Noname manuscript No.
(will be inserted by the editor)

Queue length asymptotics for the multiple server queue with heavy-tailed Weibull service times

Mihail Bazhba[1] · **Jose Blanchet**[2] · **Chang-Han Rhee**[1] · **Bert Zwart**[1]

Received: date / Accepted: date

Abstract We study the queue-length asymptotics of many-server queues with heavy-tailed Weibull service times. Our analysis hinges on a recently developed sample path large deviations principle for Lévy processes and random walks. By following a continuous mapping approach and solving the variational problem associated with the resulting large deviation principle, we identify the most probable scenario that leads to the congestion of many-server queues with Weibull service times. In contrast to the regularly varying case, we observe a number of subtle features such as the non-trivial trade-off between the number of big jobs and their sizes, and the unexpected asymmetry in job sizes.

Keywords multiple server queue · queue length asymptotics · heavy tails · Weibull service times

Mathematics Subject Classification (2010) 60K25 · 68M20

1 Introduction

The purpose of this paper is to obtain results for the challenging multiple-server queue, with unlimited waiting room and the FCFS policy, in the case that the job size distribution has a tail of the form $e^{-L(x)x^{\alpha}}$, $\alpha \in (0,1)$. Our approach exploits the large deviation theory framework and in particular the large deviations principle for random walks with heavy-tailed Weibull increments. Consequently, we can estimate the probability of a large queue length of the $G/G/d$ queue with heavy-tailed Weibull service times and extract information about "the most likely way" so that a large queue length build up occurs. The many-server queue with heavy-tailed service times has so far mainly been considered in the case of regularly

Mihail Bazhba
E-mail: bazhba@cwi.nl

Jose Blanchet
E-mail: jose.blanchet@stanford.edu

Chang-Han Rhee
E-mail: rhee@cwi.nl

Bert Zwart
E-mail: bert.zwart@cwi.nl

[1] Centrum Wiskunde & Informatica, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
[2] Management Science and Engineering, Stanford University 475 Via Ortega, Suite 310, Stanford, CA. 94305

varying service times, see [1, 2]. In this paper, we show that, for $\gamma \in (0, \infty)$,

$$\lim_{n\to\infty} \frac{\log \mathbf{P}(Q(\gamma n) \geq n)}{L(n)n^{\alpha}} = -c^*, \tag{1}$$

with c^* the value of the optimization problem

$$\begin{aligned} &\min \sum_{i=1}^{d} x_i^{\alpha} \quad \text{subject to} \\ &l\left(s; x\right) = \lambda s - \sum_{i=1}^{d} (s - x_i)^+ \geq 1 \text{ for some } s \in [0, \gamma]. \\ &x_1, ..., x_d \geq 0 \ . \end{aligned} \tag{2}$$

where λ is the arrival rate, and service times are normalized to have unit mean. Note that this problem is equivalent to an L^{α}-norm minimization problem with $\alpha \in (0, 1)$. Such problems also appear in applications such as compressed sensing, and are strongly NP hard in general, see [3] and references therein. In our particular case, we can analyze this problem exactly, and if $\gamma \geq 1/(\lambda - \lfloor \lambda \rfloor)$, the solution takes the simple form

$$c^* = \min_{l \in \{0,\ldots,\lfloor \lambda \rfloor\}} (d - l) \left(\frac{1}{\lambda - l}\right)^{\alpha}. \tag{3}$$

This simple minimization problem has at most two optimal solutions, which represent the most likely number of big jumps that are responsible for a large queue length to occur, and the most likely buildup of the queue length is through a linear path. For smaller values of γ, asymmetric solutions can occur, leading to a piecewise linear buildup of the queue length; we refer to Section 3 for more details.

Note that the intuition that the solution to (2) yields is qualitatively different from the case in which service times have a power law. In the latter case, the optimal number of big jobs equals the minimum number of servers that need to be removed to make the system unstable, which equals $\lceil d - \lambda \rceil$. In the Weibull case, there is a nontrivial trade-off between the *number* of big jobs as well as their *size*, and this trade-off is captured by (2) and (3). This essentially answers a question posed by Sergey Foss at the Erlang Centennial conference in 2009. Our result is consistent with a more refined asymptotic estimate obtained for $d = 2$ and $\lambda < 1$ in [1]. For earlier conjectures on this problem we refer to [4]. An overview of related results for the case of regularly varying service times, can be found in [2].

We derive (1) by utilizing a tail bound for $Q(t)$, which are derived in [5]. These tail bounds are given in terms of functionals of superpositions of random walks. We show these functionals are (almost) continuous in the M_1' topology, introduced in [6], making this fit into the large deviations framework introduced in the companion paper [7].

The paper is organized as follows. Section 2 provides a model description and some useful tools used in our proofs. Section 3 provides our main result and some mathematical insights associated with it. Lastly, section 4 contains the Lemmas and proofs needed to construct Theorem 1.

2 Model Description And Preliminary Results

We consider the FCFS $G/G/d$ queuing model with d servers in which inter-arrival times are independent and identically distributed (i.i.d) random variables (r.v.s) and service times are i.i.d r.v.s Let A and S be the generic inter-arrival time and the generic service time respectively, with non-negative support. In addition, we make the following assumptions:

1) There exists θ_+ such that $\mathbf{E}(e^{\theta A}) < \infty$ for every $\theta \leq \theta_+$

2) $\mathbf{P}(S \geq x) = e^{-L(x)x^\alpha}$, $\alpha \in (0,1)$ and that $L(x)x^{\alpha-1}$ is decreasing.

Let $Q(t)$ denote the queue length process at time t in the FCFS $G/G/d$ queuing system with inter-arrival times drawn i.i.d distributed on A and service times drawn i.i.d distributed on S. Our goal is to identify the limit behavior of $\mathbf{P}(Q(\gamma n) > n)$ as $n \to \infty$ in terms of the distributions A and S. Let $\lambda = 1/\mathbf{E}[A]$ and $\mu = 1/\mathbf{E}[S]$. Let M be the renewal process associated with A. That is,

$$M(t) = \inf\{s : A(s) > t\},$$

and $A(t) \triangleq A_1 + A_2 + \cdots + A_{\lfloor t \rfloor}$ where $A_1, A_2, \ldots$ are iid copies of A. Similarly, let $N^{(i)}$ be a renewal process associated with S for each $i = 1, \ldots, d$. Let $\bar{M}_n$ and $\bar{N}_n^{(i)}$ be scaled processes of M and $N^{(i)}$ in $\mathbb{D}[0,\gamma]$. That is, $\bar{M}_n(t) = M(nt)/n$ and $\bar{N}_n^{(i)}(t) = N^{(i)}(nt)/n$ for $t \geq 0$. Our analysis hinges on Theorem 3 of [5], which for the $G/G/d$ queue states

Result 1 *For all $x > 0$ and $t \geq 0$,*

$$\mathbf{P}((Q(t) - d)^+ > x) \leq \mathbf{P}\left(\sup_{0 \leq s \leq t} \left(M(s) - \sum_{i=1}^{d} N^{(i)}(s)\right) \geq x\right). \tag{4}$$

Now, from (4), we conclude that for each $\gamma, \beta \in (0,\infty)$

$$\mathbf{P}\big(Q(\gamma n) > \beta n\big) \leq \mathbf{P}\left(\sup_{0 \leq s \leq \gamma} \left(\bar{M}_n(s) - \sum_{i=1}^{d} \bar{N}_n^{(i)}(s)\right) \geq \beta\right) \tag{5}$$

and we show that (5) is an asymptotically tight upper bound in cases of big delays. It should be noted that in [5], A_1 and $S_1^{(i)}$ are defined to have the residual distribution to make M and $N^{(i)}$ equilibrium renewal processes, but such assumptions are not necessary for (5) itself.

In view of the above, a natural way to proceed is to establish LDPs for $\bar{M}_n$ and $\bar{N}_n^{(i)}$'s. We invoke a recent result in [7] on sample path large deviations for random walks with heavy-tailed Weibull increments. Let $\mathbb{D}[0,1]$ denote the Skorokhod space—space of càdlàg functions from $[0,1]$ to $\mathbb{R}$—and $\mathcal{T}_{M_1'}$ denote the M_1' Skorokhod topology on $\mathbb{D}[0,1]$. We say that $\xi \in \mathbb{D}[0,1]$ is a pure jump function if $\xi = \sum_{i=1}^{\infty} x_i \mathbb{1}_{[u_i,1]}$ for some x_i's and u_i's such that $x_i \in \mathbb{R}$ and $u_i \in [0,1]$ for each i and u_i's are all distinct. Let $\mathbb{D}_p^{\uparrow}[0,1]$ denote the subspace of $\mathbb{D}[0,1]$ consisting of non-decreasing pure jump functions which non-negative values at the origin. The following theorem is the result used in the core of our analysis,

Result 2 *Let $\bar{X}_n^{(i)}$, $i = 1, \ldots, d$, satisfy an LDP on $(\mathbb{D}, \mathcal{T}_{M_1'})$ with speed $L(n)n^\alpha$, where $\alpha \in (0,1)$. Then, $(\bar{X}_n^{(1)}, \ldots, \bar{X}_n^{(d)})$ satisfies the LDP in $\left(\prod_{i=1}^d \mathbb{D}, \prod_{i=1}^d \mathcal{T}_{M_1'}\right)$ with speed $L(n)n^\alpha$ and the rate function $I^{(d)} : \prod_{i=1}^d \mathbb{D} \to [0,\infty]$,*

$$I^{(d)}(\xi_1, \ldots, \xi_d) \triangleq \begin{cases} \sum_{i=1}^d \sum_{t \in [0,1]} (\xi_i(t) - \xi_i(t-))^\alpha & \text{if } \xi_i \in \mathbb{D}_p^{\uparrow} \text{ with } \xi_i(0) \geq 0 \text{ for each } i \in \{1, \ldots, d\}, \\ \infty & \text{o.w.} \end{cases}. \tag{6}$$

