Annals of Operations Research (2020) 293:123–140
https://doi.org/10.1007/s10479-019-03395-9

S.I.: QUEUEING THEORY AND NETWORK APPLICATIONS II

Modelling large timescale and small timescale service variability

Marco Gribaudo[1] · Illés Horváth[2] · Daniele Manini[3] · Miklós Telek[4]

Published online: 25 September 2019

Abstract

The performance of service units may depend on various randomly changing environmental effects. It is quite often the case that these effects vary on different timescales. In this paper, we consider small and large scale (short and long term) service variability, where the short term variability affects the instantaneous service speed of the service unit and a modulating background Markov chain characterizes the long term effect. The main modelling challenge in this work is that the considered small and long term variation results in randomness along different axes: short term variability along the time axis and long term variability along the work axis. We present a simulation approach and an explicit analytic formula for the service time distribution in the double transform domain that allows for the efficient computation of service time moments. Finally, we compare the simulation results with analytic ones.

Keywords Short and long term service variability · Brownian motion · Markov modulation · Performance analysis

This work is partially supported by the OTKA K-123914 and the TUDFO/51757/2019-ITM grants.

✉ Illés Horváth
horvath.illes.antal@gmail.com

Marco Gribaudo
marco.gribaudo@polimi.it

Daniele Manini
manini@di.unito.it

Miklós Telek
telek@hit.bme.hu

[1] Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy

[2] MTA-BME Information Systems Research Group, Budapest, Hungary

[3] Dipartimento di Informatica, Università di Torino, Turin, Italy

[4] Department of Networked Systems and Services, Budapest University of Technology and Economics, Budapest, Hungary

1 Introduction

Service speed variability is a problem that has been observed in many practical application scenarios. For example, in Kimber and Daly (1986), it has been observed for vehicular traffic. More recently this problem has been recognized in data centers (Guo et al. 2014). The effect of variability was also studied in Anjum and Perros (2015) with application to video-streaming. Most of the previous literature, however, focused only on large-timescale variability, where Markov-modulating models represent the random fluctuations of the environment. These set of models are commonly referred to as *reward models* and have been studied for a long time (Howard 1971).

The variation in the service speed can be modelled by dividing the jobs into "infinitesimal quantities of work to be done" and considering the "speed at which this infinitesimal work is performed", i.e., the random amount of time needed to execute the infinitesimal amount of work. Then, once a model defines how speed changes over time, the complete system can be modelled in a straight-forward way where the amount of work increases gradually along the analysis and the time required to execute the given amount of work is a random process.

If the service process depends on a time-dependent random process, e.g., on a modulating background continuous time Markov chain (CTMC) representing the environmental state, whose "clock" evolves according to the time, then the natural performance analysis is based on the gradually increasing time and the randomly varying time dependent environment state.

However, in many real applications, variability is not easily predictable and works at different timescales. Modulating CTMCs (whose "clock" evolves according to the time) works very well to model variability where the parameters of the job execution remain constant for a longer random period of time, and there are few jumps during the execution of one job. Apart from this large scale variability, in this work, we also focus on variability that occurs at much smaller timescales, where the execution speed changes thousands, if not millions, of times during the execution of the main job, and combine it with the more classical modulation that works on a larger timescale.

The remainder of this paper is structured as follows. In Sect. 2 we start by considering only the small timescale variability. In Sect. 3 we additionally introduce also the large timescale variability. Section 4 is devoted to the mathematical analysis of the obtained small and large timescale system. The effects of the considered variability is studied in Sect. 5 through numerical examples, and Sect. 6 concludes the paper.

2 Small timescale variability

In this section, we omit the large timescale variability and instead focus only on small timescale variability. So we assume that the environmental state is unchanged for now.

We introduce a second order fluid model for the short timescale variability: assuming that a job is composed of quantums of size Δx, each such quantum is served in a random amount of time with distribution $N(\mu\Delta x, \sigma^2\Delta x)$ (with $\mu > 0$). Assuming that the service times of the different quantums are independent, the progress of service is modeled by a Brownian motion $X(w)$ with parameters μ and σ^2. We emphasize that in this model, the Brownian motion corresponds to *the time required to service a job as a function of the size of the job* (see Fig. 1). A job of size x thus requires a random time T with distribution $N(\mu x, \sigma^2 x)$, whose probability density function is

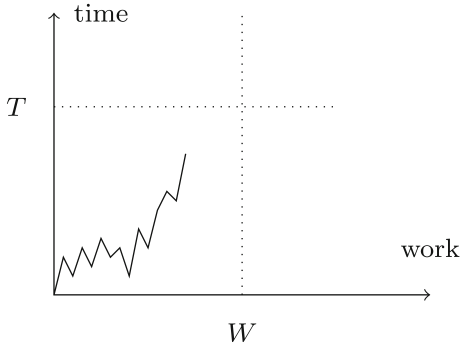


Fig. 1 The time T required to serve a job as a function of the job size W

$$f_{N(\mu x,\sigma^2 x)}(t) = \frac{e^{-\frac{(t-\mu x)^2}{2x\sigma^2}}}{\sqrt{2\pi x\sigma^2}}, \quad t \in \mathbb{R}.$$

The assumption that T may take negative values does not make sense physically. However, due to $\mu > 0$, for macroscopic job sizes, the probability of $T < 0$ is negligible, so the proposed mathematical model is a close approximation of the physical system. In the mathematical analysis, $T < 0$ does not cause any issues, and the performance measures of interest can be calculated accurately.

The moments of $N(\mu x, \sigma^2 x)$ are $x\mu$, $(x\mu)^2 + (\sqrt{x}\sigma)^2$, $(x\mu)^3 + 3(x\mu)(\sqrt{x}\sigma)^2$, $(x\mu)^4 + (x\mu)^2(\sqrt{x}\sigma)^2 + 3(\sqrt{x}\sigma)^4, \ldots$. In general, $E[N(x\mu, x\sigma^2)^k]$ can be expressed as a polynomial in $x\mu$ and $x\sigma^2$

$$E[N(x\mu, x\sigma^2)^k] = \sum_{j=0}^{k} u_{k,j}(x\mu)^j(\sqrt{x}\sigma)^{k-j}, \tag{1}$$

where the coefficients $u_{k,j}$ are such that $u_{k,j} = 0$ if $k + j$ is odd.

Note that a Brownian motion may take negative values as well, which does not make sense physically, but, since $\mu > 0$, for macroscopic values of w, the probability that T is negative is negligible.

We focus on the service of a job in a queue whose work requirement, W, is generally distributed according to probability density function $f_W(x)$.

