J. Appl. Prob. **54,** 921–942 (2017)
*doi:*10.1017/jpr.2017.42
© *Applied Probability Trust* 2017

FINITE-POOL QUEUEING WITH HEAVY-TAILED SERVICES

GIANMARCO BET,* **
REMCO VAN DER HOFSTAD * AND
JOHAN S. H. VAN LEEUWAARDEN,* *Eindhoven University of Technology*

Abstract

We consider the $\Delta_{(i)}/\mathrm{G}/1$ queue, in which a total of n customers join a single-server queue for service. Customers join the queue independently after exponential times. We consider *heavy-tailed* service-time distributions with tails decaying as $x^{-\alpha}$, $\alpha \in (1, 2)$. We consider the asymptotic regime in which the population size grows to ∞ and establish that the scaled queue-length process converges to an α-stable process with a negative quadratic drift. We leverage this asymptotic result to characterize the head start that is needed to create a long period of uninterrupted activity (a busy period). The heavy-tailed service times should be contrasted with the case of light-tailed service times, for which a similar scaling limit arises (Bet *et al.* (2015)), but then with a Brownian motion instead of an α-stable process.

Keywords: Heavy-traffic approximation; heavy-tailed distribution; functional central limit theorem; Skorokhod reflection map

2010 Mathematics Subject Classification: Primary 60K25
Secondary 90B22; 68M20

1. Introduction

In this paper we are concerned with the $\Delta_{(i)}/\mathrm{G}/1$ queue, designed to model a service system to which only a finite pool of n customers can arrive. The n potential customers in the pool each have an independent and identically distributed (i.i.d.) exponential clock and join the queue when their clock rings. Each customer joins the queue only once, and at the system level, this creates an arrival process governed by the order statistics of the clock times. The $\Delta_{(i)}/\mathrm{G}/1$ queue, in the precise form as studied in this paper, was introduced in [15] and [16], and belongs to the branch of queueing theory that deals with time-dependent or transient conditions [24]–[27], [39], [40]. Indeed, with time, the pool of potential customers (those who have not joined the queue yet) becomes smaller, and the influx of customers loses intensity.

The $\Delta_{(i)}/\mathrm{G}/1$ queue can be studied in many operating regimes; see, e.g. [16] and [23] where the focus was on overloaded regimes and [4] for a detailed overview. In [4] we introduced a new heavy-traffic regime defined by two features: the customer pool grows to ∞ and the initial (at time zero) rate of newly arriving customers is such that, on average, one new customer is expected to arrive during one service time. This gives rise to a large-scale system that (initially) operates close to full utilization, and is expected to utilize its resources efficiently. By this we mean that the server is typically busy, and that idle times are negligible. In fact, this defines our

Received 26 May 2016; revision received 14 April 2017.

* Postal address: Department of Mathematics and Computer Science, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands.
** Email address: g.bet@tue.nl

https://doi.org/10.1017/jpr.2017.42 Published online by Cambridge University Press

main goal in this paper: to characterize the conditions under which sufficiently many customers will join the queue to guarantee that the system will have a substantial backlog of customers. We focus on the first busy period, and show how to set the initial number of customers already present in the queue at time $t = 0$, referred to as the *head start*, to create a considerable first busy period during which the server can work continuously.

In [4] we have studied this heavy-traffic regime under the assumption that the variance of the service-time distribution is finite, and have shown that the queue-length process converges to a reflected Brownian motion with negative quadratic drift. The negative drift captures the effect of a pool of potential customers that diminishes with time: after some time, the activity in the queue inevitably becomes negligible. However, in the early phases, the rate of arriving customers can be high. Say you want to start a business and you estimate that a population of n persons might become customers. Then the head start can be interpreted as the persons that signed up (already) as a customer. In [4] we have shown that in our heavy-traffic regime, once the head start is of order $n^{1/3}$, and you would decide to start your business, the number of customers you will serve consecutively is of the order $n^{2/3}$. With this mental picture, our heavy-traffic regime gives insight into dimensioning rules about how large the pool n should be in comparison to the head start, how to choose the service capacity as a function of the pool size to achieve full system utilization, and how to control the first busy period, which essentially is the relevant time of operation of the system.

In the present paper we drop the finite-variance condition and study the queue-length process under the additional assumption that the service times are *heavy tailed*. More precisely, we assume that the service times follow a power-law distribution with power-law exponent $\alpha \in (1, 2)$. Under these assumptions, our model is the finite-pool analogue of the classical heavy-tailed M/G/1 queue; see below for a discussion. We establish that, in a similar heavy-traffic regime as in [4], the rescaled queue-length process converges to an α-stable process with negative quadratic drift. As in the finite-variance case, the diminishing pool effect is still there in the form of the drift term, but the oscillations of the limiting queue length are much wilder. We will also show that, as a consequence of the larger fluctuations, the desired head start and canonical busy period should scale with n, in a specific way that vitally depends on the exponent α and other more refined properties of the service-time distribution.

2. Description of the model

We now describe the $\Delta_{(i)}/\mathrm{G}/1$ model with exponential arrivals in detail. We consider a population of n potential customers that are to be served by a single server. Each customer $i \in \{1, \dots, n\} =: [n]$ is assigned a random variable T_i, representing its arrival time. We assume $(T_i)_{i=1}^n$ to be a sequence of i.i.d. exponential random variables with mean $1/\lambda$. We denote the distribution function of T_1 by $F_T(\cdot)$. When the clock T_i rings, customer i joins the queue and customers in the queue are served in a first-come–first-served manner. The arrival times are then given by the order statistics of $(T_i)_{i=1}^n$. The service requirements of customers are given by a sequence $(\bar{S}_i)_{i=1}^n$ of i.i.d. random variables. The server works with speed c_n, so that the service time of customer i is $S_i := \bar{S}_i/c_n$. We let the server speed depend on the number of customers in the population, and study how the performance of the system is affected by different choices of c_n. We denote the distribution function of $\bar{S}_1$ by $F_{\bar{S}}(\cdot)$. We say a function $\ell(\cdot)$ is *slowly varying* when $\lim_{t\to\infty} \ell(tc)/\ell(t) = 1$ for all $c > 0$. The service-time distribution is assumed to be in the domain of attraction of an α-stable law i.e. its tail decays as

$$\mathbb{P}(\bar{S} > t) = 1 - F_{\bar{S}}(t) = t^{-\alpha}\ell(t), \qquad \alpha \in (1, 2), \tag{1}$$

https://doi.org/10.1017/jpr.2017.42 Published online by Cambridge University Press

for a slowly varying function $\ell(\cdot)$. Assumption (1) implies, in particular, that $\mathbb{E}[\bar{S}^k] = \infty$ for $k > \alpha$, and $\mathbb{E}[\bar{S}^k] < \infty$ for $k < \alpha$. We further assume that the queue obeys the heavy-traffic condition

$$\max_{t\geq 0} f_T(t)\mathbb{E}[\bar{S}] = \max_{t\geq 0} f_T(t)\mathbb{E}[S]c_n = 1, \tag{2}$$

where $f_T(\cdot)$ is the density of the arrival time distribution. For exponential arrival times with rate λ, condition (2) simplifies to

$$\lambda\mathbb{E}[\bar{S}] = \lambda\mathbb{E}[S]c_n = 1. \tag{3}$$

We will assume that the speed c_n can be expressed as $c_n = 1+\varepsilon_n$, with $\varepsilon_n \ll 1$. Assumption (3) can then be interpreted as a 'critical window': the system is near the critical point $\lambda\mathbb{E}[S] = 1$ and the distance from criticality is tuned by a parameter β such that $\varepsilon_n = \beta n^{-\eta}$, for some $\eta > 0$ to be identified later, so that the queue remains asymptotically critical.

For $n \to \infty$, this gives the critical point $\lambda\mathbb{E}[S] = 1$, which can be understood as follows. Since λ represents the instantaneous (close to time $t = 0$) arrival rate of customers, (3) amounts to assuming that on average, during one service time, one customer joins the queue. We study the queue after $N_n(0)$ customers have already joined, where the head start $N_n(0)$ may depend on n and $N_n(0) \to \infty$. Since in our setting $N_n(0) \ll n$, without loss of generality we can assume there are (still) n customers in the pool. Before stating our main results, let us introduce some notation.

Denote the number of customers who arrive in the interval $[0, t]$ by $A(t) = \sum_{i=1}^{n} \mathbf{1}_{\{T_i\leq t\}}$. Let

$$\sigma(t) = \max\left\{k \geq 0 \;\middle|\; \sum_{i=1}^{k} S_i \leq t\right\} \tag{4}$$

be the renewal process associated with the service times and define the net input process as

$$X(t) = \sum_{i=1}^{A(t)} S_i - t. \tag{5}$$

The process $X_n(\cdot)$ is useful in defining the cumulative busy-time process as

$$B(t) = t - I(t) = t - \inf_{0\leq s\leq t} (X(s)^-),$$

where $f(x)^- = \min\{0, f(x)\}$ (respectively, $f(x)^+ = \max\{0, f(x)\}$), and $I_n(\cdot)$ is the cumulative idle time.