Although we have a proper framework to study (5), note that, $\bar{M}_n$ and $\bar{N}_n^{(i)}$'s depend on random number of A_j's and $S_j^{(i)}$'s, and hence may depend on arbitrarily large number of A_j's and $S_j^{(i)}$'s with strictly positive probabilities. This does not exactly correspond to the large deviations framework we mentioned

earlier. To accommodate such a context, we introduce the following maps. Fix $\gamma > 0$. Define for $\mu > 0$, $\Psi_\mu : \mathbb{D}[0, \gamma/\mu] \to \mathbb{D}[0, \gamma]$ be

$$\Psi_\mu(\xi)(t) \triangleq \sup_{s \in [0, t/\mu]} \xi(s),$$

and for each μ define a map $\Phi_\mu : \mathbb{D}[0, \gamma/\mu] \to \mathbb{D}[0, \gamma]$ as

$$\Phi_\mu(\xi)(t) \triangleq \varphi_\mu(\xi)(t) \wedge \psi_\mu(\xi)(t),$$

where

$$\varphi_\mu(\xi)(t) \triangleq \inf\{s \in [0, \gamma/\mu] : \xi(s) > t\} \qquad \text{and} \qquad \psi_\mu(\xi)(t) \triangleq \frac{1}{\mu}\Big(\gamma + \big[t - \Psi_\mu(\xi)(\gamma)\big]_+\Big).$$

Here we denoted $\max\{x, 0\}$ with $[x]_+$. In words, between the origin and the supremum of ξ, $\Phi_\mu(\xi)(s)$ is the first passage time of ξ crossing the level s; from there to the final point γ, $\Phi_\mu(\xi)$ increases linearly from γ/μ at rate $1/\mu$ (instead of jumping to ∞ and staying there). Define $\bar{A}_n \in \mathbb{D}[0, \gamma/\mathbf{E}A]$ as $\bar{A}_n(t) \triangleq A(nt)/n$ for $t \in [0, \gamma/\mathbf{E}A]$ and $\bar{S}_n^{(i)} \in \mathbb{D}[0, \gamma/\mathbf{E}S]$ as $\bar{S}_n^{(i)}(t) \triangleq S^{(i)}(nt)/n = \frac{1}{n}\sum_{j=1}^{\lfloor nt \rfloor} S_j^{(i)}$ for $t \in [0, \gamma/\mathbf{E}S]$. We will show that

- $\bar{A}_n$ and $\bar{S}_n^{(i)}$ satisfy certain LDPs (Proposition 1);
- $\Phi_{\mathbf{E}A}(\cdot)$ and $\Phi_{\mathbf{E}S}(\cdot)$ are continuous functions, and hence, $\Phi_{\mathbf{E}A}(\bar{A}_n)$ and $\Phi_{\mathbf{E}S}(\bar{S}_n^{(i)})$ satisfy the LDPs deduced by a contraction principle (Proposition 3, 4);
- $\bar{M}_n$ and $\bar{N}_n^{(i)}$ are equivalent to $\Phi_{\mathbf{E}A}(\bar{A}_n)$ and $\Phi_{\mathbf{E}S}(\bar{S}_n^{(i)})$, respectively, in terms of their large deviations (Proposition 2); so $\bar{M}_n$ and $\bar{N}_n^{(i)}$ satisfy the same LDPs (Proposition 4);
- and hence, the log asymptotics of $\mathbf{P}(Q(\gamma n) > n)$ can be derived by the solution of a quasi-variational problem characterized by the rate functions of such LDPs (Proposition 5);

and then solve the quasi-variational problem to establish the asymptotic bound.

3 Main results

In this section we will state the most important contribution of this paper.

Theorem 1 *For each $\gamma \in (0, \infty)$, for the queue length process Q, it holds that*

$$\lim_{n\to\infty} \frac{1}{L(n)n^\alpha} \log \mathbf{P}(Q(\gamma n) > n) = -c^*.$$

For $\gamma \geq \frac{1}{\lambda}$, c^* is equal with

$$\min\left\{ \inf_{0<k\leq\lfloor\lambda\rfloor:\gamma<1/(\lambda-k)} \left\{(d-k)\,\gamma^\alpha + (1-\gamma\lambda+\gamma k)^\alpha \left(k - \lfloor\lambda\rfloor \wedge \lfloor\lambda - 1/\gamma\rfloor\right)^{1-\alpha}\right\}, \right. \tag{7}$$
$$\left. \min_{l=0}^{\lfloor\lambda\rfloor\wedge\lfloor\lambda-1/\gamma\rfloor} \left\{(d-l)\left(\frac{1}{\lambda-l}\right)^\alpha\right\} \right\},$$

while for $\gamma < \frac{1}{\lambda}$, $c^* = \infty$.

We make some comments that are meant to provide some physical insight, and highlight differences with the well studied case where the job sizes follow a regularly varying distribution.

If $\gamma < 1/\lambda$, no finite number of large jobs suffice, and we conjecture that the large deviations behavior is driven by a combination of light and heavy tailed phenomena in which the light tailed dynamics involve pushing the arrival rate by exponential tilting to the critical value $1/\gamma$, followed by the heavy-tailed contribution evaluated as we explain in the following development.

If $\gamma > 1/\lambda$ the following features are contrasting with the case of regularly varying service-time tails:

1. The large deviations behavior is not driven by the smallest number of jumps which drives the queueing system to instability (i.e. $\lceil d - \lambda \rceil$). In other words, in the Weibull setting, it might be cheaper to block more servers.
2. The amount by which the servers are blocked may not be the same among all of the servers which are blocked.

To illustrate the first point, assume $\gamma > b/\left(\lambda - \lfloor \lambda \rfloor\right)$, in which case

$$\lfloor \lambda \rfloor \leq \lfloor \lambda - b/\gamma \rfloor ,$$

and the optimal solution of c^* reduces to

$$\min_{l=0}^{\lfloor \lambda \rfloor} \left\{ (d-l) \left(\frac{b}{\lambda - l} \right)^{\alpha} \right\}.$$

Let us use l^* to denote an optimizer for the previous expression; intuitively, $d - l^*$ represents the optimal number of servers to be blocked (observe that $d - \lfloor \lambda \rfloor = \lceil d - \lambda \rceil$ corresponds to the number of servers blocked in the regularly varying case). Note that if we define

$$f(t) = (d-t)(\lambda - t)^{-\alpha},$$

for $t \in [0, \lfloor \lambda \rfloor]$, then the derivative $\dot{f}(\cdot)$ satisfies

$$\dot{f}(t) = \alpha (d-t)(\lambda - t)^{-\alpha - 1} - (\lambda - t)^{-\alpha}.$$

Hence,

$$\dot{f}(t) < 0 \Longleftrightarrow t < \frac{(\lambda - \alpha d)}{(1-\alpha)}$$

and

$$\dot{f}(t) > 0 \Longleftrightarrow t > \frac{(\lambda - \alpha d)}{(1-\alpha)},$$

with $\dot{f}(t) = 0$ if and only if $t = (\lambda - \alpha d)/(1-\alpha)$. This observation allows to conclude that whenever $\gamma > b/\left(\lambda - \lfloor \lambda \rfloor\right)$ we can distinguish two cases. The first one occurs if

$$\lfloor \lambda \rfloor \leq \frac{(\lambda - \alpha d)}{(1-\alpha)},$$

in which case $l^* = \lfloor \lambda \rfloor$ (this case is qualitatively consistent with the way in which large deviations occur in the regularly varying case). On the other hand, if

$$\lfloor \lambda \rfloor > \frac{(\lambda - \alpha d)}{(1-\alpha)},$$

then we must have that l^*

$$l^* = \left\lfloor \frac{(\lambda - \alpha d)}{(1-\alpha)} \right\rfloor \text{ or } l^* = \left\lceil \frac{(\lambda - \alpha d)}{(1-\alpha)} \right\rceil,$$

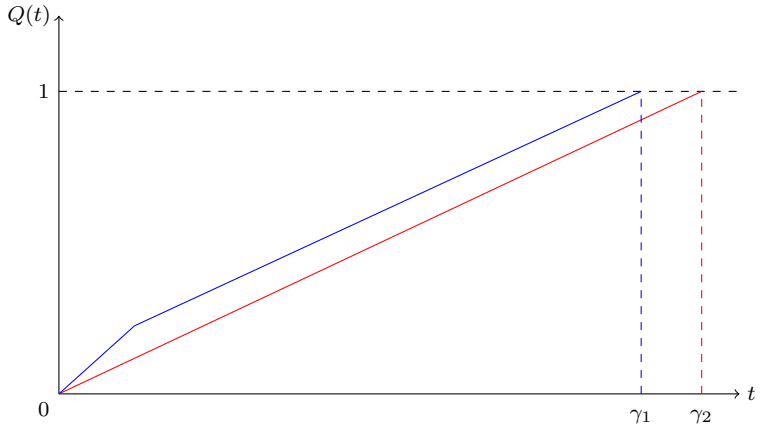

Fig. 1 Most likely path for the queue build-up upto times $\gamma_1 = \frac{1}{\lambda-1} - 0.1$ and $\gamma_2 = \frac{1}{\lambda-1}$ where the number of servers is $d = 2$, the arrival rate is $\lambda = 1.49$, and the Weibull shape parameter of the service time is $\alpha = 0.1$.

this case is the one which we highlighted in feature i) in which we may obtain $d - l^* > \lceil d - \lambda \rceil$ and therefore more servers are blocked relative to the large deviations behavior observed in the regularly varying case. Still, however, the blocked servers are symmetric in the sense that they are treated in exactly the same way.

In contrast, the second feature indicates that the most likely path to overflow may be obtained by blocking not only a specific amount to drive the system to instability, but also by blocking the corresponding servers by different loads in the large deviations scaling. To appreciate this we must assume that $\lambda^{-1} < \gamma \leq 1/(\lambda - \lfloor \lambda \rfloor)$.

In this case, the contribution of the infimum in (16) becomes relevant. In order to see that we can obtain mixed solutions, it suffices to consider the case $d = 2$, and $1 < \lambda < 2$ and

$$1/\lambda < \gamma < 1/(\lambda - 1).$$

Moreover, select $\gamma = 1/(\lambda - 1) - \delta$ and $\lambda = 2 - \delta^3$ for $\delta > 0$ sufficiently small, then

$$\gamma^\alpha + (1 - \gamma(\lambda - 1))^\alpha = 1 - \delta\alpha + \delta^\alpha + o\left(\delta^2\right) \leq 2^{1-\alpha},$$

concluding that

$$\gamma^\alpha + (1 - \gamma(\lambda - 1))^\alpha < 2\left(\frac{1}{\lambda}\right)^\alpha$$

for δ small enough and therefore we can have mixed solutions.

For example, consider the case $d = 2$, $\lambda = 1.49$, $\alpha = 0.1$ and $\gamma = \frac{1}{\lambda-1} - 0.1$. For these values, $\gamma_1^\alpha + (1 - \gamma_1(\lambda - 1))^\alpha < 2\left(\frac{1}{\lambda}\right)^\alpha$, and the most likely scenario leading to a large queue length is two big jobs arriving at the beginning and blocking both servers with different loads. On the other hand, if $\gamma = \frac{1}{\lambda-1}$ the most likely scenario is a single big job blocking one server. These two scenarios are illustrated in Figure 1.

