Quasi-stationary analysis for queues with temporary overload

S.K. Cheung*†, R.J. Boucherie† and R. Núñez-Queija‡§
*All Options, Herengracht 433, P.O. Box 11096, 1001 GB Amsterdam, The Netherlands
Email: sing.cheung@alloptions.nl
† Stochastic Operations Research group, Faculty of Electrical Engineering, Mathematics and Computer Science,
University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands
Email: r.j.boucherie@utwente.nl
‡Operations Research, Faculty of Economics and Business, University of Amsterdam, The Netherlands
Email: nunezqueija@uva.nl
§CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

***Abstract*—Motivated by the high variation in transmission rates for document transfer in the Internet and file down loads from web servers, we study the buffer content in a queue with a fluctuating service rate. The fluctuations are assumed to be driven by an independent stochastic process. We allow the queue to be overloaded in some of the server states. In all but a few special cases, either exact analysis is not tractable, or the dependence of system performance in terms of input parameters (such as the traffic load) is hidden in complex or implicit characterizations. Various asymptotic regimes have been considered to develop insightful approximations. In particular, the so-called quasi-stationary approximation has proven extremely useful under the assumption of uniform stability. We refine the quasi-stationary analysis to allow for temporary instability, by studying the "effective system load" which captures the effect of accumulated work during periods in which the queue is unstable.**

Keywords: Effective load, fluctuating rates, Markovian random environment, excess load, recovery time, quasi-stationary analysis, fluid queue

I. Introduction

Document transmissions in the Internet and file down loads from web servers commonly experience high variation in transmission rates, due to concurrence of other data traffic flows [11]. In particular, in the context of TCP-driven traffic flows, which are responsive to temporary network congestion, the transmission time is highly affected by the presence of other applications (e.g., voice, video, streaming data) that rely on unresponsive transport protocols such as UDP. In the queueing theory community, it has long been recognized that service unreliability has a decisive effect on the perceived performance [17]; this is known as Ross' conjecture. For a model similar to ours, with exponential durations of high and low service rates, [10] confirm this conjecture.

There is a rich variety of models in which the available service rate alternates between a positive value and complete absence of service, including unreliable servers, server vacations and service failures [7], [19]. These models allow for quite explicit and closed form solutions for many performance measures or structural decomposition results [9]. The situation changes completely when the service rate can vary between several positive values. For the class of Markovian queueing models with a G/M/1 structure, which gives rise to matrix-geometric stationary measures, there are efficient solutions to numerically determine these measures [13].

Various approaches have been developed to capture the essential dependence of the system performance in terms of parameters such as arrival rates, service rates, etc. One very successful line of research was the analysis through time-scale decomposition. In short, this approach consists of studying the system performance in two limiting regimes. One extreme, coined the fluid regime [5], in which the dynamics of the modulating environment is sped up to infinity, which in case of independent modulating processes, is equivalent to replacing the server by one working at constant speed equal to the original average speed. This approach in general tends to be much too optimistic and the thus obtained performance may not be approached even by far in the system with stochastic variations. The other extreme, the "quasi-stationary" regime, is obtained by assuming that capacity fluctuations are infinitely slow compared to traffic dynamics. This approach tends to be much too pessimistic and does not serve as a useful approximation in general either. A further complication is that the quasi-stationary limit has no sensible meaning if the service rates are below the arrival rate for some states of the environment.

Our work is strongly motivated by [11], who point out that in practice uniform stability (i.e., assuming that the service rate is larger than the arrival rate at all times) is not realistic. The authors conclude that no sensible stationary analysis can be done for such systems and focus on time dependent performance of the system, using a time-acceleration technique [14] similar to the time-decomposition mentioned above.

Another line of related literature concerns the investigation of the time needed to recover from a temporary overload situation [6], [16], [12]. These works focus on transient, rather than stationary, analysis. Our approach is also related to pointwise stationary fluid models, see for example [3]. The focus of our work is on the influence of the random environment on an otherwise elementary queueing system (a single server queue), whereas [3] study a more complex network scenario without random environment.

We study the buffer content in a queue with a fluctuating service rate that depends on the state of an exogenous stochastic process. This process can, for instance, model the number of transfers of unresponsive data flows in the Internet. For some states of the random environment the arrival rate of the queue may be larger than the service rate. If these overload periods are relatively long – compared with the time scale of the arrival process – performance can be very poor, manifesting itself in typically large queues and long delays, even if the load is far below the *average service capacity*. From a practical perspective, however, such a system can be thought of as being nearly unstable. With this in mind, we aim at determining the "effective" stability of the system, which incorporates the adversarial effect of slow service rate fluctuations on performance. We do so by complementing the quasi-stationary limit with a fluid queue [18] – not to be confused with the fluid limit – to capture the effect of accumulated work during periods in which the queue is unstable. Our results rely on a detailed analysis of the recovery time, i.e., the time needed to recover from the excess load after a low rate period that may include multiple stable periods.

The remainder of the paper is organized as follows. We first describe various related models in Section 2 and subsequently discuss the notion of effective load in Section 3. The quasi stationary limit with temporary instability is the subject of Section 4. We conclude in Section 5.

II. Models

We consider a queue with Poisson arrivals at rate λ and exponentially distributed service requirements with mean 1. The service rate process $\{\mu(t),\ t \geq 0\}$ fluctuates over time according to a stochastic process that is assumed to be independent of all inter-arrival times and service requirements. We will consider two cases: a Markov modulated service rate process and a high-low service process with generally distributed times of high and low service rates. For all realizations of the service rate process the sample paths are continuous and differentiable almost everywhere, except on a countable set of isolated points with measure 0.