Let $\mathcal{D} := \mathcal{D}([0, \infty))$ denote the space of càdlàg functions that are continuous from the right and admit a limit from the left at every point. All the functions that we consider are elements of $\mathcal{D}$. Let $\phi(\cdot)\colon \mathcal{D} \mapsto \mathcal{D}$ be the *reflection mapping*, defined as

$$\phi(f)(x) = f(x) + \psi(f)(x), \tag{6}$$

where $\psi(\cdot)\colon \mathcal{D} \mapsto \mathcal{D}$ is given by

$$\psi(f)(x) = -\inf_{0\leq y\leq x} (f(y)^-). \tag{7}$$

The queue-length process $Q(t)$ is given by

$$Q(t) = N_n(0) + A(t) - \sigma(B(t)),$$

where $N_n(0)$ denotes the number of customers already in the queue at the beginning of the first service.

https://doi.org/10.1017/jpr.2017.42 Published online by Cambridge University Press

For our limit result, we rescale the arrival process as $A_n(t) := A(t/n)$ so that we expect order 1 customers to arrive in a time window of length 1. Accordingly, consider the process

$$Q_n(t) = N_n(0) + A_n(t) - \sigma(B_n(t)), \tag{8}$$

where

$$B_n(t) = t - \inf_{0\le s\le t}\Bigg(\sum_{i=1}^{A_n(t)} S_i - t\Bigg)^{-}.$$

The time change $t \mapsto B_n(t)$ depends both on $(T_i)_{i=1}^{\infty}$ and $(S_i)_{i=1}^{\infty}$ and as such makes the analysis of $Q_n(t)$ challenging. An approach pioneered by Iglehart and Whitt [17] consists in studying a related queue in which the server never idles, but rather continues working according to the renewal process associated with $(S_i)_{i=1}^{\infty}$ even when the queue is empty. This is often referred to as the queue with autonomous service or the Borovkov modified system; see [38, Chapter 10.2]. It turns out that, under mild assumptions, the original queue and the Borovkov modified system are asymptotically equivalent in heavy traffic [38, Theorem 10.2.2], in the sense that the distance between the two queue-length processes converges to 0. However, for this approach to work, the service-time limit process needs to be continuous. Indeed, the distance between the two processes is bounded from above by the (scaled) maximum service time, or, equivalently, the maximum jump functional applied to the service-time process. If the service-time limit process is continuous then the maximum jump functional converges to 0. If, on the other hand, the service-time limit process is discontinuous, then the distance between the two queues cannot be shown to converge to 0.

Instead, here we adopt a different approach that allows us to deal with a discontinuous service-time limit process. This consists in expressing $Q_n(\cdot)$ as the reflection of an appropriate free process $N_n(\cdot)$. Since, after rescaling, $N_n(\cdot)$ converges and the reflection mapping is continuous almost surely in the limit point, the process $Q_n(\cdot)$ also converges by the continuous mapping theorem. The free process $N_n(\cdot)$ has the following interpretation: when the server is working, $N_n(\cdot)$ follows $Q_n(\cdot)$. When the queue is empty, $N_n(\cdot)$ decreases linearly at a rate equal to the service rate. Therefore, while in the Borovkov modified system, the server works continuously according to the service-time renewal process, and in the process $N_n(\cdot)$ when there are no customers in the system, the server provides instantaneous work with rate $1/\mathbb{E}[S]$. Consequently, the process $N_n(\cdot)$ can be seen as a fluid version of the Borovkov modified system. In Figure 1 we plot a sample path of the process $N_n(\cdot)$. The process $Q_n(\cdot)$ can then be represented as follows.

Proposition 1. (Key reformulation for the queue length.) *The queue-length process* $(Q_n(t))_{t\ge 0}$ *can be represented as*

$$Q_n(t) = \phi(N_n)(t), \qquad t \ge 0, \tag{9}$$

where $N_n(\cdot)$ *is given by*

$$N_n(t) = N_n(0) + A_n(t) - \sigma(B_n(t)) - \frac{t - B_n(t)}{\mathbb{E}[S]}. \tag{10}$$

Proof. Start from (8) and add and subtract $(t - B_n(t))/\mathbb{E}[S]$ to obtain

$$Q_n(t) = N_n(0) + A_n(t) - \sigma(B_n(t)) - \frac{t - B_n(t)}{\mathbb{E}[S]} + \frac{t - B_n(t)}{\mathbb{E}[S]}.$$

Representation (9) will follow as the solution of the so-called *Skorokhod problem.*

https://doi.org/10.1017/jpr.2017.42 Published online by Cambridge University Press

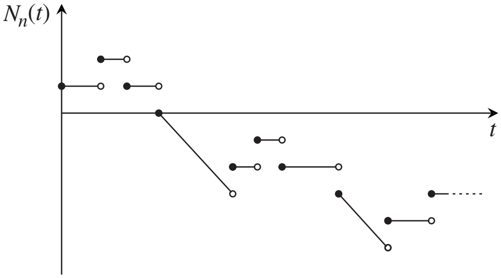

FIGURE 1: A sample path of the process $N_n(\cdot)$.

Lemma 1. (The Skorokhod problem [3, Proposition 2.2, p. 251].) *Let $X(\cdot)$ be a real-valued stochastic process such that $X(0) = 0$. Assume that $R(\cdot)$ is a nondecreasing right-continuous process such that the process $Q(\cdot)$ given by $Q(0) = q$ and $Q(t) = X(t) + R(t)$ is nonnegative for all t, and $\int_0^\infty Q(t)\,\mathrm{d}R(t) = 0$. Then*

$$R(t) = \psi(q + X(\cdot))(t) \quad \textit{and} \quad Q(t) = \phi(q + X(\cdot))(t),$$

where $\psi(\cdot)$ and $\phi(\cdot)$ are defined in (6) and (7), respectively.

To prove Proposition 1, we apply Lemma 1 with the choices

$$\begin{aligned} R(t) &= N_n(0) + \frac{t - B_n(t)}{\mathbb{E}[S]}, \qquad Q(t) = Q_n(t), \\ X(t) &= A_n(t) - \sigma(B_n(t)) - \frac{t - B_n(t)}{\mathbb{E}[S]}. \end{aligned} \tag{11}$$

Note that $R(t) = N_n(0) + I_n(t)/\mathbb{E}[S]$, where we recall that $t \mapsto I_n(t)$ is the idle-time process. In particular, $t \mapsto R(t)$ is nondecreasing. This, together with the fact that $R(t) \geq 0$, implies that $t \mapsto R(t)$ is of bounded variation. Moreover, $R(t) = N_n(0) + I_n(t)/\mathbb{E}[S]$ increases if and only if $Q(t) = 0$, that is, $\int_0^\infty Q(t)\,\mathrm{d}R(t) = 0$. Therefore, Lemma 1 implies that

$$\begin{aligned} Q(t) &= N_n(0) + X(t) - \inf_{0 \leq s \leq t} (N_n(0) + X(t))^- = \phi(N_n)(t), \\ R(t) &= - \inf_{0 \leq s \leq t} (N_n(0) + X(t))^- = \psi(N_n)(t), \end{aligned}$$

where we use the definition of $N_n(t)$ in (10) as well as (11). This completes the proof. □

We will consider the scaled versions of the processes of interest given by

$$\begin{aligned} \boldsymbol{Q}_n(t) = \phi(\boldsymbol{N}_n)(t), \qquad \tau_n(t) = t n^{\alpha/(2\alpha-1)} \ell_1(n), \\ \boldsymbol{N}_n(t) = n^{-1/(2\alpha-1)} \ell_2(n) N_n(\tau_n(t)), \end{aligned} \tag{12}$$

where $\ell_1(\cdot)$ and $\ell_2(\cdot)$ are slowly varying functions that depend on $\ell(\cdot)$. Using basic properties of slowly varying functions (see, e.g. [8, Proposition 1.3.6]), the scaling constants can be written as

$$n^{\alpha/(2\alpha-1)} \ell_1(n) = n^{(1+o(1))\alpha/(2\alpha-1)}, \qquad n^{-1/(2\alpha-1)} \ell_2(n) = n^{-(1+o(1))/(2\alpha-1)}.$$

In particular, for $\alpha = 2$, the scaling exponents are (asymptotically) the same as in the finite-variance case [4]. We can now state our main result.

https://doi.org/10.1017/jpr.2017.42 Published online by Cambridge University Press

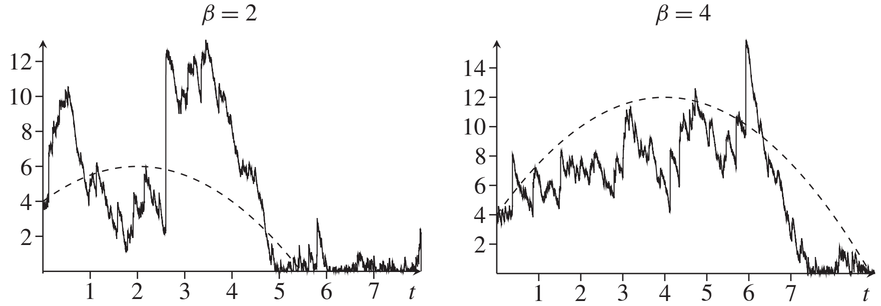

FIGURE 2: Sample paths of the process $\phi(\mathcal{N})(t)$ for $\beta = 2$ and $\beta = 4$. The dashed line is the plot of the functions $t \mapsto q_0 + \lambda\beta t - \lambda^2/2t$. In all plots, $q_0 = 4$, $\alpha = 1.8$, $\lambda = s_\alpha = 1$.

