Queueing Syst (2017) 85:269–304
DOI 10.1007/s11134-016-9511-0

Transient error approximation in a Lévy queue

Britt Mathijsen[1] · **Bert Zwart**[1,2]

Received: 19 April 2016 / Revised: 7 December 2016 / Published online: 13 January 2017

Abstract Motivated by a capacity allocation problem within a finite planning period, we conduct a transient analysis of a single-server queue with Lévy input. From a cost minimization perspective, we investigate the error induced by using stationary congestion measures as opposed to time-dependent measures. Invoking recent results from fluctuation theory of Lévy processes, we derive a refined cost function, that accounts for transient effects. This leads to a corrected capacity allocation rule for the transient single-server queue. Extensive numerical experiments indicate that the cost reductions achieved by this correction can be significant.

Keywords Single-server queue · Transient analysis · Lévy processes · Capacity allocation

Mathematics Subject Classification 60K25 · 60G51

1 Introduction

The issue of matching a service system's capacity to stochastic demand induced by its clients arises in many practical settings. Typically, the resources available to satisfy demand are scarce and hence expensive. This forces the manager to consider a trade-off between the system efficiency and the quality of service perceived by its clients.

✉ Britt Mathijsen
b.w.j.mathijsen@tue.nl

[1] Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

[2] Centrum Wiskunde & Informatica, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

In this paper, we focus on this trade-off in the context of the $M/G/1$ queue, in which the variable amenable for optimization is the server speed μ.

In general, optimizing the server speed μ in a single-server queue in a time-homogeneous environment, while trading off congestion levels against capacity allocation costs, does not pose any technical challenges. Typically, the objective function to be minimized, the total cost function, has the shape

$$\Pi_\infty(\mu) = \mathbb{E}[Q_\mu(\infty)] + \alpha\mu = \frac{\lambda\mathbb{E}[B^2]}{2(\mu - \lambda\mathbb{E}[B])} + \alpha\mu, \tag{1.1}$$

where $\mathbb{E}[Q_\mu(\infty)]$ denotes the expected steady-state amount of work given server speed μ, and B describes the service requirement per arrival. The parameter $\alpha > 0$ represents the relative capacity allocation costs incurred by deploying service rate μ. This one-dimensional optimization problem yields the optimizer

$$\mu^\star_\infty = \lambda\mathbb{E}[B] + \sqrt{\frac{\lambda\mathbb{E}[B^2]}{2\alpha}}.$$

Despite the simplicity and tractability of the problem described above, the presence of the *steady-state* measure in the cost function in (1.1) should be handled carefully. By employing this particular cost structure, one automatically agrees with the underlying assumption of the system being sufficiently close to its steady state. However, referring to the practical applications of the single-server model, system parameters rarely remain constant over time. Moreover, planning periods for the optimization problem are naturally finite. Hence, the *true* expected costs incurred, which we denote by $\Pi_T(\mu)$, in addition depend on the length of the planning period T. Consequently, the usage of steady-state models for decision making needs to be justified by a more elaborate time-dependent or *transient* analysis for these type of settings.

The time-dependent behavior of the single-server queue received much attention in queueing theory. First efforts to analyze the time-dependent properties of the $M/G/1$ queue date back to the 1950s and 1960s; for example, [7,10,17,28,29]. The analyses in these papers mostly yield implicit expressions for performance characteristics through Laplace transforms, integro-differential equations and infinite convolutions. More specifically, there is vast literature on the transient analysis of the $M/M/1$ queue, with the goal to derive explicit expressions for queue length characteristics; see, for example, [3,9,23,24]. These works provide a variety of explicit expressions for the transient dynamics, although the complexity of the resulting expressions, typically involving Bessel functions, exposes the intricate intractability of the matter. Consequently, approximation methods for insightful quantification of the dynamics based on numerical [20] or asymptotic methods have become prevalent in more recent literature. The asymptotic methods either exploit knowledge of the evolution of the queueing process as time t grows large [3,21,22], or as the arrival rate λ is increased to infinity [1,2,11]. It is noteworthy that a substantial contribution to the transient literature is made by Abate and Whitt [1–4], who exploit the existence of a decomposition of the mean transient queue length and obtain expressions for the moments of the queue

length and virtual waiting through probabilistic arguments in several queueing models. More recently, asymptotic methods have been used to justify the application of stationary performance measures in Markovian environments or to refine them; see, for example, [12,30]. Other approximative methods known as uniform acceleration expansions [19] have been developed to reveal the asymptotic behavior of the single-server queue as a function of t, which are moreover able to capture time-varying arrival rates.

The majority of the works mentioned above do reflect on the error imposed by usage of steady-state performance metrics instead of the correct time-dependent counterpart. However, no light has been shed on the accumulation of this error over a finite period of time. To the best of our knowledge, the only work that addresses this issue is the paper by Steckley and Henderson [27], who compute an approximation for the error accumulated between the steady-state and transient delay probability. Our analysis on the other hand is centered around the mean workload, which requires a different approach. In addition, the focus in [27] is on performance measures only, while the main goal of our paper is to investigate the quality of staffing rules.

Although the $M/G/1$ queue serves as the leading example in our analysis, we choose to use a more general framework for the arrival process of the queue. Namely, we let the server face a Lévy process. This gives the advantage that once we have obtained the results, we can apply them to broader queue input classes, such as Brownian motion and the Gamma process.

To shed light on the influence of the transience of the queueing process on traditional staffing questions, we will study the capacity allocation problem in the context of cost minimization in which the objective function is $\Pi_T(\mu)$, i.e., a function of both μ and T. We investigate how the invalidity of the stationary assumption is echoed through the operational cost accounting for congestion-related penalties.

Furthermore, we establish a result on the strict convexity of the function $\Pi_T(\mu)$, for almost all values of T (with a few minor exceptions for certain deterministic initial states), which is an essential property for convergence of both cost function and corresponding minimizer to their stationary counterparts.

As it will appear that an exact analysis of this disparity is intractable, we will present an explicit approximate correction to the conventional stationary objective function given by $\Psi(\mu)/T$ and prove that

$$\Pi_T(\mu) = \Pi_\infty(\mu) + \frac{\Psi(\mu)}{T} + O(1/T^2),$$

with the help of recent results from the fluctuation theory of Lévy processes. Based on this refinement, we ultimately examine how incorporating transient effects changes the optimal capacity level and propose a refinement to the steady-state capacity allocation rule,

$$\mu_T^\star = \mu_\infty^\star + \frac{\mu_\bullet}{T} + o(1/T).$$

We moreover deduce an explicit expression for $\mu_\bullet$ in terms of the initial state and the first three moments of the service requirement per arrival. It is noteworthy that similar refined square-root staffing rules have been proposed for multi-server queues in the

Halfin–Whitt regime; see, for example, [14–16,25,31]. In those cases, the relevant decision value is the number of servers and refinements are derived for $\lambda \to \infty$, whereas we consider the regime $T \to \infty$.

Building upon the insights gained through the analysis of this optimality gap, we reflect on the parameter settings of the underlying queueing process in which our refined capacity sizing rule yields significant improvement and in which cases it has little effect. Special emphasis is put on the relationship between the accuracy of the standard procedure and the length of the planning period.

The remainder of the paper is structured as follows. Section 2 is devoted to the model description and presents some preliminary results. The main result will be given in Sect. 3, and results regarding the optimization problem will be discussed in Sect. 4, followed by the validation of our novel techniques through numerical experiments in Sect. 5. We will give some concluding remarks and topics for further research in Sect. 6. We have deferred all proofs to the appendices.

2 Model description

2.1 A queueing model with Lévy input

The model that inspired our study is the standard $M/G/1$ queue starting out of equilibrium. Customers arrive to the queue according to a Poisson process with rate λ, and each arrival has iid service requirement B_i, stemming from a common random variable B. Without loss of generality, we will assume $\mathbb{E}[B] = 1$ throughout. The server is able to remove μ amounts of work from the system per time unit; a variable we will refer to as the *server speed*. For example, if $\mu = 3$ and two customers are in the system with remaining service times 4 and 2, then the queue will be empty 2 time units later, provided that no new arrivals occur in the meantime. Let $N_\lambda(t)$ denote the number of arrivals until time t. Accordingly, the total work generated by the customers is given by

$$Z_\lambda(t) = \sum_{i=1}^{N_\lambda(t)} B_i.$$

Furthermore, define $X_{\lambda,\mu}(t) := Z_\lambda(t) - \mu t$. We call $X_{\lambda,\mu}$ the *net-input process*. More generally, we assume throughout the paper that $X_{\lambda,\mu}$ is a Lévy process. Specifically, we let Z_λ be of the form $Z_\lambda(t) = U(\lambda t)$, where U is a spectrally positive Lévy process generated by the triplet (a, σ, ν) and $\mathbb{E}[U(1)] = 1$. This restriction to spectrally positive processes is equivalent to stating $\nu(-\infty, 0) = 0$ and is a vital assumption in our analysis. Subsequently, we assume the net-input process $X_{\lambda,\mu}$ to be

$$X_{\lambda,\mu}(t) = U(\lambda t) - \mu t, \qquad t \geq 0. \tag{2.1}$$

Note that by setting $a = \sigma = 0$ and $\nu = \lambda F_B$, where F_B is the cumulative distribution function of B, we retrieve the original $M/G/1$ queue. The stochastic process central to our analysis is the *workload process* $Q_{\lambda,\mu}(t), t \geq 0$, which describes the amount of work the server is facing at time t. The net-input process $X_{\lambda,\mu}$ completely determines

the trajectory of $Q_{\lambda,\mu}$, namely

$$Q_{\lambda,\mu}(t) = \max\left\{Q(0) + X_{\lambda,\mu}(t),\ \sup_{s\in[0,t]}\left[X_{\lambda,\mu}(t) - X_{\lambda,\mu}(s)\right]\right\}, \qquad t \geq 0, \quad (2.2)$$

where $Q(0)$ is the initial workload in the system. In fact, $Q_{\lambda,\mu}$ is the reflected version of $X_{\lambda,\mu}$ with reflection barrier at zero. Careful inspection of the structure also reveals that $X_{\lambda,\mu}(t) \equiv X_{\lambda/\mu,1}(\mu t) \equiv X_{1,\mu/\lambda}(\lambda t)$, so that

$$Q_{\lambda,\mu}(t) \stackrel{d}{=} Q_{\lambda/\mu,1}(\mu t) \stackrel{d}{=} Q_{1,\mu/\lambda}(\lambda t) \quad (2.3)$$

for all $\lambda, \mu, t > 0$. This identity will prove to be convenient for the numerical analysis in Sect. 5. For reasons of clarity, we omit the subscript λ in our expressions if no ambiguity is possible.

The process Q_μ is a natural indicator of the level of congestion in the system and therefore a good choice for quantifying the Quality of Service (QoS) received by a client. We remark that alternative processes characterizing congestion in the system can be deduced directly from $Q_\mu(t)$. For example, consider the virtual waiting time process $V_\mu(t)$, which is the waiting time a customer would experience if he arrives at time t. This satisfies the relation $V_\mu(t) \equiv Q_\mu(t)/\mu$ for all $t \geq 0$. Likewise, the expected number of customers in the system $L_\mu(t)$ at time $t \geq 0$ is given by Little's law:

$$\mathbb{E}[L_\mu(t)] = \lambda\,\mathbb{E}[V_\mu(t)] = \frac{\lambda}{\mu}\,\mathbb{E}[Q_\mu(t)].$$

To facilitate our investigation of the queueing model, we end this subsection by introducing some notation regarding the net-input and workload process and by stating a useful preliminary result concerning the stationary process $Q_\mu(\infty)$. Throughout the paper, we assume $\mu > \lambda$ to ensure ergodicity of the queue and convergence in distribution to the limit

$$Q_\mu(\infty) := \lim_{t\to\infty} Q_\mu(t),$$

for any initial state $Q(0) < \infty$. This random variable necessarily coincides with the stationary distribution of $Q_\mu(t)$. By $\kappa_U(\cdot)$ and $\kappa_\mu(\cdot)$, we denote the Lévy exponents of the processes U and X_μ, respectively:

$$\kappa_\mu(\theta) = \log\mathbb{E}[e^{\theta X_\mu(1)}] = \log\mathbb{E}[e^{\theta(U(\lambda)-\mu)}] = \lambda\kappa_U(\theta) - \mu\theta.$$

Furthermore, define $u_k = \mathbb{E}[\{U(1) - \mathbb{E}U(1)\}^k]$ for $k = 2, 3, \ldots$. Using this representation, we obtain the following preliminary result.

Lemma 1 *Let $\mathbb{E}|U(1)| < \infty$, $u_2, u_3 < \infty$ and $\mu > \lambda$. If $Q_\mu(\infty)$ represents the steady-state distribution of the workload process, then*

$$\mathbb{E}\left[Q_\mu(\infty)\right] = \frac{\lambda u_2}{2(\mu-\lambda)}, \qquad \mathbb{E}\left[Q_\mu^2(\infty)\right] = \frac{\lambda^2 u_2^2}{2(\mu-\lambda)^2} + \frac{\lambda u_3}{3(\mu-\lambda)}.$$

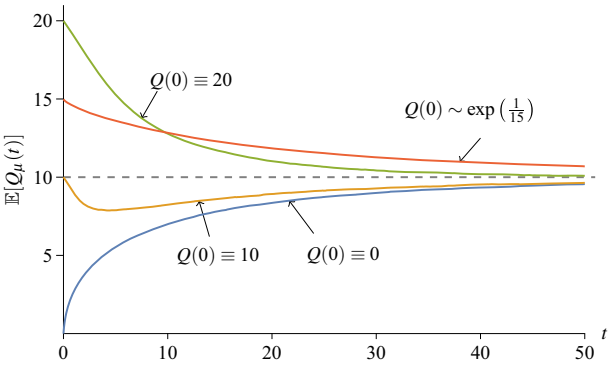

Fig. 1 Time-dependent mean workload in a $M/M/1$ queue with $\lambda = 10$ and $\mu = 11$ for different initial states $Q(0)$. The *dashed line* depicts $\mathbb{E}Q_\mu(\infty)$

The proof of Lemma 1 follows directly by differentiation of the Laplace transform of $Q_\mu(\infty)$ and is therefore straightforward.