For the *Markov modulated queue*, the service rate process $\{\mu(t),\ t \geq 0\}$ is modulated by an independent irreducible Markovian background process $\{M(t),\ t \geq 0\}$ with state space $\mathcal{M} = \{0, 1, ..., m\}$, for some $m \in \mathbb{N}$, and equilibrium distribution $\pi_i,\ i \in \mathcal{M}$. When the background process is in state i (i.e., if $M(t) = i$) the service rate at time t is $\mu(t) = \mu_i,\ i \in \mathcal{M}$. The states $i \in \mathcal{M}$ for which $r_i := \lambda - \mu_i > 0$ are called the low service rate states for which the instantaneous load λ/μ_i exceeds 1, and the queue length has a positive drift. When $r_i < 0$, the states $i \in \mathcal{M}$ are called high service rate states, and the queue length has a negative drift. It is convenient and not very restrictive to assume $r_i \neq 0$. (Extension to the case with $r_i = 0$ for one or more states i requires additional notational burden and minor technical details.) The usual stability condition *for the queue* is $\sum_{i=0}^{m} \pi_i r_i < 0$, which can be interpreted as the mean drift of the queue being negative (e.g., see [18]).

A particularly convenient special case is obtained when there is some $k \geq 0$ so that $r_i \equiv a$ for all $i \leq k$ and $r_i \equiv b$ for all $i > k$. For carefully selected rates of $\{M(t),\ t \geq 0\}$ we then have a phase-type distribution for the high and low service rate periods. This setting will allow for more explicit analysis. A further special case is the on-off model for which $b = 0$.

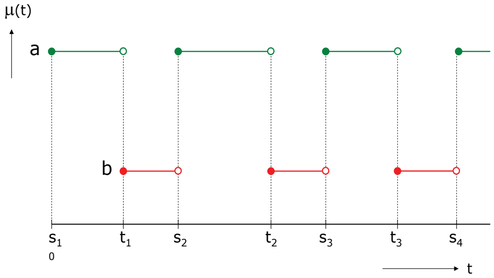

Fig. 1. Service rates in the alternating high-low model

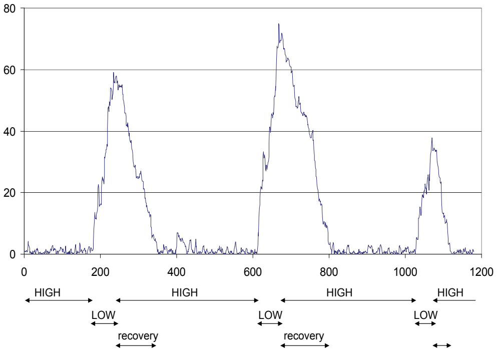

Fig. 2. A typical sample path of the queue-length process in the alternating high-low model

The high-low model will be used throughout to illustrate our results. In that case we focus on a slightly different model that allows for generally distributed periods of high and low service rates. For the *high-low model* the service rate process alternates between a high and a low value, see Figure 1. For a given realization of the service rate process, we let $\{s_i, t_i\}_{i \in \mathbb{N}}$ be the sequence of time points where the service rate switches from the low service rate to the high service rate for the i^{th} time (time s_i) and the first epoch thereafter that it switches back (time t_i). We assume that $0 = s_1 < t_1 < s_2 < t_2 < s_3 < t_3 < \cdots$, and let the time-dependent service rate be given by $\mu(t) = a$ if $t \in [s_i, t_i)$, and $\mu(t) = b$ if $t \in [t_i, s_{i+1})$, for some $i \in \mathbb{N}$ and assume that $a > \lambda > b \geq 0$. Let $A_i = t_i - s_i$ be the length of the i-th interval in which the server works at the *higher* rate a; and $B_i = s_{i+1} - t_i$ the i-th interval in which the server works at the *lower* rate b. We assume that the sequences $\{A_i\}_i$ and $\{B_i\}_i$ are two i.i.d. sequences, independent of each other. Note that this last independence assumption need not be satisfied by the above mentioned Markovian high-low model.

The usual (long-term) stability condition reads

$$\lambda < \frac{\mathbb{E}A}{\mathbb{E}A + \mathbb{E}B} a + \frac{\mathbb{E}B}{\mathbb{E}A + \mathbb{E}B} b. \tag{1}$$

We will be particularly interested in the case where λ is large, such that a very large number of arrivals occur during the typical durations of high rate and low rate periods. In Figure 2 we depict a typical realization of the queue length process for the high-low model. The service rate starts off in the higher value a and the process shows stationary behavior. As soon as the service rate switches to the lower value b the queue starts building up. The instantaneous load $\rho(t) = \lambda/\mu(t)$ then exceeds the value 1, i.e., the queue is temporarily unstable. The major trend is characterized by the linear drift $\lambda - b$, but due to the randomness in the arrival and service processes (both Poissonian) there are fluctuations around the linear trend. The top of the curve corresponds to a time instant at which the service process switches back to the high rate. With the linear trend being negative ($\lambda - a < 0$), it takes a while for the process to reach the level of the typical stationary behavior under the high service rate. Roughly speaking, this *recovery* period lasts until the linear trend hits the horizontal axis.

The main message of Figure 2 is that there are three types of periods during which the queueing dynamics are intrinsically different: (i) instability periods (when the service rate is low and the queue builds up), (ii) recovery periods (when the service rate is high, but the queue has not yet recovered from an instability period), and (iii) quasi-stationarity (the queue behaves as if the service rate is always high). These periods will be characterized via their “effective” load. It is crucial to note that some high rate periods may be too short to recover from instability, i.e., a recovery period may be interrupted by one or more instability periods.

III. Effective load

The effective load at time t captures the ability of the queue to drain the workload built up until time t, and is defined as (see [11]):

$$\rho^*(t) \equiv \sup_{0\le s<t} \frac{\int_s^t \lambda(r)dr}{\int_s^t \mu(r)dr} = \sup_{0\le s<t} \frac{(t-s)\cdot\lambda}{\int_s^t \mu(r)dr}. \qquad (2)$$

The effective load will be the basis in determining whether at a given time the queue can be characterized via the quasi-stationary limit. Note that, since the service rates $\mu(t)$ constitute a random process, the effective load itself is a random process. As we will see later, the distribution of $\rho^*(t)$ can be obtained from that of the workload in the associated Markov modulated fluid queue with constant fluid arrival rate λ and drain rate $\mu(t)$. We will say that the queue is *effectively unstable* at time t when $\rho^*(t) \ge 1$, and *effectively stable* at time t when $\rho^*(t) < 1$.

As an illustration we have depicted the effective load in Figure 3, for the alternating high-low model with high and low periods of deterministic length 1, and with $\lambda = \frac{3}{2}$, $a = 4$, $b = \frac{4}{5}$. The instantaneous load $\rho(t)$ is 0.375 during high-rate periods, and 1.875 during low-rate periods.

