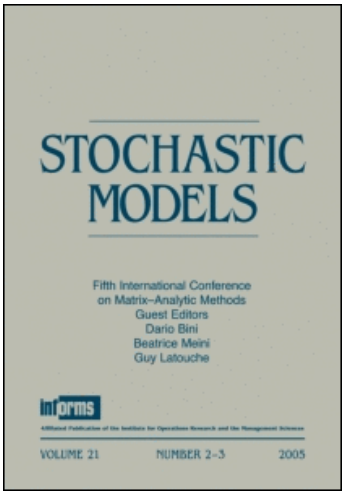
M/M/∞ transience: Tail asymptotics of congestion periods

Mandjes, M.; Roijers, F.

DOI
10.1080/15326340903291289
Publication date
2009
Document Version
Final published version
Published in
Stochastic Models



Stochastic Models

Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597301

M/M/∞ Transience: Tail Asymptotics of Congestion Periods

Michel Mandjes[abc]; Frank Roijers[ad]
[a] Korteweg-de Vries Institute for Mathematics, Amsterdam, The Netherlands [b] CWI, Amsterdam, The Netherlands [c] EURANDOM, Eindhoven, The Netherlands [d] TNO Information and Communication Technology, Delft, The Netherlands

To cite this Article Mandjes, Michel and Roijers, Frank(2009) 'M/M/∞ Transience: Tail Asymptotics of Congestion Periods', Stochastic Models, 25: 4, 614 — 647
To link to this Article: DOI: 10.1080/15326340903291289
URL: http://dx.doi.org/10.1080/15326340903291289

Stochastic Models, 25:614–647, 2009

ISSN: 1532-6349 print/1532-4214 online
DOI: 10.1080/15326340903291289

Taylor & Francis
Taylor & Francis Group

M/M/∞ TRANSIENCE: TAIL ASYMPTOTICS OF CONGESTION PERIODS

Michel Mandjes[1,2,3] and Frank Roijers[1,4]

[1] *Korteweg-de Vries Institute for Mathematics, Amsterdam, The Netherlands*
[2] *CWI, Amsterdam, The Netherlands*
[3] *EURANDOM, Eindhoven, The Netherlands*
[4] *TNO Information and Communication Technology, Delft, The Netherlands*

□ *The* c*-congestion period, defined as a time interval in which the number of customers is larger than* c *all the time, is a key quantity in the design of communication networks. Particularly in the setting of M/M/*∞ *systems, the analysis of the duration of the congestion period,* D_c*, has attracted substantial attention; related quantities have been studied as well, such as the total area* A_c *above* c*, and the number of arrived customers* N_c *during a congestion period. Laplace transforms of these three random variables being known, as well as explicit formulae for their moments, this article addresses the corresponding tail asymptotics. Our work addresses the following topics. In the so-called many-flows scaling, we show that the tail asymptotics are essentially exponential in the scaling parameter. The proof techniques stem from large-deviations theory; we also identify the most likely way in which the event under consideration occurs. In the same scaling, we approximate the model by its Gaussian counterpart. Specializing to our specific model, we show that the (fairly abstract) sample-path large-deviations theorem for Gaussian processes, viz. the generalized version of Schilder's theorem, can be written in a considerably more explicit way. Relying on this result, we derive the tail asymptotics for the Gaussian counterpart. Then we use change-of-measure arguments to find upper bounds, uniform in the model parameters, on the probabilities of interest. These change-of-measures are applied to devise a number of important sampling schemes, for fast simulation of rare-event probabilities. They turn out to yield a substantial speed-up in simulation effort, compared to naïve, direct simulations.*

Keywords Congestion period; Gaussian traffic; Importance sampling; Infinite-server queue; Large deviations.

Mathematics Subject Classification Primary 60K25; Secondary 60F10, 90B18.

Received August 2008; Accepted June 2009

Address correspondence to Michel Mandjes, Korteweg-de Vries Institute for Mathematics, Plantage Muidergracht 24, 1018 TV, Amsterdam, The Netherlands; E-mail: mmandjes@science.uva.nl

1. INTRODUCTION

Despite all research devoted to the M/M/∞ queue, this fundamental queueing system continues to pose new challenges. In this article, we study its so-called c-congestion period, which is defined as a consecutive period in which the number of customers is larger than c. Clearly, knowledge of the probabilistic characteristics of this c-congestion period is useful in several applications, for instance, when designing circuit-switched communication networks, but notably also in packet-based networks[25]. Apart from its duration D_c (in time), other interesting quantities related to the c-congestion period are: the number of customers N_c that have arrived, as well as the total amount A_c of work in excess of level c (which is often referred to as the "area above c") during the c-congestion period.

The congestion period has been the subject of several detailed studies. The seminal article by Guillemin and Simonian[14] presents closed-form expressions for the means of D_c, N_c, and A_c, but, more importantly, also obtained the Laplace transforms (LTs) of these quantities (expressed in terms of special functions) (see also Ref.[12]). They also introduce a scaling under which the M/M/∞ system actually tends to an M/M/1 system. Preater[19,20] found an alternative, elegant derivation of the LTs, and interestingly, this new approach enabled him to also derive their joint LT. Roijers et al.[22] presented iterative relations for the moments of D_c, N_c, and A_c, that is, they expressed these moments for the $(c-1)$-congestion period in terms of those for the c-congestion period. Then the resulting recursions were explicitly solved; in this approach Preater's result on the joint LT of the quantities played a crucial role, as in the recursions also "mixed expectations" appear (i.e., terms of the type $\mathbb{E}(D_c N_c)$). Progress on systems with heterogeneous servers has been reported recently in Tsybakov[24].

Procedures for determining the moments of the quantities D_c, N_c, and A_c being known, strikingly little is known about the tail probabilities $\mathbb{P}(D_c > x)$, $\mathbb{P}(A_c > x)$, and $\mathbb{P}(N_c > x)$. By majorizing the M/M/∞ queue by an appropriate M/M/1 queue, cf. Refs.[14,19], upper bounds on the tails can be derived relatively easily, but it is not *a priori* clear how tight these bounds are. We also mention here a related result by Guillemin and Pinchon[13] on the area of a busy period of an M/M/1 queue, stating that its tail distribution decays essentially in a Weibullian way.

1.1. Contribution

The goal of this article is to shed light on the tail probabilities $\mathbb{P}(D_c > x)$, $\mathbb{P}(A_c > x)$, and $\mathbb{P}(N_c > x)$. In more detail, the contributions are as follows.

1.1.1. *Asymptotics under Many-Flows Scaling*

We scale the arrival rate Λ by a parameter n, i.e., we let $\Lambda \equiv n\lambda$, but leave the mean service time μ^{-1} unchanged, so that the system load becomes $p = n\varrho$, where $\varrho := \lambda/\mu$. Starting a congestion period at level $c \equiv nc$, our aim is to find the asymptotics of the probabilities $\mathbb{P}(D_{nc} > x)$, $\mathbb{P}(A_{nc} > nx)$, and $\mathbb{P}(N_{nc} > nx)$, for n large and $x > 0$ given. We succeed in doing so by using sample-path large-deviations techniques, relying predominantly on the theory developed in Ref.[23]. It turns out that the probability $\mathbb{P}(D_{nc} > x)$ decays roughly exponentially in n (that is, we show that $-n^{-1} \cdot \log \mathbb{P}(D_{nc} > x)$ tends to a positive, finite limit); analogous results hold for $\mathbb{P}(A_{nc} > nx)$ and $\mathbb{P}(N_{nc} > nx)$. Assuming that $c > \varrho$ (which we shall do throughout the article), we explicitly identify the corresponding decay rates. As a by-product, we also identify the most likely path, which is essentially the most probable way in which the events under consideration occur: given that the rare event happens, then with overwhelming probability it does so via a path in the direct neighborhood of the most likely path. Clearly, the many-flows scaling is particularly suitable for systems with a considerable level of multiplexing.

1.1.2. *Asymptotics of the Gaussian Counterpart*

We approximate the M/M/∞-model under the many-flows scaling by an appropriate Gaussian process, the so-called Gaussian counterpart of the M/M/∞ system; for further background on this type of approximation see[1,11] (section 2). We argue that this counterpart is the so-called integrated Ornstein-Uhlenbeck (iOU) model[15]. Now we can analyze the rare events under consideration by applying sample-path large deviations results, viz. the generalized version of Schilder's theorem[2,7,15]. Owing to the fact that the iOU process has a well-defined rate process (unlike, for instance, fractional Brownian motion), the corresponding large-deviations rate function can be expressed in a considerably more explicit way than in the standard version of the generalized version of Schilder's theorem.

Relying on this explicit sample-path large-deviations result, we determine the tail asymptotics of $\mathbb{P}(D_{nc} > x)$ and $\mathbb{P}(A_{nc} > nx)$ for n large for the Gaussian counterpart; the quantity N_{nc} does not have a meaningful Gaussian counterpart. As could be expected, these Gaussian asymptotics become increasingly accurate when c approaches ϱ from above, that is, in a heavy-traffic setting. Again, we also find the corresponding most likely paths.

1.1.3. *Uniform Bounds*

All results mentioned above relate to the M/M/∞ model under the many-flows scaling, and are in terms of (relatively crude) asymptotics.

For practical purposes, however, it would be helpful to have bounds—particularly upper bounds—on the probabilities of interest, that are valid for all parameter settings (i.e., not just in an asymptotic regime). Using change-of-measure arguments, and relying on the celebrated Chernoff bound, we are able to derive such uniform upper bounds; these are in closed form.

1.1.4. *Importance Sampling Algorithms*

It is clear that estimating the probabilities $\mathbb{P}(D_c > x)$, $\mathbb{P}(A_c > x)$, and $\mathbb{P}(N_c > x)$ by direct, naïve simulation is inherently difficult, particularly for large x, because of the rarity of the event under consideration. This motivates the search for "fast-simulation" techniques[6]. The change-of-measures, mentioned above in the context of the uniform bounds, suggest parameters that can be used in importance-sampling procedures. In a numerical study, we compare the estimates (as obtained under the many-flows scaling), as well as the uniform upper bounds, with results obtained from importance-sampling-based simulations. The importance-sampling schemes turn out to yield a substantial speed-up compared to direct, naïve simulations. They are very useful for practical purposes, as the uniform upper bounds tend to overestimate the probabilities of interest.

1.2. Outline

Section 2 introduces the model, i.e., the M/M/∞ queue, the c-congestion period, and formally defines the quantities of interest, i.e., D_c, N_c, and A_c. In section 3, we present the analysis of the tail probabilities under the many-flows scaling, whereas section 4, addresses the Gaussian counterpart. Where sections 3 and 4 present logarithmic asymptotics of the scaled model, in section 5 we establish uniform, closed-form (upper) bounds on the probabilities of interest. Further, this section also describes change-of-measures that can be used in importance-sampling-based simulation schemes. In section 6, we numerically evaluate the decay rates of sections 3 and 4, and compare these with the uniform bounds, as well as with simulation results (obtained by the importance-sampling procedure sketched in section 5). Section 7 concludes.

2. MODEL AND PRELIMINARIES

2.1. Model

We consider a resource at which flows arrive according to a Poisson process with intensity Λ, and at which the jobs stay for an exponentially distributed time with mean μ^{-1}. We are thus in the setting of the

(classical) M/M/∞ model. The following properties are well-known: (i) in stationarity the number of trunks occupied has a Poisson distribution with mean $p := \Lambda/\mu$; (ii) the number of arriving flows in an interval of length t is Poisson distributed with mean Λt, and each of them has arrived on an epoch uniformly distributed over the interval $[0, t]$, independently of the other arrivals.

2.2. Congestion Periods

Let us now define the key quantities studied in this article. To this end, we first need some additional notations. First, let $X(t)$ denote the number of flows present at time t; $X(\cdot)$ constitutes a continuous-time Markov chain on $\{0, 1, \dots\}$, with upward transition rate λ, and downward transition rate (from state k) $k\mu$. $A(t)$ is the defined as the work generated by the flows in the interval $[0, t]$, which is essentially the integral of $X(\cdot)$:

$$A(t) := \int_0^t X(s)ds.$$

In this article we also need the discrete-time embedding of the above-described continuous-time process. We let Y_m be the number of flows present after m jumps, where a jump is an arrival or departure. It is clear that $(Y_m)_{m\in\mathbb{N}}$ is a discrete-time Markov chain, with upward transition probability $\lambda/(\lambda + k\mu)$ and downward transition probability $k\mu/(\lambda + k\mu)$ (from state k).