2.2 Finite horizon

For the purpose of this paper, we are interested in the dynamics of the workload process within a fixed time frame of length $T > 0$. For all $0 \leq t \leq T$, we assume that the parameters of the queue, λ, μ, u_2, u_3, remain unchanged. If at $t = 0$ the queue is not in steady-state corresponding to the specified parameters of the starting period, the process $\{Q_\mu(t) : t \in [0, T]\}$ differs from its stationary counterpart $Q_\mu(\infty)$. To illustrate this, Fig. 1 depicts the expected value Q_μ in a $M/M/1$ queue as a function of time for several initial workloads $Q(0)$ for a particular setting of λ and μ. Clearly, transient behavior of $\mathbb{E}[Q_\mu(t)]$, for $Q(0) \neq Q_\mu(\infty)$, differs significantly from the steady-state mean with the same system parameters. Note that even if $Q(0) \equiv \mathbb{E}[Q_\mu(\infty)]$, the time-dependent mean does not coincide with the steady-state mean. Moreover, $\mathbb{E}[Q_\mu(t)]$ is not even a strictly increasing nor decreasing function of time. This phenomenon is a consequence of the decomposition of the transient mean into one strictly increasing, and a strictly decreasing term for $Q(0) > 0$, as discussed in [3]. Nonetheless, $Q_\mu(t)$ converges in distribution to $Q_\mu(\infty)$ as $t \to \infty$, if $\mu > \lambda$.

Since the time horizon of our analysis is limited to $t \leq T$, the process may not approach the steady-state distribution sufficiently close to appropriately use its steady-state properties for capacity allocation. To overcome this disparity, we propose a way to include the influence of this transient phase in the capacity allocation problem.

2.3 Cost structure

As mentioned before, we are interested in balancing the QoS and efficiency of the queue by choosing the optimal server speed μ. The adjective *optimal* indicates that we intend to choose the speed according to some objective function. In our case, we

conduct our analysis based on a cost function, which consists of a part accounting for the penalty for congestion in the system and a part for staffing cost. The cost value of both parts is governed by the variable μ. The instantaneous cost incurred at time t equals

$$\mathbb{E}[Q_\mu(t)] + \alpha\mu,$$

where α is a positive constant defining the *relative staffing cost*. Hence, the cost structure we apply is a combination of the transient mean of the workload process and a linear staffing cost. Accumulated and normalized over the period $[0, T]$, the cost function on which the rest of this paper will be based equals

$$\Pi_T(\mu) := \frac{1}{T}\int_0^T \big(\mathbb{E}[Q_\mu(t)] + \alpha\mu\big)\,\mathrm{d}t = \frac{1}{T}\int_0^T \mathbb{E}[Q_\mu(t)]\,\mathrm{d}t + \alpha\mu. \tag{2.4}$$

We use shorthand notation for the normalized congestion costs:

$$C_T(\mu) := \frac{1}{T}\int_0^T \mathbb{E}[Q_\mu(t)]\mathrm{d}t, \tag{2.5}$$

and $C_\infty(\mu) := \mathbb{E}[Q_\mu(\infty)]$. In order to compare the actual costs incurred over the interval $[0, T]$ to the cost function of the queue in stationary conditions, we define

$$\Pi_\infty(\mu) := C_\infty(\mu) + \alpha\mu = \mathbb{E}[Q_\mu(\infty)] + \alpha\mu, \tag{2.6}$$

which allows an explicit expression by Lemma 1. Under mild conditions on the net-input process and the distribution of the initial state, the cost functions coincide for $T \to \infty$.

Proposition 1 *Let $\mu > \lambda$ and assume $\mathbb{E}[U(1)]$, $\mathbb{E}[Q_\mu(0)] < \infty$. Then*

$$\lim_{T\to\infty} \Pi_T(\mu) = \Pi_\infty(\mu).$$

Rewriting (2.4) gives the relation

$$\begin{aligned}\Pi_T(\mu) &= \frac{1}{T}\int_0^T \big(\mathbb{E}[Q_\mu(t)] - \mathbb{E}[Q_\mu(\infty)]\big)\,\mathrm{d}t + \mathbb{E}[Q_\mu(\infty)] + \alpha\mu \\ &= \Omega_T(\mu) + \Pi_\infty(\mu).\end{aligned} \tag{2.7}$$

Section 3 is concerned with the analysis of the correction factor $\Omega_T(\mu)$.

Ultimately, we are concerned with the additional costs incurred by choosing the server speed through minimization of $\Pi_\infty(\mu)$ instead of $\Pi_T(\mu)$. Therefore, we formulate the exact and approximate optimization problems as follows

$$\mu_T^\star := \arg\min_{\mu\geq 0} \Pi_T(\mu), \qquad \mu_\infty^\star := \arg\min_{\mu\geq 0} \Pi_\infty(\mu), \tag{2.8}$$

$$\Pi_T^\star := \Pi_T(\mu_T^\star), \qquad \Pi_\infty^\star := \Pi_T(\mu_\infty^\star). \tag{2.9}$$

In Sect. 4, we turn to the comparison of $\mu_T^\star$ and $\mu_\infty^\star$ as well as the *optimality gap* $\Pi_\infty^\star - \Pi_T^\star$.

3 Analysis of the objective function

From (2.7) it is evident that, for finding an explicit characterization of $\Pi_T(\mu)$, it suffices to study the term $\Omega_T(\mu)$ in more detail. We start by stating the main result of this section, which describes the leading order behavior of $\Omega_T(\mu)$ as T increases.

Theorem 1 *Let $X_\mu(t)$ be of the form (2.1). If $\mathbb{E}[\max(Q(0), Q_\mu(\infty))^3] < \infty$ and $u_2, u_3 < \infty$, then*

$$\begin{aligned}\Omega_T(\mu) &= \frac{\mathbb{E}[Q(0)^2] - \mathbb{E}[Q_\mu(\infty)^2]}{2T(\mu-\lambda)} + O\left(\frac{1}{T^2}\right)\\ &= \frac{1}{2T(\mu-\lambda)}\left(\mathbb{E}[Q(0)^2] - \frac{\lambda^2 u_2^2}{2(\mu-\lambda)^2} - \frac{\lambda u_3}{3(\mu-\lambda)}\right) + O\left(\frac{1}{T^2}\right),\end{aligned}$$

for $\mu > \lambda$.

Note that this expression provides an *approximation* of the actual cost function $\Pi_T(\mu)$. We elaborate on the implications of this additional information on the optimization problem in Sect. 4.

In the remainder of this section, we provide a detailed description of the steps taken to obtain this outcome. We assume a fixed service rate μ throughout the analysis in this section and therefore omit the subscript μ. Proofs of the intermediate results can be found in Appendix 2.

3.1 Constructing a coupling

Before starting our analysis of the correction term $\Omega_T(\mu)$, we introduce some auxiliary notation. By $Q^A(t)$ we denote the workload process as described in Sect. 2.1 with initial state A and $\mathbb{E}_A$ the expectation with respect to any nonnegative random variable A, which is independent of the net-input process X. To be able to compare $\mathbb{E}[Q(t)]$ and $\mathbb{E}[Q(\infty)]$ as in $\Omega_T(\mu)$, we will use a coupling technique. Observe that by the definition of the stationary distribution $Q(\infty) \stackrel{d}{=} Q^{Q(\infty)}(t)$ for all $t \geq 0$ and therefore $\mathbb{E}[Q(\infty)] = \mathbb{E}_{Q(\infty)}[Q^{Q(\infty)}(t)]$. Furthermore, $\mathbb{E}[Q(t)] = \mathbb{E}_{Q(0)}[Q^{Q(0)}(t)]$. Hence, quantifying the difference between the transient and stationary mean is equivalent to comparing the workload processes of two queues starting in two different (random) states at $t = 0$.

We starting our analysis for two queues starting in two *deterministic* states $x, y \geq 0$, respectively. At the end of our analysis, we will obtain the original form by replacing x with $Q(0)$ and y with $Q(\infty)$.

Equation (2.2) shows that all randomness in the workload process originates from the process $X(t)$. With this in mind, we couple the processes $Q^x(t)$ and $Q^y(t)$ on a

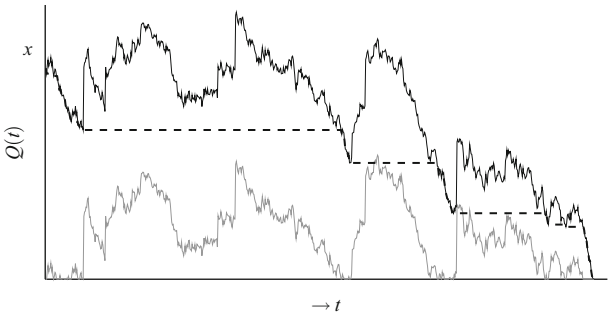

Fig. 2 Sample path visualization of the processes $Q^x(t)$ (*solid*), $Q^0(t)$ (*gray*) and $Y^{x,0}(t)$ (*dashed*)

sample path level by feeding both queues the same net-input process $X(t)$ for $t \geq 0$. This allows us to compare the processes in the same probability space, so that $\mathbb{E}[Q^x(t)] - \mathbb{E}[Q^y(t)] = \mathbb{E}[Q^x(t) - Q^y(t)]$ for all $t \geq 0$. Define

$$Y^{x,y}(t) := Q^x(t) - Q^y(t)$$

and

$$\Omega_T^{x,y} := \frac{1}{T} \int_0^T \mathbb{E}\left[Y^{x,y}(t)\right] \mathrm{d}t.$$

A possible sample path triple for $Q^x(t)$, $Q^0(t)$ and $Y^{x,0}(t)$ is depicted in Fig. 2. As we see from this figure, $Y^{x,0}(t)$ has nice structural properties which we will exploit in the next subsection.

3.2 Difference process and leading order behavior of the correction term

We further examine the *difference process* $Y^{x,y}(t)$ with $x > y$. Recall from (2.2),

$$Q^z(t) = \max\{z + X(t), \sup_{0<s\leq t}[X(t) - X(s)]\} = X(t) + \max\{z, -\inf_{0\leq s\leq t} X(s)\}, \quad (3.1)$$

for any initial state $z \geq 0$, where $X(t)$ is a Lévy process with no negative jumps. Let $\tau^z(w)$, $0 \leq w < z$, denote the first passage time of level w by the process starting in z, i.e.

$$\tau^z(w) := \inf\left\{t \geq 0 \mid Q^z(t) \leq w\right\}.$$

Then it is easily seen that for all $z \geq 0$,

$$Q^z(t) = \begin{cases} z + X(t), & \text{if } t < \tau^z(0), \\ \sup_{0<s\leq t}[X(t) - X(s)], & \text{if } t \geq \tau^z(0). \end{cases}$$

Consequently,

$$Y^{x,y}(t) = \begin{cases} x - y, & \text{if } t < \tau^y(0), \\ \inf_{0<s\le t}\{x + X(s)\}, & \text{if } \tau^y(0) \le t < \tau^x(0), \\ 0, & \text{if } t \ge \tau^x(0). \end{cases} \tag{3.2}$$

Using this representation, we can identify

$$\Omega_T^{x,y} = \frac{1}{T}\,\mathbb{E}\left[\int_0^{\tau^x(0)\wedge T} Y^{x,y}(t)\mathrm{d}t\right],$$

where $\wedge$ denotes the minimum operator, due to the fact that $Y^{x,y}(t) = 0$ for $t \ge \tau^x(0)$. Subsequently, we decompose $\Omega_T^{x,y}$ into two terms:

$$\Psi_T^{x,y} := \frac{1}{T}\int_0^{\infty} \mathbb{E}[Y^{x,y}(t)]\,\mathrm{d}t \qquad \text{and} \qquad \Delta_T^{x,y} := \Omega_T^{x,y} - \Psi_T^{x,y}. \tag{3.3}$$

Note that $\Psi_T^{x,y}$ is obtained by replacing T by ∞ only in the integration bound. It is customary in the literature, particularly in the area of stochastic simulation, to compare the truncated integral to its natural expansion of the integration range to a semi-infinite interval; see, for example, [6, Prop. 2.1].The truncated integral connects to the long-run average estimator of a certain performance metric, whereas the infinite integral reflects its exact expectation. The decomposition in (3.3) is insightful, because $\Psi_T^{x,y}$ prescribes the leading order behavior of $\Omega_T^{x,y}$, while $\Delta_T^{x,y}$ captures the smaller order error term. In this section, we only consider $\Psi_T^{x,y}$. Sect. 3.3 investigates the magnitude of $\Delta_T^{x,y}$. The next preliminary result presents a useful property of $\Psi_T^{x,y}$.

Lemma 2 *Let $x > y$. If $\mathbb{E}[\tau^x(0)] < \infty$, then*

$$\Psi_T^{x,y} = \frac{1}{T}\,\mathbb{E}[\tau^y(0)](x - y) + \Psi_T^{x-y,0}. \tag{3.4}$$

This leaves us with two unknowns: $\mathbb{E}[\tau^y(0)]$ and $\Psi_T^{x-y,0}$. The next lemma gives an equivalent form for the latter.

Lemma 3 *If $\mathbb{E}[\tau^z(0)] < \infty$, then for all $z \ge 0$,*

$$\Psi_T^{z,0} = \int_0^z \mathbb{E}[\tau^w(0)]\,\mathrm{d}w. \tag{3.5}$$

Since the term $\mathbb{E}[\tau^z(0)]$, for several values of z, appears in many of the preliminary results, we devote our attention to this in the next subsection.

First passage time

When studying the first passage time of level $0 \le w < z$, $\tau^z(w)$, of the workload process starting in z, we first observe that $\{\tau^z(z - w)\}_{w=0}^z$ is a spectrally positive Lévy process itself, also visible through Fig. 2. More precisely, it is a subordinator,

i.e., a Lévy process whose paths are almost surely non-decreasing [18]. In order to calculate $\mathbb{E}[\tau^z(z-w)]$, we use theory presented in [26, Section 46], although results presented there are valid for spectrally *negative* Lévy processes, as opposed to the absence of negative jumps in our case. Nonetheless, our setting is easily transformed into this framework by observing that $\hat{X} \equiv -X$, that is $\hat{X}(t) = -X(t)$ for all $t \geq 0$, is spectrally negative. Furthermore, let

$$\hat{\tau}^0(w) := \inf\{t \geq 0 : \hat{X}(t) \geq w\} = \inf\{t \geq 0 : z + X(t) \leq z - w\} = \tau^z(z-w). \tag{3.6}$$

For completeness, we cite [26, Thm. 46.3].