A. Markov modulated queue

For the Markov modulated queue the distribution of ρ^* can be obtained via the relation with a fluid queue as follows.

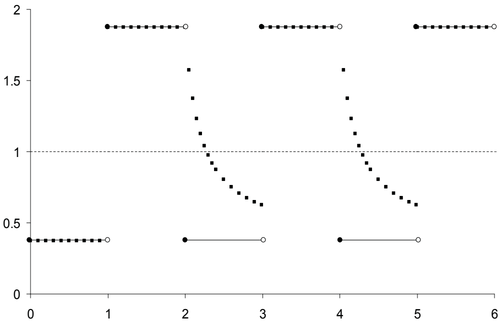

Fig. 3. Example of the effective load function $\rho^*(t)$ (marked with squares) and the instantaneous load $\rho(t)$ (solid step function).

Proposition 1: For the Markov modulated queue, for all $x > 0$, $t \ge 0$

$$P(\rho^*(t) > x) = P(W_x(t) > 0),$$

where $W_x(t)$ is the fluid content at time t in the associated Markov modulated fluid queue with arrival rate λ and service rate $x\,\mu(t)$, that is the solution of

$$\frac{dW_x(t)}{dt} = \begin{cases} 0 & \text{if } W_x(t) = 0, \quad \lambda < x\mu_{M(t)} \\ \lambda - x\mu_{M(t)}, & \text{otherwise.} \end{cases}$$

Proof From the definition of $\rho^*(t)$ in (2) we observe that the following are equivalent:

$$\begin{aligned} \rho^*(t) > x \quad &\Leftrightarrow \quad \exists s \in [0,t) : \int_s^t \lambda(r)dr - x\int_s^t \mu(r)dr > 0 \\ &\Leftrightarrow \quad \sup_{0\le s<t} \left\{ \int_s^t \lambda(r)dr - x\int_s^t \mu(r)dr \right\} > 0 \end{aligned}$$

for $x \in \mathbb{R}_+$. The supremum can be interpreted as a workload process, e.g., see [2]. In fact, $W_x(t) = \sup_{0\le s<t}\left\{\int_s^t \lambda(r)dr - x\int_s^t \mu(r)dr\right\}$ is the fluid content process at time t in the associated Markov modulated fluid queue [1], [18], where we replace the Poisson arrivals and the service times in the queue by fluid streams of rate λ (constant) and $x\,\mu(t)$, respectively. More precisely, the content $W_x(t)$ of the fluid queue (note that this process depends on x) is regulated by the background process $M(t) \in \mathcal{M}$ as specified by the differential equation for $W_x(t)$. □

Note that the fluid queue used in the proof and the original queue share exactly the same realization of the service rate process. The fluid queue, however, does not incorporate the random fluctuations in the arrival and service processes. The stability condition for the fluid queue is

$$\sum_{i=0}^{m} \pi_i(\lambda - x\mu_i) < 0, \qquad (3)$$

which is the same as that for the original queue when $x = 1$. If (3) is satisfied, the stationary distribution of the fluid queue exists and can be determined through spectral analysis, see [18].

As a special case, allowing for closed-form expressions, we consider the Markovian birth death high-low system, in which the modulating Markov process $\{M(t),\ t \ge 0\}$ is a birth-death process with constant birth rates α and constant death rates $\beta > \alpha$, and service rates $\mu_0 = b$ and $\mu_i = a > b$ for all $i \geq 1$. Note that the low-rate periods are exponentially distributed, but the high-rate periods can be fitted to a distribution with given first two moments. (High-rate periods are distributed as the busy period in an M/M/1 queue with arrival rate α and service rate β.) Also, the lengths of high-rate and low-rate periods are mutually independent. (In this model the random environment may represent a higher-priority queue that takes away a fixed amount of capacity $a - b$ when it is not empty.)

Scheinhardt [18, pp. 26–28] shows that the stationary fluid content process W_x is given by

$$\mathbb{P}(W_x > y) = p_{0;x} \cdot \exp\left\{ -\left(\frac{\alpha}{\lambda - bx} - \frac{\beta}{(a-b)x} \right) y \right\},$$

for any $y \geq 0$, where

$$p_{0;x} = \frac{1 - \alpha/\beta}{(ax - \lambda)/((a-b)x)}.$$

For $y = 0$ we obtain the stationary distribution of the effective load ρ^* as:

$$\mathbb{P}(\rho^* > x) = \mathbb{P}(W_x > 0) = p_{0;x},$$

provided that $\frac{ax-\lambda}{(a-b)x} < \frac{\alpha}{\beta} < 1$, cf. [18].

In general, the distribution of the effective load can not be obtained in closed form. Still, the effective load can be expressed explicitly in terms of the service rate process for the more general high-low model with general high-rate and low-rate periods. This is the subject of the next subsection.

B. High-low model

In this section, we fix the sequence $\{s_i, t_i\}_{i \in \mathbb{N}}$ that determines the high-rate and low-rate periods.

Proposition 2: During the i^{th} high-rate period, that is for $t \in [s_i, t_i)$, $i \geq 2$, we have

$$\rho^*(t) = \sup_{1 \leq j \leq i-1} \frac{(t - t_j)\lambda}{\int_{t_j}^t \mu(r)dr}. \tag{4}$$

During any low-rate period, $t \in [t_i, s_{i+1})$, for $i \geq 1$ we have $\rho^*(t) = \frac{\lambda}{b}$.