Using the second order fluid model assumption, the probability density function of the service time of a job, denoted by $f_T(t)$, can be computed as:

$$f_T(t) = \int_0^\infty f_W(x) \cdot \frac{e^{-\frac{(t-\mu x)^2}{2x\sigma^2}}}{\sqrt{2\pi x\sigma^2}} dx. \tag{2}$$

2.1 Moments of the scaled distribution

There are also some interesting relations between the moments of W, the moments of T and the parameters μ and σ^2. In particular, the k-th moment of T can be expressed as:

$$
\begin{aligned}
E[T^k] &= \int_0^\infty t^k f_T(t)dt = \int_0^\infty t^k \int_0^\infty f_W(x) \cdot \frac{e^{-\frac{(t-\mu x)^2}{2x\sigma^2}}}{\sqrt{2\pi x \sigma^2}} dx \cdot dt \\
&= \int_0^\infty f_W(x) \int_0^\infty t^k \cdot \frac{e^{-\frac{(t-\mu x)^2}{2x\sigma^2}}}{\sqrt{2\pi x \sigma^2}} dt \cdot dx \\
&= \int_0^\infty f_W(x) E[N(x\mu, x\sigma^2)^k] dx.
\end{aligned}
$$

Since $E[N(x\mu, x\sigma^2)^k]$ can be expressed as a polynomial in $x\mu$ and $x\sigma^2$ with coefficients $u_{k,j}$ according to (1), we can compute the moments of T as:

$$
\begin{aligned}
E[T^k] &= \int_0^\infty f_W(x) \sum_{j=0}^{k} u_{k,j} (x\mu)^j (\sqrt{x}\sigma)^{k-j} dx \\
&= \sum_{j=0}^{k} u_{k,j} \mu^j \sigma^{k-j} \int_0^\infty f_W(x) x^j (\sqrt{x})^{k-j} dx \\
&= \sum_{j=0}^{k} u_{k,j} \mu^j \sigma^{k-j} E[W^{\frac{k+j}{2}}]. \qquad (3)
\end{aligned}
$$

Since $u_{k,j} \neq 0$ only when $k + j$ is even, $E[W^{\frac{k+j}{2}}]$ is always an integer moment of W. For example, for the first and second moment of T we have:

$$
E[T] = \mu E[W], \quad E[T^2] = \mu^2 E[W^2] + \sigma^2 E[W].
$$

3 Combining large and small timescale variability

Large scale variability can be considered using a discrete state continuous time Markov modulating process (MMP), denoted by $M(t)$. We assume the MMP is a continuous time Markov chain (CTMC) on a finite state space with infinitesimal generator matrix Q. In state i, the service is characterised by rate μ_i and variance σ_i^2.

Only considering large scale variability (that is, assuming $\sigma_i \equiv 0, \forall i$) would lead to a standard Markov reward model. However, including small-scale variability makes for an interesting and complex model.

Assume that a job of size $W = u$ starts service at time $t = 0$, with the MMP $M(t)$ in state i. Then the evolution of the service time $X(w)$, $0 \leq w \leq u$ as a function of the job size is the following:

- Let a_1 denote the time of the first transition of $M(t)$. As long as $X(w)$ is smaller than a_1, $X(w)$ evolves according to a Brownian motion with parameters μ_i and σ_i^2 [denoted by BM(μ_i, σ_i^2)].
- At time a_1, $M(t)$ changes to some state j. Accordingly, assuming that the first passage of $X(w)$ to a_1 occurs at work amount w_1, for $w \geq w_1$, $X(w)$ evolves according to a BM(μ_j, σ_j) (starting from the point w_1 and from level a_1).
- This is repeated for further possible transitions of $M(t)$ at times $a_2, a_3, \ldots,$ up to the point u.

Note that in visualization, the horizontal axis denotes the job size, and the vertical axis denotes time, see Fig. 2. Thus for $X(w)$, the behaviour can be described as a type of *level-*

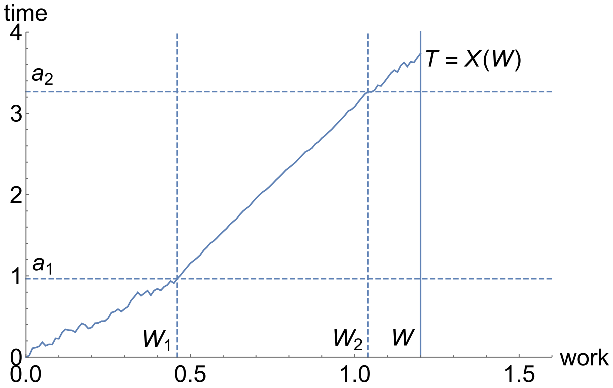


Fig. 2 A possible realization of $X(w)$ for job size $W = 1.2$

dependent Brownian motion: the parameters μ and σ of the Brownian motion change upon first passage to levels $a_1, a_2, \ldots$.

This model is essentially different from second order Markov-modulated fluid models (also referred to as Markov modulated Brownian motion) (Breuer 2012; Karandikar and Kulkarni 1995). The main difference between the two approaches is that in second order fluid models, it is *the amount of work performed per unit of time* that is assumed to have normal distribution; in the present paper, it is *the amount of time required to perform a unit of work* that is assumed to have normal distribution.

Keeping in mind that $M(t)$ is a CTMC, the entire distribution of $X(w)$ is determined by the initial points $t = 0$ and $w = 0$ and the initial state of the modulating process $M(0) = i$. The process $X(w)$ can be simulated as follows:

- If W is random, generate the value of W, denoted by u.
- $X(w)$ starts from $t = 0$, $w = 0$, with $M(0) = i$.
- Generate the first transition time a_1 of $M(t)$.
- $X(w)$ runs as a BM(μ_i, σ_i^2) until either the value of $X(w)$ reaches a_1 or w reaches u, whichever occurs first.
- If u occurred first, then the simulation is finished.
- If $X(w_1) = a_1$ for some $w_1 < W$, then we generate the next state j and also the next transition time a_2 according to the MMP $M(t)$, then continue $X(w)$ as a Brownian motion with parameters (μ_j, σ_j^2) starting from the point (w_1, a_1) until either the value of $X(w)$ reaches a_2 or reaches u, whichever occurs first.
- We keep generating new transitions and new Brownian motion sections until we reach u. The service time of the job is $T = X(u) = X(W)$.

The main question, similar to Sect. 2, is the distribution of T and performance measures derived from T. The main contribution of this paper is the analytical evaluation of the distribution of T in the double transform domain. Several related performance measures can be obtained based on this transform domain description numerically.

The analytical problem can be formulated as the cumulative distribution type functions of the service time (for fixed w)

$$G_{ij}(x, w) = \Pr\left(X(w) \le x, M(X(w)) = j | M(0) = i, W = w\right) \quad (w > 0, x \in \mathbb{R}), \qquad (4)$$

which include information about the initial and final background state of $M(t)$ along with the distribution of the service time. In accordance with the mathematical model, $G_{ij}(x, w)$ is defined for both positive and negative values of x, but $G_{ij}(0, w)$ is typically negligible.

Based on $G_{ij}(x, w)$, the corresponding cumulative distribution function in case of a random W with probability density function $f_W(w)$ is

$$\begin{aligned} G_{ij}(x) &= \Pr(X(W) \le x, M(X(W)) = j | M(0) = i) \\ &= \int_{w=0}^{\infty} G_{ij}(x, w) f_W(w) \mathrm{d}w \quad (x \in \mathbb{R}). \end{aligned} \tag{5}$$

The next section provides the mathematical analysis of $G_{ij}(x, w)$.