Theorem 2 *Let $\hat{X}(t)$ be a spectrally negative Lévy process with generating triplet $(-a, \sigma, \hat{\nu})$ and $\hat{\tau}^0(y)$ its corresponding hitting time process. Define $\Upsilon(\theta)$ for $\theta \geq 0$ as*

$$\Upsilon(\theta) = -a\theta + \tfrac{1}{2}\sigma^2\theta^2 + \int_{-\infty}^{0} \left(e^{\theta x} - 1 - \theta x \mathbf{1}_{[-1,0)}(x)\right) \hat{\nu}(\mathrm{d}x). \tag{3.7}$$

Then $\Upsilon(\theta)$ is strictly increasing and continuous, $\Upsilon(0) = 0$, and $\Upsilon(\theta) \to \infty$ as $\theta \to \infty$. For $w \geq 0$ and $0 \leq u < \infty$, we have

$$\mathbb{E}\left[\exp(-u\hat{\tau}^0(w))\right] = \exp(-w\,\Upsilon^{-1}(u)), \tag{3.8}$$

where $\theta = \Upsilon^{-1}(u)$ is the inverse function of $u = \Upsilon(\theta)$.

This immediately induces an expression for $\mathbb{E}[\tau^w(0)]$ and henceforth $\Psi^{z,0}$.

Corollary 1 *Let $X(t)$ be a spectrally positive Lévy process defined as in (2.1) with $\mu > \lambda$. Let $\Psi_T^{z,0}$ as in (3.5). Then*

$$\Psi_T^{z,0} = \frac{z^2}{2T(\mu - \lambda)}.$$

Furthermore, if $x, y \geq 0$, then

$$\Psi_T^{x,y} = \frac{x^2 - y^2}{2T(\mu - \lambda)}. \tag{3.9}$$

Randomization

As we stated before, we easily obtain the original Ω_T from $\Omega_T^{x,y}$ through substitution of x and y by $Q(0)$ and $Q(\infty)$, respectively, and taking the expectation. In the previous paragraph, we deduced an explicit expression for $\Psi_T^{x,y}$, the leading order term for $\Omega_T^{x,y}$. Therefore, we equivalently get an approximation for Ω_T, given by

$$\Psi_T := \frac{1}{T}\int_0^\infty (\mathbb{E}[Q(t)] - \mathbb{E}[Q(\infty)])\ \mathrm{d}t,$$

through randomization of x and y in $\Psi_T^{x,y}$. By combining the results in Corollary 1, Lemma 1 and Proposition 2, which is given at the end of this section, we directly prove the result in Theorem 1.

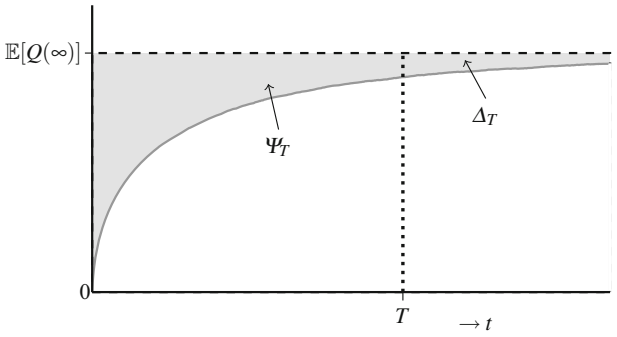

Fig. 3 Visualization of Ω_T and Ψ_T as the area between the curves $E[Q(t)]$, $\mathbb{E}[Q(\infty)]$ for $Q(0) = 0$

3.3 Truncation error

In order to get a better comprehension of the properties of Ψ_T, we depict the value in terms of the (infinite) region between the curves $\mathbb{E}[Q(t)]$, $\mathbb{E}[Q(\infty)]$ and the vertical axis for the case $Q(0) \equiv 0$ in Fig. 3. In this figure, Ω_T is given by the area enclosed by the two curves, the vertical axis and the line $t = T$. One can see that the main contribution to the correction term Ω_T is given for small t. As t increases, the difference between transient and stationary mean decreases. Hence, for moderate values of T, the contribution to the integral in (3.3) is only minor compared to the contribution over the interval $[0, T]$.

Recall the definition of $\Delta_T^{x,y}$ as in (3.3). As we alluded to in Sect. 3.2, we claim the contribution of $\Delta_T^{x,y}$ to $\Omega_T^{x,y}$ is negligible compared to $\Psi_T^{x,y}$. Also note that

$$\Delta_T := \Omega_T - \Psi_T = -\frac{1}{T}\int_T^\infty \mathbb{E}[Q(t)] - \mathbb{E}[Q(\infty)]\,\mathrm{d}t \tag{3.10}$$

can be derived through $\Delta_T^{x,y}$ in a similar manner as we did for $\Psi_T^{x,y}$ to obtain Ψ_T. To substantiate our claim, we compute an upper bound for $\Delta_T^{x,y}$ of order $1/T^2$. The existence of such an upper bound poses a limit on the error this tail integral contributed to the cost structure as a whole.

Proposition 2 *Let* $x, y \geq 0$ *and* $\mathbb{E}[\max(Q(0), Q_\mu(\infty))^3] < \infty$. *Then*

$$|\Delta_T^{x,y}| \leq \frac{1}{T^2}\left(\frac{\max(y,x)^3}{3(\mu-\lambda)^2} + \frac{u_2 \max(y,x)^2}{2(\mu-\lambda)^3}\right)$$

and

$$|\Delta_T| \leq \frac{1}{T^2}\left(\frac{\mathbb{E}\left[\max(Q(0), Q_\mu(\infty))^3\right]}{3(\mu-\lambda)^2} + \frac{u_2\mathbb{E}\left[\max(Q(0), Q_\mu(\infty))^2\right]}{2(\mu-\lambda)^3}\right).$$

Remark In the case where the net-input process X is light-tailed, that is, there exists $u > 0$ such that $\mathbb{E}[e^{uX(1)}] < \infty$, it can be shown that the truncation error is of order $e^{-\beta T}/T$ for some $\beta > 0$. See Appendix 2 for details.

4 Optimization

The result in Theorem 1, characterizing the leading order behavior of $\Omega_T(\mu)$, also reveals the behavior of $\Pi_T(\mu)$ in leading order. Namely,

$$\Pi_T(\mu) = \Pi_\infty(\mu) + \Psi_T(\mu) + O(1/T^2).$$

In fact, this representation naturally gives rise to an *approximation* of the actual cost function:

$$\hat{\Pi}_T(\mu) := \Pi_\infty(\mu) + \Psi_T(\mu). \tag{4.1}$$

Denote the corresponding minimizer of $\hat{\Pi}_T$ by

$$\hat{\mu}_T^\star := \arg\min_{\mu \geq 0} \hat{\Pi}_T(\mu), \qquad \hat{\Pi}_T^\star := \hat{\Pi}_T(\hat{\mu}_T^\star) \tag{4.2}$$

in addition to the definitions in (2.8) and (2.9). This section is devoted to the analysis of the minimizers $\mu_T^\star$, $\hat{\mu}_T^\star$ and $\mu_\infty^\star$, and the optimality gap for the two approximations.

Throughout this section, we assume that $u_2, u_3 < \infty$ and $\mathbb{E}[Q(0)^2] < \infty$.

By its definition in (2.6) and Lemma 1, we have an exact expression for the steady-state cost function:

$$\Pi_\infty(\mu) = \frac{\lambda u_2}{2(\mu - \lambda)} + \alpha\mu.$$

It is easily verified that Π_∞ is strictly convex in μ, for instance by observing that $\Pi_\infty''(\mu) > 0$ for all $\mu > \lambda$. Therefore, Π_∞ has a unique global minimizer and

$$\mu_\infty^\star = \lambda + \sqrt{\frac{\lambda u_2}{2\alpha}}, \qquad \Pi_\infty^\star = \alpha\lambda + \sqrt{2\alpha\lambda u_2}. \tag{4.3}$$

We are interested in the relation between $\mu_\infty^\star$ and $\mu_T^\star$, and between $\hat{\mu}_T^\star$ and $\mu_T^\star$. Since $\Pi_T(\mu) = \Pi_\infty(\mu) + O(1/T)$ for all $\mu > \lambda$, we have pointwise convergence of the sequence Π_T, as well as $\hat{\Pi}_T$, to Π_∞ for $T \to \infty$; we also expect $\mu_T^\star \to \mu_\infty^\star$ and $\hat{\mu}_T^\star \to \mu_\infty^\star$ for $T \to \infty$. Before proving that this convergence indeed holds, we present a result on the strict convexity of the function Π_T.

Lemma 4 *Let $\mu \geq 0$. The function $\Pi_T(\mu)$ is*

- *convex in μ, if $Q(0) \equiv x$, $T < x/\mu$ and $\sigma = 0$,*
- *strictly convex in μ, otherwise.*

Building upon strict convexity of both $\Pi_T(\mu)$ and $\Pi_\infty(\mu)$ for $\mu > \lambda$, we derive the following convergence result.

Proposition 3 *Let $\mu_T^\star$, $\hat{\mu}_T^\star$ and $\mu_\infty^\star$ be as defined in (2.8) and (4.2). Then*

$$\mu_T^\star \to \mu_\infty^\star \qquad \text{and} \qquad \hat{\mu}_T^\star \to \mu_\infty^\star,$$

for $T \to \infty$.

The next result describes a refinement of $\mu_T^\star$ in terms of $\mu_\infty^\star$.

Proposition 4 *For T sufficiently large,*

$$\mu_T^\star = \mu_\infty^\star + \frac{\mu_\bullet}{T} + o(1/T),$$

where

$$\mu_\bullet = \frac{\mathbb{E}[Q(0)^2]}{\sqrt{8\lambda u_2 \alpha}} - \frac{u_3}{3u_2} - 3\sqrt{\frac{\alpha\lambda u_2}{8}}. \tag{4.4}$$

Based on Proposition 4, we propose a *corrected staffing rule*, accounting for the finite horizon:

$$\tilde{\mu}_T^\star = \left[\mu_\infty^\star + \frac{\mu_\bullet}{T}\right]^+, \tag{4.5}$$

with $\mu_\bullet$ as in (4.4). Here $[x]^+ := \max\{x, 0\}$, which ensures the value of $\tilde{\mu}_T^\star$ is nonnegative and thus is a feasible solution of the optimization problem. This refined capacity allocation rule is expected to reduce the costs incurred in transient settings. However, the value of particular interest to us is the cost penalty for using either one of the approximations rather than the actual minimum $\mu_T^\star$, that is, the *optimality gap*. As it happens, we deduce the order of the optimality gap for $\mu_\infty^\star$ with the help of the explicit form of $\mu_\bullet$ given in (4.4), which is stated in the next proposition. The proof is given in Appendix 3.

Proposition 5 *Let $\mu_\infty^\star$ be as in (4.3). Then,*

$$\Pi_\infty^\star - \Pi_T^\star = O(1/T^2).$$

5 Numerical experiments

5.1 Influence of $\Omega_T(\mu)$

We first assess the contribution of the correction to the cost function provided by Theorem 1. In other words, we investigate whether $\hat{\Pi}_T(\mu)$ as in (2.4) yields a significantly better fit to $\Pi_T(\mu)$ than $\Pi_\infty(\mu)$ does. Note that these three functions only differ in the costs describing the congestion. Therefore, we limit our study in this subsection to the evaluation of $C_T(\mu)$ as in (2.5) with stationary equivalent $C_\infty(\mu) = \mathbb{E}[Q_\mu(\infty)]$. Our novel approximation hence reads

$$\hat{C}_T(\mu) := C_\infty(\mu) + \Omega_T(\mu),$$

with $\Omega_T(\mu)$ given in Theorem 1. We conduct our numerical experiments based on three models, namely:

1. $M/M/1$ queue: $U(t)$ is a unit rate compound Poisson process with exponentially distributed increments. We have $u_2 = 2$, $u_3 = 3$, so that

$$\hat{C}_T(\mu) = \frac{\lambda}{\mu - \lambda} + \frac{1}{T(\mu - \lambda)} \left(\frac{x^2}{2} - \frac{\lambda^2}{(\mu - \lambda)^2} - \frac{\lambda}{\mu - \lambda} \right). \tag{5.1}$$

2. M/Pareto/1 queue: $U(t)$ is a unit rate compound Poisson process with Pareto increments. The Pareto distribution deserves special attention due to its heavy tailed nature, having tail probability $\bar{F}(x) = (x/k)^{-\gamma}$, if $x \geq k$, and 1 otherwise. It is well-known that heavy-tailed service times lead to long relaxation time. For our purposes, we fix shape parameter $\gamma = 16/5$ and scale parameter $k = 11/16$, so that $\beta = 1$, $u_2 = 121/96$, $u_3 = 1331/256$ and $u_k = \infty$ for all $k > 3$. Hence,

$$\hat{C}_T(\mu) = \frac{121\lambda}{192(\mu - \lambda)} + \frac{1}{2T(\mu - \lambda)} \left(x^2 - \frac{(121\lambda/96)^2}{2(\mu - \lambda)^2} - \frac{1331\lambda/256}{2(\mu - \lambda)} \right). \tag{5.2}$$

3. Reflected Brownian motion: $U(t)$ is Brownian motion with drift 1 and infinitesimal variance σ^2. We have $u_2 = \sigma^2$, $u_3 = 0$, so that

$$\hat{C}_T(\mu) = \frac{\lambda\sigma^2}{2(\mu - \lambda)} + \frac{1}{2T(\mu - \lambda)} \left(x^2 - \frac{\lambda^2\sigma^4}{2(\mu - \lambda)^2} \right). \tag{5.3}$$

In light of the equivalence relations in (2.3), we only consider the case $\lambda = 1$. The cost values for general values of λ follow by appropriate rescaling of μ and T.

For the $M/M/1$ and M/Pareto/1 queues, we obtained the function $C_T(\mu)$ through simulation and the results are accurate up until a 95% confidence interval of width 10^{-3}. For reflected Brownian motion, we used the explicit distribution function given in [13] for double numerical integration. The results for several values of T and two different starting states are depicted in Figs. 4, 5 and 6. These plots also include the approximated functions $\hat{C}_T(\mu)$. We make a few observations based on these figures.

First, we indeed note the pointwise convergence of $\hat{C}_T(\mu)$ to $\hat{C}_\infty(\mu)$ as T grows, for all μ in all three cases. However, the difference between the stationary costs and those for small values of T can be significant. This is most clear in the plots with $x = 2.5$ and when μ is close to λ, i.e., it is in heavy traffic. In these scenarios, it is evident that refinements to the stationary cost function are needed. $\hat{C}_T(\mu)$ does a fairly good job at providing such correction, especially for moderate values of μ.