We conclude this section by presenting future directions of research. It is worth mentioning that we provide asymptotics only for the transient model of the queue length process Q. For the obverse i.e; when the queue is in steady state, more elaborate work is needed to overcome the technicalities arising with the large deviations framework. Specifically, one has to prove that, the interchange of limits as γ and n tend to infinity,

$$\lim_{\gamma\to\infty}\lim_{n\to\infty}\frac{1}{L(n)n^\alpha}\log\mathbf{P}(Q(\gamma n)\geq n) = \lim_{n\to\infty}\frac{1}{L(n)n^\alpha}\lim_{\gamma\to\infty}\log\mathbf{P}(Q(\gamma n)\geq n)$$

is valid. We conjecture that the optimal value, similar to (7), of the variational problem associated with the steady state model will consist solely of the term, $\min_{l=0}^{\lfloor\lambda\rfloor}\left\{(d-l)\left(\frac{1}{\lambda-l}\right)^{\alpha}\right\}$, obtained by taking $\gamma=\infty$ in (2). Lastly, it is possible to obtain Theorem 1 under greater generality by relaxing the assumption that A is light-tailed. For this, all one has to notice is,

$$\mathbf{P}(\sup_{s\leq\gamma}\{\bar{M}_n-\sum_{i=1}^{d}N^{(i)}(s)\geq x\})\leq\mathbf{P}(\sup_{s\leq\gamma}\{\bar{M}_n-(\lambda+\epsilon)\geq\delta x\})+\mathbf{P}(\sup_{s\leq\gamma}\{(\lambda+\epsilon)s-\sum_{i=1}^{d}N^{(i)}(s)\geq x(1-\delta)\})$$
$$=(I)+(II).$$

We expect that the logarithmic asymptotics of (II) will lead to a variational problem parameterized by δ,ϵ whose optimal value is denoted as $c^*_{\delta,\epsilon}$. Evaluating the limit of $c^*_{\delta,\epsilon}$ as ϵ,δ tend to zero and showing its equality with c^*, will lead to the same logaritmic asymptotics seen in (7). By exploiting the light-tailed traits of (I), one can reach at the same conclusion seen in Theorem 1.

4 Proof of Theorem 1

Let $\mathbb{D}^{\uparrow}_p[0,\gamma/\mu]$ be the subspace of $\mathbb{D}[0,\gamma/\mu]$ consisting of non-decreasing pure jump functions that assume non-negative values at the origin, and define $\zeta_\mu\in\mathbb{D}[0,\gamma/\mu]$ as $\zeta_\mu(t)\triangleq\mu t$. Let $\mathbb{D}^{\mu}[0,\gamma/\mu]\triangleq\zeta_\mu+\mathbb{D}^{\uparrow}_p[0,\gamma/\mu]$.

Proposition 1 *$\bar{A}_n$ satisfies the LDP on $\left(\mathbb{D}[0,\gamma/\mathbf{E}A],\,d_{M_1'}\right)$ with speed $L(n)n^{\alpha}$ and rate function*

$$I_0(\xi)=\begin{cases}0 & \text{if } \xi=\zeta_{\mathbf{E}A},\\ \infty & \text{otherwise,}\end{cases}$$

and $\bar{S}^{(i)}_n$ satisfies the LDP on $\left(\mathbb{D}[0,\gamma/\mathbf{E}S],\,d_{M_1'}\right)$ with speed $L(n)n^{\alpha}$ and the rate function

$$I_i(\xi)=\begin{cases}\sum_{t\in[0,\gamma/\mathbf{E}S]}(\xi(t)-\xi(t-))^{\alpha} & \text{if } \xi\in\mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S],\\ \infty & \text{otherwise.}\end{cases}$$

Proof Firstly, note that Mogulski's sample path large deviation principle implies that $\frac{1}{n}\sum_{j=1}^{\lfloor nt\rfloor}(A_j-\mathbf{E}A\cdot t)$ satisfies the LDP with speed $L(n)n^{\alpha}$ and with the rate function

$$I_A(\xi)=\begin{cases}0 & \text{if } \xi=0,\\ \infty & \text{otherwise,}\end{cases}.$$

On the other hand, due to Result 2, $\frac{1}{n}\sum_{j=1}^{\lfloor nt\rfloor}\left(S^{(i)}_j-\mathbf{E}S\cdot t\right)$ satisfies the LDP with the rate function

$$I_{S^{(i)}}(\xi)=\begin{cases}\sum_{t\in[0,\gamma/\mathbf{E}S]}(\xi(t)-\xi(t-))^{\alpha} & \text{if } \xi\in\mathbb{D}^{\uparrow}_p[0,\gamma/\mathbf{E}S],\\ \infty & \text{otherwise.}\end{cases}$$

Consider the map $\Upsilon_\mu:\left(\mathbb{D}[0,\gamma/\mu],\mathcal{T}_{M_1'}\right)\to\left(\mathbb{D}[0,\gamma/\mu],\mathcal{T}_{M_1'}\right)$ where $\Upsilon_\mu(\xi)\triangleq\xi+\zeta_\mu$. We prove that $\Upsilon_{\mathbf{E}A}$ is a continuous function w.r.t. the M_1' topology. Suppose that, $\xi_n\to\xi$ in $\mathbb{D}[0,\gamma/\mathbf{E}A]$ w.r.t. the M_1' topology. As a result, there exist parameterizations $(u_n(s),t_n(s))$ of ξ_n and $(u(s),t(s))$ of ξ so that,

$$\sup_{s\leq\gamma/\mathbf{E}A}\{|u_n(s)-u(s)|+|t_n(s)-t(s)|\}\to 0\qquad\text{as } n\to\infty.$$

This implies that $\max\{\sup_{t\leq\gamma/\mathbf{E}A}|u_n(s)-u(s)|,\sup_{t\leq\gamma/\mathbf{E}A}|t_n(s)-t(s)|\}\to 0$ as $n\to\infty$. Observe that, if $(u(s),t(s))$ is a parameterization for ξ, then $(u(s)+\mathbf{E}A\cdot t(s),t(s))$ is a parameterization for $\Upsilon_{\mathbf{E}A}(\xi)$. Consequently,

$$\begin{aligned}&\sup_{s\leq\gamma/\mathbf{E}A}\{|u_n(s)+\mathbf{E}A\cdot t_n(s)-u(s)-\mathbf{E}A\cdot t(s)|+|t_n(s)-t(s)|\}\\&\quad\leq\sup_{s\leq\gamma/\mathbf{E}A}\{|u_n(s)-u(s)|\}+\sup_{s\leq\gamma/\mathbf{E}A}\{(\mathbf{E}A+1)|t_n(s)-t(s)|\}\to 0.\end{aligned}$$

Thus, $\Upsilon_{\mathbf{E}A}(\xi_n)\to\Upsilon_{\mathbf{E}A}(\xi)$ in the M_1' topology, proving that the map is continuous. The same argument holds for $\Upsilon_{\mathbf{E}S}$. By the contraction principle, $\bar{A}_n$ obeys the LDP with the rate function $I_0(\zeta)\triangleq\inf\{I_A(\xi):\xi\in\mathbb{D}[0,\gamma/\mathbf{E}A],\,\zeta=\Upsilon_{\mathbf{E}A}(\xi)\}$. Observe that $I_A(\xi)=\infty$ for $\xi\neq 0$ therefore,

$$I_0(\zeta)=\begin{cases}0 & \text{if } \zeta=\Upsilon_{\mathbf{E}A}(0)=\zeta_{\mathbf{E}A},\\ \infty & \text{otherwise.}\end{cases}$$

Similarly, $I_{S^{(i)}}(\xi)=\infty$ when ξ is not a non-decreasing pure jump function. Note that $\xi\in\mathbb{D}_s^{\uparrow}$ implies that $\zeta=\Upsilon_{\mathbf{E}S}(\xi)$ belongs to $\mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S]$. Taking into account the form of $I_{S^{(i)}}$ and that $I_i(\zeta)\triangleq\inf\{I_{S^{(i)}}(\xi):\xi\in\mathbb{D}[0,\gamma/\mathbf{E}S],\zeta=\Upsilon_{\mathbf{E}S}(\xi)\}$, we conclude

$$I_i(\zeta)=\begin{cases}\sum_{t\in[0,\gamma/\mathbf{E}S]}(\zeta(t)-\zeta(t-))^{\alpha} & \text{for } \zeta\in\mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S],\\ \infty & \text{otherwise.}\end{cases}$$

Proposition 2 *$\bar{M}_n$ and $\Phi_{\mathbf{E}A}(\bar{A}_n)$ are exponentially equivalent in $\left(\mathbb{D}[0,\gamma],\mathcal{T}_{M_1'}\right)$. $\bar{N}_n^{(i)}$ and $\Phi_{\mathbf{E}S}(\bar{S}_n^{(i)})$ are exponentially equivalent in $\left(\mathbb{D}[0,\gamma],\mathcal{T}_{M_1'}\right)$ for each $i=1,\ldots,d$.*

Proof Proof We first claim that $d_{M_1'}(\bar{N}_n^{(i)},\Phi_{\mathbf{E}S}(\bar{S}_n^{(i)}))\geq\epsilon$ implies that

$$\exists\,\delta>0 \text{ such that } \gamma-\Psi_{\mathbf{E}S}(\bar{S}_n^{(i)})(\gamma)\geq\delta.$$

To see this, remember that by construction, $\bar{N}_n^{(i)}$ and $\Phi_{\mathbf{E}S}(\bar{S}_n^{(i)})$ coincide on $[0,\Psi_{\mathbf{E}S}(\bar{S}_n^{(i)})(\gamma)]$. Consequently, if the distance of $\bar{N}_n^{(i)}$ and $\Phi_{\mathbf{E}S}(\bar{S}_n^{(i)})$ is bigger than ϵ then $\Psi_{\mathbf{E}S}(\bar{S}_n^{(i)})(\gamma)$ has to be strictly smaller than γ . That is,

$$\{d_{M_1'}(\bar{N}_n^{(i)},\Phi_{\mathbf{E}S}(\bar{S}_n^{(i)}))\geq\epsilon\}\implies\{\exists\,\delta>0:\gamma-\Psi_{\mathbf{E}S}(\bar{S}_n^{(i)})(\gamma)\geq\delta\}.$$