4 Job completion in small and large timescale variable environment

Let $X(w)$ denote the time needed to service a job of fixed size w. We aim to analyse the entire process $\{X(w), w \ge 0\}$, and, based on that, derive performance measures for $X(w)$ (for a fixed job size w), and also for $X(W)$, where W is possibly random.

The system operates in a random environment characterized by the MMP $M(t), t \ge 0$, which is a Markov chain with generator Q (and the variable t denotes the time of the MMP). The process $X(w)$ starts from 0 at $w = 0$. When the MMP is in state i, the main process, $X(w)$, is a Brownian motion with parameters $\mu_i > 0$ and $\sigma_i > 0$ (given for each state i). Whenever the MMP makes a transition at time $t = a$, i.e., when $X(w)$ reaches level a, the MMP switches to a new state k and the main process continues as a Brownian motion with parameters $\mu_k > 0$ and σ_k starting at level a. Then the same procedure continues until the job of size u gets completed.

The main process $X(w)$ starts from level 0 at $w = 0$, i.e., $X(0) = 0$. We are interested in the distribution of $X(u)$, where u is the size of the workload, and introduce the notation

$$\begin{aligned} G_{ij}(x, w) &= \Pr(X(w) < x, M(x) = j | M(0) = i, X(0) = 0) \quad (w > 0, x \in \mathbb{R}), \\ g_{ij}(x, w) &= \frac{\partial}{\partial x} G_{ij}(x, w) \quad (w > 0, x \in \mathbb{R}). \end{aligned} \tag{6}$$

We aim to compute

$$g^{\star *}_{ij}(v, s) = \int_{x=-\infty}^{\infty} e^{-vx} \int_{w=0}^{\infty} e^{-sw} g_{ij}(x, w) \mathrm{d}w \mathrm{d}x, \tag{7}$$

where $\star$ refers to the double sided Laplace transform and $*$ refers to single sided Laplace transform. (7) is convergent when $\mathrm{Re}(s) > 0$ and $|v|$ is small enough (depending on $\mathrm{Re}(s)$, that is, $|v| < \varepsilon(\mathrm{Re}(s))$ for some positive function $\varepsilon(.)$. Convergence of the inner integral for $\mathrm{Re}(s) > 0$ follows directly from the fact that $g_{ij}(x, w)$ is a probability density function. Convergence in v will be addressed during the proof of Theorem 1, and the function $\varepsilon(.)$ is made explicit in (16). We remark that calculating $g^{\star *}_{ij}(v, s)$ in a region where $\mathrm{Re}(s) > 0$ and $|v|$ is small enough is sufficient for the further calculation of performance measures of interest.

Theorem 1 *Matrix* $\mathbf{G}^{\star *}(v, s) = \{g^{\star *}_{ij}(v, s)\}$ *is given by*

$$\mathbf{G}^{\star *}(v, s) = (\mathbf{Z}(s) + v\mathbf{I} - \mathbf{Q})^{-1}(\mathbf{Z}(s) - \mathbf{Q}^{\mathbf{D}} + v\mathbf{I})\mathbf{A}^{\star *}(v, s), \tag{8}$$

where $\mathbf{Q}$ *is the generator of the MMP,* $\mathbf{I}$ *is the identity matrix and* $\mathbf{A}^{\star*}(v,s) = \text{diag}\langle a_i^{\star*}(v,s)\rangle$, $\mathbf{Z}(s) = \text{diag}\langle z_i(s)\rangle$, $\mathbf{Q^D} = \text{diag}\langle q_{ii}\rangle$ *are diagonal matrices, where* q_{ii} *are the diagonal elements of* $\mathbf{Q}$*, and*

$$a_i^{\star*}(v,s) = \frac{z_i(s)+v}{z_i(s)-q_{ii}+v}\phi^{-*}(v,s,\mu_i,\sigma_i) + \phi^{+*}(v-q_{ii},s,\mu_i,\sigma_i),$$

$$z_i(s) = \frac{2s}{\mu_i + \sqrt{\mu_i^2 + 2s\sigma_i^2}}, \tag{9}$$

where furthermore

$$\phi^{-*}(v,s,\mu,\sigma) = \frac{\sigma^2}{\sqrt{\mu^2+2s\sigma^2}\left(\mu - v\sigma^2 + \sqrt{\mu^2+2s\sigma^2}\right)},$$

$$\phi^{+*}(v,s,\mu,\sigma) = \frac{\sigma^2}{\sqrt{\mu^2+2s\sigma^2}\left(-\mu + v\sigma^2 + \sqrt{\mu^2+2s\sigma^2}\right)}. \tag{10}$$

($\text{Re}(s) > 0$ *and* $|v|$ *is sufficiently small in all formulas.*)

Proof Let W_a be the first passage point along the horizontal axis, where the BM(μ_i, σ_i) starting from level 0 reaches level a ($a > 0$). The CDF and PDF of W_a are denoted by

$$F_i(a,w) = \Pr(W_a < w | M(0) = i, X(0) = 0), \quad f_i(a,w) = \frac{\partial}{\partial w}F_i(a,w).$$

$f_i(a,w)$ is given explicitly [using Girsanov's theorem and mirror principle, see e.g. Theorem 6.9 in Schilling and Partzsch (2012)] as

$$f_i(a,w) = \frac{a}{\sqrt{2\pi w^3}\sigma_i}\exp\left(-\frac{(a-\mu_i w)^2}{2w\sigma_i^2}\right), \quad 0 < a,\ 0 < w. \tag{11}$$

When the process starts in state i, two things may happen (c.f. Fig. 3): the main process will either reach level a (along the vertical axis) before w (on the horizontal axis), i.e. $W_a < w$, or not. If the main process reaches level a before w, then the MMP switches from state i to another state k at W_a, and the main process continues similarly with parameters (μ_k, σ_k), albeit starting from level a.

If the main process does not reach level a before completing w amount of work, i.e., $W_a > w$, then we need the conditional distribution of the level at w assuming that $X(w) < a$ for $\forall u < w$. To obtain it, we introduce the notation

$$B_i(x,a,w) = \Pr(X(w) < x, X(u) < a, \quad \forall u \in (0,w) | M(0) = i, X(0) = 0),$$

$$b_i(x,a,w) = \frac{\partial}{\partial x}B_i(x,a,w), \quad x < a,\ 0 < a,\ 0 < w.$$

$B_i(x,a,w)$ is a CDF type function, and it describes an incomplete distribution concentrated on $(-\infty, a)$, and it satisfies

$$F_i(a,w) + B_i(x,a,w)|_{x=a} = 1, \tag{12}$$

where the first term corresponds to the probability that the BM(μ_i, σ_i) hits level a before w ($W_a < w$), while the second term corresponds to the probability that the BM(μ_i, σ_i) hits the vertical line at w without reaching level a ($W_a > w$).