Proof For $t \in [s_i, t_i)$, $i \geq 2$, the supremum in the effective load function (2) can be split into suprema of a partition over $s \in [s_j, t_j)$, $s \in [t_j, s_{j+1})$, $j = 1, \dots, i-1$, and $[s_i, t)$. Note that

$$\sup_{s \in [s_j, t_j)} \frac{(t-s)\lambda}{\int_s^t \mu(r)dr} = \frac{(t-t_j)\lambda}{\int_{t_j}^t \mu(r)dr}, \tag{5}$$

$$\sup_{s \in [t_j, s_{j+1})} \frac{(t-s)\lambda}{\int_s^t \mu(r)dr} = \frac{(t-t_j)\lambda}{\int_{t_j}^t \mu(r)dr}, \tag{6}$$

since $\frac{(t-s)\lambda}{\int_s^t \mu(r)dr}$ is strictly increasing on $s \in [s_j, t_j)$ and strictly decreasing on $s \in [t_j, s_{j+1})$. Observe further that the expressions (5) and (6) are identical and lie in the interval $[\frac{\lambda}{a}, \frac{\lambda}{b}]$, since $b \leq \mu(t) \leq a$ for all $t > 0$. The supremum over $[s_i, t)$ equals $= \frac{\lambda}{a}$, for $t \in [s_i, t_i)$, which leads to (4). For low-rate periods we have $\rho^*(t) = \frac{\lambda}{b}$, due to $\mu(r) \geq b$ for all $r > 0$. □

Remark 1: If $a > b > 0$, then $\rho^*(t)$ is continuous and finite around $t = s_i$ (i.e., at the beginning of a high-rate period), but $\rho^*(t)$ has a jump at $t = t_i$ (i.e., at the beginning of a low-rate period).

The effective load is depicted in Figure 3. The effective load and the instantaneous load coincide during low service rate periods. The effective load at time t is strictly decreasing in t during high-rate periods (starting from the value λ/b at the beginning of a high-rate period). If the high-rate period is sufficiently long (relative to λ, a, and b), then the effective load drops below the value 1. The recovery time is the time needed (since the end of the last low-rate period) for the effective load to drop to 1. Heuristically speaking, we can say that the queue "becomes stable" at the time epoch u such that $\rho^*(u) = 1$.

The supremum in equation (4) is achieved for a certain index j^*, with $j^* \leq i - 1$. In general, if the high-rate periods are "sufficiently long", then the supremum is achieved for $j^* = i - 1$. In contrast, if the high-rate periods are too short, the supremum is achieved at a lower index $j^* < i - 1$. A characterization of "how long" a high-rate period should be, will be discussed next.

IV. The quasi-stationary limit for the high-low model

In this section we analyze instability during high-rate periods. To illustrate our goals, we first consider the on-off model with exponentially distributed high rate periods. For this model we use closed-form expressions that are available for the queue-length distribution. Second, we consider the high-low model with generally distributed high and low rate periods. Finally, we specialize those results for the high-low model with exponentially distributed high and low rate periods.

The discussion will center around a characterization of the recovery period. We will think of the existence of these recovery periods as a refinement of the usual definition of stability. Particular attention will be given to the case with exponential high-rate and low-rate periods, in which case closed-form results can readily be obtained. Ultimately, we will discuss the scaled version of the queue length in the quasi-stationary regime.

A. The on-off model

In this section we study the buffer content in a high-low queue when no service is available for some time periods (off-periods). We refine the analysis of [15] which considers the processor-sharing queue with service interruptions. In particular, based on the explicit formulas from [15] we show that the conditional queue-length distribution (given that the server is turned on) is defective in the quasi-stationary limit.

Assume that the on-periods A_i, $i \geq 1$, are iid exponentially distributed with mean α^{-1}, and the service rate during on-periods is a. The off-periods B_i, $i \geq 1$, are i.i.d. generally distributed as the random variable B with distribution function $B(t) := \mathbb{P}(B \leq t)$, $t \geq 0$, k-th moment β_k, and Laplace-Stieltjes transform $\tilde{B}(s) := \mathbb{E}e^{-sB}$, for $\operatorname{Re} s \geq 0$.

To investigate stability, we consider the fluid regime. To this end, we apply the Uniform Acceleration technique [14]. The arrival and service rates are scaled linearly with a common parameter $\eta > 0$, i.e., λ is replaced with $\eta\lambda$ and $\mu(t)$ is replaced with $\eta\mu(t)$. The scaled queue-length process is denoted by $Q^\eta(t)$, for all $\eta > 0$. Let (Q^η, μ) denote the limiting distribution of $(Q^\eta(t), \mu(t))$. From [15] we obtain the following result.

Proposition 3: The joint distribution of (Q^η, μ) has the following conditional probability generating functions

$$\begin{aligned} \mathbb{E}\left[z^{Q^\eta}|\mu = a\right] &= \frac{a - \lambda(1+\alpha\beta_1)}{a - \lambda\left(1 + \alpha\beta_1 \cdot \varphi_B(z, \eta\lambda)\right) z}, \quad (7)\\ \mathbb{E}\left[z^{Q^\eta}|\mu = 0\right] &= \varphi_B(z,\eta\lambda) \cdot \mathbb{E}\left[z^{Q^\eta}|\mu = a\right], \quad (8)\end{aligned}$$

where

$$\varphi_B(z, \eta\lambda) := \frac{1 - \widetilde{B}(\eta\lambda(1-z))}{\beta_1 \eta\lambda\,(1-z)}$$

is the pgf of the number of arrivals that occur according to a Poisson process with rate $\eta\lambda$, during the backward recurrence time of an off-period. Furthermore

$$\begin{aligned} \mathbb{E}\left[Q^\eta|\mu = a\right] &= \frac{\lambda}{p_{ON} \cdot a - \lambda} + \frac{\alpha\beta_2}{2}\frac{p_{ON} \cdot \lambda^2}{p_{ON} \cdot a - \lambda}\eta, \quad (9)\\ \mathbb{E}\left[Q^\eta|\mu = 0\right] &= \mathbb{E}\left[Q^\eta|\mu = a\right] + \eta\lambda\frac{\beta_2}{2\beta_1}, \quad (10)\end{aligned}$$

where $p_{ON} = \frac{1}{1+\alpha\beta_1}$ is the long-run fraction of time that the server is available.

Observe that the conditional mean queue-length (9) is linear in the scaling parameter η and thus tends to infinity when the scaling parameter η tends to infinity. Naturally, in the quasi-stationary limit the mean queue-length during off-periods is infinite even when the usual stability criterion (1) is satisfied. (The conditional mean queue length distribution in the quasi-stationary limit is defective.)

