial derivative $\partial^{j_1} \cdots \partial^{j_n} F(x_1, \ldots, x_n)/\partial x_1^{j_1} \cdots \partial x_n^{j_n}$ by $F^{(\vec{j})}(x_1, \ldots, x_n)$, where the vector $\vec{j} = (j_1, \ldots, j_n)$. We define the n-component unit vectors $1^k = (0, \ldots, 0, 1, 0, \ldots, 0)$, i.e. $1_k^k = 1, 1_i^k = 0$ for $i \ne k$ and the zero vector $\vec{0} = (0, \ldots, 0)$. Then the above functions D_{ij} generalise to

$$\boldsymbol{D}_{i\vec{j}}(f, g_1, \ldots, g_n)(x) = \frac{d^{i-J}}{dx^{i-J}} \left(\sum_{k=1}^{n} F^{(\vec{j}+1^k)}(g_1(x), \ldots, g_n(x))g_k^{(1)}(x) \right),$$

where $J = \sum_{i=1}^{n} j_i$. Thus, for $i \ge 0$, $\boldsymbol{D}_{i\vec{j}}(F, g_1, \ldots, g_n)(x) = \sum_{k=1}^{n} F^{(\vec{j}+1^k)}(g_1(x), \ldots, g_n(x))g_k^{(1)}(x)$ for all $\vec{j}$ with $J = i$, i.e. whose components sum to i. This expression can be computed directly from the definitions of $F, g_1, \ldots, g_n$ and their derivatives. Similarly,

$$\boldsymbol{D}_{i\vec{0}}(F, g_1, \ldots, g_n)(x) = \frac{d^{i+1} f}{dx^{i+1}}$$

since, by the chain rule,

$$\frac{dF}{dx} = \sum_{k=1}^{n} F^{(1^k)}(g_1(x), \ldots, g_n(x))g_k^{(1)}(x).$$

Using Leibnitz's rule as in the lemma above now yields the following.

Proposition 4. *For vectors of nonnegative components $\vec{j}$ such that $0 \le J \le i$,*

$$\boldsymbol{D}_{i\vec{j}}(f, g_1, \ldots, g_n)(x) = \sum_{k=1}^{n} \left(F^{(\vec{j}+1^k)}(g_1(x), \ldots, g_n(x))g_k^{(i-J+1)}(x) \right.$$
$$\left. + \sum_{m=0}^{i-J-1} \binom{i-J}{m} \boldsymbol{D}_{i-m,\vec{j}+1^k}(x)g_k^{(m+1)}(x) \right)$$

(omitting the function arguments of $\boldsymbol{D}_{i-m,\vec{j}+1^k}$ on the right-hand side).

A.2. Proof of Proposition 3

Differentiating (3) m times with respect to x, we obtain

$$\sum_{i=0}^{n} \sum_{k=0}^{m} \binom{m}{k} c_i^{(m-k)} (b^i)^{(k)} = 0. \tag{5}$$

From Lemma 2, with $f(y) \equiv y^i$, $g \equiv b$, and writing $D_{kl}(i)$ to abbreviate $D_{kl}(f, b)(x)$, we have, for $i > 0$ and $m > 0$,

$$(b^i)^{(m)} = D_{m-1,0}(i) = ib^{i-1}b^{(m)} + \sum_{j=0}^{m-2} \binom{m-1}{j} D_{m-1-j,1}(i)b^{(j+1)}. \tag{6}$$

The result now follows by substituting for $(b^i)^{(m)}$ from (6) and separating out the terms $b^{(m)}$ in (5), and then using $(b^i)^{(k)} = D_{k-1,0}(i)$.

A.3. Derivatives of $a_{sk;ij}$

These are given by direct differentiation of the cofactor matrix $\boldsymbol{C}^\top$ of $\boldsymbol{M}$ in the following proposition.

Proposition 5. *Let* $A_s(\alpha, \beta) = \boldsymbol{C}^\top / \prod_{1 \le t \ne s \le r}(\beta - b_t)^{m_t}$ *for* $1 \le s \le r$. *Then* $a_{sk;ij}^{(m)}$ *is the* (i, j)*th element of the matrix*

$$\frac{1}{(m_s - k)!} \boldsymbol{D}_{m-1,\vec{0}}\left(\left[\frac{\partial^{m_s-k} \boldsymbol{A}_s}{\partial \beta^{m_s-k}}\right]_{\beta=b_s}, I, b_1, \ldots, b_r\right)(0),$$

where I is the identity function, $\vec{0}$ is the zero vector of length r, and the function $\boldsymbol{D}$ is applied elementwise to the matrix $\boldsymbol{A}_s$.

Proof. The defining expression,

$$a_{sk;ij} = \frac{1}{(m_s - k)!}\left[\frac{\mathrm{d}^{m_s-k}}{\mathrm{d}\beta^{m_s-k}}\left(\frac{c_{ji}}{\Delta_s}\right)\right]_{\beta=b_s},$$

is differentiated m times with respect to α at $\alpha = 0$ using Proposition 4.

Note that the $m_s - k$ times partial differentiation of A_s with respect to β is straightforward since the terms V_{ij}^* are independent of β. A generalised version of Leibniz's rule can be used if symbolic differentiation is not available.