A first observation is that the process $A(t)$ is rather convenient to work with, owing to its nice structure. In particular, using elementary arguments and relying on properties (i) and (ii) mentioned above, it can be verified that, for $\vartheta < \mu$,

$$\log \mathbb{E}\big(e^{\vartheta A(t)} \mid X(0) = \mathrm{c} + 1\big) = (\mathrm{c} + 1)\log\left(\frac{\mu}{\mu - \vartheta} - \frac{\vartheta}{\mu - \vartheta}e^{-(\mu-\vartheta)t}\right) + \frac{\Lambda t\vartheta}{\mu - \vartheta} - \frac{\Lambda\vartheta}{(\mu - \vartheta)^2}\big(1 - e^{-(\mu-\vartheta)t}\big). \quad (1)$$

It is clear that $A(t)$ is smaller than the amount of work that has arrived in $[0, t]$ when the full flow would have been "injected" instantaneously. This reasoning yields that

$$\log \mathbb{E}\big(e^{\vartheta A(t)} \mid X(0) = \mathrm{c} + 1\big) \le (\mathrm{c} + 1)\log\left(\frac{\mu}{\mu - \vartheta}\right) + \frac{\Lambda t\vartheta}{\mu - \vartheta}, \quad (2)$$

which is in agreement with (1). More specifically, it is readily checked that $\mathbb{E}A(t) = pt$ and

$$\mathbb{V}\mathrm{ar}\, A(t) = \frac{2\Lambda}{\mu^3}\left(t\mu - 1 + e^{-t\mu}\right). \tag{3}$$

This article studies the tail behavior of the following three random variables:

$$\begin{aligned} D_{\mathrm{c}} &:= \inf\{t \geq 0 : X(t) = \mathrm{c} \mid X(0) = \mathrm{c}+1\};\\ A_{\mathrm{c}} &:= (A(D_{\mathrm{c}}) - \mathrm{c}D_{\mathrm{c}} \mid X(0) = \mathrm{c}+1);\\ N_{\mathrm{c}} &:= \frac{1}{2}\inf\{m \in \mathbb{N} : Y_m = \mathrm{c} \mid Y_0 = \mathrm{c}+1\} - \frac{1}{2}. \end{aligned}$$

We refer to D_{c} as the duration of the congestion period above level c. A_{c} can be interpreted as a proxy for the amount of traffic lost during a congestion period (in systems in which the number of lines is truncated at c); informally, this is the area under the graph of $X(s) - \mathrm{c}s$ during a congestion period. Furthermore, it is readily verified that N_{c} corresponds to the number of arrivals during a congestion period (which equals the number of departures during a congestion period, decreased by 1). We assume throughout that $p < \mathrm{c}$.

We recall that the LTs of the distributions of D_{c}, A_{c}, and N_{c} were found by Guillemin and Simonian[14] in terms of special functions, whereas Preater[19] elegantly derived their joint LT. Roijers et al.[22] found explicit expressions for expected values and variances of D_{c}, A_{c}, and N_{c}, and their covariances.

2.3. Performance Measures

As an alternative to deriving the distribution functions of D_{c}, A_{c}, and N_{c} from the Laplace transforms, we apply a scaling that allows explicit asymptotic analysis. In this scaling, one identifies $\Lambda \equiv n\lambda$ and $\mathrm{c} \equiv nc$, where n is large; likewise $\mathrm{p} \equiv n\varrho$. We can equivalently write that the total traffic arrival process $A^n(t)$ corresponds to the sum of n independent and identically distributed arrival processes, each distributed as the process $A(t)$ introduced above, but with Λ replaced by λ, and c replaced by c. Similarly, $X^n(t)$ is defined as the aggregate rate process, and $(Y^n_m)_{m\in\mathbb{N}}$ as the aggregate rate at jump epochs.

A first goal of this article is to asymptotically characterize the probabilities $\mathbb{P}(D_{nc} > x)$, $\mathbb{P}(A_{nc} > nx)$, and $\mathbb{P}(N_{nc} > nx)$, for n large. The scaling applied is usually referred to as the "many-flows scaling"[4,9,15], and is particularly appropriate if the level of multiplexing is reasonably large. We recall that it is assumed that $c > \varrho$, so that the events under

consideration are increasingly rare when n grows large. In this article, we rely on large-deviations theory to show that the above probabilities decay essentially exponentially in n, and to explicitly determine the corresponding exponential decay rates, i.e., for $x > 0$,

$$\delta(x) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(D_{nc} > x),$$

and likewise also the decay rate corresponding to a large area, for $x > 0$,

$$\alpha(x) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(A_{nc} > nx),$$

and the decay rate corresponding to many arriving flows per congestion period, for $x > 0$,

$$\nu(x) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(N_{nc} > nx).$$

3. LARGE DEVIATIONS ANALYSIS OF CONGESTION PERIOD

In this section, we consider the M/M/∞ model under the many-flows scaling that was described above, and apply sample-path large deviations to compute the decay rates (n large) of $\mathbb{P}(D_{nc} > x)$, $\mathbb{P}(A_{nc} > nx)$, and $\mathbb{P}(N_{nc} > nx)$. In the first subsection, we review the main results for sample-path large deviations of Markovian systems. Then we subsequently determine the decay rates $\delta(x)$, $\alpha(x)$, and $\nu(x)$.

3.1. Sample-Path Large-Deviations Theory

In our exposition, we rely extensively on the framework presented in Shwartz and Weiss[23]. In this framework a crucial role is played by the local rate function. In case of the M/M/∞ process, this function is defined as

$$I_x(u) := \sup_{\vartheta} \left(\vartheta u - \lambda(e^{\vartheta} - 1) - \mu x(e^{-\vartheta} - 1)\right).$$

In fact, the local rate function measures the "cost" of moving in direction u, when the (scaled) process is in state x, in the following sense. Suppose x flows are present. Then the position of the scaled process $X^n(\cdot)/n$ after ε time units (ε small) is, in expectation, roughly $x + (\lambda - \mu x)\varepsilon$, and hence the "most likely" derivative of moving is $u(x) := \lambda - \mu x$. Indeed, it is verified that $I_x(u(x)) = 0$: there is no "cost" involved in moving into this most likely direction. It is checked that any other direction yields strictly positive costs. We further remark that the function $I_x(u)$ can be calculated explicitly (the first-order condition being a

quadratic equation), but this is, for the purposes of the present article, not necessary.

Having the local rate function at our disposal, we can define the action functional. Informally, this action functional $\mathbb{I}(f)$ represents the "cost" of the scaled process $X^n(\cdot)/n$ following a path $f(\cdot)$:

$$\mathbb{I}(f) := \int_{-\infty}^{\infty} I_{f(s)}(f'(s))\mathrm{d}s$$

It is a matter of elementary calculus to check that, considering just the time after time 0, the path $\varphi(s) := \varrho + (\varphi_0 - \varrho)e^{-\mu s}$ (for some $\varphi_0 > 0$) yields cost 0: as $\varphi'(s) = (\lambda - \varphi_0\mu)e^{-\mu s}$,

$$\mathbb{I}(\varphi) = \int_0^{\infty} \sup_{\vartheta}(\vartheta(\lambda - \varphi_0\mu)e^{-\mu s} - \lambda(e^{\vartheta} - 1) - (\lambda + (\varphi_0\mu - \lambda)e^{-\mu s})(e^{-\vartheta} - 1))\mathrm{d}s = 0.$$

This answer makes sense, as this path is essentially the "average path" starting at φ_0 at time 0 to the system's equilibrium value ϱ.

Using this framework, the following sample-path large-deviations principle can be stated:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} X^n(\cdot) \in \mathcal{S}\right) = -\inf_{f\in\mathcal{S}} \mathbb{I}(f). \tag{4}$$

Informally, one find the most likely path f in the set $\mathcal{S}$, say $f^\star$; given that the event $\{X^n(\cdot)/n \in \mathcal{S}\}$ occurs, the realization will be close to $f^\star$. Intentionally, (4) has been stated in a slightly imprecise way: in fact one has two inequalities, respectively, for open and closed sets (in the path space). These issues are not crucial in the scope of the present article, and we refer to Ref.[23] for these and related details.

In discrete time, i.e., for the process Y_m^n, a similar framework can be set up, see, for instance, Bucklew[5]. Then the local rate function is given by

$$J_x(u) := \sup_{\vartheta} \left(\vartheta u - \log\left(\frac{\lambda}{\lambda + \mu x} e^{\vartheta} + \frac{\mu x}{\lambda + \mu x} e^{-\vartheta}\right)\right).$$

Again, this function can be evaluated in a more explicit manner, but we will refrain from doing this. Similar to before, we can define the action functional as

$$\mathbb{J}(f) := \int_{-\infty}^{\infty} J_{f(s)}(f'(s))\mathrm{d}s.$$

Again, we have a sample-path, large-deviations principle:

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\left(\frac{1}{n}Y^n(\cdot)\in\mathscr{S}\right)=-\inf_{f\in\mathscr{S}}\mathbb{J}(f). \tag{5}$$

3.2. Congestion Period

We cast our problem of identifying the decay rate of $\mathbb{P}(D_{nc}>x)$ into the large-deviations framework of the previous subsection. Immediately from the sample-path, large-deviations result (4), we have that

$$\delta(x)=-\inf_{f\in\mathscr{D}}\mathbb{I}(f),$$

with $\mathscr{D}:=\{f\,|\,\forall s\in[0,x]:f(s)\geq c, f(0)=c\}$. Heuristically reasoning, as we are looking for the "cheapest" path in $\mathscr{D}$, it cannot be that the optimal path is such that $f(x)>c$, as otherwise even a longer congestion period could be obtained "for free". Based on this argumentation, it is seen that $\inf_{f\in\mathscr{D}}\mathbb{I}(f)=\inf_{f\in\overline{\mathscr{D}}}\mathbb{I}(f)$, with

$$\overline{\mathscr{D}}:=\{f\,|\,\forall s\in[0,x]:f(s)\geq c, f(0)=f(x)=c\}.$$

Therefore we further study the following variational problem:

$$\delta(x)=-\inf_{f\in\overline{\mathscr{D}}}\int_0^x I_{f(s)}(f'(s))\mathrm{d}s.$$

Proposition 3.2.1. *For $x\geq 0$,*

$$\delta(x)=-x\delta^\star;\quad \delta^\star:=\left(\sqrt{\lambda}-\sqrt{\mu c}\right)^2.$$

Proof. We prove this result by subsequently establishing an upper bound and a lower bound. Define the path f_c through $f_c(s)=c$ for all $s\in[0,x]$. As $f_c\in\overline{\mathscr{D}}$, it follows that

$$\begin{aligned}\delta(x)&\geq-\mathbb{I}(f_c)=-x\sup_\vartheta\left(-\lambda(e^\vartheta-1)-\mu c(e^{-\vartheta}-1)\right)\\&=-x\left(\lambda-2\sqrt{\lambda\mu c}+\mu c\right)=-x\left(\sqrt{\lambda}-\sqrt{\mu c}\right)^2;\end{aligned}$$

the optimizing ϑ equals $\vartheta^\star:=\frac{1}{2}\log(\mu c/\lambda)=\frac{1}{2}\log(c/\varrho)>0$. Hence we have proven the lower bound. On the other hand,

$$\begin{aligned}\delta(x)&\leq-\inf_{f\in\overline{\mathscr{D}}}\int_0^x\left(\vartheta^\star f'(s)-\lambda(e^{\vartheta^\star}-1)-\mu f(s)(e^{-\vartheta^\star}-1)\right)\mathrm{d}s\\&\overset{\text{(i)}}{=}\sup_{f\in\overline{\mathscr{D}}}\int_0^x\left(\lambda(e^{\vartheta^\star}-1)+\mu f(s)(e^{-\vartheta^\star}-1)\right)\mathrm{d}s\end{aligned}$$

$$= \sup_{f\in\overline{\mathscr{D}}} \int_0^x \left(\sqrt{\lambda\mu c} - \lambda + f(s)\sqrt{\frac{\lambda\mu}{c}} - \mu f(s)\right)\mathrm{d}s$$

$$\overset{\text{(ii)}}{\leq} \int_0^x \left(\sqrt{\lambda\mu c} - \lambda + c\sqrt{\frac{\lambda\mu}{c}} - \mu c\right)\mathrm{d}s = -x\left(\sqrt{\lambda} - \sqrt{\mu c}\right)^2,$$

recalling for (i) that $f(0) = f(x) = c$ for all $f \in \overline{\mathscr{D}}$, and for (ii) Lemma A.1 (to be found in the Appendix). This yields the upper bound: all $f \in \overline{\mathscr{D}}$ yield at most decay rate $-x\delta^\star$, as desired. □