Proposition 4:

$$\lim_{\eta\to\infty} \mathbb{P}\left(Q^\eta = \infty|\mu = a\right) = \frac{\lambda}{a - \lambda}\alpha\beta_1. \quad (11)$$

Proof Follows directly from the fact that $\lim_{\eta\to\infty} \varphi_B(z, \eta\lambda) = 0$, so that

$$\lim_{\eta\to\infty} \mathbb{E}\left[z^{Q^\eta}|\mu = a\right] = \frac{a - \lambda(1+\alpha\beta_1)}{a - \lambda z}. \quad (12)$$

□

We can rewrite (12) as

$$\frac{\lambda\alpha\beta_1}{a - \lambda} \times 0 + \frac{a - \lambda(1+\alpha\beta_1)}{a - \lambda} \times \frac{a - \lambda}{a - \lambda z}, \quad (13)$$

which can be interpreted as follows. With probability $\frac{\lambda\alpha\beta_1}{a-\lambda}$ the queue length is infinite in the quasi-stationary limit. With the complementary distribution, the queue length is distributed as if the service rate is always a (i.e., as the queue length in the M/M/1 with load λ/a).

Let us now consider the queue length during on-periods after recovery to stability. In order to refine the quasi-stationary limit, we scale the queue length. From the linearity of the mean queue length in η we see that the proper scaling is $Q^\eta(t)/\eta$. We then have the following result.

Proposition 5: The conditional distribution of the scaled queue length $(\frac{1}{\eta}Q^\eta \mid \mu = a)$ in the quasi-stationary limiting regime is given by

$$\begin{aligned} \lim_{\eta\to\infty} \mathbb{E}\left[z^{\frac{1}{\eta}Q^\eta}|\mu = a\right] &= \frac{a - \lambda\,(1+\alpha\beta_1)}{a - \lambda\left(1 - \alpha\frac{1-\widetilde{B}(-\lambda\ln z)}{\lambda\ln z}\right)} \quad (14)\\ \lim_{\eta\to\infty} \mathbb{E}\left[z^{\frac{1}{\eta}Q^\eta}|\mu = 0\right] &= \frac{1 - \widetilde{B}\,(-\lambda\ln z)}{-\lambda\beta_1\ln z} \quad (15)\\ &\quad \times \frac{a - \lambda\,(1+\alpha\beta_1)}{a - \lambda\left(1 - \alpha\frac{1-\widetilde{B}(-\lambda\ln z)}{\lambda\ln z}\right)}.\end{aligned}$$

Furthermore

$$\begin{aligned} \lim_{\eta\to\infty} \mathbb{E}\left[\frac{1}{\eta}Q^\eta|\mu = a\right] &= \frac{\alpha\beta_2}{2}\frac{p_{ON} \cdot \lambda^2}{p_{ON} \cdot a - \lambda} \quad (16)\\ \lim_{\eta\to\infty} \mathbb{E}\left[\frac{1}{\eta}Q^\eta|\mu = 0\right] &= \frac{\alpha\beta_2}{2}\frac{p_{ON} \cdot \lambda^2}{p_{ON} \cdot a - \lambda} + \lambda\frac{\beta_2}{2\beta_1}. \quad (17)\end{aligned}$$

Proof Follows from (7) and the fact that

$$\lim_{\eta\to\infty} \varphi_B\left(z^{1/\eta}, \eta\lambda\right) = \frac{1 - \widetilde{B}\,(-\lambda\ln z)}{-\lambda\beta_1\ln z}.$$

□

This result can be interpreted as follows. From (13) we know that, in the limit, the non-scaled queue length during on-periods is non-defective with probability $\frac{a-\lambda(1+\alpha\beta_1)}{a-\lambda}$. Therefore, with that probability the scaled queue length during on-periods equals 0. With the complementary probability $\frac{\lambda\alpha\beta_1}{a-\lambda}$, the queue length "did not recover from instability" during an on-period. We therefore decompose (16) as

$$\frac{a - \lambda(1+\alpha\beta_1)}{a - \lambda} \times 0 + \frac{\lambda\alpha\beta_1}{a - \lambda} \times \frac{\beta_2}{2\beta_1}\frac{p_{ON} \cdot \lambda(a - \lambda)}{p_{ON} \cdot a - \lambda}. \quad (18)$$

Heuristically, we may say

$$\begin{aligned} &\lim_{\eta\to\infty} \mathbb{E}\left[\frac{1}{\eta}Q^\eta|\mu = a \text{ but not yet recovered}\right] \quad (19)\\ &= \frac{\beta_2}{2\beta_1}\frac{p_{ON} \cdot \lambda(a - \lambda)}{p_{ON} \cdot a - \lambda}. \quad (20)\end{aligned}$$

This decomposition of the queue length during on-periods can be done similarly for the entire distribution using the expressions for the conditional pgfs.

Remark 2: The above explains why constant-rate approximations for on-periods (high-rate periods) give poor results. The error can be made arbitrarily large by either increasing the second moment β_2 of the off-periods or the scaling parameter η.

Remark 3 (Discussion): The previous observations lead to a notion of adjusted stability as a refinement of the usual stability criterion (1). The fact that $(Q^\eta \mid \mu = a)$ is defective in the quasi-stationary limit is explained by the fact that Q^η explodes during an off-period when $\eta \to \infty$. Since the scaled system Q^η is stable in the long run, the system recovers from the explosion during an on period. The queue becomes stable again (i.e., Q^η becomes finite) during an on-period if the on-period length is "sufficiently long". If the on-period length is not sufficiently long, then, in the quasi-stationary regime, Q^η remains infinite during the on-period.

B. The high-low model

In this section we further investigate the "recovery time" and "adjusted stability" in the high-low model.

1) Recovery time: Suppose at the start of i-th high-rate period (at time s_i) for some $k \in \{1, \ldots, i-1\}$ we have $\rho^*(t_k^-) < 1$ and $\rho^*(u) \geq 1$ for all $u \in [t_k, s_i)$, i.e., the time t_k is the most recent time where the effective load increased beyond 1. Note that t_k is always the start of a low-rate period. Define the accumulated low-rate and high-rate period lengths during the interval $[t_k, s_i)$ as

$$T_{\text{low}}(t_k, s_i) = \sum_{n=k}^{i-1} B_n, \quad T_{\text{high}}(t_k, s_i) = \sum_{n=k}^{i-1} A_{n+1},$$

with $T_{\text{high}}(t_k, s_i) + T_{\text{low}}(t_k, s_i) = s_i - t_k$. Define the *recovery time* $R(t_k, s_i)$ as the time needed (after time s_i) to reduce the effective load below 1.