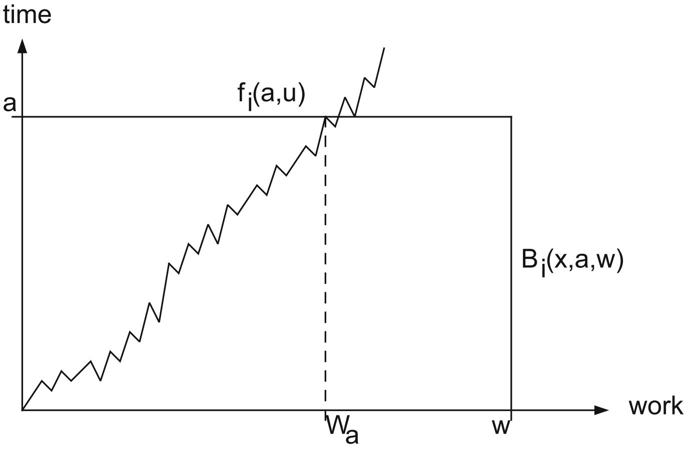


Fig. 3 A trajectory reaching level a before w ($W_a < w$)

To calculate the density $b_i(x, a, w)$, we first note that the position of a BM(μ_i, σ_i) at point w has normal distribution with parameters $(\mu_i w, \sigma_i \sqrt{w})$, so its probability density function at w is $\phi\left(x, \mu_i w, \sigma_i \sqrt{w}\right)$, where $\phi(x, \mu, \sigma)$ denotes the PDF of normal distribution with parameters μ and σ^2, i.e.,

$$\phi(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

To compute $b_i(x, a, w)$, we need to subtract the density that the BM hits level a first. We calculate it using total probability according to the first passage time at level a, W_a. Altogether, $b_i(x, a, w)$ can be calculated as

$$b_i(x, a, w) = \phi\left(x, \mu_i w, \sigma_i \sqrt{w}\right) - \int_{u=0}^{w} f_i(a, u)\phi\left(x - a, \mu_i(w-u), \sigma_i\sqrt{w-u}\right) \mathrm{d}u. \quad (13)$$

The level at which the MMP changes its state, a, is exponentially distributed with parameter $-q_{ii}$. Using that and the probability of moving from state i to state k at a state transition of the MMP, $-q_{ik}/q_{ii}$, we have

$$\begin{aligned} g_{ij}(x, w) = \delta_{ij} &\int_{a=x^+}^{\infty} b_i(x, a, w)(-q_{ii})e^{q_{ii}a}\mathrm{d}a \\ &+ \sum_{k:k\neq i} \int_{u=0}^{w} \int_{a=0}^{\infty} q_{ik}e^{q_{ii}a} f_i(a, u) g_{kj}(x - a, w - u)\mathrm{d}a\mathrm{d}u, \qquad (14) \end{aligned}$$

where $x^+ = \max(0, x)$ and δ_{ij} denotes the Kronecker delta.

Remarks:

- The first term in (14) is the probability that the main process reaches $u = w$ before hitting level a averaged out according to the distribution of a.
- In the second term, the MMP switches to state k at $u < w$, with the main process at level $X(w) = a$.
- Even though the general idea is that the main process is increasing, using a second order approach means that in the short term, the main process may decrease as well. Hence we

should care about negative values of x if possible. The formula (14) is consistent with the possibility that the process $X(w)$ may decrease and the formula is valid for negative values of x as well.

The second integral in (14) is essentially convolution in both variables x and w; to simplify it, we will take Laplace-transform in the variable w, and double-sided Laplace transform in the variable x.

We look to take the Laplace-transform of the variable w in all of the important functions in (14). Denoting the transform variable by s, the Laplace-transform of $f_i(a, w)$ in (11) is explicit:

$$\begin{aligned} f_i^*(a,s) &= \int_{w=0}^{\infty} e^{-sw} f_i(a,w)\mathrm{d}w = \exp\left(\frac{-2as}{\mu_i + \sqrt{\mu_i^2 + 2s\sigma_i^2}}\right) \\ &= e^{-z_i(s)a}, \quad a > 0,\ \mathrm{Re}(s) > 0, \end{aligned}$$

where $z_i(s)$ is given in (9).

Similarly, we have an explicit formula for the Laplace-transform of $\phi\left(x, \mu w, \sigma\sqrt{w}\right)$ with respect to w

$$\begin{aligned} \phi^*(x,s,\mu,\sigma) &= \int_{w=0}^{\infty} \phi\left(x, \mu w, \sigma\sqrt{w}\right) e^{-sw}\mathrm{d}w \\ &= \frac{1}{\sqrt{\mu^2 + 2s\sigma^2}} \exp\left(\frac{x\mu - |x|\sqrt{\mu^2 + 2s\sigma^2}}{\sigma^2}\right). \end{aligned}$$

Then from (13), we have

$$\begin{aligned} b_i^*(x,a,s) &= \int_{w=0}^{\infty} e^{-sw} b_i(x,a,w)\mathrm{d}w \\ &= \phi^*(x,s,\mu_i,\sigma_i) - \phi^*(x-a,s,\mu_i,\sigma_i) f_i^*(a,s), \end{aligned}$$

and from (14), we have

$$\begin{aligned} g_{ij}^*(x,s) &= \int_{w=0}^{\infty} e^{-sw} g_{ij}(x,w)\mathrm{d}w \\ &= \delta_{ij} \underbrace{\int_{a=x^+}^{\infty} b_i^*(x,a,s)(-q_{ii})e^{q_{ii}a}\mathrm{d}a}_{a_i^*(x,s)} + \underbrace{\sum_{k:k\neq i} q_{ik} \int_{a=0}^{\infty} e^{q_{ii}a} f_i^*(a,s) g_{kj}^*(x-a,s)\mathrm{d}a}_{h_{ij}^*(x,s)}. \qquad (15) \end{aligned}$$

To transform the level variable x as well using two-sided Laplace transform, we will use the functions $\phi^{-*}(v,s,\mu,\sigma)$ and $\phi^{+*}(v,s,\mu,\sigma)$ as defined in (10):

$$\begin{aligned}\phi^{-*}(v,s,\mu,\sigma) &= \int_{x=-\infty}^{0} e^{-vx}\phi^*(x,s,\mu,\sigma)\mathrm{d}x\\ &= \frac{\sigma^2}{\sqrt{\mu^2+2s\sigma^2}\left(\mu - v\sigma^2+\sqrt{\mu^2+2s\sigma^2}\right)},\\ \phi^{+*}(v,s,\mu,\sigma) &= \int_{x=0}^{\infty} e^{-vx}\phi^*(x,s,\mu,\sigma)\mathrm{d}x\\ &= \frac{\sigma^2}{\sqrt{\mu^2+2s\sigma^2}\left(-\mu + v\sigma^2+\sqrt{\mu^2+2s\sigma^2}\right)},\end{aligned}$$

where convergence in either integral holds when the real part of the associated denominators are positive. For $\mathrm{Re}(s) > 0$, we have $\mu < \mathrm{Re}\left(\sqrt{\mu^2+2s\sigma^2}\right)$, from which the denominators are positive when

$$|v| < \frac{\mathrm{Re}\left(\sqrt{\mu^2+2s\sigma^2}\right)-\mu}{\sigma^2};$$

from this, we have that (7) and (8) are valid when $\mathrm{Re}(s) > 0$ and

$$|v| < \varepsilon(\mathrm{Re}(s)) := \min_i \frac{\mathrm{Re}\left(\sqrt{\mu_i^2+2s\sigma_i^2}\right)-\mu_i}{\sigma_i^2}. \tag{16}$$