Theorem 1. (Scaling limit for the queue-length process.) *Assume that* $N_n(0) = q_0 n^{1/(2\alpha-1)} \times \ell_2^{-1}(n)$ *for some* $q_0 \geq 0$. *Assume further that* $c_n = 1 - \beta n^{-(\alpha-1)/(2\alpha-1)}$, *where* $\beta \in (-\infty, \infty)$. *Then*

$$\boldsymbol{N}_n(\cdot) \xrightarrow{\text{D}} \mathcal{N}(\cdot) \quad \textit{in } (\mathcal{D}, M_1),$$

where

$$\mathcal{N}(t) = q_0 + \beta\lambda t - \frac{\lambda^2}{2} t^2 + s_\alpha \mathcal{S}_\alpha(t), \tag{13}$$

with $s_\alpha = 1/\mathbb{E}[S]^{1+1/\alpha}$ *and where* $\mathcal{S}_\alpha(\cdot)$ *is a spectrally positive* α*-stable process. Moreover,*

$$\boldsymbol{Q}_n(\cdot) \xrightarrow{\text{D}} \phi(\mathcal{N})(\cdot) \quad \textit{in } (\mathcal{D}, M_1). \tag{14}$$

Convergence in $(\mathcal{D}, M_1)$ is a shorthand notation for convergence in distribution in the space of càdlàg functions $\mathcal{D}$ endowed with the M_1 topology. We elaborate on this later on. In Figure 2 we present the first passage time as a function of the linear drift β, for fixed α and q_0, and different values of the linear drift parameter β. We see that a larger β (slower server speed) corresponds to a larger busy period, as expected.

Define the first hitting time of x for a function $f(\cdot)$ as

$$H_f(x) = \inf\{t > 0\colon f(t) \leq x\}.$$

Then $H_{\boldsymbol{N}_n}(0)$ represents the first busy period of the $\Delta_{(i)}/\mathrm{G}/1$ queue. The following corollary of Theorem 1 characterizes the limiting distribution of $H_{\boldsymbol{N}_n}(0)$.

Corollary 1. (Busy period convergence.) *Under the assumptions of Theorem 1, as* $n \to \infty$,

$$n^{-\alpha/(2\alpha-1)} \ell_1(n)^{-1} H_{N_n}(0) = H_{\boldsymbol{N}_n}(0) \xrightarrow{\text{D}} H_{\mathcal{N}}(0).$$

Proof. By [19, Chapter VI, Proposition 2.11], the functional $f \mapsto H_f(0)$ is continuous in $\mathcal{N}$ with probability 1 when $\mathcal{D}$ is endowed with the M_1 topology. Note that [19, Chapter VI, Proposition 2.11] holds for the J_1 topology, but the proof is readily adapted to the M_1 topology. See also [38, Theorem 13.6.5]. The conclusion follows by an application of the continuous mapping theorem. □

Remark 1. (*Approximation of* $H_{\mathcal{N}}(0)$.) The plots in Figures 2 and 3 suggest that the quadratic drift given by

$$d_{q_0,\beta}(t) := q_0 + \beta\lambda t - \frac{\lambda^2}{2} t^2$$

https://doi.org/10.1017/jpr.2017.42 Published online by Cambridge University Press

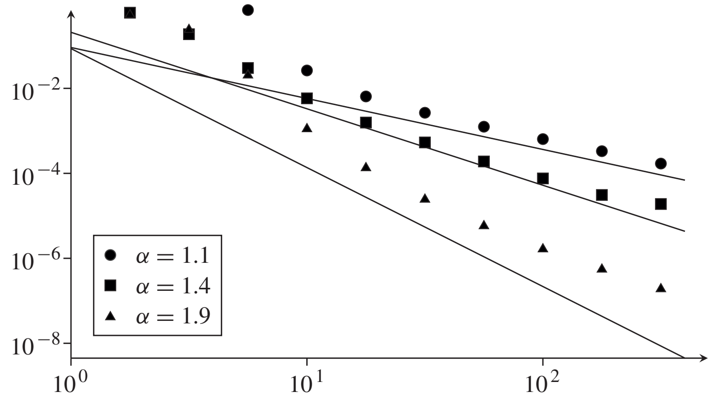

FIGURE 3: A log-log scale plot of the empirical tail distribution $\mathbb{P}(H_{\mathcal{N}}(0) > t)$ of the first busy period of $Q_n(\cdot)$ for different values of $\alpha \in (1, 2)$. The solid lines represent the asymptotic approximation (15). In all plots, $n = 1000$ and $q_0 = \beta = \lambda = s_\alpha = 1$.

yields a first-order approximation of $H_{\mathcal{N}}(0)$. In particular, the hitting time $H_{d_{q_0,\beta}(\cdot)}(0)$ is simply given by the solution of a quadratic equation, and is equal to

$$H_{d_{q_0,\beta}(\cdot)}(0) = \frac{-\beta + \sqrt{\beta^2 + 2q_0}}{\lambda},$$

where we have assumed that $q_0 \geq 0$. Note that the hitting time of zero of $\mathcal{S}_\alpha(\cdot)$ is distributed as a $(1/\alpha)$-stable random variable by [32, Theorem 46.3]; see also [33].

Remark 2. (*Tail probabilities.*) Corollary 1 allows us to estimate the tail probability for the length of the first busy period. In fact, we have the exact upper bound

$$\mathbb{P}(H_{\mathcal{N}}(0) > t) \leq \mathbb{P}\bigg(s_\alpha \mathcal{S}_\alpha(t) > -q_0 - \beta\lambda t + \frac{\lambda^2}{2}t^2\bigg),$$

where we have used the trivial inclusion of events $\{H_{\mathcal{N}}(0) > t\} \subseteq \{\mathcal{N}(t) > 0\}$. By basic properties of stable laws, we have the asymptotic relation (see [31, pp. 16–17])

$$\begin{aligned}\mathbb{P}\bigg(X_\alpha > \frac{1}{s_\alpha}(-q_0 t^{-1/\alpha} - \beta\lambda t^{1-1/\alpha} + \frac{\lambda^2}{2}t^{2-1/\alpha})\bigg) &\\ \sim \frac{C_\alpha s_\alpha^\alpha}{(-q_0 t^{-1/\alpha} - \beta\lambda t^{(\alpha-1)/\alpha} + \lambda^2 t^{(2\alpha-1)/\alpha}/2)^\alpha} &\\ \sim \frac{2^\alpha C_\alpha s_\alpha^\alpha}{\lambda^{2\alpha}} \frac{1}{t^{2\alpha-1}}, & \qquad (15)\end{aligned}$$

where X_α is distributed as a standard α-stable law, $C_\alpha = (1-\alpha)/\Gamma(2-\alpha)\cos(\pi\alpha/2)$ for $\alpha \neq 1$, and $t \mapsto \Gamma(t)$ is the standard gamma function. On the other hand, due to the strong negative drift of $\mathcal{N}(\cdot)$, it is natural to conjecture that the two events $\{H_{\mathcal{N}}(0) > t\}$ and $\{\mathcal{N}(t) > 0\}$ are of comparable measure when t is large. In Figure 3 we show that the tails of the empirical distribution of the first busy period behave like the upper bound (15) (see [1] and [36], where this was proven when $\mathcal{S}_\alpha(\cdot)$ is replaced by a more complicated *thinned* Lévy process). However, the approximation becomes less effective as $\alpha \nearrow 2$, and for $\alpha = 2$, (15) is not theoretically

https://doi.org/10.1017/jpr.2017.42 Published online by Cambridge University Press

justified. In fact, for this finite-variance case, Pittel [28] (see also [29] and [30]) showed that the tail asymptotically behaves as

$$\mathbb{P}(H_{\mathcal{N}}(0) > t) = \frac{1}{\sqrt{9\pi/8}t^{3/2}} \exp\left(\frac{t(t-2\beta)^2}{8}\right)(1+o(1)), \qquad t \to \infty.$$

Relation with the existing literature. The $\Delta_{(i)}/\mathrm{G}/1$ queue was introduced in [16] as a model for queueing systems serving only a finite pool of customers. In [15], the authors showed that the $\Delta_{(i)}/\mathrm{G}/1$ queue can be regarded as the canonical model for the study of such systems, in the sense that, under mild assumptions, every other transitory queueing model has the same asymptotic behavior. In [16], the authors assumed that $\mathbb{E}[S^2] < \infty$ and, further assuming that at some instant t the queue is overloaded, proved a functional law of large numbers and a functional central limit theorem (FCLT) for the $\Delta_{(i)}/\mathrm{G}/1$ queue. The latter turns out to be highly nontrivial. The limit process is a diffusion that switches between three regimes: a free Brownian motion, a reflected Brownian motion, and the zero process. Therefore, the results in [16] can be seen as the equivalent of the standard FCLT for light-tailed GI/G/1 queues [17], [18] in a transitory setting. However, time-varying queues exhibit a richer behavior than their ergodic counterparts. In particular, when $\rho = 1$ (where ρ is a model-dependent traffic intensity parameter) and under appropriate scaling, the queue-length process has a polynomial drift, see [24] for the $\mathrm{M}_t/\mathrm{M}_t/1$ queue and [39] for the $\mathrm{G}_t/\mathrm{G}/1$ queue.