Remark 3.2.1. Also Ref.[23] (section 13.5.6) focuses on congestion periods, albeit with a slightly different definition. They consider the random variable

$$B_c := \sup\{t \geq 0 : A(t) \geq \mathrm{c}t \mid X(0) = \mathrm{c} + 1\}.$$

Again invoking the sample-path, large-deviations result (4), the decay rate of $\mathbb{P}(B_{nc} > x)$ can be rewritten as $-\inf_{f\in\mathscr{B}} \mathbb{I}(f)$, where

$$\mathscr{B} := \left\{f \,\middle|\, \forall s \in [0,x] : \int_0^s f(r)\mathrm{d}r \geq cs, f(0) = c\right\}.$$

In Ref.[23] (equation (13.65)) it is claimed that this decay rate equals $-x\delta^\star$, i.e., $\delta(x)$. This, however, seems an error, and the correct decay rate should be[17]

$$-\sup_{\vartheta}\big(\vartheta cx - c\log\phi(\vartheta, x) - \psi(\vartheta, x)\big), \tag{6}$$

where, cf. (1),

$$\phi(\vartheta, t) := \frac{\mu}{\mu - \vartheta} - \frac{\vartheta}{\mu - \vartheta}e^{-(\mu-\vartheta)t};$$

$$\psi(\vartheta, t) := \frac{\lambda t\vartheta}{\mu - \vartheta} - \frac{\lambda\vartheta}{(\mu - \vartheta)^2}(1 - e^{-(\mu-\vartheta)t}). \tag{7}$$

The proof is based on the fact that it turns out that the most likely path in

$$\overline{\mathscr{B}} := \left\{f \,\middle|\, \int_0^x f(s)\mathrm{d}s \geq cx, f(0) = c\right\}$$

lies in $\mathscr{B}$; notice that $\overline{\mathscr{B}} \supseteq \mathscr{B}$. It is a direct implication of Cramér's theorem that the decay rate of the optimal path in $\overline{\mathscr{B}}$ indeed equals (6). Hence, the decay rate of $\mathbb{P}(B_{nc} > x)$ is (6), which is larger than $-x\delta^\star$. In other

words: the event is less rare than suggested by Ref.[23] (equation (13.65)); there is a cheaper path than $f_c(\cdot)$, namely, a path that is strictly larger than c on $(0, x)$. For additional details, we refer to Case 3 in Theorem. 3.1 in Ref.[17].

3.3. Area

We now turn our attention to the tail asymptotics of the area A_{nc}. Again, applying the sample-path, large-deviations result (4), we obtain

$$\alpha(x) = -\inf_{f\in\mathcal{A}} \mathrm{I\!I}(f), \tag{8}$$

where $\mathcal{A}$ is the set of paths that lead to an area of at least x:

$$\mathcal{A} := \left\{ f \,\middle|\, \exists t > 0 : \int_0^t f(s)ds \geq x + ct, \forall s \in [0,t] : f(s) \geq c, f(0) = c \right\}.$$

In the following lemma, we prove that the set $\overline{\mathcal{A}}$, given by

$$\overline{\mathcal{A}} := \left\{ f \,\middle|\, \exists t > 0 : \int_0^t f(s)\mathrm{d}s \geq x + ct, f(0) = c \right\},$$

which is evidently larger than $\mathcal{A}$, contains the optimal path in $\mathcal{A}$.

Lemma 3.3.1. *The following identity holds:*

$$\inf_{f\in\mathcal{A}} \mathrm{I\!I}(f) = \inf_{f\in\overline{\mathcal{A}}} \mathrm{I\!I}(f).$$

Proof. As mentioned above, $\mathcal{A} \subseteq \overline{\mathcal{A}}$. Hence, in order to prove the stated, it suffices to show that the minimizer in the larger set, $\overline{\mathcal{A}}$, is element of the smaller set, $\mathcal{A}$.

This follows directly from a reasoning analogous to section 13.2 of Ref.[23]. To this end, first observe that

$$\inf_{f\in\overline{\mathcal{A}}} \mathrm{I\!I}(f) = \inf_{t>0} \inf_{f\in\overline{\mathcal{A}}_t} \mathrm{I\!I}(f), \quad \text{where } \overline{\mathcal{A}}_t := \left\{ f \,\middle|\, \int_0^t f(s)\mathrm{d}s \geq x + ct, f(0) = c \right\}.$$

For a model intimately related to our M/M/∞ model (viz. the model with exponential on-off sources)[23] identifies, using calculus-of-variations techniques, the optimizing $t^\star$, as well as the corresponding most likely path $f^\star$ in $\overline{\mathcal{A}}_{t^\star}$. This path $f^\star$ turns out to be a symmetric hyperbolic cosine, i.e., $t^\star$ is such that $f^\star(0) = f^\star(t^\star) = c$, $f^\star(s) > c$ for all $s \in (0, t^\star)$, and

$$\int_0^{t^\star} f^\star(s)\mathrm{d}s = x + ct^\star.$$

Mimicking the analysis in Ref.[23], it is elementary to check that the same properties hold for the M/M/∞ model. This implies that $f^\star \in \mathscr{A}$, which proves the stated. □

We have reduced the problem of finding $\alpha(x)$ to finding the most likely path in $\overline{\mathscr{A}}$. Before actually computing this decay rate, which we will do in Proposition 3.3.1, we first establish another auxiliary result that reveals a relation between the decay rate corresponding to the most likely path in $\overline{\mathscr{A}}$ on one hand, and the decay rate of tail probabilities in a related queueing system.

To this end, consider a queue fed a Poisson stream of jobs (rate $n\lambda$), each staying in the system for an exponentially distributed time (mean μ^{-1}), generating traffic at a unit rate while in the system, where the buffer is emptied at a constant rate nc. Let Q^n denote the steady-state buffer content of this queue; as before, it is assumed that $\varrho < c$. The following distributional equality is well known:

$$Q^n \stackrel{d}{=} \sup_{t\geq 0} A^n(t) - nct,$$

a relation usually attributed to Reich[21]. Define, for $\vartheta < \mu$,

$$\log N_t(\vartheta) = \varrho(\phi(\vartheta, t) - 1) + \psi(\vartheta, t), \tag{9}$$

where $\phi(\vartheta, t)$ and $\psi(\vartheta, t)$ are given in (7).

Lemma 3.3.2. *The following identity holds:*

$$-\inf_{f\in\mathscr{A}} \mathbb{I}(f) - (\varrho - c) - c\log\frac{c}{\varrho} = \lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}(Q^n > nx)$$
$$= -\inf_{t\geq 0}\sup_{\vartheta>0}\big(\vartheta(x+ct) - \log N_t(\vartheta)\big). \tag{10}$$

Proof. The first equality follows from reasoning as in section 13.2 of Ref.[23]. The decay rate of the steady-state probability $\mathbb{P}(Q^n > nx)$ can be rewritten as $-\inf \mathbb{I}(f)$, where the infimum is over all f that start off in ϱ at time $-\infty$, and for which the busy period in which overflow (over level x) is reached starts at time 0; if level x is reached at some time $t > 0$, this means that

$$\int_0^t f(s)ds = b + ct,$$

and in addition $f(s) \geq c$ for all $s \in [0,t]$, and $f(0) = c$. Now an elementary splitting argument yields that this decay rate can be decomposed into

$$-\inf_{f\in\mathscr{A}^-} \mathrm{I}\!\mathrm{I}(f) - \inf_{f\in\mathscr{A}} \mathrm{I}\!\mathrm{I}(f),$$

where $\mathscr{A}^- := \{f \,|\, f(-\infty) = \varrho, f(0) = c\}$. Using arguments as in section 13.1 of Ref.[23],

$$\inf_{f\in\mathscr{A}^-} \mathrm{I}\!\mathrm{I}(f) = (\varrho - c) + c\log\frac{c}{\varrho}.$$

This, and an application of Lemma 3.3.1, proves the first equality.

The second equality follows from Botvich and Duffield[4], as follows. Let $N_t(\vartheta)$ be the moment generating function of the work generated by a single Poisson stream of jobs (that is, with rate λ), each staying in the system for an exponentially distributed time (with mean μ^{-1}):

$$\log N_t(\vartheta) = \varrho(\phi(\vartheta,t) - 1) + \psi(\vartheta,t);$$

here it is used that the number of flows present at time 0 has a Poisson distribution with mean ϱ. According to Ref.[4], the decay rate of $\mathbb{P}(Q^n > nx)$ equals

$$-\inf_{t\geq 0}\sup_{\vartheta>0}\big(\vartheta(x+ct) - \log N_t(\vartheta)\big). \tag{11}$$

This implies the second equality. □

Now the decay rate $\alpha(x)$ follows immediately from Lemma 3.3.2, in conjunction with equation (8).

Proposition 3.3.1. *For $x \geq 0$,*

$$\alpha(x) = -\inf_{t\geq 0}\sup_{\vartheta>0}\big(\vartheta(x+ct) - \log N_t(\vartheta)\big) + (\varrho - c) + c\log\frac{c}{\varrho}.$$

As opposed to $\delta(x)$ and (as we will see later) $\nu(x)$, there is no explicit, closed-form available for $\alpha(x)$. It is, however, possible to explicitly characterize $\alpha(x)$ for $x \downarrow 0$ and $x \to \infty$. We define

$$\alpha_0^\star := 2\sqrt{2}\cdot\sqrt{\lambda\left(1 - \frac{\varrho}{c} + \frac{\varrho}{c}\log\frac{\varrho}{c}\right)}; \quad \alpha_\infty^\star := \mu - \frac{\lambda}{c};$$

$$\beta_\infty^\star := \frac{(c-\varrho)^2}{\varrho} + c - \varrho - c\log\frac{c}{\varrho}.$$

Proposition 3.3.2. *The asymptotic behavior of* $\alpha(x)$ *is given by*

$$\alpha(x) = -\alpha_0^\star \sqrt{x} - O(x) \quad \text{as } x \downarrow 0;$$
$$\alpha(x) = -\beta_\infty^\star - \alpha_\infty^\star x + o(1) \quad \text{as } x \to \infty.$$

Proof. The behavior around $x = 0$ follows directly from Mandjes and Kim[16] (see the remark on the open model in section 3), in conjunction with Lemma 3.3.2. It is readily verified that, in the notation used in that remark, $\vartheta_0 = \log(c/\varrho)$, and then it is a matter of evaluating the expressions.