Furthermore, we note that $C_T(\mu)$ approaches $C_\infty(\mu)$ from below for $x = 0$ for any value of μ, while this is not strictly the case for $x > 0$. $\hat{C}_T(\mu)$ correctly captures the sign of this correction.

Finally, observe that $\hat{C}_T(\mu) \to -\infty$ as μ approaches λ from above. This divergence is clear from the expressions in (5.1)-(5.3). Our correction term relies on the premise that under the coupling scheme, the sample paths of the two queues starting from different states have hit with high probability. This is equivalent to stating that the 'largest' of the two queues is emptied at least once before time T. However, as μ approaches λ, the system enters heavy traffic, and hence the hitting time of the zero

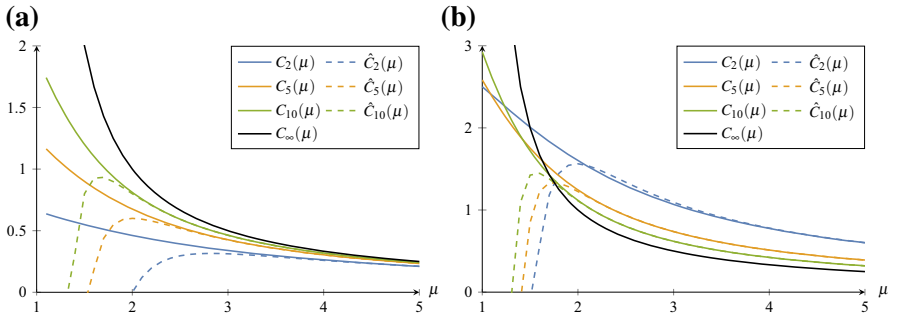

Fig. 4 Comparison of exact waiting cost function $C_T(\mu)$ against corrected cost function $\hat{C}_T(\mu)$ and PSA cost function $C_\infty(\mu)$ for $T = 2, 5$ and 10 for the $M/M/1$ queue with $\lambda = 1$. **a** $x = 0$. **b** $x = 2.5$

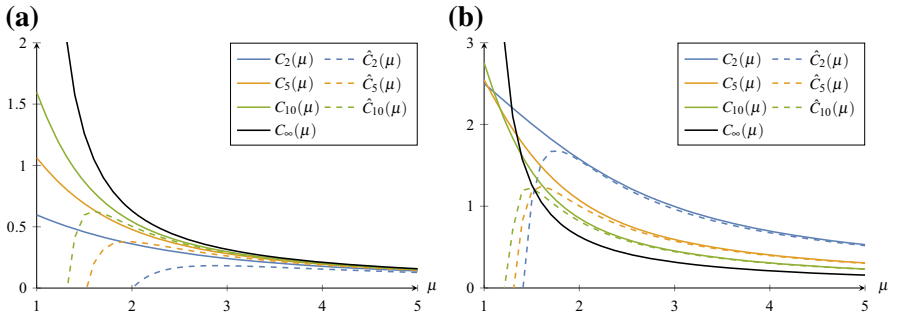

Fig. 5 Comparison of exact waiting cost function $C_T(\mu)$ against corrected cost function $\hat{C}_T(\mu)$ and PSA cost function $C_\infty(\mu)$ for $T = 2, 5$ and 10 for the $M/\text{Pareto}/1$ queue with $\lambda = 1$. **a** $x = 0$. **b** $x = 2.5$

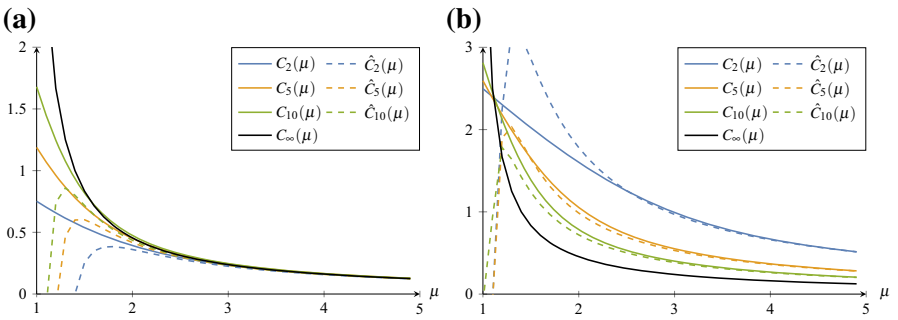

Fig. 6 Comparison of exact waiting cost function $C_T(\mu)$ against corrected cost function $\hat{C}_T(\mu)$ and PSA cost function $C_\infty(\mu)$ for $T = 2, 5$ and 10 for reflected Brownian motion with $\sigma = 1$. **a** $x = 0$. **b** $x = 2.5$

barrier is set to run off to infinity. Consequently, this causes our approximation to be inaccurate for small values of μ.

5.2 Validation of corrected staffing rule

In this section we examine whether the corrected staffing rule $\tilde{\mu}_T^\star$ as in (4.5) indeed yields a significant cost reduction over the choice of $\mu_\infty^\star$ by comparing their true costs $\Pi_T(\tilde{\mu}_T^\star)$ and $\Pi_T(\mu_\infty^\star)$. We conduct this comparison for different values of the parameters, α, T and starting state x through numerical experiments. The three models on which we do our calculations are the $M/M/1$ queue, the $M/$Pareto$/1$ queue and the reflected Brownian motion, as introduced in the previous subsection. We again focus on $\lambda = 1$ only.

For each of the three models, we adhere to the following setup. The quality of both staffing rules is assessed for $\alpha = 0.1$, 1 and 2, resembling three modes of valuation of the QoS in the system. As a benchmark, observe that the expected workload in steady-state conditions with staffing level $\mu_\infty^\star$ equals

$$C_\infty(\mu_\infty^\star) = \sqrt{\frac{\alpha\lambda u_2}{2}}.$$

For each value of α, we consider two scenarios: One in which the system starts empty, i.e., $x = 0$, and one in which the initial state is double this benchmark value, thus $x = \sqrt{2\alpha\lambda u_2}$. The numerics are presented for each model separately. We discuss general conclusions drawn from these results afterwards.

$M/M/1$ queue

As we discussed before, if U is a unit rate compound Poisson process with exponentially distributed increments, then Q_μ describes the workload process in an $M/M/1$ queue. For this setting, we get

$$\mu_\infty^\star = \lambda + \sqrt{\frac{\lambda}{\alpha}}, \qquad \tilde{\mu}_T^\star = \left[\lambda + \sqrt{\frac{\lambda}{\alpha}} + \frac{1}{T}\left(\frac{x^2}{4\sqrt{\lambda\alpha}} - 1 - \frac{3}{2}\sqrt{\lambda\alpha}\right)\right]^+.$$

Table 1 presents the actual costs corresponding to these two staffing levels for different value of x and α.

*M/*Pareto*/1 queue*

In the case where the service requirements follow a Pareto distribution with shape parameter $\gamma = 16/5$, the staffing rule becomes

$$\mu_\infty^\star = \lambda + \frac{11}{8}\sqrt{\frac{\lambda}{3\alpha}},\ \tilde{\mu}_T^\star = \left[\lambda + \frac{11}{8}\sqrt{\frac{\lambda}{3\alpha}} + \frac{1}{T}\left(\frac{2x^2}{11\sqrt{\lambda\alpha/3}} - \frac{11}{8} - \frac{11\sqrt{3\lambda\alpha}}{16}\right)\right]^+.$$

The numerical results are given in Table 2.

Table 1 Comparison of costs for the $M/M/1$ queue for steady-state and corrected staffing rules and relative cost improvement (r.c.i.)

		$x=0$					$x=2\sqrt{\alpha}$				
α	T	$\mu^\star_\infty$	$\Pi_T(\mu^\star_\infty)$	$\tilde{\mu}^\star_T$	$\Pi_T(\tilde{\mu}^\star_T)$	r.c.i.	$\mu^\star_\infty$	$\Pi_T(\mu^\star_\infty)$	$\tilde{\mu}^\star_T$	$\Pi_T(\tilde{\mu}^\star_T)$	r.c.i.
0.1	1	4.162	0.620	2.688	0.536	0.136	4.162	0.682	2.688	0.536	0.214
	2	4.162	0.669	3.425	0.641	0.041	4.162	0.700	3.425	0.641	0.085
	5	4.162	0.706	3.867	0.703	0.005	4.162	0.719	3.867	0.703	0.022
	10	4.162	0.719	4.015	0.719	0.001	4.162	0.726	4.015	0.719	0.010
1	1	2.000	2.309	0.000	0.500	0.783	2.000	3.500	0.500	2.750	0.214
	2	2.000	2.461	0.750	1.480	0.398	2.000	3.218	1.250	3.125	0.029
	5	2.000	2.675	1.500	2.400	0.103	2.000	3.043	1.700	2.968	0.025
	10	2.000	2.810	1.750	2.726	0.030	2.000	3.007	1.850	2.980	0.009
2	1	1.707	3.744	0.000	0.500	0.866	1.707	5.889	0.000	3.328	0.435
	2	1.707	3.924	0.146	1.232	0.686	1.707	5.547	0.854	4.682	0.156
	5	1.707	4.209	1.083	3.343	0.206	1.707	5.114	1.366	4.910	0.040
	10	1.707	4.424	1.395	4.108	0.071	1.707	4.945	1.536	4.868	0.016

Table 2 Comparison of costs for the M/Pareto/1 queue for steady-state and corrected staffing rules and relative cost improvement (r.c.i.)

		$x=0$					$x=11/4\cdot\sqrt{\alpha/3}$				
α	T	$\mu^\star_\infty$	$\Pi_T(\mu^\star_\infty)$	$\tilde{\mu}^\star_T$	$\Pi_T(\tilde{\mu}^\star_T)$	r.c.i.	$\mu^\star_\infty$	$\Pi_T(\mu^\star_\infty)$	$\tilde{\mu}^\star_T$	$\Pi_T(\tilde{\mu}^\star_T)$	r.c.i.
0.1	1	3.510	0.524	1.759	0.461	0.120	3.510	0.573	2.010	0.562	0.019
	2	3.510	0.555	2.635	0.539	0.029	3.510	0.580	2.760	0.574	0.010
	5	3.510	0.580	3.160	0.578	0.003	3.510	0.591	3.210	0.589	0.002
	10	3.510	0.590	3.335	0.590	0.000	3.510	0.596	3.360	0.595	0.001
1	1	1.794	2.076	0.000	0.500	0.759	1.794	2.989	0.000	2.088	0.302
	2	1.794	2.190	0.511	1.291	0.411	1.794	2.790	0.610	2.588	0.072
	5	1.794	2.345	1.281	2.108	0.101	1.794	2.638	1.320	2.607	0.012
	10	1.794	2.441	1.537	2.371	0.029	1.794	2.597	1.557	2.585	0.005
2	1	1.561	3.427	0.000	0.500	0.854	1.561	5.087	0.000	2.745	0.460
	2	1.561	3.567	0.032	1.050	0.706	1.561	4.832	0.172	3.417	0.293
	5	1.561	3.779	0.950	3.012	0.203	1.561	4.499	1.006	4.313	0.041
	10	1.561	3.935	1.255	3.356	0.147	1.561	4.351	1.284	4.304	0.011

Just as in the results for the $M/M/1$ queue, we observe a higher reduction for larger value of α and T. Also, again $\tilde{\mu}_T < \mu^\star_\infty$. Hence, the conclusions for the M/Pareto/1 queue are similar to those of the $M/M/1$ queue.

Reflected Brownian motion

Table 3 Comparison of costs for RBM with $\sigma = 1$ for steady-state and corrected staffing rules and relative cost improvement (r.c.i.)

		$x = 0$					$x = \sqrt{2\alpha}$				
α	T	$\mu^\star_\infty$	$\Pi_T(\mu^\star_\infty)$	$\tilde{\mu}^\star_T$	$\Pi_T(\tilde{\mu}^\star_T)$	r.c.i.	$\mu^\star_\infty$	$\Pi_T(\mu^\star_\infty)$	$\tilde{\mu}^\star_T$	$\Pi_T(\tilde{\mu}^\star_T)$	r.c.i.
0.1	1	3.236	0.525	2.901	0.518	0.013	3.236	0.565	3.124	0.564	0.001
	2	3.236	0.536	3.068	0.534	0.003	3.236	0.556	3.180	0.556	0.000
	5	3.236	0.543	3.169	0.542	0.000	3.236	0.551	3.214	0.551	0.000
	10	3.236	0.545	3.203	0.545	0.000	3.236	0.549	3.225	0.549	0.000
1	1	1.500	3.420	0.000	0.833	0.756	1.500	4.741	1.000	3.984	0.160
	2	1.500	3.539	0.750	2.386	0.326	1.500	4.579	1.250	4.293	0.063
	5	1.500	3.707	1.200	3.363	0.093	1.500	4.335	1.400	4.274	0.014
	10	1.500	3.820	1.350	3.705	0.030	1.500	4.190	1.450	4.175	0.004
2	1	1.500	3.420	0.000	0.833	0.756	1.500	4.741	1.000	3.984	0.160
	2	1.500	3.539	0.750	2.386	0.326	1.500	4.579	1.250	4.293	0.063
	5	1.500	3.707	1.200	3.363	0.093	1.500	4.335	1.400	4.274	0.014
	10	1.500	3.820	1.350	3.705	0.030	1.500	4.190	1.450	4.175	0.004

In the case where the input process U is Brownian motion with drift 1 and infinitesimal variance σ^2, the steady-state staffing rule and its corrected version reduce to

$$\mu^\star_\infty = \lambda + \sqrt{\frac{\lambda\sigma^2}{2\alpha}}, \qquad \tilde{\mu}^\star_T = \left[\lambda + \sqrt{\frac{\lambda\sigma^2}{2\alpha}} + \frac{1}{2\sqrt{2}\,T}\left(\frac{x^2}{\sqrt{\lambda\alpha}\sigma} - 3\sigma\sqrt{\alpha\lambda}\right)\right]^+.$$

In Tables 3 and 4, the costs obtained through numerical evaluation are presented for several values of x, T. We also vary σ to examine the influence of the volatility of the arrival process on the quality of the staffing rules.

The observations on the influence of α, x and T are similar to those for the $M/M/1$ queue and the M/Pareto/1 queue. However, here we see little improvement by the corrected staffing rule for small values of α for both values of x. The results in Tables 3 and 4 also suggest that the reduction is smaller for larger values of σ.