To compute $g_{ij}^{\star*}(v,s) = \int_{x=-\infty}^{\infty} e^{-vx} g_{ij}^*(x,s)\mathrm{d}x$, we start by investigating the transform of first term $a_i^*(x,s)$ on the right hand side in (15):

$$\begin{aligned}a_i^{\star*}(v,s) &= \int_{x=-\infty}^{\infty} e^{-vx} a_i^*(x,s)\mathrm{d}x = \int_{x=-\infty}^{\infty} e^{-vx}\int_{a=x^+}^{\infty} b_i^*(x,a,s)(-q_{ii})e^{q_{ii}a}\mathrm{d}a\mathrm{d}x\\ &= \int_{x=-\infty}^{\infty} e^{-vx}\int_{a=x^+}^{\infty}\phi^*(x,s,\mu_i,\sigma_i)(-q_{ii})e^{q_{ii}a}\mathrm{d}a\mathrm{d}x\\ &\quad -\int_{x=-\infty}^{\infty} e^{-vx}\int_{a=x^+}^{\infty}\phi^*(x-a,s,\mu_i,\sigma_i)f_i^*(a,s)(-q_{ii})e^{q_{ii}a}\mathrm{d}a\mathrm{d}x. \end{aligned}\tag{17}$$

The first term on the right hand side of (17) is

$$\begin{aligned}&\int_{x=-\infty}^{\infty} e^{-vx}\phi^*(x,s,\mu_i,\sigma_i)\int_{a=x^+}^{\infty}(-q_{ii})e^{q_{ii}a}\mathrm{d}a\mathrm{d}x\\ &= (-q_{ii})\int_{x=-\infty}^{0} e^{-vx}\phi^*(x,s,\mu_i,\sigma_i)\int_{a=0}^{\infty} e^{q_{ii}a}\mathrm{d}a\mathrm{d}x\\ &\quad + (-q_{ii})\int_{x=0}^{\infty} e^{-vx}\phi^*(x,s,\mu_i,\sigma_i)\int_{a=x}^{\infty} e^{q_{ii}a}\mathrm{d}a\mathrm{d}x\end{aligned}$$

$$
\begin{aligned}
&= (-q_{ii}) \int_{x=-\infty}^{0} e^{-vx} \phi^*(x, s, \mu_i, \sigma_i) \frac{1}{-q_{ii}} \mathrm{d}x \\
&\quad + (-q_{ii}) \int_{x=0}^{\infty} e^{-vx} \phi^*(x, s, \mu_i, \sigma_i) \frac{e^{q_{ii}x}}{-q_{ii}} \mathrm{d}x \\
&= \int_{x=-\infty}^{0} e^{-vx} \phi^*(x, s, \mu_i, \sigma_i) \mathrm{d}x + \int_{x=0}^{\infty} e^{-(v-q_{ii})x} \phi^*(x, s, \mu_i, \sigma_i) \mathrm{d}x \\
&= \phi^{-*}(v, s, \mu_i, \sigma_i) + \phi^{+*}(v - q_{ii}, s, \mu_i, \sigma_i). \qquad (18)
\end{aligned}
$$

The second term on the right hand side of (17) is

$$
\begin{aligned}
&\int_{x=-\infty}^{\infty} e^{-vx} \int_{a=x^+}^{\infty} \phi^*(x-a, s, \mu_i, \sigma_i) f_i^*(a, s)(-q_{ii}) e^{q_{ii}a} \mathrm{d}a \mathrm{d}x \\
&= (-q_{ii}) \int_{x=-\infty}^{\infty} e^{-vx} \int_{a=x^+}^{\infty} \phi^*(x-a, s, \mu_i, \sigma_i) e^{-(z_i(s)-q_{ii})a} \mathrm{d}a \mathrm{d}x \\
&= (-q_{ii}) \int_{a=0}^{\infty} e^{-(z_i(s)-q_{ii})a} \int_{x=-\infty}^{a} e^{-vx} \phi^*(x-a, s, \mu_i, \sigma_i) \mathrm{d}x \mathrm{d}a \\
&= (-q_{ii}) \int_{a=0}^{\infty} e^{-(z_i(s)-q_{ii})a} e^{-va} \int_{x=-\infty}^{a} e^{-v(x-a)} \phi^*(x-a, s, \mu_i, \sigma_i) \mathrm{d}x \mathrm{d}a \\
&= (-q_{ii}) \int_{a=0}^{\infty} e^{-(z_i(s)-q_{ii}+v)a} \mathrm{d}a \int_{x=-\infty}^{0} e^{-vx} \phi^*(x, s, \mu_i, \sigma_i) \mathrm{d}x \\
&= \frac{-q_{ii}}{z_i(s) - q_{ii} + v} \int_{x=-\infty}^{0} e^{-vx} \phi^*(x, s, \mu_i, \sigma_i) \mathrm{d}x \\
&= \frac{-q_{ii}}{z_i(s) - q_{ii} + v} \phi^{-*}(v, s, \mu_i, \sigma_i). \qquad (19)
\end{aligned}
$$

Altogether, expressing (17) as the total of (18) and (19), we get

$$
\begin{aligned}
a_i^{\star*}(v, s) &= \phi^{-*}(v, s, \mu_i, \sigma_i) + \phi^{+*}(v - q_{ii}, s, \mu_i, \sigma_i) \\
&\quad - \frac{-q_{ii}}{z_i(s) - q_{ii} + v} \phi^{-*}(v, s, \mu_i, \sigma_i) \\
&= \frac{z_i(s) + v}{z_i(s) - q_{ii} + v} \phi^{-*}(v, s, \mu_i, \sigma_i) + \phi^{+*}(v - q_{ii}, s, \mu_i, \sigma_i). \qquad (20)
\end{aligned}
$$

Now we focus on the second term on the right hand side of (15).

$$
\begin{aligned}
h_{ij}^{\star*}(v, s) &= \int_{x=-\infty}^{\infty} e^{-vx} h_{ij}^*(x, s) \mathrm{d}x \\
&= \sum_{k:k\neq i} \int_{x=-\infty}^{\infty} e^{-vx} q_{ik} \int_{a=0}^{\infty} e^{q_{ii}a} f_i^*(a, s) g_{kj}^*(x-a, s) \mathrm{d}a \mathrm{d}x \\
&= \sum_{k:k\neq i} q_{ik} \int_{x=-\infty}^{\infty} e^{-vx} \int_{a=0}^{\infty} e^{-(z_i(s)-q_{ii})a} g_{kj}^*(x-a, s) \mathrm{d}a \mathrm{d}x
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{k:k\neq i} q_{ik} \int_{a=0}^{\infty} e^{-(z_i(s)-q_{ii}+v)a} \int_{x=-\infty}^{\infty} e^{-v(x-a)} g_{kj}^{*}(x-a,s)\mathrm{d}x\mathrm{d}a \\
&= \sum_{k:k\neq i} q_{ik} \int_{a=0}^{\infty} e^{-(z_i(s)-q_{ii}+v)a}\mathrm{d}a\; g_{kj}^{\star *}(v,s) \\
&= \sum_{k:k\neq i} \frac{q_{ik}}{z_i(s)-q_{ii}+v} g_{kj}^{\star *}(v,s), \qquad (21)
\end{aligned}
$$

from which we get

$$
\begin{aligned}
g_{ij}^{\star *}(v,s) &= \delta_{ij} a_i^{\star *}(v,s) + \sum_{k,k\neq i} \frac{q_{ik}}{z_i(s)-q_{ii}+v} g_{kj}^{\star *}(v,s) \\
(z_i(s)-q_{ii}+v) g_{ij}^{\star *}(v,s) &= \delta_{ij}(z_i(s)-q_{ii}+v) a_i^{\star *}(v,s) + \sum_{k,k\neq i} q_{ij} g_{kj}^{\star *}(v,s) \\
(z_i(s)+v) g_{ij}^{\star *}(v,s) &= \delta_{ij}(z_i(s)-q_{ii}+v) a_i^{\star *}(v,s) + \sum_{k} q_{kj} g_{kj}^{\star *}(v,s). \qquad (22)
\end{aligned}
$$