A common assumption in queueing theory is that the arrival process of customers is a renewal process, a Poisson process being the classical choice. In this paper we relax this Poisson assumption and instead assume that with time more customers have passed the queue, and, hence, fewer customers can potentially join it. This is the only assumption that deviates from the classical setting and that changes an M/G/1 queue with a Poisson arrival process into the $\Delta_{(i)}/\mathrm{G}/1$ queue with *thinned* Poisson arrivals. For both the M/G/1 and the $\Delta_{(i)}/\mathrm{G}/1$ queues it is clear that the queue-length process is strongly influenced by the service times, and, in particular, depends on whether or not the service-time distribution is heavy tailed. For the M/G/1 queue, several heavy-traffic limit theorems have been established for heavy-tailed service-time distributions with infinite variance; see [3], [9], and [38] and the references therein. In this paper we pursue similar limit theorems for the heavy-tailed $\Delta_{(i)}/\mathrm{G}/1$ queue, although the thinned arrival process leads to vastly different results. A connection, however, with the classical work on the $M/\mathrm{G}/1$ queue [3], [9], [38] is that also in the case of the $\Delta_{(i)}/\mathrm{G}/1$ queue stable laws play a crucial role. For the M/G/1 queue, and in queueing theory in general, one typically distinguishes between light-tailed and heavy-tailed service-time distributions. As such, this paper should be regarded as the heavy-tailed extension of the light-tailed setting studied in [4, Theorem 1].

Remark 3. (*Exponential arrival times.*) In this paper we restrict ourselves to exponentially distributed arrival times $(T_i)_{i=1}^{\infty}$. However, in [4, Theorem 6] it was shown that, in the finite-variance case, when the density function has its maximum in $\bar{t} = 0$ and is sufficiently smooth around $\bar{t}$, the first excursion of the limit process only depends on the value of the density function of the arrival epochs, and of its first nonzero derivative in $\bar{t}$, and, thus, is essentially independent from the distribution of T. This insensitivity is further supported by the relation

$$T_i \overset{\mathrm{D}}{=} F_T^{-1}(1-\exp(-E_i)),$$

where '$\overset{\mathrm{D}}{=}$' denotes equality in distribution. This suggests that results for the exponential arrivals $(E_i)_{i=1}^{\infty}$ can be extended to general arrival times $(T_i)_{i=1}^{\infty}$ by application of an appropriate functional. Indeed, this is how [4, Theorem 6] was proven.

Remark 4. (*Limiting processes with quadratic drift.*) Our previous work [4] exploited previous results on Brownian motion with quadratic drift to give asymptotically exact approximation formulas for the first passage time density and the tail of the busy period. In fact, the process $W(t) = B(t) - ct^2$, where $B(\cdot)$ is the standard Brownian motion and $c > 0$ a constant, has been studied by several authors. In [13], [14], and [20], analytic expressions were derived for the joint density of the maximum and location of the maximum of $t \mapsto W(t)$, and tail estimates were derived in [35]. In [2] it was shown that the length of the excursions of $t \mapsto W(t)$ above its past minima can be ordered. To the best of the authors' knowledge, similar results for α-stable processes for $\alpha \in (0, 2)$ with quadratic drift as in (13) are not known.

Remark 5. (*Connections with random graphs.*) There is an interesting connection between the $\Delta_{(i)}/\mathrm{G}/1$ queue and a class of *random graphs*. Indeed, we can associate a (rooted) random forest to the queueing process, as follows. Customers are the vertices, the first customer in the system is the root, and when customer i joins the queue during the service of customer j, an edge is placed between i and j. The queue-length process then corresponds to the *exploration* of the random tree constructed this way. The exploration process encodes useful information on the underlying random graph. For example, excursions above past minima are the sizes of the connected components. Therefore, Theorem 1 should be compared with analogous results for other random graph models; see [5], [6], [11], [12], [21], and [37].

Outline. In Section 3 we give some background on α-stable laws and a derivation of the scaling constants in (12). In Section 4 we provide the proof of Theorem 1. We prove, separately, the convergence to the stable motion, which follows from classical arguments, and the convergence to the parabolic drift, which is achieved through a novel coupling argument. Finally, in Section 5 we present some conclusions and open problems.

Notation. A common topology on $\mathcal{D}$, which extends the uniform topology U, is the so-called J_1 topology, defined by Skorokhod [34]. However, when dealing with limit processes with *unmatched jumps*, the coarser M_1 topology is needed. When dealing with vector-valued functions (taking values, say, in $\mathbb{R}^k$), we make use of the *weak* M_1 topology M_1^{W}, which coincides with the product topology on $\mathcal{D} \times \mathcal{D} \times \cdots \times \mathcal{D} = \mathcal{D}^k$. For an in-depth discussion on the various Skorokhod topologies, see [38]. For most results in this paper, convergence of processes means convergence in distribution in the space $\mathcal{D}$ endowed with the M_1 topology. Recall that convergence $X_n(\cdot) \xrightarrow{\mathrm{D}} X(\cdot)$ in $\mathcal{D}([0, \infty))$ is equivalent to convergence in $\mathcal{D}([0, T])$ for all T that are continuity points of $X(\cdot)$. For a sequence of real-valued random variables $(X_n)_{n=1}^{\infty}$, we say that X_n converges to X in probability, and denote it by $X_n \xrightarrow{\mathbb{P}} X$, if, for each $\varepsilon > 0$,

$$\mathbb{P}(|X_n - X| \geq \varepsilon) \to 0 \quad \text{as } n \to \infty.$$

With $X_n = o_{\mathbb{P}}(Y_n)$ we mean that $X_n/Y_n \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$. Given two functions $f(\cdot)$ and $g(\cdot)$ (either on the real numbers or on the integers), the notation $f \sim g$ means $\lim_{x\to\infty} f(x)/g(x) = 1$, where $x \in \mathbb{R}$ or $x \in \mathbb{N}$.

3. Preliminaries

In this section we introduce some results that will be useful for the proof of Theorem 1. In Section 3.1 we present an FCLT for the service-time process $\sigma(\cdot)$. In Section 3.2 we derive an alternative characterization of the arrival process of the $\Delta_{(i)}/\mathrm{G}/1$ queue which reveals a connection with the Poisson process. Finally, in Section 3.3 we give a heuristic argument that motivates the scaling constants appearing in Theorem 1.

https://doi.org/10.1017/jpr.2017.42 Published online by Cambridge University Press

3.1. FCLT for a renewal process

We will present an FCLT for the renewal process $\sigma(\cdot)$ in (4). To do so we exploit the well-known equivalence between the FCLT for partial sums and counting processes. Let $(S_i)_{i=1}^n$ be a sequence of nonnegative random variables and let

$$\mathbf{\Sigma}_n(t) = \frac{\Sigma_{\lfloor nt \rfloor} - \mathbb{E}[S]nt}{c_n} \tag{16}$$

be its rescaled partial sum, where $\Sigma_k = S_1 + \cdots + S_k$ and $(c_n)_{n=1}^\infty$ will be chosen appropriately. Hence, $\sigma(t) = \max\{k \geq 0 \mid \Sigma_k \leq t\}$. Let $\boldsymbol{\sigma}_n(\cdot)$ denote the corresponding rescaled process

$$\boldsymbol{\sigma}_n(t) = \frac{\sigma(nt) - \mathbb{E}[S]^{-1}nt}{c_n}. \tag{17}$$

The relation between the scaling limits of $\mathbf{\Sigma}_n(\cdot)$ and $\boldsymbol{\sigma}_n(\cdot)$ is described in the following theorem.

Theorem 2. (FCLT equivalence [38, Theorem 7.3.2].) *Assume that $(S_i)_{i\geq 1}$ is a sequence of nonnegative random variables, and $(c_n)_{n\geq 1}$ is such that $c_n \to \infty$, $n/c_n \to \infty$. Then*

$$\mathbf{\Sigma}_n(\cdot) \xrightarrow{\mathrm{D}} \mathcal{S}_\alpha(\cdot) \quad \textit{in } (\mathcal{D}, M_1) \tag{18}$$

for some process $\mathcal{S}_\alpha(\cdot)$ if and only if

$$\boldsymbol{\sigma}_n(\cdot) \xrightarrow{\mathrm{D}} -\mathbb{E}[S]^{-1}\mathcal{S}_\alpha \circ \mathbb{E}[S]^{-1}\,\mathrm{id}(\cdot) \quad \textit{in } (\mathcal{D}, M_1), \tag{19}$$

where $\mathrm{id}(\cdot)$ *is the identity function.*

The topology M_1 plays a crucial role in Theorem 2. Indeed, it can be seen that while (18) holds in most cases in the J_1 topology, the convergence (19) can only take place in the M_1 topology when the limit process has positive jumps; see [38, Chapter 7.3.2] for a more detailed explanation. By assumption (1), the sequence $(S_i)_{i\geq 1}$ is in the domain of attraction of an α-stable motion, that is, (18) holds, and $\mathcal{S}_\alpha(\cdot)$ is a centered, spectrally positive α-stable motion. By Theorem 2, the process $\boldsymbol{\sigma}_n(\cdot)$ is then also in the domain of attraction of an α-stable motion. Note that the space scaling constants c_n in (16) and (17) are the same.