Remark 4: In the associated fluid queue, the period $T_{\text{high}}(t_k, s_i)$ is not long enough to remove the backlog accumulated in the period $T_{\text{low}}(t_k, s_i)$, and $R(t_k, s_i)$ is the time to drain the queue starting at s_i.

We now investigate under which conditions the system becomes effectively stable during the i-th high-period A_i.

Proposition 6: Let the queue be effectively unstable during the period $[t_k, s_i)$, $k \leq i-1$. The queue becomes effectively stable during the i-th high-period, if and only if

$$\frac{\lambda - b}{a - \lambda} \sum_{j=k}^{i-1} B_j < \sum_{j=k}^{i-1} A_{j+1}. \tag{21}$$

Proof If $R(t_k, s_i) \geq A_i$, the queue does not become effectively stable. If $R(t_k, s_i) < A_i$, then the effective load drops below 1 during the i-th high-period, and it must be that

$$\begin{aligned} &\lambda \left[T_{\text{high}}(t_k, s_i) + T_{\text{low}}(t_k, s_i)\right] + \lambda R(t_k, s_i) \\ &= \left[a \cdot T_{\text{high}}(t_k, s_i) + b \cdot T_{\text{low}}(t_k, s_i)\right] + a \cdot R(t_k, s_i), \end{aligned}$$

so that

$$R(t_k, s_i) = \frac{\lambda - b}{a - \lambda} T_{\text{low}}(t_k, s_i) - T_{\text{high}}(t_k, s_i) \geq 0.$$

□

Remark 5: Note that the term $(\lambda - b)$ in (21) is the growth rate of the fluid queue during low-rate periods, and $(a - \lambda)$ is the (potential) decrease rate during high-rate periods.

The distribution of *the number of high-rate periods needed for recovery*, N, is obtained as follows. Without loss of generality, let $k = 1$, i.e., t_1 is the most recent moment when the system became effectively unstable. If $\{N = n\}$, for $n \geq 1$, then each of the first $n - 1$ high-periods are not long enough to stabilize the queue. As a consequence, N is the *first ladder epoch* in the random walk

$$S_0 = 0, \quad S_n = \sum_{i=1}^{n} V_i, \quad n = 1, 2, \ldots, \tag{22}$$

with $V_i = A_{i+1} - cB_i$ and $c = \frac{\lambda - b}{a - \lambda}$, i.e.,

$$N = \inf\{n \geq 1 | S_n > 0\}. \tag{23}$$

Note that for special cases such as for exponentially distributed A_i or exponentially distributed B_i the distribution of N can be obtained in closed form.

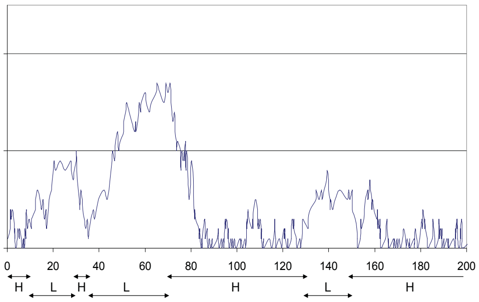

Fig. 4. A sample path of the scaled queue-length process $\frac{1}{\eta} Q^\eta(t)$, for $\eta = 1$, in the high-low model with $\lambda = 1$, $a = 2$, $b = \frac{1}{2}$

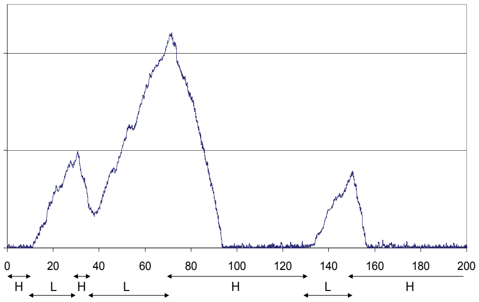

Fig. 5. A sample path of the scaled queue-length process $\frac{1}{\eta} Q^\eta(t)$, for $\eta = 10$, in the high-low model with $\lambda = 1$, $a = 2$, $b = \frac{1}{2}$

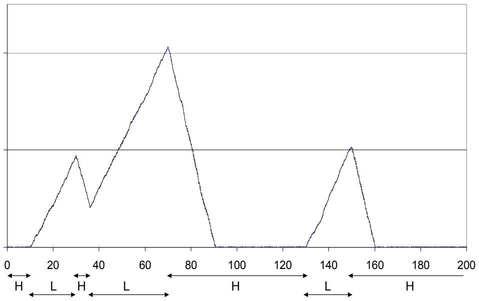

Fig. 6. A sample path of the scaled queue-length process $\frac{1}{\eta} Q^\eta(t)$, for $\eta = 100$, in the high-low model with $\lambda = 1$, $a = 2$, $b = \frac{1}{2}$

2) Adjusted stability: In Figures 4-6 we have depicted three different realizations of the scaled queue-length process $\frac{1}{\eta} Q^\eta(t)$, for $\eta = 1$, $\eta = 10$ and $\eta = 100$, respectively. The realization for the high and low period lengths are the same in Figures 4-6 for comparison purposes. The service rate starts off in the higher value $a = 2$ and the process shows stationary behavior, since the instantaneous load $\rho(t)$ is less than 1 during (the first) high rate period(s). As soon as the service rate switches to the lower rate $b = \frac{1}{2}$, the queue starts building up. Whenever the service rate switches back to the higher service rate, then the queue starts decreasing again. The fluctuations around the linear trend get smaller as η grows. From these figures, stationary behavior

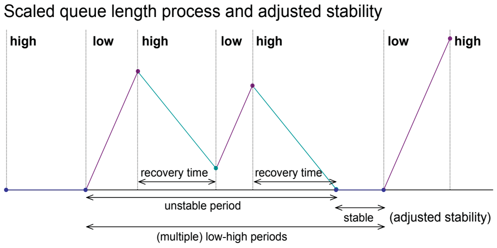

Fig. 7. Scaled queue length process and recovery periods.

during high rate periods is observed when the queue has decreased "sufficiently". Ultimately, in the quasi-stationary limit $\eta \to \infty$, stationary behavior is observed when the negative drift hits the horizontal axis, which is also the time epoch where the buffer content in the associated fluid queue becomes empty. In the figures we also observe that, in this example, the second high rate period is too short to recover from the excess load of the first low rate period. In contrast, the third high rate period is sufficiently long to recover from the excess load from the first two low rate periods. (Heuristically, the queue becomes stable again during the third high rate period.)