Introducing matrix $\mathbf{G}^{\star *}(v,s) = \{g_{ij}^{\star *}(v,s)\}$, and diagonal matrices $\mathbf{A}^{\star *}(v,s) = \mathrm{diag}\langle a_i^{\star *}(v,s)\rangle$, $\mathbf{Z}(s) = \mathrm{diag}\langle z_i(s)\rangle$, $\mathbf{Q}^{\mathbf{D}} = \mathrm{diag}\langle q_{ii}\rangle$, (22) can be written as

$$
(\mathbf{Z}(s) + v\mathbf{I})\mathbf{G}^{\star *}(v,s) = (\mathbf{Z}(s) - \mathbf{Q}^{\mathbf{D}} + v\mathbf{I})\mathbf{A}^{\star *}(v,s) + \mathbf{Q}\mathbf{G}^{\star *}(v,s),
$$

whose solution is (8). □

Theorem 1 provides an explicit expression (involving a matrix inversion) for the double transform domain description of the service time distribution. For the $s \to w$ one-sided inverse Laplace transformation, we applied different approaches depending on the distribution of the work requirement W. To simplify calculations, we decided to avoid doing a $v \to x$ two-sided inverse Laplace transformation, instead calculating the moments of $X(w)$ explicitly, which can be obtained from

$$
E(X(w)^k) = (-1)^k \mathcal{L}_{s\to w}^{-1} \left(\frac{\partial^k}{\partial v^k} g_{ij}^{\star *}(v,s)\Big|_{v=0} \right), \qquad (23)
$$

where $\mathcal{L}_{s\to w}^{-1}$ denotes inverse Laplace transformation from parameter s to parameter w. To compute these moments it is important to note that the eigenvalues of $\mathbf{Z}(s) - \mathbf{Q}$ are all positive, which provides a symbolic derivative according to v also for the matrix inverse, e.g., to compute the mean of $X(w)$ we have

$$
\begin{aligned}
\frac{\partial}{\partial v} \mathbf{G}^{\star *}(v,s)\Big|_{v=0} &= -(\mathbf{Z}(s)-\mathbf{Q})^{-2}(\mathbf{Z}(s)-\mathbf{Q}^{\mathbf{D}})\mathbf{A}^{\star *}(0,s) \\
&\quad + (\mathbf{Z}(s)-\mathbf{Q})^{-1}\mathbf{A}^{\star *}(0,s) \\
&\quad + (\mathbf{Z}(s)-\mathbf{Q})^{-1}(\mathbf{Z}(s)-\mathbf{Q}^{\mathbf{D}}) \frac{\partial}{\partial v}\mathbf{A}^{\star *}(v,s)\Big|_{v=0}.
\end{aligned}
$$

The explicit formulas for $\frac{\partial^k}{\partial v^k}\mathbf{A}^{\star *}(v,s)\Big|_{v=0}$ with higher k values are omitted here, but symbolic mathematical packages can compute them easily.

For deterministic work requirement ($W = w$) we applied the numerical inverse Laplace transformation method from Horváth et al. (2018) with order 24. This numerical inverse

Laplace transform procedure evaluates the Laplace transform function only in points with a positive real part.

For exponentially distributed work requirement (W is exponentially distributed with rate ϑ) we use an explicit inversion formula based on

$$g_{ij}(x) = \int_{w=0}^{\infty} g_{ij}(x, w) f_W(w) \mathrm{d}w = \vartheta \int_{w=0}^{\infty} g_{ij}(x, w) e^{-\vartheta w} \mathrm{d}w = \vartheta g_{ij}^{*}(x, s)\Big|_{s=\vartheta}.$$

Consequently, the kth moment of the service time for an exponentially distributed work requirement (W) with rate ϑ can be computed explicitly as

$$E(X(W)^k | M(0) = i) = (-1)^k \vartheta \frac{\partial^k}{\partial v^k} \sum_j g_{ij}^{\star *}(v, s)\Big|_{v=0, s=\vartheta}. \tag{24}$$

5 Numerical examples

5.1 Simulation results

To study the effects of variability, we have applied the procedure outlined in Sect. 3 to simulate the behaviour of the queue with short and long scale variability. In particular, to find the intersection between the Brownian motion and the level determined by the time at which the modulating process changes state, we have discretised the work with a quantum Δx, and during the period when the MMP stays in state i, for each quantum we have set the evolution of the time according to a normal distribution $N(\mu_i \Delta x, \sigma_i^2 \Delta x)$ (following the procedure outlined at the beginning of Sect. 2). The MMP leaves state i at the first time instant in which the discretised BM crosses the level T_n, where T_n is the time of the nth state transition of the MMP. When the nth state transition occurs in state i, then $T_n = T_{n-1} + \tau_i$, where T_{n-1} is the time of the previous state transition and τ_i is exponentially distributed with parameter $-q_{ii}$ (the ith diagonal element of the generator matrix $\mathbf{Q}$ of the MMP). This simulation approach is indeed an approximation, but it can be made arbitrarily precise by choosing appropriately small values of Δx (at the cost of simulation time). Simulations were run for several choices of Δx to examine the error of this approximation.

In our numerical experiment, we have considered a two-state modulating process with jump rates q_{12} and q_{21}, and studied the effects of different service speed and variability parameters μ_i and σ_i ($i = 1, 2$). Apart from computing performance measures related to the service time distribution, we also included simulation results for the response time in an M/G/1 queue, where jobs arrive according to a Poisson process of rate λ and are served by a single server subject to short and long term variability according to a first-come-first-served discipline. Job sizes may be either deterministic or random. λ is set so that the queue is stable.

For the first batch of simulations, we examine the effect of short and long term variability of the server by changing μ and σ while leaving the other parameters fixed:

$$\lambda = \frac{1000}{350}\,\text{job/s}, \quad W = E[W] = 100\,\text{ms}, \quad \frac{1}{q_{12}} = 1.25\,\text{s}, \quad \frac{1}{q_{21}} = 0.8\,\text{s}.$$

We compare the following cases:

- *Base* no variability, $\mu_1 = \mu_2 = 2.4848$ and $\sigma_1 = \sigma_2 = 0$.
- *Small (fixed)* small term variability with $\mu_1 = \mu_2 = 2.4848$ and $\sigma_1 = \sigma_2 = 0.98773$.