3.2. Poissonian representation of the arrival process

In order to further simplify the representation of $Q_n(t)$, we now introduce an alternative characterization of the arrival process as a thinned, marked Poisson process. It is constructed as follows. Given $\Pi(t)$, a rate λ homogeneous Poisson process, assign to each of its points a mark chosen uniformly in $[n] := \{1, \ldots, n\}$. We then discard a point if it has a mark that has already been observed in the past. Therefore, conditioned on the marks $M_1, \ldots, M_{k-1}$, the next point of $\Pi(t)$ will be accepted with probability $(n - |\{M_1, \ldots, M_{k-1}\}|)/n$. We denote this thinned process as $A_n^m(t)$. Then $A_n^m(t)$ can be represented as

$$A_n^m(t) = \Pi(t) - R_n(t), \tag{20}$$

where $R_n(t)$ counts the number of *repeated* marks. We emphasize that $\Pi(\cdot)$ and $R_n(\cdot)$ are *not* independent. The arrival process just defined is closely related with the i.i.d. sampling in the $\Delta_{(i)}/\mathrm{G}/1$ queue, as we discuss now.

https://doi.org/10.1017/jpr.2017.42 Published online by Cambridge University Press

In the $\Delta_{(i)}/\mathrm{G}/1$ queue, arrivals are given by an i.i.d. sequence of arrival clocks $(T_i)_{i=1}^n$, where it is assumed that customer $i \in [n]$ joins the queue at time T_i. This definition departs from the usual queueing assumption of a renewal process, which entails i.i.d. *inter*-arrival times rather than *arrival* times. However, the above characterization as a marked Poisson process, which holds when the arrival clocks are exponentially distributed, is closer to the usual renewal setting. In what follows we show that the two are equivalent. First, let us introduce some preliminary notation and results. Given a sequence of random variables $(X_i)_{i=1}^n$, let $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ denote their order statistics. When $(X_i)_{i=1}^n$ are i.i.d. exponential random variables, the distribution of the order statistics are well known.

Lemma 2. (Order statistics of exponentials.) *Let* $E_1, \ldots, E_n$ *be independent exponentially distributed random variables with mean* 1. *Then*

$$(E_{(j)})_{j=1}^n \overset{\mathrm{D}}{=} \left(\sum_{s=1}^{j} \frac{E_s}{n-s+1}\right)_{j=1}^n .$$

See, e.g. [10, Section 2.5] for a proof. Lemma 2 allows us to relate the process $A_n^m(\cdot)$ we have just defined to the arrival process in the $\Delta_{(i)}/\mathrm{G}/1$ queue.

Lemma 3. *For all* $t \ge 0$,

$$A_n^m(t) \overset{\mathrm{D}}{=} A_n(t). \tag{21}$$

Proof. The (ordered) arrival times in the $\Delta_{(i)}/\mathrm{G}/1$ queue are precisely the order statistics of $(T_i)_{i=1}^n$ and the interarrival times are the differences between the order statistics. By Lemma 2, the distributions of the interarrival times are

$$\frac{1}{\lambda}(E_{(k)} - E_{(k-1)}) \overset{\mathrm{D}}{=} \frac{E_k/\lambda}{n-k+1}, \qquad k \ge 1,$$

where we set $E_{(0)} = 0$ for convenience. Multiplying both sides by n, and noting that $E_i/\lambda = T_i$, yields

$$n(T_{(k)} - T_{(k-1)}) \overset{\mathrm{D}}{=} \frac{E_k}{1-(k-1)/n}\frac{1}{\lambda}. \tag{22}$$

Now consider the arrival process defined in (20). Conditioned on the process up to the arrival $k-1$, the next point of $\Pi(\cdot)$ is accepted with probability $1-(k-1)/n$, where $k-1$ is also equal to the number of distinct marks. Then, since $\Pi(\cdot)$ is a rate λ Poisson process, the time at which the next point of $A_n^m(\cdot)$ occurs is distributed as an exponential random variable with rate $\lambda(1-(k-1)/n)$. Equation (22) then implies that the interarrival times in the process $t \mapsto A_n^m(t)$ are equal (in distribution) to the interarrival times of $A_n(t) = A(t/n)$. □

Representation (21) motivates us to consider the rescaled arrival process $A_n(t) = A(t/n)$. As we show below, the limit result is not influenced by this time rescaling, as long as the scaling constants are defined appropriately.

3.3. Determining the scaling constants

We now derive the space and time scaling that allows us to obtain the limit process $\mathcal{N}(\cdot)$ in (13). We first derive the scaling of time denoted by $k = k(n)$. It is well known that, whenever the limit $\mathcal{S}_\alpha(\cdot)$ in (18) is an α-stable motion, the fluctuations of $\sum_{i=1}^{\lfloor kt\rfloor} S_i$ around its mean are of the order $c_k = \ell_0(k)k^{1/\alpha}$ (see, e.g. [38, Theorem 4.5.1]), where $\ell_0(\cdot)$ is a slowly varying function that is *a priori* different from $\ell(\cdot)$ in (1) (but can be determined from it). Moreover,

https://doi.org/10.1017/jpr.2017.42 Published online by Cambridge University Press

in (36) below we show that the highest-order contribution to the drift component $R_n(kt)$ is $\Pi(kt)^2/(2n) = O_{\mathbb{P}}(k^2/n)$, all the other terms being negligible. In the process $\mathcal{N}(\cdot)$ both a drift and a random component appear, so that we must have

$$\ell_0(k)k^{1/\alpha} = \frac{k^2}{n}. \tag{23}$$

Equivalently,

$$\ell_0(k)^{-\alpha/(2\alpha-1)}k = n^{\alpha/(2\alpha-1)}, \tag{24}$$

where $\ell_0(\cdot)^{-\alpha/(2\alpha-1)}$ is, by basic properties of slowly varying functions, again slowly varying. On the left-hand side of (24), we recognize a regularly varying function with index 1. By [8, Theorem 1.5.12], each regularly varying function with index γ admits an (asymptotic) inverse that is itself regularly varying, with index $1/\gamma$. Therefore, there exists a slowly varying function $\rho(\cdot)$ so that we must have

$$k = n^{\alpha/(2\alpha-1)}\rho(n^{\alpha/(2\alpha-1)}). \tag{25}$$

Any sequence $(k(n))_{n\geq 1}$ that satisfies condition (25) is suitable for our purposes, so that we simply take $k(n) = n^{\alpha/(2\alpha-1)}\ell_1(n)$, where $\ell_1(n) = \rho(n^{\alpha/(2\alpha-1)})$. Note that $n \mapsto \ell_1(n)$ is again slowly varying. Therefore, the rescaled time parameter is defined as

$$\tau_n(t) := tn^{\alpha/(2\alpha-1)}\ell_1(n). \tag{26}$$

We shall denote the time scaling factor by $\tau_n(1) = n^{\alpha/(2\alpha-1)}\ell_1(n)$. In order to obtain the space-scaling sequence $(s_n)_{n\geq 1}$, it is enough to insert $k = n^{\alpha/(2\alpha-1)}\ell_1(n)$ into $f(k) := k^2/n$. Therefore, we define s_n as

$$s_n = \left(\frac{(n^{\alpha/(2\alpha-1)}\ell_1(n))^2}{n}\right)^{-1} = \ell_1(n)^{-2}n^{-1/(2\alpha-1)} =: \ell_2(n)n^{-1/(2\alpha-1)}, \tag{27}$$

where $\ell_2(n) = \ell_1(n)^{-2}$ is again slowly varying.

4. Proof of Theorem 1

In this section we carry out the proof of Theorem 1. We first prove Theorem 1 for $\beta = 0$, and later show how to extend it to the general $\beta \neq 0$ case. Rewriting (10) using (20) yields

$$\begin{aligned} N_n(t) &\overset{\mathrm{D}}{=} N_n(0) + \left(A_n^m(t) - \frac{t}{\mathbb{E}[S]}\right) + \left(\frac{B_n(t)}{\mathbb{E}[S]} - \sigma(B_n(t))\right) \\ &= N_n(0) + \left(\Pi(t) - \frac{t}{\mathbb{E}[S]}\right) + \left(\frac{B_n(t)}{\mathbb{E}[S]} - \sigma(B_n(t))\right) - R_n(t). \end{aligned} \tag{28}$$

For simplicity, we introduce the scaled version of the arrival and service processes, and of the busy time, as

$$\begin{aligned} \boldsymbol{\Pi}_n(t) &= n^{-1/(2\alpha-1)}\ell_2(n)\left(\Pi(\tau_n(t)) - \frac{\tau_n(t)}{\mathbb{E}[S]}\right), \\ \boldsymbol{R}_n(t) &= n^{-1/(2\alpha-1)}\ell_2(n)R_n(\tau_n(t)), \\ \boldsymbol{\sigma}_n(t) &= n^{-1/(2\alpha-1)}\ell_2(n)\left(\frac{\tau_n(t)}{\mathbb{E}[S]} - \sigma(\tau_n(t))\right), \\ \hat{\boldsymbol{B}}_n(t) &= \frac{B_n(\tau_n(t))}{\tau_n(1)}. \end{aligned}$$

https://doi.org/10.1017/jpr.2017.42 Published online by Cambridge University Press

Assume that $N_n(0) = q_0 n^{1/(2\alpha-1)}\ell_1(n)$ for some $q_0 \geq 0$. After rescaling, (28) becomes

$$\boldsymbol{N}_n(\tau_n(t)) = q_0 + \boldsymbol{\Pi}_n(t) + \boldsymbol{\sigma}_n(\hat{\boldsymbol{B}}_n(t)) - \boldsymbol{R}_n(t). \tag{29}$$

The proof of Theorem 1 proceeds as follows. First, the term $\boldsymbol{\Pi}_n(\cdot)$ is shown to be negligible in the limit. Second, $\boldsymbol{\sigma}_n(\cdot)$ converges to an α-stable motion by (1) and Theorem 2. Third, $\boldsymbol{R}_n(\cdot)$ is shown to converge to the parabolic drift $-\lambda^2/2t^2$ in Section 4.2. Finally, $\hat{\boldsymbol{B}}_n(\cdot)$ is shown to converge to the identity function. All these results are then pieced together in Section 4.3. Convergence of the above processes is proven in $\mathcal{D}([0,T])$ for a fixed $T > 0$. Since T is arbitrary, this implies convergence in $\mathcal{D}([0,\infty])$ by [7, Lemma 3, p. 174].