The behavior for $t \to \infty$ follows immediately from the expression for $N_t(\vartheta)$ for t large, and Theorem 3 of Botvich and Duffield[4]. The latter result states that the decay rate of $\mathbb{P}(Q^n > nx)$ equals $-\bar{\beta}_\infty^\star - \alpha_\infty^\star x + o(1)$ for x large, where $\alpha_\infty^\star$ solves

$$\lim_{t\to\infty} \frac{1}{t} \log N_t(\vartheta) = c\vartheta,$$

i.e., $\alpha_\infty^\star = \mu - \lambda/c = \mu(1 - \varrho/c)$, and

$$\bar{\beta}_\infty^\star := -\lim_{t\to\infty} \left(\log N_t(\alpha_\infty^\star) - c\alpha_\infty^\star t \right) = \frac{(c-\varrho)^2}{\varrho}.$$

Now an application of Lemma 3.3.2 yields the stated. □

3.4. Number of Flows

We cast our problem of identifying the decay rate of $\mathbb{P}(N_{nc} > nx)$ into the large-deviations framework introduced earlier. We now use the discrete-time, sample-path large deviations. Immediately from (5),

$$v(x) = -\inf_{f\in\mathcal{N}} \mathbb{J}(f),$$

with $\mathcal{N} := \{f \mid \forall s \in [0, 2x] : f(s) \geq c, f(0) = c\}$. Analogously to the duration of the congestion period, we are looking for the "cheapest" path, that cannot be optimal if $f(2x) > c$, as otherwise even a longer congestion period could be obtained "for free." Hence $\inf_{f\in\mathcal{N}} \mathbb{J}(f) = \inf_{f\in\bar{\mathcal{N}}} \mathbb{J}(f)$, with

$$\bar{\mathcal{N}} := \{f \mid \forall s \in [0, 2x] : f(s) \geq c, f(0) = f(2x) = c\}.$$

Therefore we further study the following variational problem:

$$v(x) = -\inf_{f\in\bar{\mathcal{N}}} \int_0^{2x} J_{f(s)}(f'(s))\, ds.$$

Proposition 3.4.1. *For $x > 0$,*

$$v(x) = -xv^\star; \quad v^\star := 2\log\frac{\lambda+\mu c}{2\sqrt{\lambda\mu c}} = \log\frac{(\lambda+\mu c)^2}{4\lambda\mu c}.$$

Proof. We prove this result by subsequently establishing a lower bound and an upper bound. Define the path f_c through $f_c(s) = c$ for all $s \in [0, 2x]$. As $f_c \in \overline{\mathcal{N}}$, it follows that

$$\begin{aligned} v(x) \geq -\mathbb{J}(f_c) &= -2x \cdot \sup_\vartheta\left(-\log\left(\frac{\lambda}{\lambda+\mu c}e^{\vartheta} + \frac{\mu c}{\lambda+\mu c}e^{-\vartheta}\right)\right) \\ &= 2x \cdot \log\left(\frac{\lambda\sqrt{\mu c/\lambda} + \mu c/\sqrt{\mu c/\lambda}}{\lambda+\mu c}\right) = -2x \cdot \log\frac{\lambda+\mu c}{2\sqrt{\lambda\mu c}}; \end{aligned}$$

the optimizing ϑ equals $\vartheta^\star := \frac{1}{2}\log(\mu c/\lambda) = \frac{1}{2}\log(c/\varrho) > 0$. Hence we have proven the upper bound. On the other hand,

$$\begin{aligned} v(x) &\leq -\inf_{f\in\overline{\mathcal{D}}}\int_0^{2x}\left(\vartheta^\star f'(s) - \log\left(\frac{\lambda}{\lambda+\mu f(s)}e^{\vartheta^\star} + \frac{\mu f(s)}{\lambda+\mu f(s)}e^{-\vartheta^\star}\right)\right)\mathrm{d}s \\ &\overset{\text{(i)}}{=} -\inf_{f\in\overline{\mathcal{D}}}\int_0^{2x}\left(-\log\left(\frac{\lambda}{\lambda+\mu f(s)}e^{\vartheta^\star} + \frac{\mu f(s)}{\lambda+\mu f(s)}e^{-\vartheta^\star}\right)\right)\mathrm{d}s \\ &= \sup_{f\in\overline{\mathcal{D}}}\int_0^{2x}\log\left(\frac{1}{\lambda+\mu f(s)}\left(\lambda e^{\vartheta^\star} + \mu f(s)e^{-\vartheta^\star}\right)\right)\mathrm{d}s \\ &= \sup_{f\in\overline{\mathcal{D}}}\int_0^{2x}\log\left(\frac{\sqrt{\lambda\mu c}\,(1+f(s)/c)}{\lambda+\mu f(s)}\right)\mathrm{d}s \\ &\overset{\text{(ii)}}{\leq} 2x\log\left(\frac{2\sqrt{\lambda\mu c}}{\lambda+\mu c}\right). \end{aligned}$$

Here (i) is due to the fact that $f(0) = f(x) = c$ for all $f \in \overline{\mathcal{N}}$, and (ii) due to Lemma A.2. This yields the lower bound: all $f \in \overline{\mathcal{N}}$ yield at most decay rate $-xv^\star$, as desired. □

4. LARGE DEVIATIONS ANALYSIS OF THE GAUSSIAN COUNTERPART

So far, we have considered the asymptotics of $\mathbb{P}(D_{nc} > x)$, $\mathbb{P}(A_{nc} > nx)$, and $\mathbb{P}(N_{nc} > nx)$, using sample-path, large-deviations. In this section, we approximate $A^n(\cdot)$ by its so-called Gaussian counterpart $\overline{A}^n(\cdot)$, that is,

the superposition of n Gaussian processes, each with mean and variance given through

$$\mathbb{E}\overline{A}(t) = \varrho t, \quad v(t) := \mathbb{V}\mathrm{ar}\ \overline{A}(t) = \frac{2\lambda}{\mu^3}(t\mu - 1 + e^{-t\mu}),$$

cf. (3). This specific Gaussian process is known as the integrated Ornstein–Uhlenbeck (iOU) process. The procedure of replacing stochastic processes by their Gaussian counterpart was proposed and extensively motivated by Addie et al.[1]; for a further justification in the M/M/∞ case, also see Refs.[11,?] (section 2).

With $\overline{A}(\cdot)$ corresponding to a single iOU process with the mean and variance define above, it is observed that $\overline{A}(\cdot)$ is a genuine Gaussian counterpart of our original Markovian system, in the sense that the following two properties hold:

- In the first place, the "rate process" $\overline{X}(t) := \overline{A}'(t)$ is well defined (which is not the case for several other Gaussian processes such as fractional Brownian motion). This is a stationary Gaussian process (where $\overline{A}(\cdot)$ was a Gaussian process with stationary increments). It is readily verified that $\mathbb{E}\overline{X}(t) = \varrho$, and

$$\mathbb{V}\mathrm{ar}(\overline{X}(t)) = \lim_{\varepsilon \downarrow 0} \frac{v(t+\varepsilon) - v(t)}{\varepsilon^2} = \varrho.$$

  These results are in agreement with the fact that in the original (that is, non-Gaussian) model $X(t)$ has a Poisson distribution with mean (and hence also variance) ϱ.
- In the second place the Gaussian process has a Markovian structure, in the sense that, for $0 < u < T$ and $s > 0$,

$$\big(\overline{A}(T, T+s) \mid \overline{X}(T), \overline{A}(0,u)\big) \stackrel{\mathrm{d}}{=} \big(\overline{A}(T, T+s) \mid \overline{X}(T)\big).$$

  This follows by showing that both sides of the previous display have the same mean and variance. We briefly present the procedure for the mean; the variance can be done analogously. To this end, recall

$$\mathbb{E}\big(Y_1 \mid Y_2 = y_2, Y_3 = y_3\big) = \mathbb{E}Y_1 + \begin{pmatrix} \mathbb{C}\mathrm{ov}(Y_1, Y_2) \\ \mathbb{C}\mathrm{ov}(Y_1, Y_3) \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} y_2 - \mathbb{E}Y_2 \\ y_3 - \mathbb{E}Y_3 \end{pmatrix};$$

$$\mathbb{V}\mathrm{ar}\big(Y_1 \mid Y_2 = y_2, Y_3 = y_3\big) = \mathbb{V}\mathrm{ar}(Y_1) + \begin{pmatrix} \mathbb{C}\mathrm{ov}(Y_1, Y_2) \\ \mathbb{C}\mathrm{ov}(Y_1, Y_3) \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} \mathbb{C}\mathrm{ov}(Y_1, Y_2) \\ \mathbb{C}\mathrm{ov}(Y_1, Y_3) \end{pmatrix}$$

where

$$\Sigma := \begin{pmatrix} \mathbb{V}\mathrm{ar}(Y_2) & \mathbb{C}\mathrm{ov}(Y_2, Y_3) \\ \mathbb{C}\mathrm{ov}(Y_2, Y_3) & \mathbb{V}\mathrm{ar}(Y_3) \end{pmatrix}.$$

As a consequence, $\mathbb{E}(\overline{A}(T, T+s) \mid \overline{X}(T) = x, \overline{A}(0,u) = y)$ equals

$$\varrho s + \begin{pmatrix} \mathbb{C}\mathrm{ov}(\overline{A}(T, T+s), \overline{X}(T)) \\ \mathbb{C}\mathrm{ov}(\overline{A}(T, T+s), \overline{A}(0,u)) \end{pmatrix}^T \times \begin{pmatrix} \varrho & \mathbb{C}\mathrm{ov}(\overline{X}(T), \overline{A}(0,u)) \\ \mathbb{C}\mathrm{ov}(\overline{X}(T), \overline{A}(0,u)) & v(u) \end{pmatrix}^{-1} \begin{pmatrix} x - \varrho \\ y - \varrho u \end{pmatrix}.$$

It is a matter of straightforward computations to verify that

$$\mathbb{C}\mathrm{ov}\big(\overline{A}(T, T+s), \overline{X}(T)\big) = \frac{\lambda}{\mu^2}(1 - e^{-\mu s});$$

$$\mathbb{C}\mathrm{ov}\big(\overline{A}(T, T+s), \overline{A}(0,u)\big) = \frac{\lambda}{4\mu^3} e^{-\mu T}(1 - e^{-\mu s})(e^{\mu u} - 1);$$

$$\mathbb{C}\mathrm{ov}\big(\overline{X}(T), \overline{A}(0,u)\big) = \frac{\lambda}{4\mu^2}\, e^{-\mu T}(e^{\mu u} - 1).$$

Now tedious calculus yields that

$$\mathbb{E}\big(\overline{A}(T, T+s) \mid \overline{X}(T) = x, \overline{A}(0,u) = y\big) = \varrho s + \frac{1 - e^{-\mu s}}{\mu} \cdot (x - \varrho),$$

which is in agreement with $\mathbb{E}(\overline{A}(T, T+s) \mid \overline{X}(T) = x)$ (observe that, in particular, u and y cancel). Similarly, $\mathbb{V}\mathrm{ar}(\overline{A}(T, T+s) \mid \overline{X}(T), = x, \overline{A}(0,u) = y)$ coincides with $\mathbb{V}\mathrm{ar}(\overline{A}(T, T+s) \mid \overline{X}(T) = x)$ and equals $v(s) + (\lambda/\mu^3)(1 - e^{-\mu s})^2 = (2\lambda/\mu^3)(s\mu - 3/2 + 2e^{-s\mu} - e^{-2\mu s}/2)$.

Useful Relations. We give a number of additional useful relations:

$$\mathbb{E}\big(\overline{A}(0,t) \mid \overline{X}(0) = x\big) = \varrho t + \frac{(1 - e^{-t\mu})}{\mu}(x - \varrho).$$

$$\mathbb{V}\mathrm{ar}\big(\overline{A}(0,t) \mid \overline{X}(0) = x\big) = \frac{2\lambda}{\mu^3}\left(t\mu - \frac{3}{2} + 2e^{-t\mu} - \frac{1}{2}e^{-2t\mu}\right).$$

Now

$$\mathbb{E}\big(\overline{X}(t) \mid \overline{X}(0) = x\big) = \mathbb{E}\left(\lim_{\epsilon \downarrow 0} \frac{\overline{A}(0, t+\epsilon) - \overline{A}(0,t)}{\epsilon} \,\middle|\, \overline{X}(0) = x\right) = \varrho + e^{-t\mu}(x - \varrho)$$

entails that $\mathbb{E}(\overline{X}(\epsilon) \mid \overline{X}(0) = x) = x + \epsilon(\lambda - \mu x) + O(\epsilon^2)$, for $\epsilon \downarrow 0$, and likewise

$$\mathbb{V}\mathrm{ar}\big(\overline{X}(t) \mid \overline{X}(0) = x\big) = \mathbb{V}\mathrm{ar}\left(\lim_{\epsilon \downarrow 0} \frac{\overline{A}(0, t+\epsilon) - \overline{A}(0,t)}{\epsilon} \,\middle|\, \overline{X}(0) = x \right)$$
$$= \varrho(1 - e^{-2t\mu})$$

leads to $\mathbb{V}\mathrm{ar}(\overline{X}(\epsilon) \mid \overline{X}(0) = x) = 2\lambda\epsilon + O(\epsilon^2)$.