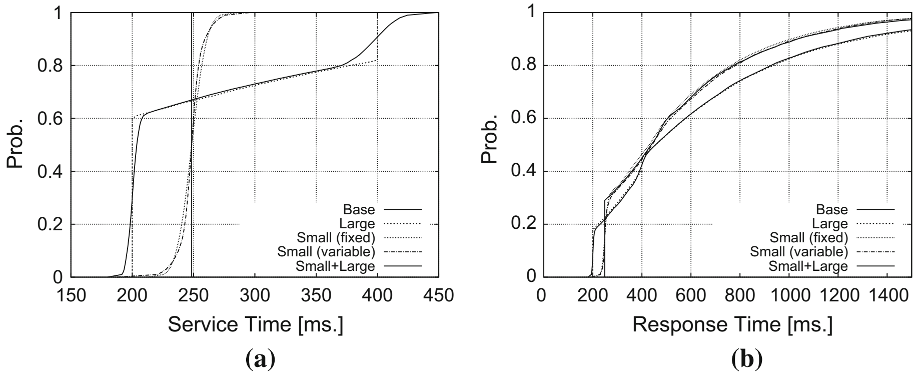


Fig. 4 Considering different small scale and large scale variability configurations for a fixed job length: **a** service time distribution, **b** response time distribution

- *Small (variable)* small term variability with $\mu_1 = \mu_2 = 2.4848$ and $\sigma_1 = 0.4$ and $\sigma_2 = 1.5$.
- *Large* Long term variability is present, but no short term variability: $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_1 = \sigma_2 = 0$.
- *Small* + *Large* both effects are combined: $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_1 = 0.4$, $\sigma_2 = 1.5$.

(For this set of simulation results, the discretization interval Δx is set to 0.05 ms so that on average, the BM for each job requires 2000 samples. Each simulation considers the execution of $N = 10{,}000$ jobs.)

Figure 4a shows the service time distribution for the different server variability configurations. For the *Base* case, as it is expected, all the probability mass is centered along $\mu_1 E[W] = \mu_2 E[W] = 248.48$. For the *Small (fixed)* case, the introduction of small term variability destroys the deterministic behaviour, resulting in a smooth distribution still concentrated near $\mu_1 E[W] = \mu_2 E[W] = 248.48$. For the *Small (variable)* case, the distribution is similar, with larger tails due to the long term variability in σ. For the *Large* case, in state 1, service time is exactly $\mu_1 E[W] = 200$ ms, and in state 2, service time is exactly $\mu_2 E[W] = 400$ ms. The probability masses in Fig. 4a at 200 ms and 400 ms are associated with the cases when the MMP stays in state 1 (2, respectively) for the whole period of the service. The cases when the MMP experiences state transition during the service are represented by the continuously increasing part of the *Large* curve. The case that combines both small and large scale variability (*Small* + *Large*) further smooths the curves, and the effect is more evident near the two probability masses at 200 ms, and 400 ms.

Figure 4b shows the response time distribution of the corresponding queuing models. It is interesting to see that in the cases where small variability is considered there are no jumps due to its perturbation effect.

The second batch of simulations focuses on the MMP by changing the speed of the background MMP. We set

$$\lambda = \frac{1000}{350}\,\text{job/s}, \quad W = E[W] = 100\,\text{ms}, \quad \Delta x = 0.05\,\text{ms}$$

as before, along with

$$\mu_1 = 2, \quad \mu_2 = 4, \quad \sigma_1 = 0.4, \quad \sigma_2 = 1.5,$$

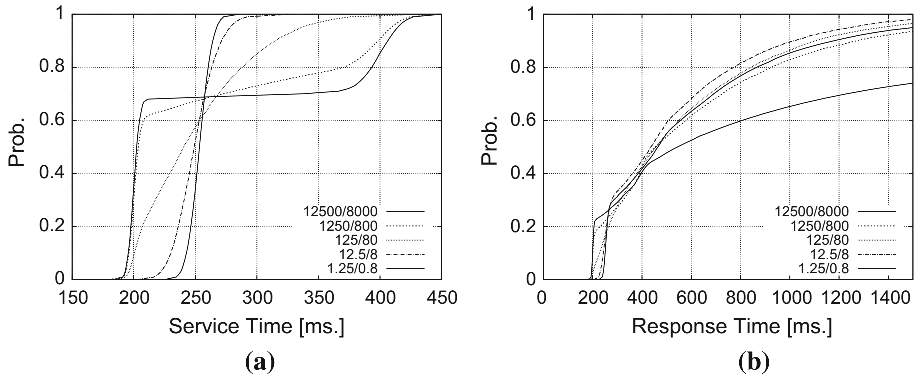


Fig. 5 Considering different durations in the modulating process for a fixed job length: **a** service time distribution, **b** response time distribution

but the values of the average sojourn times $1/q_{12}$ and $1/q_{21}$ change from

$$\frac{1}{q_{12}} = 12.5\,\text{s}, \quad \frac{1}{q_{21}} = 8\,\text{s}$$

all the way through to

$$\frac{1}{q_{12}} = 1.25\,\text{ms}, \quad \frac{1}{q_{21}} = 0.8\,\text{s}$$

with each step corresponding to a factor of 10.

Figure 5 shows simulation results for the second batch of simulations. When the sojourn times are very large, the service time distribution tends to concentrate the probability mass near the times required in either state (200 ms and 400 ms). On the other hand, when the MMP changes rapidly, the distribution tends to concentrate on the average case, producing results very similar to the one seen in Fig. 4 for the cases with small variability only: in this case, there is almost no difference between large scale and small scale variability, because the quick alternation of the modulating process eliminates the large scale effect. As a final remark, to consider the case with sojourn times 1.25 ms and 0.8 ms, the sampling time was reduced to $\Delta x = 0.01$ to allow a sufficient number of samples during the sojourn in a modulating state. As for the response time (Fig. 5c), longer sojourn times create bursts when the MMP remains in a single state, considerably decreasing the performance of the system.

The third batch of simulations focuses on the effect of variability on different job length distributions. In particular,

$$\begin{aligned}
&\lambda = \frac{1000}{350}\,\text{job/s}, \quad E[W] = 100\,\text{ms}, \quad \Delta x = 0.05\,\text{ms}\\
&\mu_1 = 2, \quad \mu_2 = 4, \quad \sigma_1 = 0.4, \quad \sigma_2 = 1.5,\\
&\frac{1}{q_{12}} = 1.25\,\text{ms}, \quad \frac{1}{q_{21}} = 0.8\,\text{s}
\end{aligned} \tag{25}$$

and we examine the following job size distributions for W:

- Deterministic $W = 100\,\text{ms}$,
- Exponential with mean 100 ms,
- Erlang with 4 stages with mean 100 ms,

- Hyper-exponential with probability density function

$$f_W(x) = \frac{1}{2}\lambda_1 e^{-\lambda_1 x} + \frac{1}{2}\lambda_2 e^{-\lambda_2 x} \quad x > 0$$

  with parameters $\lambda_1 = 1/(100(1+\sqrt{0.6}))$, $\lambda_2 = 1/100((1-\sqrt{0.6}))$
- Pareto with probability density function

$$f_W(x) = \begin{cases} \frac{\frac{5}{4}\cdot 20^{\frac{5}{4}}}{x^{\frac{9}{4}}} & x > 20, \\ 0 & x < 20 \end{cases}$$

(To make the results easily comparable, $E[W] = 100$ ms is identical in each case.)