4.1. Stable limit

We start by showing that the process $\boldsymbol{\Pi}(\cdot)$ vanishes in the limit.

Lemma 4. *As $n \to \infty$,*

$$\sup_{t\leq T} |\boldsymbol{\Pi}(\tau_n(t))| \xrightarrow{\mathbb{P}} 0.$$

Proof. By the FCLT for the Poisson process,

$$\frac{\Pi(\tau_n(\cdot)) - \tau_n(\cdot)\lambda}{\sqrt{\tau_n(1)}} \xrightarrow{\text{D}} B(\cdot) \quad \text{in } (\mathcal{D}, U),$$

where $B(\cdot)$ is a standard Brownian motion, since $1/\mathbb{E}[S] = \lambda$ by the heavy-traffic assumption. By the Skorokhod representation theorem, this implies that we can couple $\Pi(\tau_n(\cdot))$ and $B(\cdot)$ such that

$$\sup_{t\leq T}\left|\frac{\Pi(\tau_n(t)) - \tau_n(\cdot)/\mathbb{E}[S]}{\sqrt{\tau_n(1)}} - B(t)\right| \xrightarrow{\mathbb{P}} 0.$$

Moreover, for any $C > 0$ and large enough n,

$$C\sqrt{\tau_n(1)} = Cn^{\alpha/2(2\alpha-1)}\ell_1(n)^{1/2} \leq n^{1/(2\alpha-1)}\ell_2(n)^{-1},$$

so that $k_n := n^{1/(2\alpha-1)}\ell_2(n)/\sqrt{\tau_n} \to \infty$ and

$$\sup_{t\leq T}\left|\frac{\Pi(\tau_n(t)) - \tau_n(\cdot)/\mathbb{E}[S]}{n^{1/(2\alpha-1)}\ell_2(n)^{-1}}\right| \leq \frac{1}{k_n}\sup_{t\leq T}\left|\frac{\Pi(\tau_n(t)) - \tau_n(t)/\mathbb{E}[S]}{\sqrt{\tau_n(1)}} - B(t)\right| + \sup_{t\leq T}\left|\frac{B(t)}{k_n}\right|. \tag{30}$$

Since the right-hand side of (30) converges in probability to 0 as $n \to \infty$, the stated claim follows. □

Next, we show convergence of the rescaled service process $\sigma(\cdot)$ to an α-stable motion.

Lemma 5. (Stable limit.) *As $n \to \infty$,*

$$\boldsymbol{\sigma}_n(\cdot) \xrightarrow{\text{D}} s_\alpha \mathcal{S}_\alpha(\cdot) \quad \textit{in } (\mathcal{D}, M_1), \tag{31}$$

where $s_\alpha = \mathbb{E}[S]^{-(\alpha+1)/\alpha}$ and $\mathcal{S}_\alpha(\cdot)$ is a spectrally positive α-stable motion.

Proof. By classical results, the rescaled partial sums of $(S_i)_{i\geq 1}$ converge to a spectrally positive α-stable motion; see, e.g. [19] and [38, Theorem 4.5.3]. In particular, (18) is satisfied. Theorem 2 implies (19), that is,

$$\boldsymbol{\sigma}_n(\cdot) \xrightarrow{\text{D}} \frac{1}{\mathbb{E}[S]}\mathcal{S}_\alpha\left(\frac{\cdot}{\mathbb{E}[S]}\right) \quad \text{in } (\mathcal{D}, M_1).$$

By the standard properties of stable motion, $(\mathcal{S}_\alpha(ct))_{t\geq 0} \overset{\text{D}}{=} (c^{1/\alpha}\mathcal{S}_\alpha(t))_{t\geq 0}$ for $c > 0$ and so the claim follows. □

https://doi.org/10.1017/jpr.2017.42 Published online by Cambridge University Press

Remark 6. Although our results do not directly hold for $\alpha = 2$ (finite-variance case), it is still possible to substitute $\alpha = 2$ in the formulas that we obtain, and what is obtained should be consistent with the previously found results for the finite-variance case. This is true, e.g. for the coefficient of the stable motion in (31). Indeed, in [4, Theorem 1] it was proven that if $\mathbb{E}[S^2] = 1$, the standard deviation of the limiting Brownian motion is $\lambda^{3/2} = \mathbb{E}[S]^{-3/2}$.

4.2. Drift limit

The most difficult task in proving Theorem 1 is to deal with the complicated drift $\boldsymbol{R}_n(\cdot)$ in (29). We will prove the following result.

Proposition 2. (Drift limit.) *Under the same assumptions as in Theorem 1, as $n \to \infty$ and for any $T > 0$,*

$$\sup_{t \le T} \left| \boldsymbol{R}_n(t) - \frac{\lambda^2}{2} t^2 \right| \xrightarrow{\mathbb{P}} 0.$$

The proof will use upper and lower bounds for a distributionally equivalent characterization of $R_n(\cdot)$. First, note that the probability of extracting a mark that has already appeared at time $i \ge 1$ is $D_n(i-1)/n$, where $D_n(i)$ denotes the number of *different* marks seen up to the ith arrival epoch in $\Pi(\cdot)$. Therefore, conditionally on $D_n(i-1)$, the thinning procedure can be represented by a Bernoulli random variable with parameter $D_n(i-1)/n$. Since at time t there have been a total of $\Pi(t)$ points, we have

$$R_n(t) \overset{\mathrm{D}}{=} \sum_{i \le \Pi(t)} \mathbf{1}_{\{U_i \le D_n(i-1)/n\}},$$

where $(U_i)_{i \ge 1}$ are uniformly distributed (on $[0, 1]$) random variables, independent of all other randomness, and $\mathbf{1}_{\{U_i \le x\}}$ is distributed as a Bernoulli random variable with parameter x. Moreover, $D_n(i)$ can be written as

$$D_n(i) = i - Z_n(i),$$

where $Z_n(i)$ is the number of *repeated* marks seen up to the time of the ith arrival. In other words, we have the crucial relation

$$D_n(i) \overset{\mathrm{D}}{=} i - R_n(\Pi^{-1}(i)),$$

where $\Pi^{-1}(i)$ is the arrival time of the ith customer; see Figure 4.

Exploiting these ideas, we can recursively construct a process $(\tilde{R}_n(k))_{k \ge 0}$ with $\tilde{R}_n(0) := 0$ and

$$\tilde{R}_n(k) := \tilde{R}_n(k-1) + \mathbf{1}_{\{U_k \le (k-1-\tilde{R}_n(k-1))/n\}}, \qquad k \ge 1.$$

Unraveling the recursion, we obtain

$$\tilde{R}_n(k) := \sum_{i=1}^{k} \mathbf{1}_{\{U_i \le (i-1-\tilde{R}_n(i-1))/n\}}, \qquad k \ge 1.$$

Then

$$R_n(t) \overset{\mathrm{D}}{=} \tilde{R}_n(\Pi(t)). \tag{32}$$

As already mentioned, the processes $R_n(\cdot)$ and $\Pi(\cdot)$ are *not* independent. The distributional equality (32) reveals the dependency of $R_n(\cdot)$ on the process $\Pi(\cdot)$.

https://doi.org/10.1017/jpr.2017.42 Published online by Cambridge University Press

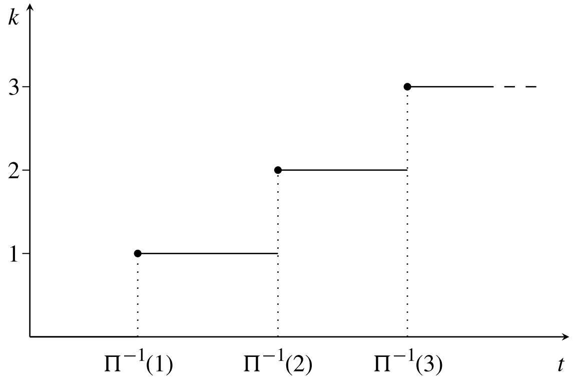

FIGURE 4: A sample path of the process $\Pi(\cdot)$.

The next step is to construct an upper and a lower bound on $\tilde{R}_n(k)$. Since $\tilde{R}_n(k) \geq 0$, the upper bound is trivially

$$\mathbf{1}_{\{U_i \leq (i-1-\tilde{R}_n(i-1))/n\}} \leq \mathbf{1}_{\{U_i \leq (i-1)/n\}},$$

so that, almost surely,

$$\tilde{R}_n(k) \leq \tilde{R}_n^{(\mathrm{up})}(k) := \sum_{i=1}^{k} \mathbf{1}_{\{U_i \leq (i-1)/n\}}. \tag{33}$$

The lower bound is more involved. By (33),

$$\mathbf{1}_{\{U_i \leq (i-1-\tilde{R}_n(i-1))/n\}} \geq \mathbf{1}_{\{U_i \leq (i-1-\tilde{R}_n^{(\mathrm{up})}(i-1))/n\}},$$

so that

$$\tilde{R}_n(k) \geq \tilde{R}_n^{(\mathrm{low})}(k) := \sum_{i=1}^{k} \mathbf{1}_{\{U_i \leq (i-1-\tilde{R}_n^{(\mathrm{up})}(i-1))/n\}}.$$

Note that U_i is independent of $\tilde{R}_n^{(\mathrm{up})}(i-1)$. We have then constructed a coupling such that, *for all* $t \geq 0$, almost surely,

$$R_n^{(\mathrm{low})}(t) \leq R_n(t) \leq R_n^{(\mathrm{up})}(t), \tag{34}$$

where $R_n^{(\mathrm{low})}(t) := \tilde{R}_n^{(\mathrm{low})}(\Pi(t))$ and $R_n^{(\mathrm{up})}(t) := \tilde{R}_n^{(\mathrm{up})}(\Pi(t))$. For the next and final step we now prove uniform convergence of the upper and lower bounds to the same limit.