4.1. Sample-Path, Large-Deviations Theory

The computation of the decay rates $\bar{\delta}(x)$ of the congestion period, $\bar{\alpha}(x)$ of the area, and $\bar{\nu}(x)$ of the number of customers, can, as before, be done relying on a sample-path, large-deviations result. In the setting of Gaussian processes, this result is known as (the generalized version of) Schilder's theorem[2,7,15]. It is noted that this result is of a rather implicit nature, in that there is, in general, no closed-form expression for the action functional (that is, we do not have an explicit expression for the "cost" of a given path f). Owing to the fact that the iOU process has a well-defined rate process, however, the corresponding action functional can, for this specific Gaussian process, be expressed explicitly. The goal of this subsection is to identify this action functional—we do so by first heuristically deriving the sample-path, large-deviations result, which will be rigorized in the second part of this subsection.

Heuristic Approach. With $\overline{X}^n(t) := (\overline{A}^n)'(t)$, we focus on the likelihood that the sample mean $n^{-1}\overline{X}^n(\cdot)$ follows the function $f(\cdot)$ on the interval $[0, T]$, given the initial condition $n^{-1}\overline{X}^n(0) = x$. The function $f(\cdot)$ is evidently such that $f(0) = x$. Then we require, after discretizing time for $k = 1, \ldots, T/\triangle t$, that

$$\frac{1}{n}\overline{X}^n(k\triangle t) = f(k\triangle t)$$

for all k; the finer the grid, the better the approximation. Hence, we consider for $\triangle t$ small

$$\mathbb{P}\left(\frac{1}{n}\overline{X}^n(t) \approx f(t), \forall t \in [0,T] \,\middle|\, \frac{1}{n}\overline{X}^n(0) = x \right)$$
$$\approx \mathbb{P}\left(\frac{1}{n}\overline{X}^n(k\triangle t) \approx f(k\triangle t), \forall k \in \{1, \ldots, T/\triangle t\} \,\middle|\, \frac{1}{n}\overline{X}^n(0) = x \right).$$

By the Markovian property of the rate process, the previous display reads

$$\prod_{k=1}^{T/\triangle t} \mathbb{P}\left(\frac{1}{n}\overline{X}^n(k\triangle t) \approx f(k\triangle t) \,\middle|\, \frac{1}{n}\overline{X}^n((k-1)\triangle t) \approx f((k-1)\triangle t)\right)$$
$$= \prod_{k=1}^{T/\triangle t} \mathbb{P}\left(\frac{1}{n}\overline{X}^n(\triangle t) \approx f(k\triangle t) \,\middle|\, \frac{1}{n}\overline{X}^n(0) \approx f((k-1)\triangle t)\right),$$

for paths $f(\cdot)$ with $f(0) = x$. Relying on standard large-deviations results for the normal distribution, we thus obtain the decay rate

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\overline{X}^n(t) \approx f(t), \forall t \in [0,T] \,\middle|\, \frac{1}{n}\overline{X}^n(0) = x\right)$$
$$= \lim_{\triangle t \downarrow 0} \sum_{k=1}^{T/\triangle t} \frac{(f(k\triangle t) - \mathbb{E}(\overline{X}(\triangle t) \mid \overline{X}(0) = f((k-1)\triangle t)))^2}{2\mathbb{V}\mathrm{ar}\overline{X}(\triangle t) \mid \overline{X}(0) = f((k-1)\triangle t)}.$$

Applying the approximations of $\mathbb{E}(\overline{X}(\epsilon) \mid \overline{X}(0) = x)$ and $\mathbb{V}\mathrm{ar}(\overline{X}(\epsilon) \mid \overline{X}(0) = x)$ for small ϵ, as given above, this further reduces to

$$\lim_{\triangle t \downarrow 0} \frac{1}{2} \sum_{k=1}^{T/\triangle t} \frac{(f(k\triangle t) - f((k-1)\triangle t) - \triangle t(\lambda - \mu f((k-1)\triangle t)))^2}{2\lambda \triangle t}$$
$$= \lim_{\triangle t \downarrow 0} \frac{1}{4\lambda} \sum_{k=1}^{T/\triangle t} \triangle t \left(\frac{(f(k\triangle t) - f((k-1)\triangle t))}{\triangle t} - (\lambda - \mu f((k-1)\triangle t))\right)^2$$
$$= \lim_{\triangle t \downarrow 0} \frac{1}{4\lambda} \sum_{k=1}^{T/\triangle t} \triangle t (f'((k-1)\triangle t) - \lambda + \mu f((k-1)\triangle t))^2$$
$$= \frac{1}{4\lambda} \int_0^T (f'(t) - \lambda + \mu f(t))^2 \mathrm{d}t.$$

Hence, the candidate rate function of a path f is

$$\overline{\mathbb{I}}(f) = \int_0^T \overline{I}_{f(s)}(f'(s))\mathrm{d}s, \quad \text{where } \overline{I}_x(u) := \frac{(u - \lambda + \mu x)^2}{4\lambda}.$$

So far we have considered paths on $[0,T]$, that start in x at time 0. Extending the argument to paths on $(-\infty, \infty)$, the candidate for the rate function would become

$$\overline{\mathbb{I}}(f) = \int_{-\infty}^{\infty} \overline{I}_{f(s)}(f'(s))\mathrm{d}s. \tag{12}$$

The remainder of this subsection is devoted to a formal approach to establishing (12) by applying the generalized version of Schilder's theorem.

Sample-Path, Large-Deviations Principle. For any Gaussian process with stationary increments $\overline{A}(\cdot)$, the generalized version of Schilder's theorem states that, under mild conditions,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \overline{A}^n(\cdot) \in \mathcal{S} \right) = - \inf_{F\in\mathcal{S}} \mathbb{K}(F); \tag{13}$$

as before, we should formally distinguish between open and closed sets $\mathcal{S}$. In general, the action functional $\mathbb{K}(F)$ is only explicitly given for paths $F(\cdot)$ that are mixtures of covariance functions: if, for $\alpha_i, s_i \in \mathbb{R}$, and $\Gamma(s,t) := \mathbb{C}\mathrm{ov}(\overline{A}(s), \overline{A}(t))$, the path $F(s)$ is of the form $\sum_{i=1}^{d} \alpha_i \Gamma(s, s_i)$. Then

$$\mathbb{K}(F) = \sum_{i=1}^{d} \sum_{j=1}^{d} \alpha_i \alpha_j \Gamma(s_i, s_j).$$

Also for $F(\cdot)$ that are not given as a mixture of covariance functions, one can determine $\mathbb{K}(F)$ by approximating $F(\cdot)$ by a mixture of covariance functions, and by using a limiting procedure—we leave out details here.

In case $\overline{A}(\cdot)$ has a derivative, then one could also consider large deviations probabilities that relate to the *rate process* $\overline{X}^n(\cdot)$ rather than the cumulative traffic process $\overline{A}^n(\cdot)$. With $f(s) := F'(s)$,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \overline{X}^n(\cdot) \in \mathcal{S} \right) = - \inf_{f\in\mathcal{S}} \mathbb{K}(F). \tag{14}$$

In order to rigorize (12), we show that for $F(\cdot)$ being a linear combination of covariance functions, we have that indeed

$$\mathbb{K}(F) = \frac{1}{4\lambda} \int_{-\infty}^{\infty} \big(f'(t) - \lambda + \mu f(t) \big)^2 \mathrm{d}t. \tag{15}$$

It is elementary to show that, using the shorthand notation $\Gamma_i(t) := \Gamma(t, s_i)$, the right-hand side of the previous display equals

$$\frac{1}{2\lambda} \sum_{i=1}^{d} \sum_{j=1}^{d} \alpha_i \alpha_j \int_{-\infty}^{\infty} \big(\Gamma_i''(t) + \mu \Gamma_i'(t) \big) \big(\Gamma_j''(t) + \mu \Gamma_j'(t) \big) \mathrm{d}t.$$

Then (15) indeed follows from the next lemma.

Lemma 4.1.1. *For all $i, j = 1, \ldots, d$,*

$$\int_{-\infty}^{\infty} \big(\Gamma_i''(t) + \mu\Gamma_i'(t)\big)\big(\Gamma_j''(t) + \mu\Gamma_j'(t)\big)\mathrm{d}t = 2\lambda\Gamma(s_i, s_j).$$

Proof. See Appendix. □

4.2. Congestion Period

The decay rate of the congestion period can be found in the same fashion as in section 3.2, the only major difference being that now we should rely on the sample-path, large-deviations result for iOU traffic.

Proposition 4.2.1. *For $x \geq 0$,*

$$\bar{\delta}(x) = -x\bar{\delta}^\star; \quad \bar{\delta}^\star := \frac{(\mu c - \lambda)^2}{4\lambda}.$$

Proof. We prove this result by first establishing a lower bound. Clearly, $\bar{\delta}(x) \geq \overline{\mathrm{I\!I}}(f_c)$, recalling that $f_c(s) = c$ for all $s \in [0, x]$. Evaluating $\overline{\mathrm{I\!I}}(f_c)$, we find the lower bound.

Now focus on the upper bound. Straightforward algebra yields that

$$\bar{\delta}(x) = -\inf_{f\in\overline{\mathscr{D}}} \int_0^x \left(\frac{(f'(s))^2}{4\lambda} - \frac{1}{2}f'(s) + \frac{1}{4}\frac{f(s)f'(s)}{\lambda} + \frac{\lambda}{4} + \frac{\mu^2}{4\lambda}f^2(s) - \frac{1}{2}\mu f(s)\right)\mathrm{d}s.$$

Evidently, the fact that $f(0) = f(x) = c$ (for all $f \in \overline{\mathscr{D}}$) entails that

$$\int_0^x \frac{1}{2}f'(s)\mathrm{d}s = \int_0^x \frac{1}{4}\frac{f(s)f'(s)}{\lambda}\mathrm{d}s = 0,$$

which immediately leads to

$$\bar{\delta}(x) \leq -\inf_{f\in\overline{\mathscr{D}}} \int_0^x \left(\frac{\mu^2}{4\lambda}f^2(s) - \frac{1}{2}\mu f(s)\right)\mathrm{d}s - \frac{\lambda}{4}x.$$

The right-hand side of the previous display is smaller than $-x\bar{\tau}^\star$, as follows, after elementary algebra, from the inequality

$$\frac{\mu^2}{4\lambda}(y^2 - c^2) = \frac{\mu^2}{4\lambda}(y + c)(y - c) \geq \frac{\mu^2}{4\lambda} \cdot 2c \cdot (y - c) = \frac{\mu}{2} \cdot \frac{c}{\varrho}(y - c) \geq \frac{\mu}{2}(y - c),$$

for all $y \geq c$. This completes the upper bound. □

Remark 4.2.1. We now consider the so-called heavy-traffic regime $c = \varrho + \epsilon$ for ϵ small, and we show that $\delta^\star$ and $\bar{\delta}^\star$ are very much alike. In other words, in heavy-traffic the Gaussian approximation is particularly accurate. The formal calculation is as follows. It is easily checked that

$$\delta^\star = \mu\left(\frac{\epsilon^2}{4\varrho} + \frac{\epsilon^3}{8\varrho^2}\right) + O(\epsilon^4), \quad \text{and} \quad \bar{\delta}^\star = \mu\left(\frac{\epsilon^2}{4\varrho}\right),$$

as $\epsilon \downarrow 0$. For related results also see Ref.[11] (section 5.2).

4.3. Area

We now consider the decay rate $\bar{\alpha}(x)$ of the area exceeding an amount x, which is obtained in a similar manner as was done for the M/M/∞ model in section 3.3. Again exploiting the relation with the tail probabilities of an appropriately chosen queueing system, and the large-deviations results by Botvich and Duffield[4], we obtain the following proposition. As its proof is identical to that of Proposition 3.3.1, we leave it out. Realize that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\overline{X}^n(0) \geq c\right) = -\frac{(c-\varrho)^2}{2\varrho}.$$

Proposition 4.3.1. *For $x \geq 0$,*

$$\bar{\alpha}(x) = -\inf_{t\geq 0} \frac{(x + (c-\varrho)t)^2}{2v(t)} + \frac{(c-\varrho)^2}{2\varrho}.$$

There is no explicit, closed-form available for $\alpha(x)$. It is, however, possible to explicitly characterize $\alpha(x)$ in the asymptotic regimes $x \downarrow 0$ and $x \to \infty$.