Therefore,

$$\limsup_{n\to\infty}\frac{\log\mathbf{P}\Big(d_{M_1'}(\bar{N}_n^{(i)},\Phi_{\mathbf{E}S}(\bar{S}_n^{(i)}))\geq\epsilon\Big)}{L(n)n^{\alpha}}\leq\limsup_{n\to\infty}\frac{\log\mathbf{P}\Big(\exists\,\delta>0:\gamma-\Psi_{\mathbf{E}S}(\bar{S}_n^{(i)})(\gamma)\geq\delta\Big)}{L(n)n^{\alpha}}$$

and we are done for the exponential equivalence between $\bar{N}_n^{(i)}$ and $\Phi_{\mathbf{E}A}(\bar{S}_n^{(i)})$ if we prove that

$$\limsup_{n\to\infty}\frac{\log\mathbf{P}\Big(\exists\,\delta>0:\gamma-\Psi_{\mathbf{E}S}(\bar{S}_n^{(i)})(\gamma)\geq\delta\Big)}{L(n)n^{\alpha}}=-\infty.\tag{8}$$

Observe that for every $\xi \in \mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S]$, it holds that, $\xi(t) \geq \mathbf{E}S \cdot t$ for every $t \in [0,\gamma/\mathbf{E}S]$. That is, $\Psi_{\mathbf{E}S}(\xi)(\gamma) \geq \gamma$. Now, for (8), the event $\{\exists \delta > 0 : \Psi_\mu(\bar{S}_n^{(i)})(\gamma) \leq \gamma - \delta\}$ implies $\{d_{M_1'}(\bar{S}_n^{(i)}, \mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S]) \geq \delta\}$. Thus,

$$\begin{aligned}\limsup_{n\to\infty} \frac{\log \mathbf{P}\Big(\exists \delta > 0 : \gamma - \Psi_\mu(\bar{S}_n^{(i)})(\gamma) \geq \delta\Big)}{L(n)n^\alpha} &\leq \limsup_{n\to\infty} \frac{\log \mathbf{P}\Big(d_{M_1'}(\bar{S}_n^{(i)}, \mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S]) > 0\Big)}{L(n)n^\alpha} \\ &\leq -\inf_{\xi \in (\mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S])^c} I_i(\xi) = -\infty,\end{aligned}$$

where the second inequality is due to the LDP upper bound for $\bar{S}_n^{(i)}$ in Proposition 1.

This concludes the proof for the exponential equivalence between $\bar{N}_n^{(i)}$ and $\Phi_{\mathbf{E}S}(\bar{S}_n^{(i)})$. The exponential equivalence between $\bar{M}_n$ and $\Phi_{\mathbf{E}A}(\bar{A}_n)$ is essentially identical, and hence, omitted.

Let $\mathcal{D}_{\Phi_\mu} \triangleq \{\xi \in \mathbb{D}[0,\gamma/\mu] : \Phi_\mu(\xi)(\gamma) - \Phi_\mu(\xi)(\gamma-) > 0\}$.

Proposition 3 *For each $\mu \in \mathbb{R}$, $\Phi_\mu : (\mathbb{D}[0,\gamma/\mu], \mathcal{T}_{M_1'}) \to (\mathbb{D}[0,\gamma], \mathcal{T}_{M_1'})$ is continuous on $\mathcal{D}_{\Phi_\mu}^{\mathsf{c}}$.*

Proof Note that $\Phi_\mu = \Phi_\mu \circ \Psi$ and Ψ is continuous, so we only need to check the continuity of Φ_μ over the range of Ψ, in particular, non-decreasing functions. Let ξ be a non-decreasing function in $\mathbb{D}[0,\gamma/\mu]$. We consider two cases separately: $\Phi_\mu(\xi)(\gamma) > \gamma/\mu$ and $\Phi_\mu(\xi)(\gamma) \leq \gamma/\mu$.

We start with the case $\Phi_\mu(\xi)(\gamma) > \gamma/\mu$. Pick $\epsilon > 0$ such that $\Phi_\mu(\xi)(\gamma) > \gamma/\mu + 2\epsilon$ and $\xi(\gamma/\mu) + 2\epsilon < \gamma$. For such an ϵ, it is straightforward to check that $d_{M_1'}(\zeta,\xi) < \epsilon$ implies $\Phi_\mu(\zeta)(\gamma) > \mu/\gamma$ and ζ never exceeds γ on $[0,\gamma/\mu]$. Therefore, the parametrizations of $\Phi_\mu(\xi)$ and $\Phi_\mu(\zeta)$ consist of the parametrizations—with the roles of space and time interchanged—of the original ξ and ζ concatenated with the linear part coming from ψ_μ. More specifically, suppose that $(x,t) \in \Gamma(\xi)$ and $(y,r) \in \Gamma(\zeta)$ are parametrizations of ξ and ζ. If we define on $s \in [0,1]$,

$$\begin{aligned}
x'(s) &\triangleq \begin{cases} t(2s) & \text{if } s \leq 1/2 \\ \frac{1}{\mu}\big(t'(s) - \Psi(\xi)(\gamma) + \gamma\big) & \text{if } s > 1/2 \end{cases}, \\
t'(s) &\triangleq \begin{cases} x(2s) & \text{if } s \leq 1/2 \\ \big(\gamma - \Psi(\xi)(\gamma)\big)(2s-1) + \Psi(\xi)(\gamma) & \text{if } s > 1/2 \end{cases}, \\
y'(s) &\triangleq \begin{cases} r(2s) & \text{if } s \leq 1/2 \\ \frac{1}{\mu}\big(r'(s) - \Psi(\zeta)(\gamma) + \gamma\big) & \text{if } s > 1/2 \end{cases}, \\
r'(s) &\triangleq \begin{cases} y(2s) & \text{if } s \leq 1/2 \\ \big(\gamma - \Psi(\zeta)(\gamma)\big)(2s-1) + \Psi(\zeta)(\gamma) & \text{if } s > 1/2 \end{cases},
\end{aligned}$$

then $(x',t') \in \Gamma(\Phi_\mu(\xi))$, $(y',r') \in \Gamma(\Phi_\mu(\zeta))$. Noting that

$$\begin{aligned}
&\|x' - y'\|_\infty + \|t' - r'\|_\infty \\
&= \sup_{s\in[0,1/2]} |t(2s) - r(2s)| \vee \sup_{s\in(1/2,1]} |x'(s) - y'(s)| + \sup_{s\in[0,1/2]} |x(2s) - y(2s)| \vee \sup_{s\in(1/2,1]} |t'(s) - r'(s)| \\
&= \|t - r\|_\infty \vee \mu^{-1}|\Psi(\zeta)(\gamma) - \Psi(\xi)(\gamma)| + \|x - y\|_\infty \vee |\Psi(\zeta)(\gamma) - \Psi(\xi)(\gamma)| \\
&\leq \mu^{-1}\|t - r\|_\infty \vee \|x - y\|_\infty + \|x - y\|_\infty \leq (1 + \mu^{-1})(\|x - y\|_\infty + \|t - r\|_\infty),
\end{aligned}$$

and taking the infimum over all possible parametrizations, we conclude that $d_{M_1'}(\Phi_\mu(\xi), \Phi_\mu(\zeta)) \leq (1 + \mu^{-1}) d_{M_1'}(\xi,\zeta) \leq (1 + \mu^{-1})\epsilon$, and hence, Φ_μ is continuous at ξ.

Turning to the case $\Phi_\mu(\xi)(\gamma) \le \gamma/\mu$, let $\epsilon > 0$ be given. Due to the assumption that $\Phi_\mu(\xi)$ is continuous at γ, there has to be a $\delta > 0$ such that $\varphi_\mu(\xi)(\gamma) + \epsilon < \varphi_\mu(\xi)(\gamma - \delta) \le \varphi_\mu(\xi)(\gamma + \delta) \le \varphi_\mu(\xi)(\gamma) + \epsilon$. We will prove that if $d_{M_1'}(\xi, \zeta) < \delta \wedge \epsilon$, then $d_{M_1'}(\Phi_\mu(\xi), \Phi_\mu(\zeta)) \le 8\epsilon$. Since the case where $\Phi_\mu(\zeta)(\gamma) \ge \mu/\gamma$ is similar to the above argument, we focus on the case $\Phi_\mu(\zeta)(\gamma) < \mu/\gamma$; that is, ζ also crosses level γ before γ/μ. Let $(x,t) \in \Gamma(\xi)$ and $(y,r) \in \Gamma(\zeta)$ be such that $\|x-y\|_\infty + \|t-r\|_\infty < \delta$. Let $s_x \triangleq \inf\{s \ge 0 : x(s) > \gamma\}$ and $s_y \triangleq \inf\{s \ge 0 : y(s) > \gamma\}$. Then it is straightforward to check $t(s_x) = \varphi_\mu(\xi)(\gamma)$ and $r(s_y) = \varphi_\mu(\zeta)(\gamma)$. Of course, $x(s_x) = \gamma$ and $y(s_y) = \gamma$. If we set $x'(s) \triangleq t(s \wedge s_x)$, $t'(s) \triangleq x(s \wedge s_x)$, and $y'(s) \triangleq r(s \wedge s_y)$, $r'(s) \triangleq y(s \wedge s_y)$, then

$$\begin{aligned}
\|x'-y'\|_\infty &\le \|t-r\|_\infty + \sup_{s \in [s_x \wedge s_y, s_x \vee s_y]} \big\{ |t(s_x) - r(s)| \vee |t(s) - r(s_y)| \big\} \\
&\le \|t-r\|_\infty + \sup_{s \in [s_x \wedge s_y, s_x \vee s_y]} \big\{ \big(|t(s_x) - t(s)| + |t(s) - r(s)| \big) \vee \big(|t(s) - t(s_y)| + |t(s_y) - r(s_y)| \big) \big\} \\
&\le \|t-r\|_\infty + \big(|t(s_x) - t(s_y)| + \|t-r\|_\infty \big) \vee \big(|t(s_y) - t(s_x)| + \|t-r\|_\infty \big) \\
&\le 2\|t-r\|_\infty + 2|t(s_x) - t(s_y)|.
\end{aligned}$$

Now we argue that $t(s_x) - \epsilon \le t(s_y) \le t(s_x) + \epsilon$. To see this, note first that $x(s_y) < x(s_x) + \delta = \gamma + \delta$, and hence,

$$t(s_y) \le \varphi_\mu(\xi)(x(s_y)) \le \varphi_\mu(\xi)(\gamma + \delta) \le \varphi_\mu(\xi)(\gamma) + \epsilon = t(s_x) + \epsilon.$$