4.2.1. *Upper bound.* Define the quantity to be estimated as

$$U_n(T) := \sup_{t \leq T} \left| n^{-1/(2\alpha-1)} \ell_2(n) R_n^{(\mathrm{up})}(\tau_n(t)) - \frac{\lambda^2}{2} t^2 \right|. \tag{35}$$

We will prove the following.

Lemma 6. (Upper bound converges to 0.) *Under the assumptions of Theorem 1, as $n \to \infty$,*

$$U_n(T) \xrightarrow{\mathbb{P}} 0 \quad \textit{for every fixed } T > 0.$$

Proof. The absolute value in (35) can be split as

$$
\begin{aligned}
U_n(T) &\le \left| n^{-1/(2\alpha-1)}\ell_2(n) \sum_{i\le \Pi(\tau_n(t))} \left(\mathbf{1}_{\{U_i\le (i-1)/n\}} - \frac{i-1}{n} \right) \right| \\
&\quad + \left| n^{-1/(2\alpha-1)}\ell_2(n) \sum_{i\le \Pi(\tau_n(t))} \left(\frac{i-1}{n} \right) - \frac{\lambda^2}{2}t^2 \right| \\
&\le \left| n^{-1/(2\alpha-1)}\ell_2(n) \sum_{i\le \Pi(\tau_n(t))} \left(\mathbf{1}_{\{U_i\le (i-1)/n\}} - \frac{i-1}{n} \right) \right| \\
&\quad + \left| \frac{\Pi(\tau_n(t))^2}{2n^{2\alpha/(2\alpha-1)}\ell_2^{-1}(n)} - \frac{\lambda^2}{2}t^2 \right| + \varepsilon_n,
\end{aligned} \tag{36}
$$

where $\varepsilon_n = |\Pi(\tau_n(t))/2n|$ is an error term. By the functional strong law of large numbers (LLN) for the Poisson process,

$$
\frac{\Pi(tn^{\alpha/(2\alpha-1)}\ell_1(n))}{n^{\alpha/(2\alpha-1)}\ell_2(n)^{-1/2}} \overset{\text{a.s.}}{\to} \lambda t \quad \text{in } (\mathcal{D}, U). \tag{37}
$$

It is worth noting that we have made explicit use of the specific form of the scaling functions $\ell_1(\cdot)$ and $\ell_2(\cdot)$ as determined above in (23)–(27). More specifically, by definition $\ell_1(n)^{-2} = \ell_2(n)$. Moreover, the functional $x \mapsto x^2$ from $\mathcal{D}([0, T])$ to itself is almost surely continuous in $f(t) = \lambda t$ in the uniform topology. This implies that the second and third terms in (36) converge to 0 uniformly for $t \le T$ as $n \to \infty$.

By the LLN for the Poisson process, we have $\Pi(s) \le (\lambda + \varepsilon)s$ with high probability for $s = O(n^{\alpha/(2\alpha-1)})$. The sum in the first term in (36) can then be bounded on the event $\{\Pi(s) \le (\lambda+\varepsilon)s\}$ as

$$
\sup_{s\le \tau_n(T)} \left| \sum_{i\le \Pi(s)} \left(\mathbf{1}_{\{U_i\le (i-1)/n\}} - \frac{i-1}{n} \right) \right| \le \sup_{s\le (\lambda+\varepsilon)\tau_n(T)} \left| \sum_{i\le \lfloor s \rfloor} \left(\mathbf{1}_{\{U_i\le (i-1)/n\}} - \frac{i-1}{n} \right) \right|. \tag{38}
$$

This can be recognized as the supremum of a martingale. In the following and future computations, we shall denote $\bar{T} := T(\lambda+\varepsilon)$. Then, an application of Doob's L^2 martingale inequality [22, Theorem 11.2] yields

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{s\le \bar{T} n^{\alpha/(2\alpha-1)}\ell_1(n)} \left| \sum_{i\le \lfloor s \rfloor} \left(\mathbf{1}_{\{U_i\le (i-1)/n\}} - \frac{i-1}{n} \right) \right| \ge \varepsilon n^{1/(2\alpha-1)}\ell_2^{-1}(n) \right) \\
&\quad \le \sum_{i\le \bar{T} n^{\alpha/(2\alpha-1)}\ell_1(n)} \frac{\mathbb{E}[(\mathbf{1}_{\{U_i\le (i-1)/n\}} - (i-1)/n)^2]}{\varepsilon^2 n^{2/(2\alpha-1)}\ell_2^{-2}(n)} \\
&\quad = \frac{1}{\varepsilon^2 n^{2/(2\alpha-1)}\ell_2^{-2}(n)} \sum_{i\le \bar{T} n^{\alpha/(2\alpha-1)}\ell_1(n)-1} \left(\frac{i}{n} - \frac{i^2}{n^2} \right) \\
&\quad \le \frac{\bar{T}^2 n^{2\alpha/(2\alpha-1)}\ell_1^2(n)}{\varepsilon^2 n^{(2\alpha+1)/(2\alpha-1)}\ell_2^{-2}(n)} \\
&\quad = O(n^{-1/(2\alpha-1)}\ell_2(n)),
\end{aligned}
$$

and this implies that the right-hand side of (38) is $o_{\mathbb{P}}(n^{1/(2\alpha-1)}\ell_2^{-1}(n))$. □

https://doi.org/10.1017/jpr.2017.42 Published online by Cambridge University Press

4.2.2. *Lower bound.* By (34), we also have

$$R_n(t) \succeq R_n^{(\mathrm{low})} = \sum_{i \le \Pi(t)} \mathbf{1}_{\{U_i \le (i-1-\tilde{R}_n^{(\mathrm{up})}(i-1))/n\}}.$$

Consequently, we now estimate

$$L_n(T) := \sup_{t \le T} \bigg| n^{-1/(2\alpha-1)} \ell_2(n) R_n^{(\mathrm{low})}(\tau_n(t)) - \frac{\lambda^2}{2} t^2 \bigg|.$$

Lemma 7. (Lower bound converges to 0.) *Under the assumptions of Theorem 1, as $n \to \infty$,*

$$L_n(T) \xrightarrow{\mathbb{P}} 0 \quad \textit{for every fixed } T > 0.$$

Proof. Similarly as before, conditioned on the event $\{\Pi(s) \le (\lambda + \varepsilon)s\}$,

$$\begin{aligned} L_n(T) \le \sup_{s \le \tau_n(\bar{T})} &\bigg| n^{-1/(2\alpha-1)} \ell_2(n) \sum_{i \le \lfloor s \rfloor} \bigg(\mathbf{1}_{\{U_i \le (i-1-\tilde{R}_n^{(\mathrm{up})}(i-1))/n\}} - \frac{i-1-\tilde{R}_n^{(\mathrm{up})}(i-1)}{n} \bigg) \bigg| \\ &+ \sup_{t \le T} \bigg| n^{-1/(2\alpha-1)} \ell_2(n) \sum_{i \le \Pi(\tau_n(t))} \frac{i-1}{n} - \frac{\lambda^2}{2} t^2 \bigg| \\ &+ \sup_{s \le \tau_n(\bar{T})} \bigg| n^{-1/(2\alpha-1)} \ell_2(n) \sum_{i \le \lfloor s \rfloor} \frac{\tilde{R}_n^{(\mathrm{up})}(i-1)}{n} \bigg|. \end{aligned} \tag{39}$$

The first term in (39) can also be bounded as before, since it is again the supremum of a martingale. Denote $Y_n(i) := (i - 1 - \tilde{R}_n^{(\mathrm{up})}(i-1))/n$ for convenience. By Doob's L^2 martingale inequality,

$$\begin{aligned} &\varepsilon^2 n^{2/(2\alpha-1)} \ell_2^{-2}(n) \mathbb{P}\bigg(\sup_{s \le \tau_n(\bar{T})} \bigg| \sum_{i \le \lfloor s \rfloor} (\mathbf{1}_{\{U_i \le Y_n(i)\}} - Y_n(i)) \bigg| \ge \varepsilon n^{1/(2\alpha-1)} \ell_2^{-1}(n) \bigg) \\ &\quad \le \mathbb{E}\bigg[\bigg(\sum_{i \le \tau_n(\bar{T})} \mathbf{1}_{\{U_i \le Y_n(i)\}} - Y_n(i) \bigg)^2 \bigg] \\ &\quad = \sum_{i \le \tau_n(\bar{T})} \mathbb{E}[(\mathbf{1}_{\{U_i \le Y_n(i)\}} - Y_n(i))^2]. \end{aligned}$$

Since the variance of a Bernoulli random variable with parameter p is $p(1-p)$, we obtain

$$\mathbb{E}[(\mathbf{1}_{\{U_i \le Y_n(i)\}} - Y_n(i))^2] = \mathbb{E}[Y_n(i) - Y_n(i)^2] \le \mathbb{E}[Y_n(i)] \le \frac{i}{n}.$$