Proposition 4.3.2. *The asymptotic behavior of $\alpha(x)$ is given by*

$$\bar{\alpha}(x) = -\sqrt{2\lambda/3} \cdot \left(\frac{c-\varrho}{\varrho}\right)^{3/2} \cdot \sqrt{x} + O(x) \quad \text{as } x \downarrow 0;$$

$$\bar{\alpha}(x) = -\frac{\mu(c-\varrho)}{\rho} x - \frac{(c-\varrho)^2}{2\rho} \quad \text{as } x \to \infty.$$

Proof. First consider the regime $x \downarrow 0$. The infimum will be reached for $t(x)$ close to 0, in which case the variance function $v(t)$ behaves as

$$\varrho t^2 - \lambda t^3/3 + O(t^4).$$

By virtue of equation (6.6) in Ref.[15],

$$\inf_{t\geq 0}\frac{(x+(c-\varrho)t)^2}{2v(t)}=\frac{(c-\varrho)^2}{2\varrho}+\sqrt{2\lambda/3}\cdot\left(\frac{c-\varrho}{\varrho}\right)^{3/2}\cdot\sqrt{x}+O(x).$$

Now focus on $x\to\infty$. We again use Theorem 3 of Botvich and Duffield[4], which implies that

$$\inf_{t\geq 0}\frac{(x+(c-\varrho)t)^2}{2v(t)}=\frac{(c-\varrho)^2}{\rho}+\frac{\mu(c-\varrho)}{\rho}x+o(1)$$

for x large. Now an application of Lemma 3.3.2 yields the stated.

Remark 4.3.1. There is not a Gaussian equivalent of an "arrival". We therefore do not have a Gaussian counterpart of the decay rate $v(x)$. See, however, the appendix in Ref.[11], where it is pointed out how an "artificial" Gaussian counterpart can be constructed (which lacks a straightforward interpretation).

5. UNIFORM BOUNDS

In the previous sections, we computed the decay rates of the probabilities of interest, but these do not provide us with estimates of the probabilities themselves. In particular, a statement of the type $n^{-1}\log f(n)\to-\zeta$ just says that $f(n)=g(n)\exp(-\zeta n)$ for some subexponential function $g(\cdot)$, that is, $\log g(n)=o(n)$ as $n\to\infty$; the function $g(n)$ can still be of the form $\exp(n^{1-\delta})$ for a small positive constant δ. For practical purposes, conservative (but preferably tight) estimates of the probabilities of interest are useful. In this section such approximations are derived. They indicate that the logarithmic estimates are rather precise.

5.1. Congestion Period

We consider the original, that is, unscaled, model. The exponential part in the following bound should be compared with the logarithmic asymptotics found in section 3.2.

Proposition 5.1.1. *Uniformly in $x\geq 0$,*

$$\mathbb{P}(D_c>x)\leq\left(\sqrt{\frac{c}{p}}\right)^{c+1}\cdot\exp\left(-(\sqrt{\Lambda}-\sqrt{c\mu})^2x\right).$$

Proof. It is clear that $\mathbb{P}(D_c > x) \le \mathbb{P}(A(x) > \mathrm{c}x \mid X(0) = \mathrm{c}+1)$. The Markov inequality yields that, for any $\vartheta > 0$,

$$\mathbb{P}\big(A(x) > \mathrm{c}x \mid X(0) = \mathrm{c}+1\big) \le \mathbb{E}\big(e^{\vartheta A(x)} \mid X(0) = \mathrm{c}+1\big) e^{-\vartheta \mathrm{c}x}.$$

Applying (2), we obtain

$$\mathbb{P}\big(A(x) > \mathrm{c}x \mid X(0) = \mathrm{c}+1\big) \le \left(\frac{\mu}{\mu-\vartheta}\right)^{\mathrm{c}+1} \exp\left(\vartheta\left(\frac{\Lambda}{\mu-\vartheta} - \mathrm{c}\right)x\right).$$

Now plug in $\vartheta = \vartheta^\star := \mu - \sqrt{\Lambda\mu/\mathrm{c}} > 0$. □

A slightly different bound can be found as follows. We include it here, as it gives us insight into the way importance-sampling algorithms might be devised. Suppose we wish to estimate $\mathbb{P}(D_c > x)$ by simulation, applying importance sampling with arrival rate $\Lambda^\star := \sqrt{\Lambda\mu\mathrm{c}}$ and service rate $\mu^\star := \sqrt{\Lambda\mu/\mathrm{c}}$, irrespective of the number of flows present; call the new measure $\mathbb{Q}$. It is elementary that, in self-evident notation,

$$\mathbb{P}(D_c > x) = \mathbb{E}_{\mathbb{Q}} LI,$$

where L is the so-called likelihood ratio, and I the indicator function of the event under consideration. In more detail, the likelihood ratio can be expressed as follows. Let τ_i denote the ith jump of the congestion period, i.e., τ_i is 1 if the ith jump is upward and 0 if it is downward. (With τ_0 we mean the jump to level $\mathrm{c}+1$ that starts the congestion period.) Let Z_i denote the state (i.e., the number of flows present) between the ith and $(i+1)$st jump, and S_i the time between these jumps. Then, with N denoting the last jump before time x, realizing that the $(N+1)$st jump epoch is the first jump epoch at which we are certain that it can decided whether indeed $D_c > x$,

$$L = \prod_{i=0}^{N} \frac{(\Lambda+\mu Z_i)\exp(-(\Lambda+\mu Z_i)S_i)}{(\Lambda^\star+\mu^\star Z_i)\exp(-(\Lambda^\star+\mu^\star Z_i)S_i)} \prod_{i=0}^{N-1} (p(Z_i))^{\tau_{i+1}} (q(Z_i))^{1-\tau_{i+1}},$$

where

$$p(k) := \left(\frac{\Lambda}{\Lambda+\mu k}\right) \Big/ \left(\frac{\Lambda^\star}{\Lambda^\star+\mu^\star k}\right), \quad q(k) := \left(\frac{\mu x}{\Lambda+\mu k}\right) \Big/ \left(\frac{\mu^\star x}{\Lambda^\star+\mu^\star k}\right).$$

Elementary calculus yields that this likelihood equals

$$\frac{\Lambda+\mu Z_N}{\sqrt{\Lambda\mu\mathrm{c}}+\sqrt{\Lambda\mu/\mathrm{c}}\,Z_N} \times \exp\left(-\sum_{i=0}^{N}\left(\Lambda+\mu Z_i-\sqrt{\Lambda\mu\mathrm{c}}-\sqrt{\Lambda\mu/\mathrm{c}}\,Z_i\right)S_i\right)\left(\sqrt{\frac{p}{\mathrm{c}}}\right)^{Z_N-(\mathrm{c}+1)}.$$

Relying on Lemma A.1, it is elementary to show that, as long as $Z_i > \mathrm{c}$,

$$\Lambda + \mu Z_i - \sqrt{\Lambda\mu\mathrm{c}} - \sqrt{\Lambda\mu/\mathrm{c}}Z_i \geq \Lambda - 2\sqrt{\Lambda\mu\mathrm{c}} + \mu\mathrm{c}.$$

Now observe that during a run in which $I = 1$, $Z_i > \mathrm{c}$ for all $i \in \{0, \ldots, N\}$, and $\sum_{i=0}^{N} S_i > x$ as well as $Z_N > \mathrm{c}$. Also, due to $p < \mathrm{c}$,

$$\frac{\Lambda + \mu Z_N}{\sqrt{\Lambda\mu\mathrm{c}} + \sqrt{\Lambda\mu/\mathrm{c}}Z_N} \leq \sqrt{\frac{\mathrm{c}}{p}}.$$

We thus find the upper bound

$$\mathbb{P}(D_{\mathrm{c}} > x) \leq \frac{\mathrm{c}}{p} \cdot \exp\bigl(-(\sqrt{\Lambda} - \sqrt{\mathrm{c}\mu})^2 x\bigr). \tag{16}$$

Note that for $\mathrm{c} > 2$, this bound is sharper than the one we presented in Proposition 5.1.1.

5.2. Area

A similar argument can be used to find a uniform upper bound on $\mathbb{P}(A_{\mathrm{c}} > x)$.

Proposition 5.2.1. *Uniformly in* $x \geq 0$,

$$\mathbb{P}(A_{\mathrm{c}} > x) \leq \left(\frac{\mathrm{c}}{p}\right)^2 \cdot \exp\left(-\left(\mu - \frac{\Lambda}{\mathrm{c}}\right)x\right).$$

Proof. First observe that

$$\mathbb{P}(A_{\mathrm{c}} > x) \leq \mathbb{P}(\exists t \geq 0 : A(t) > x + \mathrm{c}t).$$

Let us find an upper bound for the right-hand side of the previous display. Suppose we perform importance sampling under a measure $\mathbb{Q}$, that is, such that the arrival rate is $\Lambda^\star := \mu\mathrm{c}$ and service rate $\mu^\star := \Lambda/\mathrm{c}$. It is clear that the resulting system is such that under the new measure, $A(t)$ indeed crosses level $x + \mathrm{c}t$ with probability 1, as the mean rate under $\mathbb{Q}$ is $\Lambda^\star/\mu^\star = \mathrm{c}^2\mu/\Lambda = \mathrm{c}^2/p > \mathrm{c}$.

A fundamental equality is, with $\mathbb{E}_{\mathbb{Q}}$ denoting expectation under $\mathbb{Q}$,

$$\mathbb{P}(\exists t \geq 0 : A(t) > x + \mathrm{c}t) = \mathbb{E}_{\mathbb{Q}} L,$$

where L is the so-called likelihood ratio. In more detail, the likelihood ratio can be expressed as follows.

Using the same definitions as in the previous section,

$$L = \prod_{i=0}^{N} \frac{(\Lambda + \mu Z_i)\exp(-(\Lambda + \mu Z_i)S_i)}{(\Lambda^\star + \mu^\star Z_i)\exp(-(\Lambda^\star + \mu^\star Z_i)S_i)} \prod_{i=0}^{N-1} (p(Z_i))^{I_{i+1}} (q(Z_i))^{1-I_{i+1}},$$

elementary calculus yields that

$$L = \frac{\Lambda + \mu Z_N}{\mu \mathrm{c} + \Lambda Z_N/\mathrm{c}} \exp\left(-\left(\mu - \frac{\Lambda}{\mathrm{c}}\right)\left(A\left(\sum_{i=0}^{N} S_i\right) - \mathrm{c}\cdot\sum_{i=0}^{N} S_i\right)\right)\left(\frac{p}{\mathrm{c}}\right)^{Z_N-(\mathrm{c}+1)}.$$

Due to $p < \mathrm{c}$, it holds that

$$\frac{\Lambda + \mu Z_N}{\mu \mathrm{c} + \Lambda Z_N/\mathrm{c}} \le \frac{\mathrm{c}}{p}.$$

As we know that $Z_N \ge \mathrm{c}$, and by definition of N,

$$A\left(\sum_{i=0}^{N} S_i\right) - \mathrm{c}\cdot\sum_{i=0}^{N} S_i \ge x,$$

the upper bound follows. □

5.3. Number of Flows

Finally, we use the change-of-measure technique to find a uniform upper bound on $\mathbb{P}(N_{\mathrm{c}} > m)$. We start, however, by a result that can be proven in a more elementary way.

Proposition 5.3.1. *Uniformly in* $x \in \mathbb{N}$,

$$\mathbb{P}(N_{\mathrm{c}} > x) \le \left(\frac{4\Lambda\mu\mathrm{c}}{(\Lambda + \mu\mathrm{c})^2}\right)^x.$$

Proof. First observe that, stochastically, $Y_{2x} \le Y'_{2x} := \mathrm{c} + \sum_{i=1}^{2x} Z_i$, where the Z_i are i.i.d., and $Z_i = 1$ with probability $\Lambda/(\Lambda + \mu\mathrm{c})$ and -1 otherwise. By the Markov inequality, it follows that, for any $\vartheta \ge 0$,

$$\mathbb{P}(N_{\mathrm{c}} > m) \le \mathbb{P}\left(\sum_{i=1}^{2x} Z_i \ge 0\right) \le \left(\mathbb{E}e^{\vartheta Z}\right)^{2x},$$

with Z distributed as the Z_i. Now minimize the last expression over all $\vartheta \ge 0$, and the desired follows. □

As previously mentioned, a similar bound can be found by an importance-sampling argumentation. Simulate the discrete-time process (i.e., the jump process) that results after changing Λ into $\Lambda^\star := \sqrt{\Lambda\mu c}$ and $\mu^\star := \sqrt{\Lambda\mu/c}$ until either the process drops below the value c, or $2x$ transitions have been performed. It is readily checked that the likelihood at this stopping epoch equals

$$L = \left(\sqrt{\frac{p}{c}}\right)^{Z_N-(c+1)} \prod_{i=0}^{N} \frac{\sqrt{\Lambda\mu c} + \sqrt{\Lambda\mu/c}\, Z_i}{\Lambda + \mu Z_i}.$$

Applying Lemma A.2, it is elementary to show that

$$\frac{\sqrt{\Lambda\mu c} + \sqrt{\Lambda\mu/c}\, Z_i}{\Lambda + \mu Z_i} \le 2\frac{\sqrt{\Lambda\mu c}}{\Lambda + \mu c}.$$

Using that, if $I = 1$, then $N \ge 2x$ and $Z_N > c$, we find the same upper bound as above, but now multiplied with $\sqrt{c/p}$, i.e., slightly weaker.