On the other hand,

$$t(s_x) - \epsilon = \varphi_\mu(\xi)(\gamma) - \epsilon \le \varphi_\mu(\gamma - \delta) \le t(s_y),$$

where the last inequality is from $\xi(t(s_y)) \ge x(s_y) > x(s_x) - \delta = \gamma - \delta$ and the definition of φ_μ. Therefore, $\|x'-y'\|_\infty \le 2\delta + 2\epsilon < 4\epsilon$. Now we are left with showing that $\|t'-r'\|_\infty$ can be bounded in terms of ϵ.

$$\begin{aligned}
\|t'-r'\|_\infty &\le \|x-y\|_\infty + \sup_{s \in [s_x \wedge s_y, s_x \vee s_y]} \{ |x(s_x) - y(s)| \vee |x(s) - y(s_y)| \} \\
&\le \|x-y\|_\infty + \sup_{s \in [s_x \wedge s_y, s_x \vee s_y]} \big\{ \big(|x(s_x) - x(s)| + |x(s) - y(s)| \big) \vee \big(|x(s) - x(s_y)| + |x(s_y) - y(s_y)| \big) \big\} \\
&\le \|x-y\|_\infty + \big(|t(s_x) - t(s_y)| + \|x-y\|_\infty \big) \vee \big(|x(s_x) - x(s_y)| + \|x-y\|_\infty \big) \\
&\le 2\|x-y\|_\infty + 2|x(s_x) - x(s_y)| = 2\|x-y\|_\infty + 2|y(s_y) - x(s_y)| \le 4\|x-y\|_\infty < 4\epsilon.
\end{aligned}$$

Therefore, $d_{M_1'}\big(\Phi_\mu(\xi), \Phi_\mu(\zeta)\big) \le \|x'-y'\|_\infty + \|t'-r'\|_\infty < 8\epsilon$.

Let $\check{\mathbb{C}}^\mu[0,\gamma] \triangleq \{\zeta \in \mathbb{C}[0,\gamma] : \zeta = \varphi_\mu(\xi) \text{ for some } \xi \in \mathbb{D}^\mu[0,\gamma/\mu]\}$ where $\mathbb{C}[0,\gamma]$ is the subspace of $\mathbb{D}[0,\gamma]$ consisting of continuous paths, and define $\tau_s(\xi) \triangleq \max\Big\{0, \sup\{t \in [0,\gamma] : \xi(t) = s\} - \inf\{t \in [0,\gamma] : \xi(t) = s\}\Big\}$.

Proposition 4 *$\Phi_{\mathbf{E}A}(\bar{A}_n)$ and $\bar{M}_n$ satisfy the LDP with speed $L(n)n^\alpha$ and the rate function*

$$I_0'(\xi) \triangleq \begin{cases} 0 & \text{if } \xi(t) = t/\mathbf{E}A, \\ \infty & \text{otherwise,} \end{cases}$$

and for $i = 1, ..., d$, $\Phi_{\mathbf{E}S}(\bar{S}_n^{(i)})$ and $\bar{N}_n^{(i)}$ satisfy the LDP with speed $L(n)n^\alpha$ and the rate function

$$I_i'(\xi) \triangleq \begin{cases} \sum_{s \in [0, \gamma/\mathbf{E}S]} \tau_s(\xi)^\alpha & \text{if } \xi \in \check{\mathbb{C}}^{1/\mathbf{E}S}[0,\gamma], \\ \infty & \text{otherwise.} \end{cases}$$

Proof Let $\hat{I}'_0(\zeta) \triangleq \inf\{I_0(\xi) : \xi \in \mathbb{D}[0,\gamma/\mathbf{E}A], \zeta = \Phi_{\mathbf{E}A}(\xi)\}$ and $\hat{I}'_i(\zeta) \triangleq \inf\{I_i(\xi) : \xi \in \mathbb{D}[0,\gamma/\mathbf{E}S], \zeta = \Phi_{\mathbf{E}S}(\xi)\}$ for $i = 1,\dots,d$. From Proposition 1, 2, 3, and the extended contraction principle (p.367 of [6], Theorem 2.1 of [8]), it is enough to show that $I'_i = \hat{I}'_i$ for $i = 0,\dots,d$.

Starting with $i = 0$, note that $I_0(\xi) = \infty$ if $\xi \neq \zeta_{\mathbf{E}A}$, and hence,

$$\hat{I}'_0(\zeta) = \inf\{I_0(\xi) : \xi \in \mathbb{D}[0,\gamma/\mathbf{E}A], \zeta = \Phi_{\mathbf{E}A}(\xi)\} = \begin{cases} 0 & \text{if } \zeta = \Phi_{\mathbf{E}A}(\zeta_{\mathbf{E}A}) \\ \infty & \text{o.w.} \end{cases}. \tag{9}$$

Also, since $\Psi(\zeta_{\mathbf{E}A})(\gamma/\mathbf{E}A) = \gamma$, $\psi_{\mathbf{E}A}(\zeta_{\mathbf{E}A})(t) = \frac{1}{\mathbf{E}A}(\gamma + [t-\gamma]_+) = \gamma/\mathbf{E}A$. Therefore, $\Phi_{\mathbf{E}A}(\zeta_{\mathbf{E}A})(t) = \inf\{s \in [0,\gamma/\mathbf{E}A] : s > t/\mathbf{E}A\} \wedge (\gamma/\mathbf{E}A) = (t/\mathbf{E}A) \wedge (\gamma/\mathbf{E}A) = t/\mathbf{E}A$ for $t \in [0,\gamma]$. With (9), this implies $I'_0 = \hat{I}'_0$.

Turning to $i = 1,\dots,d$, note first that since $I_i(\xi) = \infty$ for any $\xi \notin \mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S]$,

$$\hat{I}'_i(\zeta) = \inf\{I_i(\xi) : \xi \in \mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S], \zeta = \Phi_{\mathbf{E}S}(\xi)\}.$$

Note also that $\Phi_{\mathbf{E}S}$ can be simplified on $\mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S]$: it is easy to check that if $\xi \in \mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S]$, $\psi_{\mathbf{E}S}(\xi)(t) = \gamma/\mathbf{E}S$ and $\varphi_{\mathbf{E}S}(\xi)(t) \leq \gamma/\mathbf{E}S$ for $t \in [0,\gamma]$. Therefore, $\Phi_{\mathbf{E}S}(\xi) = \varphi_{\mathbf{E}S}(\xi)$, and hence,

$$\hat{I}'_i(\zeta) = \inf\{I_i(\xi) : \xi \in \mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S], \zeta = \varphi_{\mathbf{E}S}(\xi)\}.$$

Now if we define $\varrho_{\mathbf{E}S} : \mathbb{D}[0,\gamma/\mathbf{E}S] \to \mathbb{D}[0,\gamma/\mathbf{E}S]$ as

$$\varrho_{\mathbf{E}S}(\xi)(t) \triangleq \begin{cases} \xi(t) & t \in [0, \varphi_{\mathbf{E}S}(\xi)(\gamma)) \\ \gamma + (t - \varphi_{\mathbf{E}S}(\xi)(\gamma))\mathbf{E}S & t \in [\varphi_{\mathbf{E}S}(\xi)(\gamma), \gamma/\mathbf{E}S] \end{cases},$$

then it is straightforward to check that $I_i(\xi) \geq I_i(\varrho_{\mathbf{E}S}(\xi))$ and $\varphi_{\mathbf{E}S}(\xi) = \varphi_{\mathbf{E}S}(\varrho_{\mathbf{E}S}(\xi))$ whenever $\xi \in \mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S]$. Moreover, $\varrho_{\mathbf{E}S}(\mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S]) \subseteq \mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S]$. From these observations, we see that

$$\hat{I}'_i(\zeta) = \inf\{I_i(\xi) : \xi \in \varrho_{\mathbf{E}S}(\mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S]), \zeta = \varphi_{\mathbf{E}S}(\xi)\}. \tag{10}$$

Note that $\xi \in \varrho_{\mathbf{E}S}(\mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S])$ and $\zeta = \varphi_{\mathbf{E}S}(\xi)$ implies that $\zeta \in \check{\mathbb{C}}^{1/\mathbf{E}S}[0,\gamma]$. Therefore, in case $\zeta \notin \check{\mathbb{C}}^{1/\mathbf{E}S}[0,\gamma]$, no $\xi \in \mathbb{D}[0,\gamma/\mathbf{E}S]$ satisfies the two conditions simultaneously, and hence,

$$\hat{I}'_i(\zeta) = \inf \emptyset = \infty = I'_i(\zeta). \tag{11}$$

Now we prove that $\hat{I}'_i(\zeta) = I'_i(\zeta)$ for $\zeta \in \check{\mathbb{C}}^{1/\mathbf{E}S}[0,\gamma]$. We claim that if $\xi \in \varrho_{\mathbf{E}S}(\mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S])$,

$$\tau_s(\varphi_{\mathbf{E}S}(\xi)) = \xi(s) - \xi(s-)$$

for all $s \in [0,\gamma/\mathbf{E}S]$. The proof of this claim will be provided at the end of the proof of the current proposition. Using this claim,

$$\begin{aligned} \hat{I}_i(\zeta) &= \inf\left\{\textstyle\sum_{s\in[0,\gamma/\mathbf{E}S]}(\xi(s)-\xi(s-))^\alpha : \xi \in \varrho_{\mathbf{E}S}(\mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S]), \zeta = \varphi_{\mathbf{E}S}(\xi)\right\} \\ &= \inf\left\{\textstyle\sum_{s\in[0,\gamma/\mathbf{E}S]}\tau_s(\varphi_{\mathbf{E}S}(\xi))^\alpha : \xi \in \varrho_{\mathbf{E}S}(\mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S]), \zeta = \varphi_{\mathbf{E}S}(\xi)\right\} \\ &= \inf\left\{\textstyle\sum_{s\in[0,\gamma/\mathbf{E}S]}\tau_s(\zeta)^\alpha : \xi \in \varrho_{\mathbf{E}S}(\mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S]), \zeta = \varphi_{\mathbf{E}S}(\xi)\right\}. \end{aligned}$$

Note also that $\zeta \in \check{\mathbb{C}}^{1/\mathbf{E}S}[0,\gamma]$ implies the existence of ξ such that $\zeta = \varphi_{\mathbf{E}S}(\xi)$ and $\xi \in \varrho_{\mathbf{E}S}(\mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S])$. To see why, note that there exists $\xi' \in \mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S]$ such that $\zeta = \varphi_{\mathbf{E}S}(\xi')$ due to the definition of $\check{\mathbb{C}}^{1/\mathbf{E}S}[0,\gamma]$. Let $\xi \triangleq \varrho_{\mathbf{E}S}(\xi')$. Then, $\zeta = \varphi_{\mathbf{E}S}(\xi)$ and $\xi \in \varrho_{\mathbf{E}S}(\mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S])$. From this observation, we see that

$$\left\{\textstyle\sum_{s\in[0,\gamma/\mathbf{E}S]}\tau_s(\zeta)^\alpha : \xi \in \varrho_{\mathbf{E}S}(\mathbb{D}^{\mathbf{E}S}[0,\gamma/\mathbf{E}S]), \zeta = \varphi_{\mathbf{E}S}(\xi)\right\} = \left\{\textstyle\sum_{s\in[0,\gamma/\mathbf{E}S]}\tau_s(\zeta)^\alpha\right\},$$

and hence,

$$\hat{I}_i(\zeta) = \sum_{s\in[0,\gamma/\mathbf{E}S]} \tau_s(\zeta)^\alpha = I_i'(\zeta) \tag{12}$$

for $\zeta \in \check{\mathbb{C}}^{1/\mathbf{E}S}[0,\gamma]$. From (11) and (12), we conclude that $I_i' = \hat{I}_i$ for $i = 1,\dots,d$.