The situation does not arise when the roots of the equation $\Delta = 0$ are distinct, whereupon $a_{s1;ij}^{(m)}$ simplifies to the (i, j)th element of the matrix:

$$\boldsymbol{D}_{m-1,\vec{0}}(A_s, I, b_1, \ldots, b_r)(0).$$

A.4. Derivatives of G_{ij}

Proposition 6. *For* $i \in \mathcal{F}$ *and* $j \in \mathcal{E}$,

$$G_{ij}^{(m)}(\alpha, \phi(\alpha)) = \sum_{k \in \mathcal{F}} p_{ik} H_{ik}^*(\alpha + \phi r_i) G_{kj}^{(m)}(\alpha, \phi) + p_{ij} H_{ij}^{*(m)}(\alpha + \phi r_i)$$

$$+ \sum_{k \in \mathcal{F}} p_{ik} \sum_{l=1}^{m} \binom{m}{l} H_{ik}^{*(l)}(\alpha + \phi r_i) G_{kj}^{(m-l)}(\alpha, \phi).$$

Proof. Differentiating (1) with respect to α, again using Leibniz's rule, we obtain

$$G_{ij}^{(m)}(\alpha, \phi(\alpha)) = p_{ij} H_{ij}^{*(m)}(\alpha + \phi r_i) + \sum_{k \in \mathcal{F}} p_{ik} \sum_{l=0}^{m} \binom{m}{l} H_{ik}^{*(l)}(\alpha + \phi r_i) G_{kj}^{(m-l)}(\alpha, \phi).$$

For $m = 1$, this is a collection of $n_e \times n_f$ linear equations in the same number of variables $G_{ij}^{(1)}(\alpha, \phi(\alpha))$, functions of α only, which can be solved numerically at any given α: we will only need $\alpha = 0$ to compute moments. For $m = 2, 3, \ldots$, the same applies to the variables $G_{ij}^{(m)}$ since the lower derivatives $G_{ij}^{(n)}$ for $n < m$ are known from previous steps. However, we do need to compute $H_{ij}^{*(l)}(\alpha + \phi r_i)$ for $l = 1, 2, \ldots$. The functions H_{ij}^* are given, so all we need to do is apply Proposition 2, with $f \equiv H_{ij}^*$, $g(\alpha) \equiv \alpha + r_i \phi(\alpha)$ for the specified ϕ, giving $H_{ij}^{*(l)}(\alpha + \phi r_i) = D_{l-1,0}$.

A.5. Moments of V_{ij}

For the case of distinct roots b_s, Theorem 1 gives $V_{ij}^*(\alpha) = \sum_{k \in \mathcal{E}} \sum_{s=1}^{n_e} a_{s1;kj} G_{ik}(\alpha, -b_s)$. Leibnitz's rule then immediately gives the following result.

Theorem 2. *The mth moment of a busy period that starts in state $i \in \mathcal{F}$ and ends in state $j \in \mathcal{E}$ is*

$$V_{ij}^{*(m)}(0) = (-1)^m \sum_{k \in \mathcal{E}} \sum_{s=1}^{n_e} \sum_{l=0}^{m} \binom{m}{l} a_{s1;kj}^{(m-l)} G_{ik}^{(l)}(0, -b_s(0)).$$

All the derivatives in the summations on the right-hand side can be computed using the results of the previous sections. A MATHEMATICA implementation of these results has been used to solve the numerical examples given earlier.

Acknowledgement

References

[1] ABATE, J. AND VALKÓ, P. (2004). Multi-precision Laplace transform inversion. *Internat. J. Numer. Meth. Eng.* **60,** 979–993.

[2] BARBOT, N. AND SERICOLA, B. (2003). Exact stationary solution to tandem fluid queues. *Internat. J. Simul. Systems Sci. Tech.* **4,** 12–20.

[3] BARBOT, N., SERICOLA, B. AND TELEK, M. (2001). Distribution of busy period in stochastic fluid models. *Stoch. Models* **17,** 407–427.

[4] BOXMA, O. J. AND DUMAS, V. (1998). The busy period in the fluid queue. In *ACM SIGMETRICS Performance Evaluation Review*, ACM, New York, pp. 100–110.

[5] FIELD, A. J. AND HARRISON, P. G. (2007). An approximate compositional approach to the analysis of fluid queue networks. *Performance Evaluation* **64,** 1137–1152.

[6] GUILLEMAN, F. AND SERICOLA, B. (2007). Stationary analysis of a fluid queue driven by some countable state space Markov chain. *Methodology Comput. Appl. Prob.* **9,** 521–540.

[7] GUPTA, V. AND HARRISON, P. G. (2008). Fluid level in a reservoir with on-off source. In *Proc. 10th Workshop Math. Performance Modeling Analysis* (MAMA, 2008), ACM, pp. 128–130.

[8] HARRISON, P. G., PATEL, N. M. AND ZERTAL, S. (2010). Response time distribution of flash memory accesses. *Performance Evaluation* **67,** 248–259.

[9] MANDJES, M., MITRA, D. AND SCHEINHART, W. (2003). Simple models of network access, with applications to the design of joint rate and admission control. *Comput. Networks* **41,** 489–504.

[10] MITRA, D. (1988). Stochastic theory of a fluid model of producers and consumers coupled by a buffer. *Adv. Appl. Prob.* **20,** 646–676.

[11] RABEHASAINA, L. (2006). Moments of a Markov-modulated, irreducible network of fluid queues. *J. Appl. Prob.* **43,** 510–522.

[12] SERICOLA, B. AND TUFFIN, B. (1999). A fluid queue driven by a Markovian queue. *Queueing Systems* **31,** 253–264.

[13] VAN FOREEST, N., MANDJES, M. AND SCHEINHARDT, W. (2003). Analysis of a feedback fluid model for heterogeneous TCP sources. *Stoch. Models* **19,** 299–324.

[14] VIRTAMO, J. AND NORROS, I. (1994). Fluid queue driven by an M/M/1 queue. *Queueing Systems* **16,** 373–386.