5.3 Discussion

Based upon these numerical results in Tables 1, 2, 3 and 4, we make a few remarks. The three models roughly exhibit similar behavior as T, x and α are varied.

Non-surprisingly, we note that $\tilde{\mu}_T$ approaches $\mu^\star_\infty$ with increasing T, which also implies that the cost reduction achieved by the corrected staffing rule vanishes as $T \to \infty$. Also, we observe that in all scenarios examined, the cost reduction increases with α. This can be explained through investigation of the objective function Π_T as a function of μ. Namely, for α small, the curve is relatively flat around the true

Table 4 Comparison of costs for RBM with $\sigma = 2$ for steady-state and corrected staffing rules and relative cost improvement (r.c.i.)

		$x = 0$					$x = 2\sqrt{2\alpha}$				
α	T	$\mu^\star_\infty$	$\Pi_T(\mu^\star_\infty)$	$\tilde{\mu}^\star_T$	$\Pi_T(\tilde{\mu}^\star_T)$	r.c.i.	$\mu^\star_\infty$	$\Pi_T(\mu^\star_\infty)$	$\tilde{\mu}^\star_T$	$\Pi_T(\tilde{\mu}^\star_T)$	r.c.i.
0.1	1	5.472	0.950	4.801	0.936	0.015	5.472	1.030	5.249	1.029	0.001
	2	5.472	0.972	5.137	0.968	0.003	5.472	1.012	5.360	1.012	0.000
	5	5.472	0.985	5.338	0.985	0.000	5.472	1.002	5.427	1.002	0.000
	10	5.472	0.990	5.405	0.990	0.000	5.472	0.998	5.450	0.998	0.000
1	1	2.414	3.176	0.293	1.546	0.513	2.414	4.633	1.707	4.228	0.087
	2	2.414	3.356	1.354	2.690	0.199	2.414	4.375	2.061	4.247	0.029
	5	2.414	3.573	1.990	3.411	0.045	2.414	4.094	2.273	4.073	0.005
	10	2.414	3.689	2.202	3.646	0.012	2.414	3.966	2.344	3.962	0.001
2	1	2.000	4.839	0.000	1.339	0.723	2.000	7.481	1.000	5.967	0.202
	2	2.000	5.078	0.500	2.773	0.454	2.000	7.158	1.500	6.585	0.080
	5	2.000	5.414	1.400	4.726	0.127	2.000	6.670	1.800	6.549	0.018
	10	2.000	5.639	1.700	5.409	0.041	2.000	6.380	1.900	6.349	0.005

optimum $\mu^\star_T$. Hence, in this case a moderate deviation from $\mu^\star_T$ will likely not lead to a significant cost increase. However, as α becomes larger, i.e., server efficiency is valued more than minimization of congestion, the curve becomes more sharp around $\mu^\star_T$, and hence more accurate approximations of $\mu^\star_T$ are required to achieve an acceptable cost level. Hence, the corrected staffing rule (4.5) proves particularly useful in these cases.

Another point we highlight is that the relative improvement is higher for $x = 0$ than for $x = \sqrt{2\alpha\lambda u_2}$. Moreover, even though the initial state of the system is above the optimal equilibrium, $\tilde{\mu}_T$ is smaller than $\mu^\star_\infty$. This is somewhat counter-intuitive. In fact, from (4.4) it follows that $\mu_\bullet$ positively contributes to the corrected staffing function if

$$\mathbb{E}[Q^2(0)] > 3\alpha\lambda u_2 + \frac{2u_2}{3u_3}\sqrt{2\alpha\lambda u_2}.$$

6 Conclusion and further research

Motivated by the time-varying nature of queues in practical applications, we studied the impact that the transient phase has on traditional capacity allocation questions. By defining a cost minimization problem in which the objective function contains a correction accounting for the transient period, we identified the leading and second-order behavior of the cost function as a function of the interval length T. As a by-product, this result yields an approximation for the actual cost function, which is a refinement to its stationary counterpart. Our numerical experiments in Sect. 5.1 demonstrate the improved accuracy achieved by this approximation in a number of settings. By perturbation analysis of the optimization problem, this furthermore gives

rise to a correction to the steady-state optimal capacity allocation of order $1/T$. The necessity of the refined capacity allocation level is substantiated by the numerics in Sect. 5.2, which show the cost reduction that can be achieved in a number of settings, compared to settings in which stationary metrics are used. Particularly for small values of T and large values of α, this reduction is significant. Additionally, these results also indicate that it is relatively safe to use the stationary cost when T is moderate, or α is small. The latter reflects the scenario in which QoS is much more valued than service efficiency. This observation links to the flat nature of the cost function around its optimal value for α small, a statement on the optimality gap that we formally proved in Proposition 4.

Besides the validation of our theoretical results of Sects. 3 and 4, the numerical results also reveal some phenomena that require more investigation.

As noted, our corrected capacity allocation level $\tilde{\mu}_T^\star$ is in most cases studied less than the steady-state optimal value $\mu_\infty^\star$. This implies that congestion levels tend to be higher under our staffing scheme then under stationary staffing. A possible explanation for this may be the fact that the planning period under consideration is finite. Clearly, in the setting we analyzed, anything that happens after time T is neglected. Therefore, it might be beneficial from the cost perspective to end the period with a higher expected congestion level, as it does not need to be canceled out in the future. Related to this observation, it would be interesting to look at the setting in which staffing decisions need to be made in consecutive periods of equal length, in which the arrival rate changes at the start of each period. This case requires careful consideration of the correlation among the staffing decisions within the separate periods.

Another question that arises concerns the translation of our (qualitative) findings to more general queues, in particular the $M/G/s$ queue. Whereas in our analysis the central decision variable is the server speed μ, the variable of interest in multi-server queues is typically the number of servers. It may well be that similar explicit corrections to staffing levels can be deduced to account for transience. Since our analysis heavily relies on the comparability of the sample paths of two single-server queues, which is due to the equal negative drift for the two processes, another approach must be taken to tackle this extension.

The analysis and findings for the single-server queue with Lévy input presented in this paper may serve as a stepping stone for investigation of these more elaborate problems.

Appendix 1: Proofs of Section 2

Proof of Proposition 1

Proof We prove the limit by showing that the difference

$$\Pi_T(\mu) - \Pi_\infty(\mu) = \frac{1}{T}\int_0^T \mathbb{E}[Q_\mu(t)] - \mathbb{E}[Q_\mu(\infty)]\,\mathrm{d}t$$

converges to zero as $T \to \infty$ for $\mu > \lambda$ fixed. The assumption $\mathbb{E}[U(1)], \mathbb{E}[Q(0)] < \infty$ implies by [4, Prop. 1] that $\mathbb{E}[Q_\mu(t)] < \infty$ for all $t \geq 0$. Following [4], we use the decomposition

$$\mathbb{E}[Q_\mu(t)] = \mathbb{E}[Q_\mu^0(t)] + \left\{\mathbb{E}[Q_\mu(t)] - \mathbb{E}[Q_\mu^0(t)]\right\},$$

where $Q_\mu^0(t)$ represents the workload process if the system starts empty. From this decomposition, it is revealed that $\mathbb{E}[Q_\mu^0(t)]$ and $\left\{\mathbb{E}[Q_\mu(t)] - \mathbb{E}[Q_\mu^0(t)]\right\}$ are nonnegative monotonically increasing and decreasing functions of t, respectively; see [4, Prop. 2,Thm. 11]. Recall that $\mathbb{E}[Q_\mu(t)] \to \mathbb{E}[Q_\mu(\infty)]$ for $t \to \infty$ by ergodicity of the workload process for any initial state $\mathbb{E}[Q(0)] < \infty$, if $\mu > \lambda$. Hence,

$$\begin{aligned}\mathbb{E}[Q_\mu(t)] &\leq \sup_t \mathbb{E}[Q_\mu^0(t)] + \sup_t \left\{\mathbb{E}[Q_\mu(t)] - \mathbb{E}[Q_\mu^0(t)]\right\} \\ &= \mathbb{E}[Q_\mu(\infty)] + \left\{\mathbb{E}[Q_\mu(0)] - \mathbb{E}[Q_\mu^0(0)]\right\} = \mathbb{E}[Q_\mu(\infty)] + \mathbb{E}[Q(0)],\end{aligned}$$

for all $t \geq 0$, which proves that the expected workload is bounded. Fix $\varepsilon > 0$. By convergence of $\mathbb{E}[Q_\mu(t)]$ for $t \to \infty$, there exists a value $t^* := t^*(\varepsilon)$ such that for all $t \geq t^*$

$$\left|\mathbb{E}[Q_\mu(t)] - \mathbb{E}[Q_\mu(\infty)]\right| < \varepsilon/2. \tag{7.1}$$

Next, set

$$T^* := T^*(\varepsilon) = \frac{2\,t^*(\varepsilon)}{\varepsilon}\left(2\mathbb{E}[Q_\mu(\infty)] + \mathbb{E}[Q(0)]\right).$$

Then, for $T \geq \hat{T} := \max\{t^*, T^*\}$, we have

$$\begin{aligned}\left|\frac{1}{T}\int_0^T \mathbb{E}[Q_\mu(t)] - \mathbb{E}[Q_\mu(\infty)]\,\mathrm{d}t\right| &\leq \frac{1}{T}\int_0^{t^*} \left|\mathbb{E}[Q_\mu(t)] - \mathbb{E}[Q_\mu(\infty)]\right|\,\mathrm{d}t \\ &\quad + \frac{1}{T}\int_{t^*}^T \left|\mathbb{E}[Q_\mu(t)] - \mathbb{E}[Q_\mu(\infty)]\right|\,\mathrm{d}t \\ &\leq \frac{1}{T}\int_0^{t^*} \mathbb{E}[Q_\mu(t)] + \mathbb{E}[Q_\mu(\infty)]\,\mathrm{d}t + \frac{1}{T}\int_{t^*}^T \frac{\varepsilon}{2}\,\mathrm{d}t \\ &< \frac{t^*}{T}\left(2\mathbb{E}[Q_\mu(\infty)] + \mathbb{E}[Q(0)]\right) + \frac{T - t^*}{T}\,\frac{\varepsilon}{2} \\ &< \frac{t^*}{T^*}\left(2\mathbb{E}[Q_\mu(\infty)] + \mathbb{E}[Q(0)]\right) + \frac{\varepsilon}{2} = \varepsilon.\end{aligned}$$

Hence, for any choice of $\varepsilon > 0$ we can find a value $\hat{T}$ such that $\Pi_{\hat{T}}(\mu)$ approaches $\Pi_\infty(\mu)$ within distance ϵ, which proves the limit. □

Appendix 2: Proofs of Section 3

Proof of Lemma 2

Proof Using the representation in (3.2), we write

$$\begin{aligned}
\Psi_T^{x,y} &= \frac{1}{T}\int_0^\infty \mathbb{E}[Y^{x,y}(t)]\mathrm{d}t \\
&= \frac{1}{T}\,\mathbb{E}\left[\int_0^{\tau^y(0)} Y^{x,y}(t)\right]\mathrm{d}t + \frac{1}{T}\,\mathbb{E}\left[\int_{\tau^y(0)}^{\tau^x(0)} Y^{x,y}(t)\mathrm{d}t\right] \\
&\quad + \frac{1}{T}\,\mathbb{E}\left[\int_{\tau^y(0)}^{\infty} Y^{x,y}(t)\mathrm{d}t\right], \\
&= \frac{1}{T}\,\mathbb{E}\left[\int_0^{\tau^y(0)} (x-y)\mathrm{d}t\right] + \frac{1}{T}\,\mathbb{E}\left[\int_{\tau^y(0)}^{\tau^x(0)} Y^{x,y}(t)\mathrm{d}t\right] \\
&= \frac{1}{T}\,\mathbb{E}[\tau^y(0)](x-y) + \frac{1}{T}\,\mathbb{E}\left[\int_{\tau^y(0)}^{\tau^x(0)} Y^{x,y}(t)\mathrm{d}t\right].
\end{aligned}$$

By (3.2) and the strong Markov property holding for Lévy processes [5], observe that $Y^{x-y,0}(t) \overset{d}{=} Y^{x,y}(\tau^y(0)+t)$, whereby

$$\frac{1}{T}\,\mathbb{E}\left[\int_{\tau^y(0)}^{\tau^x(0)} Y^{x,y}(t)\,\mathrm{d}t\right] = \frac{1}{T}\,\mathbb{E}\left[\int_0^{\tau^{x-y}(0)} Y^{x-y,0}(t)\mathrm{d}t\right] = \Psi_T^{x-y,0},$$

which completes the proof. □

Proof of Lemma 3

Proof Note that $Y^{z,0}(t)$ and $\tau^z(w)$ are intimately related. Namely, due to the fact that X has no negative jumps,

$$\{\tau^z(w) \le t\} = \{Y^{z,0}(t) \le w\}.$$

In fact, $Y^{z,0}(\tau^z(w)) = w$, which implies that τ^z is a right inverse for $Y^{z,0}(t)$. Therefore, the following equality holds:

$$\int_0^{\tau^z(0)} Y^{z,0}(t)\,\mathrm{d}t = \int_0^z \tau^z(w)\,\mathrm{d}w,$$

which implies with the help of Fubini's theorem that

$$\Psi_T^{z,0} = \frac{1}{T}\int_0^z \mathbb{E}[\tau^z(w)]\,\mathrm{d}w = \frac{1}{T}\int_0^z \mathbb{E}[\tau^{z-w}(0)]\,\mathrm{d}w = \frac{1}{T}\int_0^z \mathbb{E}[\tau^w(0)]\,\mathrm{d}w.$$

□

Proof of Corollary 1

Proof From (3.8),

$$\mathbb{E}\left[\hat{\tau}^0(w)\right] = -\tfrac{\mathrm{d}}{\mathrm{d}u}\,\mathbb{E}\left[\exp(-u\,\hat{\tau}^0(w))\right]\Big|_{u=0} = w\,\frac{\mathrm{d}}{\mathrm{d}u}\Upsilon^{-1}(u)\Big|_{u=0}. \tag{8.1}$$

Since $\Upsilon(\theta)$ is strictly increasing and $\Upsilon(0) = 0$, we get $\Upsilon^{-1}(0) = 0$ and

$$\tfrac{\mathrm{d}}{\mathrm{d}u}\Upsilon^{-1}(u)\Big|_{u=0} = \frac{1}{\Upsilon'(\Upsilon^{-1}(0))} = \{\Upsilon'(0)\}^{-1}.$$