Now we are done if we prove the claim. We consider the cases $s > \varphi_{\mathbf{E}S}(\xi)(\gamma)$ and $s \le \varphi_{\mathbf{E}S}(\xi)(\gamma)$ separately. First, suppose that $s > \varphi_{\mathbf{E}S}(\xi)(\gamma)$. Since $\varphi_{\mathbf{E}S}(\xi)$ is non-decreasing, this means that $\varphi_{\mathbf{E}S}(\xi)(t) < s$ for all $t \in [0,\gamma]$, and hence, $\{t \in [0,\gamma] : \varphi_{\mathbf{E}S}(t) = s\} = \emptyset$. Therefore,

$$\tau_s(\varphi_{\mathbf{E}S}(\xi)) = 0 \vee \big(\sup\{t \in [0,\gamma] : \varphi_{\mathbf{E}S}(t) = s\} - \inf\{t \in [0,\gamma] : \varphi_{\mathbf{E}S}(t) = s\}\big) = 0 \vee (-\infty - \infty) = 0.$$

On the other hand, since ξ is continuous on $[\varphi_{\mathbf{E}S}(\xi)(\gamma), \gamma/\mathbf{E}S]$ by its construction,

$$\xi(s) - \xi(s-) = 0.$$

Therefore,

$$\tau_s(\varphi_{\mathbf{E}S}(\xi)) = 0 = \xi(s) - \xi(s-)$$

for $s > \varphi_{\mathbf{E}S}(\xi)(\gamma)$.

Now we turn to the case $s \le \varphi_{\mathbf{E}S}(\xi)(\gamma)$. Since $\varphi_{\mathbf{E}S}(\xi)$ is continuous, this implies that there exists $u \in [0,\gamma]$ such that $\varphi_{\mathbf{E}S}(\xi)(u) = s$. From the definition of $\varphi_{\mathbf{E}S}(\xi)(u)$, it is straightforward to check that

$$u \in [\xi(s-), \xi(s)] \qquad \Longleftrightarrow \qquad s = \varphi_{\mathbf{E}S}(\xi)(u). \tag{13}$$

Note that $[\xi(s-), \xi(s)] \subseteq [0,\gamma]$ for $s \le \varphi_{\mathbf{E}S}(\xi)(\gamma)$ due to the construction of ξ. Therefore, the above equivalence (13) implies that $[\xi(s-), \xi(s)] = \{u \in [0,\gamma] : \varphi_{\mathbf{E}S}(\xi)(u) = s\}$, which in turn implies that $\xi(s-) = \inf\{u \in [0,\gamma] : \varphi_{\mathbf{E}S}(\xi)(u) = s\}$ and $\xi(s) = \sup\{u \in [0,\gamma] : \varphi_{\mathbf{E}S}(\xi)(u) = s\}$. We conclude that

$$\tau_s(\varphi_{\mathbf{E}S}(\xi)) = \xi(s) - \xi(s-)$$

for $s \le \varphi_{\mathbf{E}S}(\xi)(\gamma)$.

Now we are ready to characterize an asymptotic evaluation for $\mathbf{P}\big(Q(\gamma n) \ge n\big)$. Recall that $\tau_s(\xi) \triangleq \max\Big\{0, \sup\{t \in [0,\gamma] : \xi(t) = s\} - \inf\{t \in [0,\gamma] : \xi(t) = s\}\Big\}$.

Proposition 5

$$\lim_{n\to\infty} \frac{1}{L(n)n^\alpha} \log \mathbf{P}\big(Q(\gamma n) \ge n\big) = -c^*$$

where c^* is the solution of the following quasi-variational problem:

$$\begin{aligned} &\inf_{\xi_1,\dots,\xi_d} \quad \sum_{i=1}^{d} \sum_{s\in[0,\gamma/\mathbf{E}S]} \tau_s(\xi_i)^\alpha \\ &\textit{subject to} \quad \sup_{0\le s\le\gamma} \Big(\frac{s}{\mathbf{E}A} - \sum_{i=1}^{d} \xi_i(s)\Big) \ge 1; \\ &\qquad\qquad \xi_i \in \check{\mathbb{C}}^{1/\mathbf{E}S}[0,\gamma] \quad \textit{for } i = 1,\dots,d. \end{aligned} \tag{14}$$

Proof Note first that for any $\epsilon > 0$,

$$\begin{aligned}
&\mathbf{P}\left(\sup_{0\le s\le\gamma}\left(\bar{M}_n(s)-\sum_{i=1}^{d}\bar{N}_n^{(i)}(s)\right)\ge 1\right)\\
&=\mathbf{P}\left(\sup_{0\le s\le\gamma}\left(\bar{M}_n(s)-\frac{s}{\mathbf{E}A}\right)+\sup_{0\le s\le\gamma}\left(\frac{s}{\mathbf{E}A}-\sum_{i=1}^{d}\bar{N}_n^{(i)}(s)\right)\ge 1\right)\\
&\le\mathbf{P}\left(\sup_{0\le s\le\gamma}\left(\bar{M}_n(s)-\frac{s}{\mathbf{E}A}\right)\ge\epsilon\right)+\mathbf{P}\left(\sup_{0\le s\le\gamma}\left(\frac{s}{\mathbf{E}A}-\sum_{i=1}^{d}\bar{N}_n^{(i)}(s)\right)\ge 1-\epsilon\right).
\end{aligned}$$

Since $\limsup_{n\to\infty}\frac{1}{L(n)n^\alpha}\log\mathbf{P}\left(\sup_{0\le s\le\gamma}\left(\bar{M}_n(s)-\frac{s}{\mathbf{E}A}\right)\ge\epsilon\right)=-\infty$,

$$\limsup_{n\to\infty}\frac{\mathbf{P}\left(\sup_{0\le s\le\gamma}\left(\bar{M}_n(s)-\sum_{i=1}^{d}\bar{N}_n^{(i)}(s)\right)\ge 1\right)}{L(n)n^\alpha}=\limsup_{n\to\infty}\frac{\mathbf{P}\left(\sup_{0\le s\le\gamma}\left(\frac{s}{\mathbf{E}A}-\sum_{i=1}^{d}\bar{N}_n^{(i)}(s)\right)\ge 1-\epsilon\right)}{L(n)n^\alpha}.$$

To bound the right hand side, we proceed to deriving an LDP for $\sup_{0\le s\le\gamma}\left(\frac{s}{\mathbf{E}A}-\sum_{i=1}^{d}\bar{N}_n^{(i)}(s)\right)$. Due to Proposition 4, $(\bar{N}_n^{(i)},\dots,\bar{N}_n^{(d)})$ satisfy the LDP in $\prod_{i=1}^{d}\mathbb{D}[0,\gamma]$ (w.r.t. the d-fold product topology of $\mathcal{T}_{M_1'}$) with speed $L(n)n^\alpha$ and rate function

$$I'(\xi_1,\dots,\xi_d)\triangleq\sum_{i=1}^{d}I_i'(\xi_i).$$

Let $\mathbb{D}^\uparrow[0,\gamma]$ denote the subspace of $\mathbb{D}[0,\gamma]$ consisting of non-decreasing functions. Since $\bar{N}_n^{(i)}\in\mathbb{D}^\uparrow[0,\gamma]$ with probability 1 for each $i=1,\dots,d$, we can apply Lemma 4.1.5 (b) of [9] to deduce the same LDP for $(\bar{N}_n^{(i)},\dots,\bar{N}_n^{(d)})$ in $\prod_{i=1}^{d}\mathbb{D}^\uparrow[0,\gamma]$. If we define $f:\prod_{i=1}^{d}\mathbb{D}^\uparrow[0,\gamma]\to\mathbb{D}[0,\gamma]$ as

$$f(\bar{N}_n^{(1)},\dots,\bar{N}_n^{(d)})\triangleq\xi_{\mathbf{E}A}-\sum_{i=1}^{d}\bar{N}_n^{(i)}$$

where $\xi_{\mathbf{E}A}(t)\triangleq t/\mathbf{E}A$, then f is continuous since all the jumps are in one direction, and hence, we can apply the contraction principle to deduce the LDP for $\xi_{\mathbf{E}A}-\sum_{i=1}^{d}\bar{N}_n^{(i)}$, which is controlled by the rate function

$$I''(\zeta)\triangleq\inf_{\{(\xi_1,\dots,\xi_d):\,\xi_{\mathbf{E}A}-\sum_{i=1}^{d}\xi_i=\zeta\}}I'(\xi_0,\dots,\xi_d).$$

Now, applying the contraction principle again with the supremum functional to $\xi_{\mathbf{E}A}-\sum_{i=1}^{d}\bar{N}_n^{(i)}$, we get the LDP for $\sup_{t\in[0,\gamma]}\left(\xi_{\mathbf{E}A}-\sum_{i=1}^{d}\bar{N}_n^{(i)}(t)\right)$, which is controlled by the rate function

$$I'''(x)=\inf_{\{\zeta:\sup_{t\in[0,\gamma]}\zeta(t)=x\}}I''(\zeta)=\inf_{\left\{(\xi_1,\dots,\xi_d):\sup_{t\in[0,\gamma]}\left(\xi_{\mathbf{E}A}(t)-\sum_{i=1}^{d}\xi_i(t)\right)=x\right\}}I'(\xi_0,\dots,\xi_d).$$

The conclusion of the proposition follows from considering the upper bound of this LDP for the closed set $[1,\infty)$ and taking $\epsilon\to 0$.