6. NUMERICAL RESULTS

In this section, we demonstrate our asymptotics and bounds through a number of numerical experiments. In these experiments, we choose $\mu = 1$ and $c = 1$, and we compare the situation $\lambda = 0.5$ with $\lambda = 0.9$. The primary goal of this section is to present a comparison between the rough asymptotics of sections 3 and 4, the bounds of section 5, and the "real" values.

A number of remarks need to be made here.

- The results in sections 3 and 4 are in terms of decay rates, and in order to compare them we do as if the decay is "purely exponential." For instance for the congestion duration, the resulting approximation, based on Proposition 3.2.1, is, with as before $\Lambda = n\lambda$ and $c = nc$,

$$\mathbb{P}(D_c > x) \approx \exp\bigl(-(\sqrt{\Lambda} - \sqrt{\mu c})^2 x\bigr), \tag{17}$$

  cf. Proposition 5.1.1. In case of the area, this approximation is somewhat trickier to derive; we now sketch how the approximation for $\mathbb{P}(A_c > x)$ can be found. Focusing for the moment on the regime $x \to \infty$, Proposition 3.3.2 entails that

$$\mathbb{P}(A_{nc} > nx) \approx \exp\bigl(-n\beta^\star_\infty - n\alpha^\star_\infty x\bigr).$$

Noticing that

$$\beta^{\star}_{\infty} = \frac{1}{n}\left(\frac{(c-p)^2}{p} + c - p - c\log\frac{c}{p}\right); \quad \alpha^{\star}_{\infty} = \mu - \frac{\Lambda}{c},$$

we obtain the approximation

$$\mathbb{P}(A_c > x) \approx \exp\left(-\frac{(c-p)^2}{p} - c + p + c\log\frac{c}{p} - \left(\mu - \frac{\Lambda}{c}\right)x\right);$$

cf. Proposition 5.2.1. In the regime $x \downarrow 0$ an analogous argumentation yields

$$\mathbb{P}(A_c > x) \approx \exp\left(-2\sqrt{2}\cdot\sqrt{x\Lambda\left(1 - \frac{p}{c} + \frac{p}{c}\log\frac{p}{c}\right)}\right).$$

The Gaussian counterparts can be dealt with similarly.

- To obtain "real" values of the probabilities of interest, we used importance sampling-based simulations, with the change-of-measures suggested in section 5. We also performed direct simulations (that is, simulations under the original measure), where we empirically observed that under importance sampling, substantially less simulation effort is needed to obtain an estimate of given precision; for higher values of x direct simulation becomes prohibitively time-consuming. The outcomes of these direct simulations (in the graphs corresponding to the label "M/M/∞") coincide with the importance sampling-based estimates, as should be the case.

For the congestion duration D_c we consider the tail probabilities for $n = 20, 50, 100$ in Figure 1. In the numerical results we compare Proposition 3.2.1 for the M/M/∞ process, Proposition 4.2.1 for the Gaussian counterpart, the uniform upper bound given by Expression (16), and results from importance-sampling simulations as well as direct simulations of the M/M/∞ process.

In each of the graphs, we see that the uniform upper bound of Expression (16) is slightly larger than Proposition 3.2.1, namely, a factor c/p. The result from the Gaussian counterpart (Proposition 4.2.1) is in between the results for the M/M/∞ and the simulation results for low load, but for high load it is close to Proposition 3.2.1 and Expression (16). Observe that for higher loads (right graphs), the probabilities of a long congestion duration are higher, which is evident as these occurrences become less rare. By comparing the graphs of Figure 1 it is seen that increasing the scaling parameter n indeed leads to smaller probabilities.

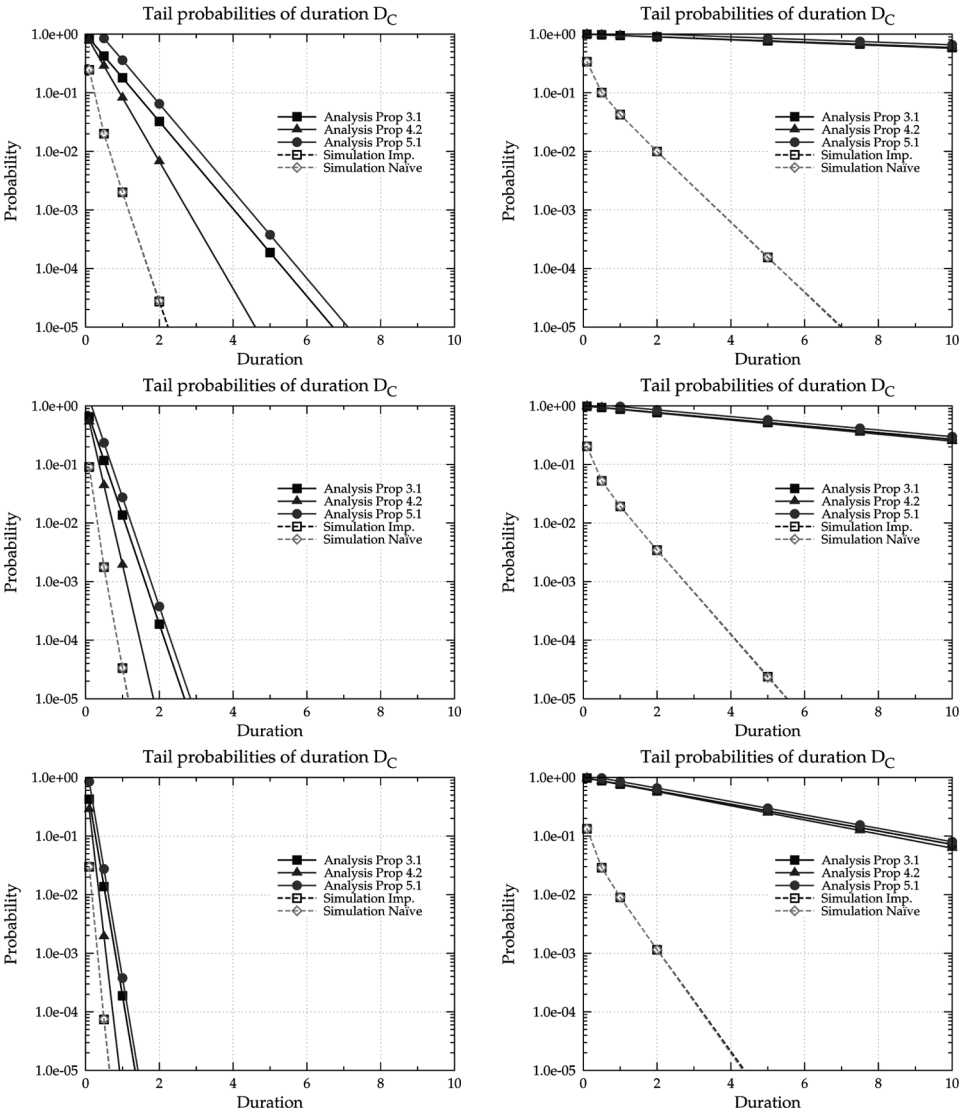


FIGURE 1 Congestion duration for $\mu = 1$, and $c = 1$. Left: $\lambda = 0.5$; right: $\lambda = 0.9$; top: $n = 20$; middle: $n = 50$; bottom: $n = 100$.

Given our analytical results, the curve with simulated probabilities should eventually be (that is, for n large) parallel to the curves obtained from Proposition 3.2.1 and Expression (16), which is evidently not yet the case for $n = 100$. A similar slow convergence has been observed for the tail asymptotics of the sojourn time in processor-sharing (PS) queues in, e.g., Ref.[18]. It may also play a role that, just as is the case for the sojourn-time distribution in the M/M/1 PS queue, the asymptotics are likely to

be not of a "purely exponential" form (as suggested by (17)); instead there may be in addition a polynomial factor $\delta x^{-\gamma}$ (for some $\gamma, \delta > 0$), and potentially also a Weibullian factor $\exp(-\alpha x^{\beta})$ (for some $\alpha > 0$ and $\beta \in (0, 1)$), cf. Refs.[3,8].

The figures show that the results obtained from Proposition 3.2.1 and Expression (16) can, in practical situations, only be used as (very rough) indications of the probability of interest. In case quick, reliable estimates are required (for instance, for dimensioning purposes), we advise relying on the described (efficient) importance sampling scheme.

The results for the area A_c are displayed in Figure 2; we only present the result for $n = 20$, as the effect of increasing n is similar as for the duration. The graphs compare the results of Proposition 3.3.2 for the M/M/∞ process, Proposition 4.3.2 for the Gaussian counterpart, the uniform upper bound of Proposition 5.2.1 and the simulation results, both from direct simulations and importance sampling. Recall that Propositions 3.3.2 and 4.3.2 include both the behavior of x close to 0, and x large, respectively; therefore in the graphs there are two curves for each proposition, and it is emphasized that these curves are not valid for the entire range of x. The uniform upper bound corresponds to the "highest" curve, as expected. For low loads, all curves are relatively close, and the simulation results are in-between the other mentioned results; the latter property is in contrast with the results for the duration and the number of arrivals, for which the probabilities from the simulation are always the smallest. It can be seen that in the low-load case, the part of Proposition 3.3.2 corresponding to $x \to \infty$ is already highly accurate for moderate x.

In Figure 3 the tail probabilities of the number of arrivals N_c are considered, again for $n = 20$. We compare the results from Proposition 3.4.1, the uniform upper bound from Proposition 5.3.1,

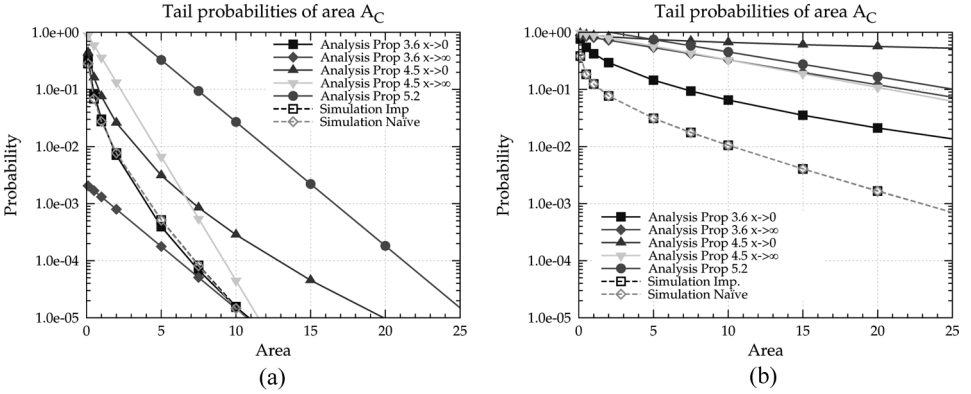


FIGURE 2 Area for $n = 20$, $\mu = 1$, and $c = 1$. (a) $\lambda = 0.5$ and (b) $\lambda = 0.9$.