Furthermore,

$$\begin{aligned}\Upsilon'(\theta) &= -a + \sigma^2\theta + \int_{-\infty}^0 (x\,e^{\theta x} - x\mathbf{1}_{[-1,0)}(x))\hat{\nu}(\mathrm{d}x)\\ &= -a + \sigma^2\theta - \int_0^\infty (y\,e^{-\theta y} - y1_{(0,1]}(y))\nu(\mathrm{d}y).\end{aligned}$$

Thus, $\Upsilon'(0) = -\mathbb{E}[X(1)] = \mu - \lambda$ and $\mathbb{E}[\hat{\tau}^0(w)] = w/(\mu-\lambda)$. By (3.5) and (3.6), we deduce that

$$\Psi_T^{z,0} = \frac{1}{T}\int_0^z \mathbb{E}[\tau^w(0)]\,\mathrm{d}w = \frac{1}{T}\int_0^z \mathbb{E}[\hat{\tau}^0(w)]\mathrm{d}w = \frac{z^2}{2T(\mu-\lambda)}.$$

For $x > y$, we use Lemma 2 to conclude

$$\Psi_T^{x,y} = \frac{y(x-y)}{T(\mu-\lambda)} + \frac{(x-y)^2}{2T(\mu-\lambda)} = \frac{x^2-y^2}{2T(\mu-\lambda)}.$$

The result for $x < y$ follows directly by the observation $\Psi_T^{x,y} = -\Psi_T^{x,y}$. □

Proof of Proposition 2

Proof To derive the upper bound for $\Delta_T^{x,y}$, we apply the same coupling argument as described in Sect. 3. Let us assume without loss of generality that $x > y$. In this case,

$$|\Delta_T^{x,y}| = \frac{1}{T}\int_T^\infty \mathbb{E}[Q^x(t) - Q^y(t)]\,\mathrm{d}t \le \frac{1}{T}\int_T^\infty \mathbb{E}[Q^x(t) - Q^0(t)]\,\mathrm{d}t.$$

By the decomposition in (3.2),

$$\begin{aligned}\int_T^\infty \mathbb{E}\left[Q^x(t) - Q^0(t)\right] \mathrm{d}t &= \int_T^\infty \mathbb{E}\left[\left(x + \inf_{s\le t} X(s)\right)\mathbf{1}_{\{\tau^x(0)>t\}}\right] \mathrm{d}t \\ &= \int_T^\infty \int_0^x P(x-u+\inf_{s\le t} X(s) > 0)\,\mathrm{d}u\,\mathrm{d}t \\ &= \int_T^\infty \int_0^x P(\tau^{x-u}(0) > t)\,\mathrm{d}u\,\mathrm{d}t \\ &\le \int_T^\infty \int_0^x \frac{\mathbb{E}[\tau^{x-u}(0)^2]}{t^2}\,\mathrm{d}u\,\mathrm{d}t \\ &= \int_0^x \int_T^\infty \frac{\mathbb{E}[\tau^{x-u}(0)^2]}{t^2}\,\mathrm{d}t\,\mathrm{d}u = \int_0^x \frac{\mathbb{E}[\tau^w(0)^2]}{T}\,\mathrm{d}w.\end{aligned}$$

We obtain $\mathbb{E}[\tau^w(0)^2]$ with the help of its Laplace transform in (3.8). Namely,

$$\begin{aligned}\mathbb{E}[\tau^w(0)^2] &= \tfrac{\mathrm{d}^2}{\mathrm{d}u^2}\mathbb{E}[\exp(-u\tau^w(0))]\Big|_{u=0} \\ &= w^2\left(\tfrac{\mathrm{d}}{\mathrm{d}u}\Upsilon^{-1}(u)\Big|_{u=0}\right)^2 - w\,\tfrac{\mathrm{d}^2}{\mathrm{d}u^2}\Upsilon^{-1}(u)\Big|_{u=0}.\end{aligned}$$

As in the previous subsection, we have $\frac{\mathrm{d}}{\mathrm{d}u}\Upsilon^{-1}(u)\big|_{u=0} = (\mu-\lambda)^{-1}$, and

$$\tfrac{\mathrm{d}^2}{\mathrm{d}u^2}\Upsilon^{-1}(u)\Big|_{u=0} = -\frac{\Upsilon''(\Upsilon^{-1}(0))}{\Upsilon'(\Upsilon^{-1}(0))^3} = -\frac{\Upsilon''(0)}{\Upsilon'(0)^3}.$$

Since $\Upsilon'(0) = \mu - \lambda$ and

$$\Upsilon''(0) = \sigma^2 + \int_0^\infty x^2\,\nu(\mathrm{d}x) = u_2,$$

we conclude

$$\mathbb{E}[\tau^w(0)^2] = \frac{w^2}{(\mu-\lambda)^2} + \frac{u_2 w}{(\mu-\lambda)^3},$$

so that

$$|\Delta_T^{x,y}| \le \frac{1}{T^2}\int_0^x \frac{w^2}{(\mu-\lambda)^2} + \frac{u_2 w}{(\mu-\lambda)^3}\mathrm{d}w = \frac{1}{T^2}\left(\frac{x^3}{3(\mu-\lambda)^2} + \frac{u_2 x^2}{2(\mu-\lambda)^3}\right). \tag{8.2}$$

For general $x, y \ge 0$,

$$|\Delta_T^{x,y}| \le \frac{1}{T^2}\left(\frac{\max(y,x)^3}{3(\mu-\lambda)^2} + \frac{u_2\max(y,x)^2}{2(\mu-\lambda)^3}\right).$$

As a direct consequence,

$$|\Delta_T| \leq \frac{1}{T^2}\left(\frac{\mathbb{E}\left[\max(Q(0), Q_\mu(\infty))^3\right]}{3(\mu-\lambda)^2} + \frac{u_2\mathbb{E}\left[\max(Q(0), Q_\mu(\infty))^2\right]}{2(\mu-\lambda)^3}\right).$$

□

Remark Observe that if X is light-tailed, that is $\mathbb{E}[\exp\{-\theta X(1)\}] = \mathbb{E}[\exp\{\kappa(\theta)\}] < \infty$ for some $\theta < 0$, then $\Upsilon(\theta)$ as in (3.8) has an analytic continuation in the negative half-plane, and in this region $\Upsilon(\theta) < 0$. Consequently, we can replace the upper bound on the tail probability of $\tau^{x-u}(0)$ by

$$\mathbb{P}\left(\tau^{x-u}(0) > t\right) = \mathbb{P}\left(e^{\beta\tau^{x-u}(0)} > e^{\beta t}\right) \leq e^{-\beta t}\, e^{(x-u)\Upsilon^{-1}(-\beta)},$$

for some $\beta > 0$, so that

$$\int_T^\infty \mathbb{E}[Q^x(t) - Q^0(t)]\,\mathrm{d}t \leq e^{-\beta T}\,\frac{e^{x\Upsilon^{-1}(-\beta)} - 1}{\beta\,\Upsilon^{-1}(-\beta)}.$$

Along similar lines, we deduce

$$|\Delta_T^{x,y}| \leq \frac{e^{-\beta T}}{T}\,\frac{e^{x\Upsilon^{-1}(-\beta)} + e^{y\Upsilon^{-1}(-\beta)} - 2}{\beta\,\Upsilon^{-1}(-\beta)}$$

and

$$|\Delta_T| \leq \frac{e^{-\beta T}}{T}\,\frac{\mathbb{E}\left[e^{Q(0)\Upsilon^{-1}(-\beta)}\right] + \mathbb{E}\left[e^{Q_\mu(\infty)\Upsilon^{-1}(-\beta)}\right] - 2}{\beta\,\Upsilon^{-1}(-\beta)},$$

assuming that $\mathbb{E}[\mathrm{e}^{-yQ(0)}] < \infty$ for all $y > 0$. The condition $\mathbb{E}[\mathrm{e}^{Q_\mu(\infty)\Upsilon^{-1}(-\beta)}] < \infty$ follows from Lemma 1. Hence, the error decays exponentially fast for light-tailed input processes.

Appendix 3: Proofs of Section 4

Proof of Lemma 4

Proof Since the term $\alpha\mu$ is convex, the strictness should come from the term $\frac{1}{T}\int_0^T \mathbb{E}[Q_\mu(t)]\,\mathrm{d}t$. Furthermore, observe that if a function $f_\mu(t)$ is convex for all $t \geq 0$, and strictly convex for all $t \geq \varepsilon$ for some $\varepsilon \in [0, T)$, i.e., for any $\mu_1, \mu_2 > 0$ and $a \in (0, 1)$

$$a\, f_{\mu_1}(t) + (1-a) f_{\mu_2}(t) > f_{a\mu_1+(1-a)\mu_2}(t),$$

then

$$a\int_0^T f_{\mu_1}(t)\,\mathrm{d}t + (1-a)\int_0^T f_{\mu_2}(t)\mathrm{d}t = \int_0^T af_{\mu_1}(t) + (1-a)f_{\mu_2}(t)\mathrm{d}t$$
$$= \int_0^\varepsilon af_{\mu_1}(t) + (1-a)f_{\mu_2}(t)\mathrm{d}t + \int_\varepsilon^T af_{\mu_1}(t) + (1-a)f_{\mu_2}(t)\mathrm{d}t$$
$$> \int_0^\varepsilon f_{a\mu_1+(1-a)\mu_2}(t)\,\mathrm{d}t + \int_\varepsilon^T f_{a\mu_1+(1-a)\mu_2}(t)\,\mathrm{d}t = \int_0^T f_{a\mu_1+(1-a)\mu_2}(t)\,\mathrm{d}t.$$

Hence, it suffices to prove the convexity of $\mathbb{E}[Q_\mu(t)]$ as a function of μ for all $t \geq 0$, and strict convexity for $t \geq \varepsilon$ for some $\varepsilon \in [0, T)$.

Let $\tau_\mu^x(0)$ denote the first passage time of level 0 in the process Q_μ with $Q(0) = x$. Then,

$$Q_\mu(t) = U(t) - \mu t + \max\left\{x, -\inf_{s\leq t}[U(s) - \mu s]\right\} \tag{9.1}$$
$$= \begin{cases} x + U(t) - \mu t, & \text{if } t < \tau_\mu^x(0), \\ U(t) - \mu t - \inf_{s\leq t}[U(s) - \mu s], & \text{if } t \geq \tau_\mu^x(0), \end{cases} \tag{9.2}$$

where

$$\tau_\mu^x(0) := \inf\{t \geq 0 : x + U(t) - \mu t \leq 0\}$$

and $U(t)$ is a spectrally positive Lévy process. Fix $\mu_1, \mu_2 > 0$ and $a \in (0, 1)$. Define $\mu_3 := a\mu_1 + (1-a)\mu_2$, and

$$D(t) := aQ_{\mu_1}(t) + (1-a)Q_{\mu_2}(t) - Q_{\mu_3}(t).$$

In order to prove strict convexity, we have to show that $D(t) \geq 0$ for all $t \geq 0$, thereby implying $\mathbb{E}[D(t)] \geq 0$, i.e., convexity, for all $t \geq 0$, and $D(t) > 0$ with positive probability for $t \in [\varepsilon, T]$, for some $\varepsilon \in [0, T)$.

We distinguish two cases: $x > 0$ and $x = 0$.

Case $x > 0$. We start by noticing that if Q_{μ_1}, Q_{μ_2} and Q_{μ_3} experience the same input process $U(t)$, then by absence of negative jumps in $U(t)$, it holds that

$$\tau_{\mu_2}^x(0) < \tau_{\mu_3}^x(0) < \tau_{\mu_1}^x(0). \tag{9.3}$$

We use shorthand notation

$$I_k(t) := \inf_{0\leq s\leq t}[U(s) - \mu_k s],$$

for $k = 1, 2, 3$. Using representation (9.2) of the workload process, we obtain

$$D(t) = \begin{cases} 0, & \text{if } t < \tau_{\mu_2}^x(0), \\ -(1-a)\,(x + I_2(t))\,, & \text{if } \tau_{\mu_2}^x(0) \leq t < \tau_{\mu_3}^x(0), \\ ax - (1-a)I_2(t) + I_3(t), & \text{if } \tau_{\mu_3}^x(0) \leq t < \tau_{\mu_1}^x(0), \\ -aI_1(t) - (1-a)I_2(t) + I_3(t), & \text{if } t \geq \tau_{\mu_1}^x(0). \end{cases}$$

This partition allows us to spot when strict convexity can occur. Note that by definition $t \geq \tau^x_{\mu_2}(0)$, $I_2(t) = \inf_{0\leq s\leq t}[U(s) - \mu_2 s] \leq -x$, so that $D(t) \geq 0$ if $\tau^x_{\mu_2}(0) \leq t < \tau^x_{\mu_3}(0)$. Moreover, by subadditivity of the infimum,

$$\begin{aligned} I_3(t) &= \inf_{0\leq s\leq t}[U(s) - \mu_3 s] = \inf_{0\leq s\leq t}[a(U(s) - \mu_1 s) + (1-a)(U(s) - \mu_2 s)] \\ &\geq a \inf_{0\leq s\leq t}[U(s) - \mu_1 s] + (1-a)\inf_{0\leq s\leq t}[U(s) - \mu_2 s] \\ &= aI_1(t) + (1-a)I_2(t), \end{aligned}$$

and hence $D(t) \geq 0$ for $t \geq \tau^x_{\mu_1}(0)$. Using the same argument, we deduce

$$ax - (1-a)I_2(t) + I_3(t) \geq ax - (1-a)I_2(t) + aI_1(t) + (1-a)I_2(t) = a(x + I_1(t)).$$

In particular, for $t < \tau^x_{\mu_1}(0)$ this value is strictly positive. As a result, $D(t) \geq 0$ for all $t \geq 0$. On top of that $D(t) > 0$ for $t \in [\tau^x_{\mu_3}(0), \tau^x_{\mu_1}(0))$. Accordingly, the latter implies strict positivity of $\mathbb{E}D(t)$, and therefore strict convexity of $\mathbb{E}Q_\mu(t)$, if the event $\{\tau^x_{\mu_3}(0) \leq t < \tau^x_{\mu_1}(0)\}$ occurs with positive probability. That is,

$$\begin{aligned} P(D(t) > 0) &\geq P\left(a(x + I_1(t))\mathbf{1}_{\{\tau^x_{\mu_3}(0)\leq t<\tau^x_{\mu_1}(0)\}} > 0\right) \\ &= P\left(x + I_1(t) > 0, \tau^x_{\mu_3}(0) \leq t < \tau^x_{\mu_1}(0)\right) \\ &= P\left(x + I_1(t) > 0 | \tau^x_{\mu_3}(0) \leq t < \tau^x_{\mu_1}(0)\right) P\left(\tau^x_{\mu_3}(0) \leq t < \tau^x_{\mu_1}(0)\right) \\ &= P\left(\tau^x_{\mu_3}(0) \leq t < \tau^x_{\mu_1}(0)\right) = P(\tau^x_{\mu_3}(0) \leq t) - P(\tau^x_{\mu_1}(0) \leq t) > 0, \end{aligned} \tag{9.4}$$

by the stochastic dominance in (9.3). To ensure the strict inequality in (9.4), we have to enforce the condition

$$P(\tau^x_{\mu_1}(0) < T) > 0. \tag{9.5}$$

Remark An example illustrating the need for this condition is the case in which $U(t)$ is a compound Poisson process and $T < x/\mu_2 < x/\mu_1$. Then

$$Q_{\mu_k}(t) = x + U(t) - \mu_k t,$$

for all $t \in [0, T]$, since $U(t) \geq 0$ and therefore $\tau^x_{\mu_1}(0) > T$. Consequently, for all $a \in (0, 1)$,

$$a\, Q_{\mu_1} + (1-a)\, Q_{\mu_2}(t) = Q_{\mu_3}(t),$$

proving only convexity of $\mathbb{E}Q_\mu(t)$ and subsequently $\int_0^T \mathbb{E}[Q_\mu(t)]\,\mathrm{d}t$. In the case $\sigma > 0$, the probability in (9.5) is necessarily positive.