Figure 7 schematically represents the typical evolution of the workload process for $\eta \to \infty$ after linear scaling, and specifies the effectively instable, effectively stable, and recovery periods. Let π_{low} and π_{high} be the fraction of time that the system serves at *low* and *high* service rate, i.e., $\pi_{\text{low}} = \frac{\mathbb{E}B}{\mathbb{E}A+\mathbb{E}B}$, and $\pi_{\text{high}} = 1 - \pi_{\text{low}}$. Let π_{stable} and π_{unstable}, denote the fractions of time that the system is effectively stable and unstable, respectively, and let π_{recovery} be the fraction of time that the system is in a recovery period: $\pi_{\text{unstable}} = \pi_{\text{low}} + \pi_{\text{recovery}}$, $\pi_{\text{stable}} = \pi_{\text{high}} - \pi_{\text{recovery}}$. We may interpret π_{stable} as a measure for adjusted stability: instability is due to periods with a positive drift, i.e., π_{low}, which would be a first measure for instability. From a practical perspective, however, the system is also unstable during recovery periods. We now determine these fractions.

Proposition 7:

$$
\begin{aligned}
\pi_{\text{stable}} &= \frac{\mathbb{E}A - \frac{\lambda-b}{a-\lambda}\mathbb{E}B}{\mathbb{E}A + \mathbb{E}B} \\
\pi_{recovery} &= \frac{\lambda - b}{a - \lambda}\pi_{low}
\end{aligned}
$$

Proof Recall that N is the first ladder epoch of the random walk $\{S_n\}_n$, see (22). The *first ladder height* $S_N = \sum_{i=1}^{N}(A_{i+1} - cB_i)$ is exactly the time length that the queue is stable within the total period $\sum_{i=1}^{N}(A_{i+1} + B_i)$. Note that the ladder epochs are regeneration points for the queue length process. As a consequence, invoking renewal theory, and Wald's theorem,

$$
\begin{aligned}
\pi_{\text{stable}} &= \frac{\mathbb{E}S_N}{\mathbb{E}\sum_{i=1}^{N}(A_{i+1} + B_i)} \\
&= \frac{\mathbb{E}A - \frac{\lambda-b}{a-\lambda}\mathbb{E}B}{\mathbb{E}A + \mathbb{E}B}.
\end{aligned}
$$

□

C. Recovery time and adjusted stability for exponential distributions of high and low service durations

In this section we specify the distribution of N when A_i and B_i have exponential distributions with means $1/\alpha$ and $1/\beta$, respectively. To simplify the formulas in this section, we set $c := \frac{\lambda-b}{a-\lambda} = 1$. The distribution of N in the exponential case is given by the following proposition taken from [4].

Proposition 8: Let $p := \frac{(a-\lambda)\mathbb{E}A}{(a-\lambda)\mathbb{E}A + (\lambda-b)\mathbb{E}B}$. The distribution of the number of high-periods needed for recovery (re-stabilizing the system) is given by

$$
\mathbb{P}(N = n) = C_{n-1}p^n q^{n-1}, \quad \text{for } n \geq 1,
$$

where

$$
C_n = \frac{1}{n+1}\binom{2n}{n} = \frac{(2n)!}{n!(n+1)!}
$$

are Catalan numbers. The pgf $P_N(z) = \mathbb{E}z^N$ is given by

$$
P_N(z) = \sum_{n=1}^{\infty} z^n \mathbb{P}(N = n) = \frac{1 - \sqrt{1 - 4pqz}}{2q}.
$$

In particular, the probability that the queue length process recovers from instability is $P_N(1) = \frac{2p}{1+|2p-1|} = \frac{p \wedge q}{q}$, where $p \wedge q = \min\{p, q\}$. Indeed, if $p \geq \frac{1}{2}$ then N is finite with probability 1. However, if $p = \frac{1}{2}$ then we have $\mathbb{E}N = \infty$ (see Proposition 9; and also see relation with the symmetric Bernoulli walk [8]). The next proposition summarizes the mean and variance of N.

Proposition 9: The expected number of high-rate periods needed for recovery is given by

$$
\mathbb{E}N = \frac{\mathbb{E}A}{\mathbb{E}A - \frac{\lambda-b}{a-\lambda}\mathbb{E}B}, \quad \text{if } \mathbb{E}A > \frac{\lambda - b}{a - \lambda}\mathbb{E}B,
$$

otherwise, if $\mathbb{E}A \leq \frac{\lambda-b}{a-\lambda}\mathbb{E}B$, then $\mathbb{E}N = \infty$. In addition, the variance is given by

$$
\mathbb{V}arN = \frac{pq}{(p-q)^3} = \frac{\frac{\lambda-b}{a-\lambda}\mathbb{E}A\mathbb{E}B\left(\mathbb{E}A + \frac{\lambda-b}{a-\lambda}\mathbb{E}B\right)}{\left(\mathbb{E}A - \frac{\lambda-b}{a-\lambda}\mathbb{E}B\right)^3},
$$

which only depends on the means of the high and low periods.