In particular, Fig. 6a shows the service time distribution for each job size distribution. The effect of service variability is more evident on job length distributions with a lower coefficient of variation. Figure 6b shows the effect on response time: indeed, combining the effect of service variability with a heavy tailed distribution, as for the Pareto case, can create very long queues which can lead to extremely long response times.

5.2 Comparison of analytical and simulation results

For the last batch of simulations, we compare empirical moments from the simulation to moments calculated using the double transform method of Sect. 4.

The system parameters are the same as in (25). Two different job size distributions are examined: deterministic and exponential. To test the inaccuracy of the simulation with finite discretization steps, we run each simulation with two different choices of Δx: $\Delta x = 0.05$ ms and $\Delta x = 0.005$ ms.

Table 1 presents the moments of the service time distribution obtained from the simulator and the transform domain description. $\Delta x = 0.05$ ms corresponds to sim. 1 and $\Delta x = 0.005$ ms corresponds to sim. 2.

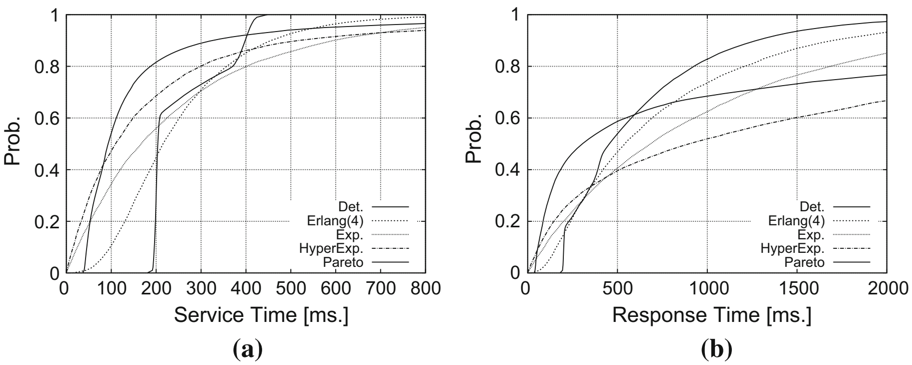
From Table 1, we observe increasing relative error for higher moments.

For the mean, the relative error is around or smaller than 2%. The relative error of the mean decreases as Δx is refined from $\Delta x = 0.05$ ms (sim. 1) to $\Delta x = 0.005$ ms (sim. 2). We note that for the exponential case, the service time moments were calculated using an analytic formula, while for deterministic job size, some inaccuracy might also come from the numerical inverse Laplace transformation method.

6 Conclusions

In this work, we have introduced a queue with a service model where a modulating background Markov process models the large timescale variability, and a second-order fluid process models the service capacity on small timescale. The resulting service model can be interpreted as a certain type of level-dependent Brownian motion.

We have presented both a simulation approach for the service time and response time of a job for various job size distributions and a double Laplace transform domain analysis of the service time distribution. A numerical example illustrates the effect of small and large scales of service variability. Using that example, we compared the results obtained from simulation and the Laplace transform domain analytical description.



(a)

(b)

Fig. 6 Considering small scale and large scale variability for different job length distributions: **a** service time distribution, **b** response time distribution

Table 1 Comparison of the numerical analysis and the simulation results

	Deterministic job size		
	sim. 1	sim. 2	Transform
$E(X(W)\|M(0)=1)$	215.5	215.8	213.0
$E(X(W)\|M(0)=2)$	367.1	364.9	359.3
$E(X(W)^2\|M(0)=1)$	4.689×10^4	4.835×10^4	4.818×10^4
$E(X(W)^2\|M(0)=2)$	1.383×10^5	1.368×10^5	1.332×10^5
$E(X(W)^3\|M(0)=1)$	1.133×10^7	1.140×10^7	1.082×10^7
$E(X(W)^3\|M(0)=2)$	5.303×10^7	5.23×10^7	5.054×10^7
$E(X(W)^4\|M(0)=1)$	2.846×10^9	2.87×10^9	2.656×10^9
$E(X(W)^4\|M(0)=2)$	2.060×10^{10}	2.03×10^{10}	1.950×10^{10}
	Exponential job size		
	sim. 1	sim. 2	Transform
$E(X(W)\|M(0)=1)$	226.1	224.9	219.3
$E(X(W)\|M(0)=2)$	322.5	330.0	339.7
$E(X(W)^2\|M(0)=1)$	1.081×10^5	1.108×10^5	1.055×10^5
$E(X(W)^2\|M(0)=2)$	2.030×10^5	2.089×10^5	2.165×10^5
$E(X(W)^3\|M(0)=1)$	8.171×10^7	8.820×10^7	8.226×10^7
$E(X(W)^3\|M(0)=2)$	1.892×10^8	1.981×10^8	2.007×10^8
$E(X(W)^4\|M(0)=1)$	8.680×10^{10}	9.810×10^{10}	9.044×10^{10}
$E(X(W)^4\|M(0)=2)$	2.352×10^{11}	2.580×10^{11}	2.443×10^{11}

Acknowledgements Open access funding provided by Budapest University of Technology and Economics (BME).

References

Anjum, B., & Perros, H. (2015). Bandwidth estimation for video streaming under percentile delay, jitter, and packet loss rate constraints using traces. *Computer Communications*, *57*, 73–84. https://doi.org/10.1016/j.comcom.2014.08.018.

Breuer, L. (2012). Occupation times for Markov-modulated Brownian motion. *Journal of Applied Probability*, *49*(2), 549–565. https://doi.org/10.1239/jap/1339878804.

Guo, J., Liu, F., Huang, X., Lui, J. C., Hu, M., Gao, Q., & Jin, H. (2014). On efficient bandwidth allocation for traffic variability in datacenters. In: *Proceedings—IEEE INFOCOM*, pp. 1572–1580.

Horváth, I., Talyigás, Z., & Telek, M. (2018). An optimal inverse Laplace transform method without positive and negative overshoot—An integral based interpretation. *Electronic Notes in Theoretical Computer Science*, *337*, 87–104.

Howard, R. (1971). *Dynamic probabilistic systems, Volume II: Semi-Markov and decision processes*. New York: Wiley.

Karandikar, R. L., & Kulkarni, V. G. (1995). Second-order fluid flow models: Reflected Brownian motion in a random environment. *Operations Research*, *43*(1), 77–88. https://doi.org/10.1287/opre.43.1.77.

Kimber, R., & Daly, P. (1986). Time-dependent queueing at road junctions: Observation and prediction. *Transportation Research Part B: Methodological*, *20*(3), 187–203. https://doi.org/10.1016/0191-2615(86)90016-0.

Schilling, R., & Partzsch, L. (2012). *Brownian motion: An introduction to stochastic processes*. Berlin: De Gruyter.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.