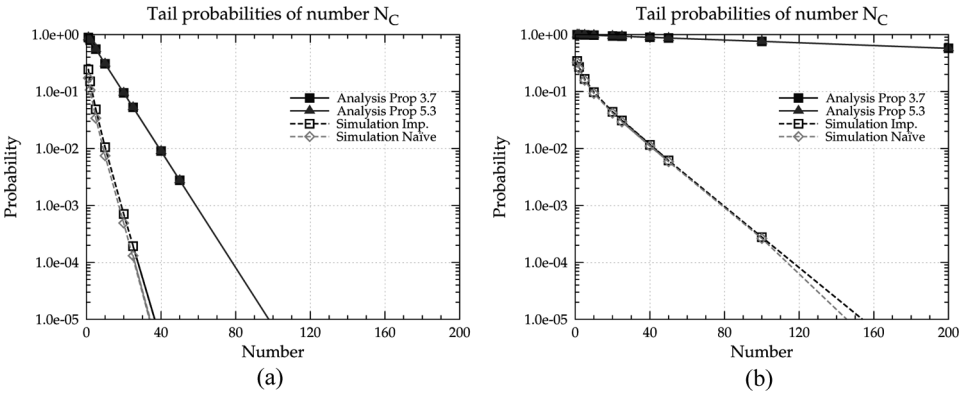


FIGURE 3 Number of arrivals for $n = 20$, $\mu = 1$, and $c = 1$. (a) $\lambda = 0.5$ and (b) $\lambda = 0.9$.

and simulation results; recall that in this case there is no meaningful Gaussian counterpart. Propositions 3.4.1 and 5.3.1 lead to the same expression, viz.,

$$\mathbb{P}(N_c > x) \approx \left(\frac{4\Lambda\mu c}{(\Lambda + \mu c)^2}\right)^x,$$

as is easily checked. Furthermore, observe that the propositions grossly overestimate the real (that is, simulated) probabilities, as is immediately clear after inspection of the simulation results (just as was the case for the congestion duration). We again advise using the proposed importance sampling scheme to quickly generate reliable estimates of the probability of interest.

7. DISCUSSION AND CONCLUDING REMARKS

This article considered tail asymptotics of congestion period-related quantities. Large-deviations theory is applied to explicitly calculate the exponential decay rates under the many-flows scaling, both for the actual M/M/∞ model and its Gaussian counterpart. Then uniform upper bounds on the tail probabilities are derived, which also reveal an efficient change-of-measure to be used in importance-sampling simulations. There are several directions for future research, of which we now mention a few.

We derived tail asymptotics under the many-flows scaling (x fixed, n large). These are presumably tightly related to the asymptotics of $\mathbb{P}(D_c > x)$, $\mathbb{P}(A_c > x)$, and $\mathbb{P}(N_c > x)$ for x large. As the LTs of these three random variables are known[14], one may attempt to obtain from these the corresponding tail asymptotics, cf. also[11,12]. In addition, we saw that our asymptotic results and bounds not necessarily lead, for given n, x, to

accurate approximations of the tail probabilities of interest, and therefore one may investigate techniques to improve on this (for values of n, x of practical interest), cf. the results in Ref.[10] (section 5).

We saw that for iOU Gaussian processes, the generalized-Schilder-based large deviations rate function of a path f could be computed explicitly. One may wonder for which class within the family of Gaussian processes similar explicit expressions can be derived; one may expect that these should be such that, like is the case for iOU, the corresponding rate process is well defined, but it is not *a priori* clear what additional conditions should be imposed.

Empirically, we observed that the proposed importance-sampling algorithms led to a substantial speed up: in order to obtain estimates with a predefined level of precision, the simulation time needed was reduced significantly. We expect that the proposed change-of-measures are actually asymptotically efficient. A proof of this property is beyond the scope of the present article.

A. APPENDIX

A.1. A Few Elementary Inequalities

The following, useful lemmas are straightforward to prove.

Lemma A.1. *For all $\alpha, \beta > 0$ with $\alpha < \beta$, and $y > c$,*

$$-\sqrt{\alpha\beta} + \alpha - \frac{y}{c}\sqrt{\alpha\beta} + \frac{y}{c}\beta \geq \left(\sqrt{\alpha} - \sqrt{\beta}\right)^2.$$

Lemma A.2. *For all $\alpha, \beta > 0$ with $\alpha < \beta$, and $y > c$,*

$$\frac{1 + y/c}{\alpha + \beta y/c} \leq \frac{2}{\alpha + \beta}.$$

A.1. Rate Function

We here present the proof of Lemma 4.1.1.

Proof of Lemma 4.1.1.

$$\Gamma_i(t) = \frac{\lambda}{\mu^3} \times \begin{cases} 1 - e^{-|t|\mu} + e^{-(|t|+s_i)\mu} - e^{-s_i\mu} & \text{for } t \leq 0, \\ e^{-t\mu} + e^{-s_i\mu} - e^{-(s_i-t)\mu} - 1 + 2t\mu & \text{for } t \in (0, s_i), \\ e^{-t\mu} + e^{-s_i\mu} - e^{-(t-s_i)\mu} - 1 + 2s_i\mu & \text{for } t \geq s_i. \end{cases}$$

$$\Gamma_i'(t) = \frac{d}{dt}\Gamma_i(t) = \frac{\lambda}{\mu^3} \times \begin{cases} -\mu e^{-|t|\mu} + \mu e^{-(|t|+s_i)\mu} & \text{for } t \leq 0, \\ -\mu e^{-t\mu} - \mu e^{-(t_i-s)\mu} + 2\mu & \text{for } t \in (0, s_i), \\ \mu e^{-t\mu} - \mu e^{-(t-s_i)\mu} & \text{for } t \geq s_i. \end{cases}$$

$$\Gamma_i''(t) = \frac{d^2}{dt^2}\Gamma_i(t) = \frac{\lambda^2}{\mu^3} \times \begin{cases} -\mu^2 e^{-|t|\mu} + \mu^2 e^{-(|t|+s_i)\mu} & \text{for } t \le 0, \\ \mu^2 e^{-t\mu} - \mu^2 e^{-(s_i-t)\mu} & \text{for } t \in (0, s_i), \\ -\mu^2 e^{-t\mu} + \mu^2 e^{-(t-s_i)\mu} & \text{for } t \ge s_i. \end{cases}$$

Integrating by parts assuming $0 < s_i < s_j$ yields

$$\begin{aligned} &\int_{-\infty}^{0} (\Gamma_i''(t) + \mu\Gamma_i'(t))(\Gamma_j''(t) + \mu\Gamma_j'(t))\,dt \\ &\quad = \frac{\lambda^2}{\mu^2} \int_{-\infty}^{0} \left(-2e^{-|t|\mu} + 2e^{-(|t|+s_i)\mu}\right)\left(-2e^{-|t|\mu} + 2e^{-(|t|+s_j)\mu}\right) dt = \frac{2\lambda^2}{\mu^3}; \\ &\int_{0}^{s_i} (\Gamma_i''(t) + \mu\Gamma_i'(t))(\Gamma_j''(t) + \mu\Gamma_j'(t))\,dt \\ &\quad = \frac{\lambda^2}{\mu^2} \int_{0}^{s_i} \left(-2e^{-(s_i-t)\mu} + 2\right)\left(-2e^{-(s_j-t)\mu} + 2\right) dt \\ &\quad = \frac{2\lambda^2}{\mu^3}\left(2s_i\mu - 2 - e^{(s_i-s_j)\mu} + 2e^{-s_i\mu} + 2e^{-s_j\mu} - e^{-(s_i+s_j)\mu}\right). \end{aligned}$$

Observe that $\Gamma_i''(t) + \mu\Gamma_i'(t) = 0$ for $t > s_i$; hence,

$$\begin{aligned} &\int_{s_i}^{s_j} (\Gamma_i''(t) + \mu\Gamma_i'(t))(\Gamma_j''(t) + \mu\Gamma_j'(t))\,dt \\ &\quad = \int_{s_j}^{\infty} (\Gamma_i''(t) + \mu\Gamma_i'(t))(\Gamma_j''(t) + \mu\Gamma_j'(t))\,dt = 0. \end{aligned}$$

Finally, upon combining the above, it is straightforward that

$$\int_{-\infty}^{\infty} (\Gamma_i''(t) + \mu\Gamma_i'(t))(\Gamma_j''(t) + \mu\Gamma_j'(t))\,dt = 2\lambda\Gamma(s_i, s_j).$$

□

ACKNOWLEDGMENTS

The authors thank Ilkka Norros (VTT, Finland), Fabrice Guillemin (France Télécom, France), and Hans van den Berg (TNO Information and Communication Technology, The Netherlands) for stimulating discussions and useful remarks.

REFERENCES

1. Addie, R.; Mannersalo, P.; Norros, I. Most probable paths and performance formulae for buffers with Gaussian input traffic. European Transactions on Telecommunications **2002**, *13*, 183–196.
2. Azencott, R. *Ecole d'Et'e de Probabilités de Saint-Flour VIII-1978*; Lecture Notes in Mathematics, Springer: Berlin, 1980; Vol. 774, 1–176.

3. Borst, S.; Boxma, O.; Morrison, J.; Núñez Queija, R. The equivalence between processor sharing and service in random order. Operations Research Letters **2003**, *31*, 254–262.
4. Botvich, D.; Duffield, N. Large deviations, the shape of the loss curve, and economies of large scale multiplexers. Queueing Systems **1995**, *20*, 293–320.
5. Bucklew, J. *Large Deviation Techniques in Decision, Simulation and Estimation*; Wiley: New York, NY, 1990.
6. Bucklew, J. *Introduction to Rare Event Simulation*; Springer: New York, NY, 2004.
7. Deuschel, J.-D.; Stroock, D. *Large Deviations*; Academic Press: Boston, MA, 1989.
8. Flatto, L. The waiting time distribution for the random order service M/M/1 queue. Annals of Applied Probability **1997**, *7*, 382–409.
9. Ganesh, A.; O'Connell, N.; Wischik, D. *Big Queues*; Lecture Notes in Mathematics. Springer: Berlin, Germany, 2004, Vol. 1838.
10. Guillemin, F.; Boyer, J. Analysis of the M/M/1 queue with processor sharing via spectral theory. Queueing Systems **2001**, *39*, 377–397.
11. Guillemin, F.; Mazumdar, R.; Simonian, A. On heavy-traffic approximations for transient characteristics of M/M/∞ queues. Journal of Applied Probability **1996**, *33*, 490–506.
12. Guillemin, F.; Pinchon, D. Continued fraction analysis of the duration of an excursion in an M/M/∞ system. Journal of Applied Probability **1998**, *35*, 165–183.
13. Guillemin, F.; Pinchon, D. On the area swept under the occupation process of an M/M/1 queue in a busy period. Queueing Systems **1998**, *29*, 383–398.
14. Guillemin, F.; Simonian, A. Transient characteristics of an M/M/∞ system. Advances in Applied Probability **1995**, *27*, 862–888.
15. Mandjes, M. *Large Deviations for Gaussian Queues*; Wiley: Chichester, UK, 2007.
16. Mandjes, M.; Kim, J.-H. Large deviations for small buffers: an insensitivity result. Queueing Systems **2001**, *37*, 349–362.
17. Mandjes, M.; Mannersalo, P. Queueing systems fed by many exponential on-off sources: An infinite-intersection approach. Queueing Systems **2006**, *54*, 5–20.
18. Mandjes, M.; Zwart, B. Large deviations for sojourn times in processor sharing queues. Queueing Systems **2006**, *52*, 237–250.
19. Preater, J. M/M/∞ transience revisited. Journal of Applied Probability **1997**, *34*, 1061–1067.
20. Preater, J. On the severity of M/M/∞ congested periods. Journal of Applied Probability **2002**, *39*, 228–230.
21. Reich, E. On the integrodifferential equation of Takács I. Annals of Mathematical Statistics **1958**, *29*, 563–570.
22. Roijers, F.; Mandjes, M.; Van Den Berg, H. Analysis of congestion periods of an M/M/∞-queue. Performance Evaluation **2007**, *64*, 737–754.
23. Shwartz, A.; Weiss, A. *Large Deviations for Performance Analysis. Queues, Communications, and Computing*; Chapman & Hall: London, UK, 1995.
24. Tsybakov, B. Busy periods in M/M/∞ systems with heterogeneous servers. Queueing Systems **2006**, *52*, 153–156.
25. Van Den Berg, H.; Mandjes, M.; Van De Meent, R.; Pras, A.; Roijers, F.; Venemans, P. QoS-aware bandwidth provisioning of IP links. Computer Networks **2006**, *50*, 631–647.