The case $x = 0$. By the fact that $\tau_\mu(0) = 0$ for all $\mu > 0$, proving that $D(t) > 0$ in the case $x = 0$ reduces to showing that the probability of

$$D(t) = aI_1(t) + (1-a)I_2(t) - I_3(t) > 0$$

happening is positive for all $t > 0$. Define

$$t_0 := \inf\{t > 0 \,:\, U(t) > 0\},$$

and

$$\tilde{\tau}_\mu := \inf\{t > t_0 \,:\, U(t) - \mu t \le 0\}.$$

We note that t_0, as defined above, also defines the epoch of the start of a new excursion of the reflection Q_μ for all $\mu > 0$. Namely,

$$\begin{aligned} U(s) \le 0 \quad &\Rightarrow \quad U(s) - \mu s \le -\mu s \qquad \text{for all } 0 \le s < t_0 \\ &\Rightarrow \inf_{0 \le s < t_0} [U(s) - \mu s] \le -\mu t_0 \\ &\Rightarrow U(t_0) - \mu t_0 - \inf_{0 \le s < t_0} [U(s) - \mu s] \ge U(t_0) > 0. \end{aligned}$$

Then $Q_\mu(t_0-) = 0$ for all $\mu > 0$. By virtue of the strong Markov Property, note that $Q_\mu(t_0 + t) \stackrel{d}{=} Q_\mu(t)$. Hence, we assume without loss of generality that $t_0 = 0$. Again, we have a stochastic dominance relation similar to (9.3):

$$\tilde{\tau}_{\mu_2} < \tilde{\tau}_{\mu_3} < \tilde{\tau}_{\mu_1},$$

for all $\mu_1 < \mu_3 < \mu_2$. Then

$$D(t) \stackrel{d}{=} \begin{cases} 0, & \text{if } t < \tilde{\tau}_{\mu_2}, \\ -(1-a)I_2(t), & \text{if } \tilde{\tau}_{\mu_2} \le t < \tilde{\tau}_{\mu_3}, \\ (1-a)I_2(t) + I_3(t), & \text{if } \tilde{\tau}_{\mu_3} \le t < \tilde{\tau}_{\mu_1}, \\ -aI_1(t) - (1-a)I_2(t) + I_3(t), & \text{if } t \ge \tilde{\tau}_{\mu_1}. \end{cases}$$

Clearly, $D(t) \ge 0$ for all $t \ge 0$ and

$$-(1-a)I_2(t) + I_3(t) \ge aI_1(t) > 0,$$

for $\tilde{\tau}_{\mu_3} \le t < \tilde{\tau}_{\mu_1}$. Hence, in a similar manner to (9.4),

$$\begin{aligned} P(D(t) > 0) &\ge P\left(aI_1(t)\mathbf{1}_{\{\tilde{\tau}_{\mu_3} \le t < \tilde{\tau}_{\mu_1}\}} > 0\right) \\ &= P\left(I_1(t) > 0, \tilde{\tau}_{\mu_3} \le t < \tilde{\tau}_{\mu_1}\right) \\ &= P\left(I_1(t) > 0 | \tilde{\tau}_{\mu_3} \le t < \tilde{\tau}_{\mu_1}\right) P\left(\tilde{\tau}_{\mu_3} \le t < \tilde{\tau}_{\mu_1}\right) \\ &= P\left(\tilde{\tau}_{\mu_3} \le t < \tilde{\tau}_{\mu_1}\right) = P(\tilde{\tau}_{\mu_3} \le t) - P(\tilde{\tau}_{\mu_1} \le t) > 0, \end{aligned} \tag{9.6}$$

The last inequality is satisfied if $P(\tilde{\tau}_{\mu_1} < T) > 0$, which is equivalent to $P(U(T) - \mu T \leq 0) > 0$, a condition that is clearly true for all our choices of U. In conclusion, for $x = 0$, $\mathbb{E}[D(t)] > 0$ and therefore $\mathbb{E}[Q_\mu(t)]$ is a strictly convex function of μ. □

Proof of Proposition 3

The proof of the proposition relies on the following auxiliary lemma, for which we include the proof for completeness.

Lemma 5 *Consider the sequence for functions $f_n : [x_0, \infty) \to \mathbb{R}$ and let $f : [x_0, \infty) \to \mathbb{R}$ be the pointwise limit for some $x_0 \in \mathbb{R}$. Assume f and f_n are strictly convex for all n. Furthermore, let $f(y) \to \infty$ for both $y \to x_0^+$ and $y \to \infty$. If x_n and x are the minimizers for f_n and f, respectively, then $x_n \to x$ for $n \to \infty$.*

Proof We start by showing that the sequence x_n is bounded. Fix u_l, u_r such that $x_0 < u_l < x < u_r$. We claim that there exists a $N \in \mathbb{N}$ such that $x_n \in [u_l, u_r]$ for all $n \geq N$. First, we prove the upper bound on x_n. For any strictly convex function h with minimizer x_h, the following statement holds true:

$$x_h < u_r \quad \Leftrightarrow \quad h \text{ is strictly increasing at } u_r. \tag{9.7}$$

The first implication follows from observing that $h(x_h) < h(y)$ for all $y > x^*$ and the definition of convexity:

$$0 < \frac{h(u_r) - h(x_h)}{u_r - x_h} \leq \frac{h(u_r + \delta) - h(u_r)}{\delta},$$

for all $\delta > 0$, so that $h(u_r) < h(u_r + \delta)$, i.e., h is increasing at u_r. The converse follows immediately by observing that $h(u_r) < h(u_r + \delta)$ for all $\delta > 0$, so that $x_h < u_r$. Next, we show that f_n must be increasing at u_r for n sufficiently large. By pointwise convergence of f_n, we have

$$\lim_{n \to \infty} [f_n(u_r + \delta) - f_n(u_r)] = f(u_r + \delta) - f(u_r).$$

Let $w_r := f(u_r + \delta) - f(u_r) > 0$. Then

$$\exists N_r \in \mathbb{N} : \forall n \geq N_r : |[f_n(u_r + \delta) - f_n(u_r)] - [f(u_r + \delta) - f(u_r)]| < w_r/2.$$

Hence for $n \geq N_r$,

$$\begin{gathered} f(u_r + \delta) - f(u_r) - w_r/2 < f_n(u_r + \delta) - f_n(u_r) < f(u_r + \delta) - f(u_r) + w_r/2 \\ \Rightarrow 0 < w_r/2 < f_n(u_r + \delta) - f_n(u_r). \end{gathered}$$

Hence, by (9.7), $x_n < u_r$ for sufficiently large n. Similarly, we argue

$$x_h > u_l \quad \Leftrightarrow \quad h \text{ is strictly decreasing at } u_l,$$

for any strictly convex function h with minimizer x_h. Note that $x_h > u_l$ implies $h(x_h) - h(u_l) < 0$ and for all $\delta > 0$ we get by strict convexity

$$\frac{h(u_l) - h(u_l - \delta)}{\delta} < \frac{h(x_h) - h(u_l)}{x_h - u_l} < 0,$$

by which $h(u_l - \delta) > h(u_l)$, i.e., h is decreasing in u_l. Moreover, if h is decreasing at u_l, then it is decreasing for all $y < u_l$, by arguments similar to the above. Therefore, $h(u_l - \delta) > h(u_l)$ for all $\delta > 0$ and it must hold that $x_h > u_l$. Define $f(u_l) - f(u_l - \delta) := w_l < 0$, then, again by pointwise convergence, we have that

$$\exists N_l \in \mathbb{N} : \forall n \geq N_l : |[f_n(u_l) - f_n(u_l - \delta)] - [f(u_l) - f(u_l - \delta)]| < w_l,$$

whereupon

$$f_n(u_l) - f_n(u_l - \delta) < f(u_l) - f(u_l - \delta) + w_l = 2w_l < 0.$$

Hence, for sufficiently large n, we also have $x_n > u_l$. Fix $N = \max\{N_l, N_r\}$. Then, for $n \geq N$, $x_n \in (u_l, u_r)$. That is, the sequence x_n is bounded. Therefore, by the theorem of Bolzano–Weierstrass, x_n has to have a convergent subsequence. That is, there exists a sequence n_k such that $n_k \to \infty$ and $x_{n_k} \to a$ as $k \to \infty$ for some $a \in [u_l, u_r]$. We prove that every subsequence must converge to x by contradiction. Suppose there exists a subsequence n_k such that $x_{n_k} \to a \neq x$. Since, $x_n \in [u_l, u_r]$ for $n \geq N$, we may restrict our attention on the sequence of functions $\hat{f}_n : [u_l, u_r] \to \mathbb{R}^+$, consisting of the original function f_n restricted to the domain $[u_l, u_r]$. To be precise, $x_n = \arg\min_y f_n(y) = \arg\min_y \hat{f}_n(y)$ for $n \geq N$. Because $\hat{f}_n$ and $\hat{f}$ are bounded, we furthermore have $\hat{f}_n \to \hat{f}$ uniformly.

Fix $\varepsilon > 0$. By uniform convergence, there exists $K_0 \in \mathbb{N}$ such that

$$|\hat{f}_{n_k}(y) - \hat{f}(y)| < \varepsilon/2, \quad \forall k \geq K_0, \ y \in [u_l, u_r].$$

Also, because $\hat{f}$ is convex, it is continuous, so that there exists $\delta := \delta(\varepsilon)$ such that

$$|z - y| < \delta \quad \Rightarrow \quad |\hat{f}(z) - \hat{f}(y)| < \varepsilon/2.$$

Let K_1 be such that $|x_{n_k} - a| < \delta$ for all $k \geq K_1$. Then for $k \geq K = \max\{K_0, K_1\}$ this implies

$$\begin{aligned} |f_{n_k}(x_{n_k}) - f(a)| &= |\hat{f}_{n_k}(x_{n_k}) - \hat{f}(a)| \\ &\leq |\hat{f}_{n_k}(x_{n_k}) - \hat{f}(x_{n_k})| + |\hat{f}(x_{n_k}) - f(a)| < \varepsilon/2 + \varepsilon/2 = \varepsilon. \end{aligned}$$

Hence, we conclude $\lim_{k\to\infty} \hat{f}_{n_k}(x_{n_k}) = f(a)$. Therefore,

$$\limsup_{n\to\infty} f_n(x_n) \geq f(a) > f(x),$$

by minimality of x. However, $f_n(x_n) \leq f_n(x)$, which implies $\limsup_{n\to\infty} f_n(x_n) \leq \lim_{n\to\infty} f_n(x) = f(x)$, contradicting the strict inequality above. Hence, we deduce $x = a$. Consequently, every subsequence of x_n converges to x and therefore $x_n \to x$ as $n \to \infty$. □

Applying Lemma 5 to the functions Π_T and Π_∞ with $x_0 = \lambda$, together with Lemma 4, we obtain the result immediately. □

Proof of Proposition 4

Proof Note that Π_∞ is a smooth function. By the first optimality condition, $\Pi'_\infty(\mu^\star_\infty) = 0$. We first prove that also $\Pi_T(\mu)$ is differentiable with respect to μ for all $\mu \geq 0$. Recall (2.4), which defines the cost function as a combination of the accumulated expected transient queue length, and linear staffing costs. The latter term is clearly differentiable; hence, it remains to be proved that

$$C_T(\mu) = \frac{1}{T}\int_0^\infty \mathbb{E}[Q_\mu(t)]\,\mathrm{d}t$$

admits a derivative for all $\mu \geq 0$ with T fixed. This holds if and only if $\mathbb{E}[Q_\mu(t)]$ is differentiable for all $t \geq 0$. Let $Q(0) = x \geq 0$. Following (2.2),

$$\begin{aligned}\mathbb{E}[Q_\mu(t)] &= \mathbb{E}[X_\mu(t)] + \mathbb{E}\left[\max\{x, \sup_{s\in[0,t]}\{-X_\mu(s)\}\}\right] \\ &= (\lambda-\mu)t + \mathbb{E}\left[\max\{x, \sup_{s\in[0,t]}\{-X_\mu(s)\}\}\right],\end{aligned}$$

where the first term is differentiable. Furthermore,

$$\begin{aligned}\mathbb{E}[\max\{x, \sup_{s\in[0,t]}\{-X_\mu(s)\}\}] &= x + \int_x^\infty P(\sup_{s\in[0,t]}\{-X_\mu(s)\} > u)\,\mathrm{d}u \\ &= x + \int_x^\infty P(\hat{\tau}^0(u) \leq t)\,\mathrm{d}u,\end{aligned}$$

with $\hat{\tau}^0(u)$ as defined in (3.6).