Proof By induction on n it follows that:

$$
\frac{d^n}{dz^n}P(z) = n!C_{n-1}\frac{p^n q^{n-1}}{(1 - 4pqz)^{(2n-1)/2}}.
$$

Then, use the fact that $\frac{d}{dz}P(z)\Big|_{z=1} = \mathbb{E}N$ and $\frac{d^2}{dz^2}P(z)\Big|_{z=1} = \mathbb{E}N(N-1)$. □

D. Scaling of the queue-length for the high-low model

We now extend the analysis to the high-low model. Here, we focus on the case where the A_i and B_i have exponential distributions with means $1/\alpha$ and $1/\beta$, respectively. The stationary distribution of Q^η is then known explicitly [13]:

$$\mathbb{P}(Q = i; \mu = j) = c_j p^i + d_j q^i, \tag{24}$$

for $j \in \{a, b\}$, where c_j and d_j are such that p and q are the two roots within the unit disc of the following equations

$$\begin{aligned} c_a p(\lambda + a + \alpha) &= c_a p^2 a + c_a \lambda + c_b p \beta, \\ c_b p(\lambda + b + \beta) &= c_b p^2 b + c_b \lambda + c_a p \alpha, \end{aligned}$$

and

$$\begin{aligned} d_a q(\lambda + a + \alpha) &= d_a q^2 a + d_a \lambda + d_b q \beta, \\ d_b q(\lambda + b + \beta) &= d_b q^2 b + d_b \lambda + d_a q \alpha. \end{aligned}$$

The precise form of these coefficients is not essential (they are characterized through the solution to a cubic equation). We are primarily interested in the queue length as $\eta \to \infty$. With standard algebra it follows that p and q tend to λ/a and 1 respectively. (Although p, q, c_j and d_j depend on η when applying uniform acceleration, we will not reflect this in the notation.) The corresponding constants then follow from the equations above and we get after convenient rewriting:

$$\begin{aligned} \lim_{\eta\to\infty} \mathbb{P}(Q^\eta > x \mid \mu = a) =& \frac{\lambda(\alpha+\beta) - b\alpha}{a\beta} + \\ & \left(1 - \frac{\lambda(\alpha+\beta) - b\alpha}{a\beta}\right) \left(\frac{\lambda}{a}\right)^x. \end{aligned} \tag{25}$$

Naturally, we find $\lim_{\eta\to\infty} \mathbb{P}(Q^\eta > x \mid \mu = b) = 1$ for all x. The term $\frac{\lambda(\alpha+\beta)-b\alpha}{a\beta}$ can be interpreted as the fraction of high-rate service time that is needed for recovery. It can be shown that this indeed coincides with the probability that the associated fluid queue is non-empty.

If we scale the queue length with the parameter η, it can be shown that

$$\lim_{\eta\to\infty} q^{\frac{1}{\eta}} = \frac{\beta}{b - \lambda} - \frac{\alpha}{\lambda - a} =: \delta.$$

Substituting this into the distribution for Q we get

$$\lim_{\eta\to\infty} \mathbb{P}(\frac{1}{\eta} Q^\eta > x \mid \mu = a) = \frac{\lambda(\alpha+\beta) - b\alpha}{a\beta} \delta^x,$$

and hence

$$\lim_{\eta\to\infty} \mathbb{P}(\frac{1}{\eta} Q^\eta > x \mid \mu = b) = \delta^x.$$

V. Conclusion and extensions

In this paper we considered the quasi-stationary regime for a single server queue with service rate fluctuation driven by an independent Markov process, where the arrival rate is allowed to temporarily exceed the service rate. For the system with service rate alternating between high and low rate, we have discussed notions of effective load and adjusted stability that allow us to characterize the fraction of the high service rate period during which the queue is recovering from instability due to the low service rate period. We have characterized the distribution of the effective load via a related fluid queue, and have obtained the distribution of the number of high service periods required for recovery to stability. This allows us to obtain the distribution of the number of customers during a high service period.

References

[1] D. Anick, D. Mitra, and M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell Systems Technical J.*, 61:1871–1894, 1982.

[2] S. Asmussen. *Applied probability and queues, 2nd revised and extended ed., Source: Applications of Mathematics. 51.* Springer, New York, NY, 2003.

[3] A. Bassamboo, J.M. Harrison, and A. Zeevi. Pointwise stationary fluid models for stochastic processing networks. *Manufacturing and Service Operations Management*, 11(1):70–89, 2009.

[4] S.K. Cheung *Processor-sharing queues and resource sharing in wireless LANs.* PhD thesis. University of Twente, 2007

[5] F. Delcoigne, A. Proutière, and G. Régnié. Modeling integration of streaming and data traffic. *Performance Evaluation*, 55:185–209, 2004.

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

[7] A. Federgruen and L. Green. Queueing systems with service interruptions. *Operations Research*, 34:752–768, 1986.

[8] W. Feller. *An Introduction to Probability Theory and Its Applications, volume I.* Wiley, third edition, New York, NY, 1966.

[9] S. Fuhrmann and R. Cooper. Stochastic decompositions in the M/G/1 queue with generalized vacations. *Operations Research*, 33(5):1117–1129.

[10] V. Gupta, M. Harchol-Balter, A. Wolf, and U. Yechiali. Fundamental characteristics of queues with fluctuating load. In *Sigmetrics/Performance '06*. Saint Malo, France, June 2006.

[11] R. Hampshire, M. Harchol-Balter, and W. Massey. Fluid and diffusion limits for transient sojourn times of processor sharing queues with time varying rates. *Queueing Systems*, 53(1-2):19–30, June 2006.

[12] M.T.S. Jonckheere, R. Núñez-Queija, and B.J. Prabhu. Performance analysis of traffic surges in multi-class communication networks. *Proc. ITC 22*, this volume.

[13] G. Latouche and V. Ramaswami. *An Introduction to Matrix Analytic Methods in Stochastic Modeling*. Cambridge.

[14] W. Massey and W. Whitt. Uniform acceleration expansions for Markov chains with time-varying rates. *Annals of Applied Probability*, 8(4):1130–1155, 1998.

[15] R. Núñez-Queija. Sojourn times in a processor sharing queue with service interruptions. *Queueing Systems*, 34(1-4):351–386, 2000.

[16] O. Perry, and W. Whitt. Responding to unexpected overloads in large-scale service systems. *Management Science*, 55(8):1353–1367, 2009.

[17] S. Ross. Average delay in queues with non-stationary poisson arrivals. *Journal of Applied Probability*, 15:602–609, 1978.

[18] W. Scheinhardt. *Markov-modulated and feedback fluid queues*. Ph.D. thesis, University of Twente, Enschede, The Netherlands, 1998.

[19] H. Takagi. *Queueing Analysis, Vacations and Priority Systems, Part 1, vol. 1.* Elsevier Science Publishers B.V., The Netherlands, 1991.