We conclude with the proof for the matching lower bound in case $\gamma>1/\lambda$. Considering the obvious coupling between Q and $(M,N^{(1)},\cdots,N^{(d)})$, one can see that $M(s)-\sum_{i=1}^{d}N^{(i)}(s)$ can be interpreted as (a lower bound of) the length of an imaginary queue at time s where the servers can start working on the jobs that have not arrived yet. Therefore, $\mathbf{P}(Q((a+s)n)>n)\ge\mathbf{P}(Q((a+s)n)>n|Q(a)=0)\ge$

$\mathbf{P}(\bar{M}_n(s) - \sum_{i=1}^d \bar{N}_n^{(i)}(s) > 1)$ for any $a \geq 0$. Let s^* be the level crossing time of the optimal solution of (14). Then, for any $\epsilon > 0$,

$$\begin{aligned}
\mathbf{P}(Q(\gamma n) > n) &\geq \mathbf{P}\Big(\bar{M}_n(s^*) - \sum_{i=1}^d \bar{N}_n^{(i)}(s^*) > 1\Big) \\
&\geq \mathbf{P}\Big(\bar{M}_n(s^*) - s^*/\mathbf{E}A > -\epsilon \;\text{ and }\; s^*/\mathbf{E}A - \sum_{i=1}^d \bar{N}_n^{(i)}(s^*) > 1 + \epsilon\Big) \\
&\geq \mathbf{P}\Big(s^*/\mathbf{E}A - \sum_{i=1}^d \bar{N}_n^{(i)}(s^*) > 1 + \epsilon\Big) - \mathbf{P}\Big(\bar{M}_n(s^*) - s^*/\mathbf{E}A \leq -\epsilon\Big)
\end{aligned}$$

Since $\mathbf{P}(\bar{M}_n(s^*) - s^*/\mathbf{E}A \leq -\epsilon)$ decays exponentially fast w.r.t. n,

$$\liminf_{n\to\infty} \frac{\log \mathbf{P}(Q(\gamma n) > n)}{L(n)n^\alpha} \geq \liminf_{n\to\infty} \frac{\log \mathbf{P}(s^*/\mathbf{E}A - \sum_{i=1}^d \bar{N}_n^{(i)}(s^*) > 1 + \epsilon)}{L(n)n^\alpha} \geq - \inf_{(\xi_1,\ldots,\xi_d)\in A^\circ} I'(\xi_1, \ldots, \xi_d)$$

where $A = \{(\xi_1, \ldots, \xi_d) : s^*/\mathbf{E}A - \sum_{i=1}^d \xi_i(s^*) > 1 + \epsilon\}$. Note that the optimizer $(\xi_1^*, \ldots, \xi_d^*)$ of (14) satisfies $s^*/\mathbf{E}A - \sum_{i=1}^d \xi_i^*(s^*) \geq 1$. Consider $(\xi_1', \ldots, \xi_d')$ obtained by increasing one of the job size of $(\xi_1^*, \ldots, \xi_d^*)$ by $\delta > 0$. One can always find a small enough such δ since $\gamma > 1/\lambda$. Note that there exists $\epsilon > 0$ such that $s'/\mathbf{E}A - \sum_{i=1}^d \xi'(s') > 1 + \epsilon$. Therefore,

$$\liminf_{n\to\infty} \frac{\log \mathbf{P}(Q(\gamma n) > n)}{L(n)n^\alpha} \geq -I'(\xi_1', \ldots, \xi_d') \geq -c^* - \delta^\alpha$$

where the second inequality is from the subadditivity of $x \mapsto x^\alpha$. Since δ can be chosen arbitrarily small, letting $\delta \to 0$, we arrive at the matching lower bound.

To obtain more insight in the way the rare event $\{Q(\gamma n) \geq n\}$ occurs, we now simplify the expression of c^* given in Proposition 5. To ease notation, we assume from now on that $\mathbf{E}[S] = \mu^{-1} = 1$.

Proposition 6 *If $\gamma < 1/\lambda$, $c^* = \infty$. If $\gamma \geq 1/\lambda$, c^* can be computed via*

$$\begin{aligned}
&\min \sum_{i=1}^d x_i^\alpha \quad s.t. \\
&l(s; x) = \lambda s - \sum_{i=1}^d (s - x_i)^+ \geq 1 \text{ for some } s \in [0, \gamma]. \\
&x_1, ..., x_d \geq 0\ ,
\end{aligned} \tag{15}$$

which in turn equals

$$\begin{aligned}
\min \Bigg\{ &\inf_{0<k\leq\lfloor\lambda\rfloor : \gamma < 1/(\lambda-k)} \Big\{(d-k)\gamma^\alpha + (1 - \gamma\lambda + \gamma k)^\alpha (k - \lfloor\lambda\rfloor \wedge \lfloor\lambda - 1/\gamma\rfloor)^{1-\alpha}\Big\}, \\
&\min_{l=0}^{\lfloor\lambda\rfloor\wedge\lfloor\lambda-1/\gamma\rfloor} \left\{(d-l)\left(\frac{1}{\lambda - l}\right)^\alpha\right\}\Bigg\}.
\end{aligned} \tag{16}$$

Proof We want to show that c^* is equal to

$$\begin{aligned} &\inf \sum_{i=1}^{d} \sum_{s\in[0,\gamma]} \tau_s \left(\zeta_i\right)^{\alpha} \\ &\text{s.t.} \\ &\sup_{0\leq s\leq \gamma} \left\{ \lambda s - \sum_{i=1}^{d} \zeta_i \left(s\right) \right\} \geq 1 \\ &\zeta_i = \varphi_\mu \left(\xi_i\right), \, \xi_i \in \mathbb{D}^{\mu}[0,\gamma/\mu] \text{ for } i \in 1,...,d \, . \end{aligned} \tag{17}$$

After a simple transformation, we might assume that $\mu = 1$. For simplicity in the exposition we will assume the existence of an optimizer. The argument that we present can be carried out with $\bar{\varepsilon}$-optimizers. In the end, the representation that we will provide will show the existence of an optimizer. First, we will argue that without loss of generality we may assume that if $(\zeta_1, ..., \zeta_d)$ is an optimal solution then the corresponding functions $\xi_1, ..., \xi_d$ have at most one jump which occurs at time zero. To see this suppose that $(\zeta_1, ..., \zeta_d)$ is an optimal solution and consider the corresponding functions $(\xi_1, ..., \xi_d)$ such that $\zeta_i = \varphi_\mu \left(\xi_i\right)$. By feasibility, we must have that at least one of the ξ_i's exhibit at least one jump in $[0, \gamma]$. Assume that ξ_i exhibits two or more jumps and select two jump times, say $0 \leq u_0 < u_1 \leq \gamma$, with corresponding jump sizes x_0 and x_1, respectively. Let

$$\bar{\xi}_i \left(\cdot\right) = \xi_i \left(\cdot\right) - x_1 \mathbb{I}_{[u_1,\gamma]} \left(\cdot\right) + x_1 \mathbb{I}_{[u_0,\gamma]} \left(\cdot\right);$$

in simple words, $\bar{\xi}_i \left(\cdot\right)$ is obtained by merging the jump at time u_1 with the jump at time u_0. It is immediate (since $x_0, x_1 > 0$) that for each t

$$\bar{\xi}_i \left(t\right) \geq \xi_i \left(t\right)$$

and, therefore, letting $\bar{\zeta}_i = \varphi_\mu \left(\bar{\xi}_i\right)$ we obtain (directly from the definition of the functional $\bar{\zeta}_i$ as a generalized inverse) that for every s

$$\bar{\zeta}_i \left(s\right) \leq \zeta_i \left(s\right).$$

Therefore, we conclude that the collection $\zeta_1, ..., \bar{\zeta}_i, ..., \zeta_d$ is feasible. Moreover, since

$$\sum_{s\in[0,\gamma]} \tau_s \left(\bar{\zeta}_i\right)^{\alpha} = \sum_{s\in[0,\gamma]} \tau_s \left(\zeta_i\right)^{\alpha} + \left(x_0 + x_1\right)^{\alpha} - x_0^{\alpha} - x_1^{\alpha}$$

and, by strict concavity,

$$\left(x_0 + x_1\right)^{\alpha} < x_0^{\alpha} + x_1^{\alpha},$$

we conclude that $\zeta_1, ..., \bar{\zeta}_i, ..., \zeta_d$ improves the objective function, thus violating the optimality of $\zeta_1, ..., \zeta_d$. So, we may assume that $\xi_i \left(\cdot\right)$ has a single jump of size $x_i > 0$ at some time u_i and therefore

$$\zeta_i \left(s\right) = \min \left(s, u_i\right) + \left(s - x_i - u_i\right)^{+}. \tag{18}$$

Now, define $t = \inf\{s \in [0, \gamma] : \lambda s - \sum_{i=1}^{d} \zeta_i \left(s\right) \geq 1\}$, then

$$\lambda t - \sum_{i=1}^{d} \zeta_i \left(t\right) = 1 \tag{19}$$

and we must have that $t \geq x_i + u_i$; otherwise, if $x_i + u_i > t$ then we might reduce the value of the objective function while preserving feasibility (this can be seen from the form of $\zeta_i \left(\cdot\right)$), thus contradicting optimality. Now, suppose that $u_i > 0$, choose $\varepsilon \in (0, \min(u_i, x_i))$ and define

$$\bar{\xi}_i \left(s\right) = \xi_i \left(s\right) - x_i \mathbb{I}_{[u_i,\gamma]} \left(s\right) + x_i \mathbb{I}_{[u_i-\varepsilon,\gamma]} \left(s\right).$$

In simple words, we just moved the first jump slightly earlier (by an amount ε). Once again, let $\bar{\zeta}_i = \varphi_\mu\left(\bar{\xi}_i\right)$, and we have that

$$\bar{\zeta}_i(s) = \min(s, u_i - \varepsilon) + (s - x_i - u_i + \varepsilon)^+ \leq \zeta_i(s).$$

Therefore, we preserve feasibility without altering the objective function. As a consequence, we may assume that $u_i = 0$ and using expression (18) we then obtain that (14) takes the form

$$\begin{aligned} &\min \sum_{i=1}^{d} x_i^\alpha \quad \text{s.t.} \qquad (20)\\ &l(s;x) = \lambda s - \sum_{i=1}^{d} (s - x_i)^+ \geq 1 \text{ for some } s \in [0, \gamma],\\ &x_1, ..., x_d \geq 0\ . \end{aligned}$$