Since $-X_\mu$ is a process with no positive jumps, we may apply Corollary VII3 of [8], which states that the following equivalence between measures holds:

$$s\,P(\hat{\tau}^0(u) \in \mathrm{d}s)\mathrm{d}u = u\,P(-X_\mu(s) \in \mathrm{d}u)\mathrm{d}s, \tag{9.8}$$

so that

$$\int_x^\infty P(\hat{\tau}^0(u) \leq t)\,\mathrm{d}u = \int_x^\infty \int_0^t P(\hat{\tau}^0(u) \in \mathrm{d}s)\mathrm{d}u$$

$$= \int_x^\infty \int_0^t s^{-1} u \; P(-X_\mu(s) \in \mathrm{d}u)\mathrm{d}s \tag{9.9}$$

$$= \int_x^\infty \int_0^t s^{-1} u \; P(X_\mu(s) \in \mathrm{d}u)\mathrm{d}s$$

$$= \int_0^t s^{-1}\mathbb{E}[\max\{x, X_\mu(s)\}]\,\mathrm{d}s \tag{9.10}$$

$$= \int_0^t \int_{x/s}^\infty P(X_\mu(s)/s > v)\,\mathrm{d}v\,\mathrm{d}s$$

$$= \int_0^t \int_{x/s}^\infty P(U(\lambda s)/s > v+\mu)\,\mathrm{d}v\,\mathrm{d}s \tag{9.11}$$

$$= \int_0^t \int_{x/s+\mu}^\infty P(U(\lambda s)/s > w)\,\mathrm{d}w\,\mathrm{d}s, \tag{9.12}$$

where the interchange of integrals is justified by Fubini's theorem and this last form is differentiable with respect to μ. Substituting $Q(0)$ for x straightforwardly yields differentiability of the complete cost function Π_T for all T.

Consequently, we invoke the first optimality condition for $\mu_T^\star$ to find

$$\begin{aligned}
0 = \Pi_T'(\mu_T^\star) &= \Pi_\infty'(\mu_T^\star) + \Psi_T'(\mu_T^\star) + O(1/T^2) \\
&= \Pi_\infty'(\mu_\infty^\star) + \Psi_T'(\mu_\infty^\star) + (\mu_T^\star - \mu_\infty^\star)\left[\Pi_\infty''(\mu_\infty^\star) + \Psi_T''(\mu_\infty^\star)\right] \\
&\quad + \frac{1}{2}(\mu_T - \mu_\infty^\star)^2\left[\Pi_T'''(\xi) + \Psi_T'''(\xi)\right] + O(1/T^2) \\
&= \Psi_T'(\mu_\infty^\star) + (\mu_T^\star - \mu_\infty^\star)\left[\Pi_\infty''(\mu_\infty^\star) + \Psi_T''(\mu_\infty^\star)\right] \\
&\quad + \frac{1}{2}(\mu_T - \mu_\infty^\star)^2\left[\Pi'''(\xi) + \Psi_T'''(\xi)\right] + O(1/T^2),
\end{aligned}$$

for some $\xi \in [\mu_T^\star, \mu_\infty^\star]$. Rearranging this gives

$$\begin{aligned}
\mu_T^\star - \mu_\infty^\star &= \frac{-\Psi_T'(\mu_\infty^\star)}{\Pi_\infty''(\mu_\infty^\star) + \Psi_T''(\mu_\infty^\star) + \frac{1}{2}(\mu_T^\star - \mu_\infty^\star)(\Pi_\infty'''(\mu_T^\star) + \Psi_T'''(\xi))} + O(1/T) \\
&= -\frac{\Psi_T'(\mu_\infty^\star)}{\Pi_\infty''(\mu_\infty^\star)}\left[1 - \frac{\Psi_T''(\mu)}{\Pi_\infty''(\mu_\infty^\star)} - \frac{1}{2}(\mu_T^\star - \mu_\infty)\frac{\Pi_\infty'''(\mu_\infty^\star) + \Psi_T'''(\mu_\infty^\star)}{\Pi_\infty''(\mu_\infty^\star)}\right] + O(1/T) \\
&= -\frac{\Psi_T'(\mu_\infty^\star)}{\Pi_\infty''(\mu_\infty^\star)}[1 + o(1)]
\end{aligned}$$

for $T \to \infty$, since both $\mu_T - \mu_\infty$ and $\Psi_T''(\mu_\infty^\star)$ are $o(1)$. Let

$$\mu_\bullet := \lim_{T\to\infty} \frac{T\Psi_T'(\mu_\infty^\star)}{\Pi_\infty''(\mu_\infty^\star)}.$$

By (3.9) we have

$$T\Psi_T'(\mu) = -\frac{\mathbb{E}[Q(0)^2]}{2(\mu-\lambda)^2} + \frac{\lambda u_3}{3(\mu-\lambda)^3} + \frac{3\lambda^2 u_2^2}{4(\mu-\lambda)^4}.$$

Together with

$$\Pi_\infty''(\mu) = \frac{\lambda u_2}{(\mu-\lambda)^3}$$

and (4.3) we obtain the expression for $\mu_\bullet$ in (4.4). □

Proof of Proposition 5

Proof We upper bound the optimality gap by using the decomposition in (4.1).

$$\begin{aligned}|\Pi_\infty^\star - \Pi_T^\star| &= \left|\hat{\Pi}_T(\mu_\infty) + \Delta_T(\mu_\infty^\star) - \hat{\Pi}_T(\mu_T^\star) - \Delta_T(\mu_T^\star)\right| \\ &\le |\hat{\Pi}_T(\mu_\infty^\star) - \hat{\Pi}_T(\mu_T^\star)| + |\Delta_T(\mu_\infty^\star)| + |\Delta_T(\mu_T^\star)| \\ &= |\hat{\Pi}_T(\mu_\infty^\star) - \hat{\Pi}_T(\mu_T^\star)| + O(1/T^2), \end{aligned} \tag{9.13}$$

since $\Delta_T(\mu) = O(1/T^2)$ by Proposition 2. Next, we find an upper bound for $|\hat{\Pi}_T(\gamma) - \hat{\Pi}_T(\beta)|$, with $\hat{\Pi}_T(\cdot)$ as in (4.1), in terms of the difference between γ and β. For simplicity, denote $\hat{\gamma} = \gamma - \lambda$ and $\hat{\beta} = \beta - \lambda$, implying $\hat{\gamma} - \hat{\beta} = \gamma - \beta$. Then, using (3.9), we get

$$\begin{aligned}|\hat{\Pi}_T(\mu_\infty^\star) - \hat{\Pi}_T(\mu_T^\star)| = \Bigg|&\alpha(\hat{\gamma} - \hat{\beta}) + \left(\frac{\lambda u_2}{2} + \frac{\mathbb{E}[Q(0)^2]}{2T}\right)\left(\frac{1}{\hat{\gamma}} - \frac{1}{\hat{\beta}}\right) \\ &- \frac{\lambda^2 u_2^2}{4T}\left(\frac{1}{\hat{\gamma}^3} - \frac{1}{\hat{\beta}^3}\right) - \frac{\lambda u_3}{6T}\left(\frac{1}{\hat{\gamma}^2} - \frac{1}{\hat{\beta}^2}\right)\Bigg|.\end{aligned}$$

Furthermore, we have

$$\begin{aligned}\frac{1}{\hat{\gamma}} - \frac{1}{\hat{\beta}} &= -\frac{\hat{\gamma} - \hat{\beta}}{\hat{\beta}^2} + \frac{(\hat{\gamma} - \hat{\beta})^2}{\hat{\beta}^3} + O\left((\gamma - \beta)^3\right), \\ \frac{1}{\hat{\gamma}^2} - \frac{1}{\hat{\beta}^2} &= -\frac{2(\hat{\gamma} - \hat{\beta})}{\hat{\beta}^3} + \frac{3(\hat{\gamma} - \hat{\beta})^2}{\hat{\beta}^4} + O\left((\gamma - \beta)^3\right), \\ \frac{1}{\hat{\gamma}^3} - \frac{1}{\hat{\beta}^3} &= -\frac{3(\hat{\gamma} - \hat{\beta})}{\hat{\beta}^4} + \frac{6(\hat{\gamma} - \hat{\beta})^2}{\hat{\beta}^5} + O\left((\gamma - \beta)^3\right).\end{aligned}$$

Substituting these yields

$$|\hat{\Pi}_T(\gamma) - \hat{\Pi}_T(\beta)| = \left| (\gamma-\beta)\left[\alpha - \frac{\lambda u_2}{2\hat{\beta}^2} + \frac{1}{2T\hat{\beta}^2}\left(\mathbb{E}[Q(0)^2] + \frac{3\lambda^2 u_2^2}{2\hat{\beta}^2} + \frac{2\lambda u_3}{3\hat{\beta}}\right)\right] \right.$$
$$\left. -(\gamma-\beta)^2\left[\frac{\lambda u_2}{2\hat{\beta}^3} + \frac{1}{2T\hat{\beta}^3}\left(\mathbb{E}[Q(0)^2] - \frac{3\lambda^2 u_2^2}{\hat{\beta}^2} - \frac{\lambda u_3}{\hat{\beta}}\right)\right]\right|$$
$$+ O\left((\gamma-\beta)^3\right).$$

Given that $\mu_T^\star = \mu_\infty^\star + \mu_\bullet/T + o(1/T)$, we find

$$|\hat{\Pi}_T(\mu_\infty^\star) - \hat{\Pi}_T(\mu_T^\star)| = \frac{|\mu_\bullet|}{T}\left(\alpha - \frac{\lambda u_2}{2(\mu_\infty^\star - \lambda)^2}\right) + O(1/T^2)$$
$$= \frac{|\mu_\bullet|}{T}\left(\alpha - \frac{\lambda u_2}{2(\sqrt{\lambda u_2/2\alpha})^2}\right) + O(1/T^2) = O(1/T^2),$$

which concludes the proof. □

References

1. Abate, J., Whitt, W.: Transient behavior of regulated Brownian motion, I: starting at the origin. Adv. Appl. Probab. **19**(3), 560–598 (1987)
2. Abate, J., Whitt, W.: Transient behavior of regulated Brownian motion, II: non-zero initial conditions. Adv. Appl. Probab. **19**(3), 599–631 (1987)
3. Abate, J., Whitt, W.: Transient behavior of the M/M/1 queue: starting at the origin. Queueing Syst. Theory Appl. **2**(1), 41–65 (1987)
4. Abate, J., Whitt, W.: Transient behavior of the M/G/1 workload process. Oper. Res. **42**(4), 750–764 (1994)
5. Asmussen, S.: Applied Probability and Queues, 2nd edn. Springer, New York (2003)
6. Awad, H.P., Glynn, P.W.: On the theoretical comparison of low-bias steady-state estimators. ACM Trans. Model. Comput. Simul. **17**(1), 4 (2007)
7. Benes, V.E.: On queues with Poisson arrivals. Ann. Math. Stat. **28**(3), 670–677 (1957)
8. Bertoin, J.: Lévy processes. Cambridge University Press, Cambridge (1996)
9. Cohen, J.W.: The Single Server Queue. North-Holland Pub. Co., Amsterdam (1969)
10. Gaver, D.P.: Imbedded Markov chain analysis of a waiting-line process in continuous time. Ann. Math. Stat. **30**(3), 698–720 (1959)
11. Gaver, D.P.: Diffusion approximations and models for certain congestion problems. J. Appl. Probab. **5**(3), 607–623 (1968)
12. Green, L.V., Kolesar, P.: The pointwise stationary approximation for queues with non-stationary arrivals. Manag. Sci. **37**(1), 84–97 (1991)
13. Harrison, J.M.: Brownian Motion and Stochastic Flow Systems. Wiley, New York (1985)
14. Janssen, A.J.E.M., van Leeuwaarden, J.S.H., Mathijsen, B.W.J.: Novel heavy-traffic regimes for large-scale service systems. SIAM J. Appl. Math. **75**(2), 787–812 (2015)
15. Janssen, A.J.E.M., van Leeuwaarden, J.S.H., Zwart, A.P.: Gaussian expansions and bounds for the Poisson distribution applied to the Erlang B formula. Adv. Appl. Prob. **40**(1), 122–143 (2008)
16. Janssen, A.J.E.M., van Leeuwaarden, J.S.H., Zwart, A.P.: Refining square-root safety staffing by expanding Erlang C. Oper. Res. **59**(6), 1512–1522 (2011)
17. Kendall, D.G.: Some problems in the theory of queues. J. R. Stat. Soc. **113**(2), 151–185 (1951)

18. Kyprianou, A.E.: Introductory Lectures on Fluctuations of Lévy Processes with Applications. Springer, Berlin (2006)
19. Massey, W.A., Whitt, W.: Uniform acceleration expansions for Markov chains with time-varying rates. Ann. Appl. Prob. **8**(4), 1130–1155 (1998)
20. Neuts, M.F.: The single server queue with Poisson input and semi-Markov service times. J. Appl. Probab. **3**(1), 202–230 (1966)
21. Newell, G.F.: Applications of Queueing Theory. Chapman and Hall, London (1982)
22. Odoni, A.R., Roth, E.: An empirical investigation of the transient behavior of stationary queueing systems. Oper. Res. Int. J. **31**(3), 432–455 (1983)
23. Pegden, C.D., Rosenshine, M.: Some new results for the M/M/1 queue. Manag. Sci. **28**(7), 821–828 (1982)
24. Prabhu, N.U.: Time-dependent results in storage theory. J. Appl. Probab. **1**(1), 1–46 (1964)
25. Randhawa, R.S.: Optimality gap of symptotically derived prescriptions in queueing systems: $o(1)$-optimality. Queueing Syst. Theory Appl. **83**(1), 131–155 (2016)
26. Sato, K.-I.: Lévy Processes and Infinitely Divisible Distributions. Cambridge University Press, Cambridge (1999)
27. Steckley, S.G., Henderson, S.G.: The error in steady-state approximations for the time-dependent waiting time distribution. Stoch. Models **23**(2), 307–332 (2007)
28. Takaćs, L.: Investigation of waiting time problems by reduction to Markov processes. Acta Math. Acad. Sci. Hung. **6**(1), 101–129 (1955)
29. Takaćs, L.: The time dependence of a single-server queue with Poisson input and general service times. Ann. Math. Stat. **33**(4), 1340–1348 (1962)
30. Whitt, W.: The pointwise stationary approximation is apsymptotically correct as the rates increase. Management Science, (1991)
31. Zhang, B., van Leeuwaarden, J.S.H., Zwart, A.P.: Staffing call centers with impatient customers: refinements to many-server asymptotics. Oper. Res. **60**(2), 461–474 (2012)