Let $x = (x_1, ..., x_d)$ be any optimal solution, we may assume without loss of generality that $0 \leq x_1 \leq ... \leq x_d$. We claim that x satisfies the following features. First, $x_d \leq \gamma$, this is immediate from the fact that we are minimizing over the x_i's and if $x_d > \gamma$ we can reduce the value of x_d without affecting the feasibility of x, thereby improving the value of (20). The same reasoning allows us to conclude that $\inf\{s : l(s;x) \geq 1\} = x_d$. Consequently, letting $x_i = a_1 + ... + a_i$, (20) is equivalent to

$$\begin{aligned} &\min \sum_{i=1}^{m} \left(\sum_{j=1}^{i} a_j \right)^\alpha \quad \text{s.t.}\\ &\lambda(a_1 + ... + a_d) - \sum_{i=1}^{d} (a_1 + ... + a_d - \sum_{j=1}^{i} a_1) \geq 1\\ &a_1 + ... + a_d \leq \gamma\ ,\ a_1, ..., a_d \geq 0. \end{aligned}$$

This problem can be simplified to

$$\begin{aligned} &\min \sum_{i=1}^{m} \left(\sum_{j=1}^{i} a_j \right)^\alpha \quad \text{s.t.}\\ &\lambda a_1 + (\lambda - 1) a_2 + ... + (\lambda - d + 1) a_d \geq 1\\ &a_1 + ... + a_d \leq \gamma\ ,\ a_1, ..., a_d \geq 0. \end{aligned}$$

In turn, we know that $0 < \lambda < d$, then it suffices to consider

$$\begin{aligned} &\min \sum_{i=1}^{\lfloor \lambda \rfloor} \left(\sum_{j=1}^{i} a_j \right)^\alpha + (d - \lfloor \lambda \rfloor) \left(\sum_{j=1}^{\lfloor \lambda \rfloor + 1} a_j \right)^\alpha \quad \text{s.t.} \qquad (21)\\ &\lambda a_1 + (\lambda - 1) a_2 + ... + (\lambda - \lfloor \lambda \rfloor) a_{\lfloor \lambda \rfloor + 1} = 1 \qquad (22)\\ &a_1 + ... + a_{\lfloor \lambda \rfloor + 1} \leq \gamma\ , \qquad (23)\\ &a_1, ..., a_{\lfloor \lambda \rfloor + 1} \geq 0\ , \qquad (24) \end{aligned}$$

because $(\lambda - m) < 0$ implies $a_{\lambda - m + 1} = 0$ (otherwise we can reduce the value of the objective function).

We first consider the case $\lambda > \lfloor \lambda \rfloor$. Moreover, observe that if $\gamma \geq 1/(\lambda - \lfloor \lambda \rfloor))$ then any solution satisfying (22) and (24) automatically satisfies (23), so we can ignore the constraint (23) if assume that

$\gamma \geq 1/(\lambda - \lfloor\lambda\rfloor)$. If λ is an integer we will simply conclude that $a_{\lfloor\lambda\rfloor+1} = 0$ and if we only assume $\gamma > 1/\lambda$ we will need to evaluate certain extreme points, as we shall explain later.

Now, the objective function is clearly concave and lower bounded inside the feasible region, which in turn is a compact polyhedron. Therefore, the optimizer is achieved at some extreme point in the feasible region (see [10]). Under our simplifying assumptions, we only need to characterize the extreme points of (22), (24), which are given by $a_i = 1/(\lambda - i + 1)$ for $i = 1, ..., \lfloor\lambda\rfloor + 1$.

So, the solution, assuming that $\gamma \geq 1/(\lambda - \lfloor\lambda\rfloor)$, is given by

$$\begin{aligned} &\min\{da_1^\alpha, (d-1)\, a_2^\alpha, ..., (d - \lfloor\lambda\rfloor)\, a_{\lfloor\lambda\rfloor+1}^\alpha\} \\ &= \min_{i=1}^{\lfloor\lambda\rfloor+1} \left\{ (d-i+1) \left(\frac{1}{\lambda - i + 1} \right)^\alpha \right\}. \end{aligned}$$

In the general case, that is, assuming $\gamma > \lambda^{-1}$ and also allowing the possibility that $\lambda = \lfloor\lambda\rfloor$, our goal is to show that the additional extreme points which arise by considering the inclusion of (23) might potentially give rise to solutions in which large service requirements are not equal across all the servers. We wish to identify the extreme points of (22), (23), (24) which we represent as

$$\begin{aligned} &\lambda a_1 + (\lambda - 1)\, a_2 + ... + (\lambda - \lfloor\lambda\rfloor)\, a_{\lfloor\lambda\rfloor+1} = 1\ , \\ &a_0 + a_1 + ... + a_{\lfloor\lambda\rfloor+1} = \gamma\ , \\ &a_0, a_1, ..., a_{\lfloor\lambda\rfloor+1} \geq 0\ . \end{aligned}$$

Note the introduction of the slack variable $a_0 \geq 0$. From elementary results in polyhedral combinatorics, we know that extreme points correspond to basic feasible solutions. Choosing $a_{i+1} = 1/(\lambda - i)$ and $a_0 = \gamma - a_{i+1}$ recover basic solutions which correspond to the extreme points identified earlier, when we ignored (23). If $\lambda = \lfloor\lambda\rfloor$ we must have, as indicated earlier, that $a_{\lfloor\lambda\rfloor+1} = 0$; so we can safely assume that $\lambda - i > 0$. We observe that $\gamma \geq 1/(\lambda - i)$ implies that $a_{i+1} = 1/(\lambda - i)$ and $a_j = 0$ for $j \neq i+1$ is a basic feasible solution for the full system (i.e. including (23)). Additional basic solutions (not necessarily feasible) are obtained by solving (assuming that $0 \leq l < k < \lambda$)

$$\begin{aligned} 1 &= (\lambda - k)\, a_{k+1} + (\lambda - l)\, a_{l+1}, \\ \gamma &= a_{k+1} + a_{l+1}. \end{aligned}$$

This system of equations always has a unique solution because the equations are linearly independent if $l \neq k$. The previous pair of equations imply that

$$\lambda\gamma - 1 = k a_{k+1} + l a_{l+1}.$$

Therefore, we obtain the solution $(\bar{a}_{k+1}, \bar{a}_{l+1})$ is given by

$$\begin{aligned} (k-l)\, \bar{a}_{k+1} &= (\lambda - l)\, \gamma - 1, \\ (k-l)\, \bar{a}_{l+1} &= 1 - \gamma\, (\lambda - k)\, . \end{aligned}$$

So, for the solution to be both basic and feasible we must have that $1/(\lambda - l) \leq \gamma \leq 1/(\lambda - k)$ (with strict inequality holding on one side).

If we evaluate the solution $a_{k+1} = \bar{a}_{k+1}$, $a_{l+1} = \bar{a}_{l+1}$, $a_{i+1} = 0$ for $i \notin \{k, l\}$ in the objective function we obtain

$$\begin{aligned} &\sum_{i=1}^{\lfloor\lambda\rfloor} \left(\sum_{j=1}^{i} a_j \right)^\alpha + (d - \lfloor\lambda\rfloor) \left(\sum_{j=1}^{\lfloor\lambda\rfloor+1} a_j \right)^\alpha \\ &= \bar{a}_{l+1}^\alpha\, (k-l) + (\lfloor\lambda\rfloor - k)\, (\bar{a}_{k+1} + \bar{a}_{l+1})^\alpha + (d - \lfloor\lambda\rfloor)\, (\bar{a}_{k+1} + \bar{a}_{l+1})^\alpha \\ &= \bar{a}_{l+1}^\alpha\, (k-l) + (d-k)\, (\bar{a}_{k+1} + \bar{a}_{l+1})^\alpha\, . \end{aligned}$$

Note that in the case $\gamma = 1/(\lambda - k)$ we have that $a_{k+1} = 1/(\lambda - k)$ and $a_i = 0$ for $i \neq k+1$ is a feasible extreme point with better performance than the solution involving $\bar{a}_{k+1}$ and $\bar{a}_{l+1}$,

$$\bar{a}_{l+1}^{\alpha}(k-l) + (d-k)(\bar{a}_{k+1} + \bar{a}_{l+1})^{\alpha} > (d-k)\, a_{k+1}^{\alpha}.$$

Consequently, we may consider only cases $1/(\lambda - l) \leq \gamma < 1/(\lambda - k)$ and we conclude that the general solution is given by

$$\min\left\{ \min_{0<k\leq \lfloor\lambda\rfloor : \gamma < 1/(\lambda-k)} \left\{ (d-k)\gamma^{\alpha} + (1-\gamma(\lambda-k))^{\alpha} \min_{0\leq l < \lfloor\lambda\rfloor : 1/(\lambda-l)\leq\gamma} \left(\frac{1}{k-l}\right)^{\alpha}(k-l) \right\}, \right.$$
$$\left. \min_{l=0}^{\lfloor\lambda\rfloor \wedge \lfloor\lambda - 1/\gamma\rfloor} \left\{ (d-l)\left(\frac{1}{\lambda-l}\right)^{\alpha} \right\} \right\}.$$

Simplifying, we obtain (16).

References

1. Sergey Foss and Dmitry Korshunov. Heavy tails in multi-server queue. *Queueing Systems: Theory and Applications*, 52(1):31–48, January 2006.
2. Sergey Foss and Dmitry Korshunov. On large delays in multi-server queues with heavy tails. *Mathematics of Operations Research*, 37(2):201–218, 2012.
3. Dongdong Ge, Xiaoye Jiang, and Yinyu Ye. A note on the complexity of L_p minimization. *Math. Program.*, 129(2, Ser. B):285–299, 2011.
4. Ward Whitt. The impact of a heavy-tailed service-time distribution upon the $M/GI/s$ waiting-time distribution. *Queueing Systems Theory Appl.*, 36(1-3):71–87, 2000.
5. David Gamarnik and David A Goldberg. Steady-state $gi/g/n$ queue in the halfin–whitt regime. *The Annals of Applied Probability*, 23(6):2382–2419, 2013.
6. Anatolii A Puhalskii and Ward Whitt. Functional large deviation principles for first-passage-time processes. *The Annals of Applied Probability*, pages 362–381, 1997.
7. M. Bazhba, J. Blanchet, C.-H. Rhee, and B. Zwart. Sample-path large deviations for L\'evy processes and random walks with Weibull increments. *ArXiv e-prints*, October 2017.
8. Anatolii Puhalskii. Large deviation analysis of the single server queue. *Queueing Systems*, 21(1):5–66, 1995.
9. Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*. Springer-Verlag, Berlin, 2010.
10. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970.