This implies that

$$\sup_{i \le \bar{T} n^{\alpha/(2\alpha-1)} \ell_1(n)} \mathbb{E}[(\mathbf{1}_{\{U_i \le Y_n(i)\}} - Y_n(i))^2] \le \bar{T} n^{(1-\alpha)/(2\alpha-1)} \ell_1(n).$$

In particular,

$$\begin{aligned} \sum_{i \le \tau_n(\bar{T})} \mathbb{E}[(\mathbf{1}_{\{U_i \le Y_n(i)\}} - Y_n(i))^2] &\le \tau_n(\bar{T}) \bar{T} n^{(1-\alpha)/(2\alpha-1)} \ell_1^2(n) \\ &= \bar{T}^2 n^{1/(2\alpha-1)} \ell_1^2(n) \\ &= o(n^{2/(2\alpha-1)}). \end{aligned}$$

https://doi.org/10.1017/jpr.2017.42 Published online by Cambridge University Press

The second term in (39) has been shown to converge in (36) and (37). The third term can be bounded, using the fact that $t \mapsto \tilde{R}_n^{(\mathrm{up})}(t)$ is nondecreasing, as

$$\sup_{s \le \tau_n(\bar{T})} \bigg| \sum_{i \le \lfloor s \rfloor} \frac{\tilde{R}_n^{(\mathrm{up})}(i-1)}{n} \bigg| \le \bar{T} n^{(1-\alpha)/(2\alpha-1)} \ell_1(n) \tilde{R}_n^{(\mathrm{up})}(\tau_n(\bar{T})).$$

Note that $\bar{T} n^{(1-\alpha)/(2\alpha-1)} \ell_1(n) \to 0$ as $n \to \infty$. Since $n^{-1/(2\alpha-1)} \ell_2(n) \tilde{R}_n^{(\mathrm{up})}(\tau_n(\bar{T})) \xrightarrow{\mathbb{P}} 0$ by Lemma 6,

$$\begin{aligned} n^{-1/(2\alpha-1)} \ell_2(n) & \sup_{s \le \tau_n(\bar{T})} \bigg| \sum_{i \le \lfloor s \rfloor} \frac{\tilde{R}_n^{(\mathrm{up})}(i-1)}{n} \bigg| \\ & \le (\bar{T} n^{(1-\alpha)/(2\alpha-1)} \ell_1(n)) n^{-1/(2\alpha-1)} \ell_2(n) \tilde{R}_n^{(\mathrm{up})}(\tau_n(\bar{T})) \\ & \xrightarrow{\mathbb{P}} 0 \quad \text{as } n \to \infty. \end{aligned}$$

This concludes the proof of Lemma 7. □

Proof of Proposition 2. Since

$$\begin{aligned} \sup_{t \le T} & \big| n^{-1/(2\alpha-1)} \ell_2(n) R_n(t n^{\alpha/(2\alpha-1)} \ell_1(n)) - \tfrac{1}{2} t^2 \big| \\ & = \sup_{t \le T} \big(n^{-1/(2\alpha-1)} \ell_2(n) R_n(t n^{\alpha/(2\alpha-1)} \ell_1(n)) - \tfrac{1}{2} t^2 \big)^+ \\ & \quad + \sup_{t \le T} \big(n^{-1/(2\alpha-1)} \ell_2(n) R_n(t n^{\alpha/(2\alpha-1)} \ell_1(n)) - \tfrac{1}{2} t^2 \big)^-, \end{aligned}$$

we obtain

$$\sup_{t \le T} \big| n^{-1/(2\alpha-1)} \ell_2(n) R_n(t n^{\alpha/(2\alpha-1)} \ell_1(n)) - \tfrac{1}{2} t^2 \big| \le U_n(T) \vee L_n(T),$$

and both $U_n(T)$ and $L_n(T)$ converge in probability to 0 by Lemmas 6 and 7. This completes the proof of Proposition 2. □

4.3. Busy-time process limit

For the final step, we prove that the cumulative busy-time process converges to the identity function.

Lemma 8. (Cumulative idle time is negligible.) *As* $n \to \infty$,

$$\hat{\boldsymbol{B}}_n(t) \xrightarrow{\mathrm{D}} \mathrm{id}(\cdot) \quad \textit{in } (\mathcal{D}, U).$$

where $\mathrm{id}(\cdot) \cdot \mathbb{R}^+ \mapsto \mathbb{R}^+$ *is the identity function.*

Proof. Since $B_n(t) = t - I_n(t)$, we equivalently prove that $I_n(t) = \inf_{0 \le s \le t}(X_n(s)^-)$ converges uniformly to 0, where $X_n(t)$ is the net-input process defined in (5). By continuity of the map ψ given by $\psi \colon f(\cdot) \to \inf_{0 \le s \le \cdot}(f(s)^-)$, it is sufficient to prove that $X_n(\cdot)$ converges

https://doi.org/10.1017/jpr.2017.42 Published online by Cambridge University Press

uniformly to 0, when appropriately rescaled. By manipulating (5), we immediately obtain

$$
\begin{aligned}
&\frac{1}{\tau_n(1)}\sup_{t\le T}|X_n(\tau_n(t))|\\
&\quad=\sup_{t\le T}\bigg|\frac{A_n(\tau_n(t))}{\tau_n(1)}\frac{1}{A_n(\tau_n(t))}\sum_{i=1}^{A_n(\tau_n(t))}S_i-1\bigg|\\
&\quad\le\sup_{t\le T}\bigg|\frac{A_n(\tau_n(t))}{\tau_n(1)}-\frac{1}{\mathbb{E}[S]}\bigg|\bigg|\frac{1}{A_n(\tau_n(t))}\sum_{i=1}^{A_n(\tau_n(t))}S_i+\sup_{t\le T}\bigg|\frac{1}{\mathbb{E}[S]}\frac{1}{A_n(\tau_n(t))}\sum_{i=1}^{A_n(\tau_n(t))}S_i-1\bigg|.
\end{aligned}
$$

Note that $\tau_n(t)\to\infty$ and $A_n(\tau_n(t))\xrightarrow{\mathbb{P}}\infty$ as $n\to\infty$. Then the second term converges to 0 in probability by the LLN and the first one converges to 0 by the LLN for the Poisson process. Indeed, $A_n(\tau_n(t))=\Pi(\tau_n(t))-R_n(\tau_n(t))$, so that

$$
\begin{aligned}
&\sup_{t\le T}\bigg|\frac{A_n(\tau_n(t))}{\tau_n(1)}-\frac{1}{\mathbb{E}[S]}\bigg|\\
&\quad\le\sup_{t\le T}\bigg|\frac{\Pi(\tau_n(t))}{\tau_n(1)}-\frac{1}{\mathbb{E}[S]}\bigg|+\frac{1}{\tau_n(1)}\sup_{t\le T}|R_n(\tau_n(t))|\\
&\quad=\sup_{t\le T}\bigg|\frac{\Pi(\tau_n(t))}{\tau_n(1)}-\frac{1}{\mathbb{E}[S]}\bigg|+\frac{n^{1/(2\alpha-1)}\ell_n(n)^{-1}}{\tau_n(1)}\sup_{t\le T}\frac{|R_n(\tau_n(t))|}{n^{1/(2\alpha-1)}\ell_n(n)^{-1}}. \qquad (40)
\end{aligned}
$$

As shown above in Proposition 2, $n^{-1/(2\alpha-1)}\ell_2(n)R_n(\tau_n(t))$ converges uniformly to $-\lambda^2/2t^2$, and since $n^{1/(2\alpha-1)}\ell_2(n)^{-1}/\tau_n(1)\to0$, the second term in (40) is negligible. By the heavy-traffic assumption (3) and the LLN for the Poisson process, the first term also converges to 0, completing the proof. □

4.4. Proof of Theorem 1

We now conclude the proof of Theorem 1 by collecting various results from the previous sections. First, we split the process $\boldsymbol{N}_n(\cdot)$ in its martingale and drift components as in (29) to obtain

$$\boldsymbol{N}_n(t)=q_0+\boldsymbol{\Pi}_n(t)+\boldsymbol{\sigma}_n(\hat{\boldsymbol{B}}_n(t))-\boldsymbol{R}_n(t).$$

Since $\boldsymbol{\Pi}_n(\cdot)$ and $\boldsymbol{\sigma}_n(\cdot)$ are independent, and $\hat{\boldsymbol{B}}_n(\cdot)$ and $\boldsymbol{R}_n(\cdot)$ converge to deterministic limits in $\mathcal{D}$, we have

$$(\boldsymbol{\Pi}_n(\cdot),\boldsymbol{\sigma}_n(\cdot),\hat{\boldsymbol{B}}_n(\cdot),\boldsymbol{R}_n(\cdot))\xrightarrow{\mathrm{D}}\bigg(0,s_\alpha\mathcal{S}_\alpha(\cdot),\mathrm{id}(\cdot),\frac{\lambda^2}{2}t^2\bigg)\quad\text{in }(\mathcal{D}^4,M_1^{\mathrm{W}}).$$

This, together with a time-change theorem for processes with discontinuous sample paths (see, e.g. [38, Theorem 13.2.3]) implies that

$$(\boldsymbol{\Pi}_n(\cdot),\boldsymbol{\sigma}_n(\hat{\boldsymbol{B}}_n(\cdot)),\boldsymbol{R}_n(\cdot))\xrightarrow{\mathrm{D}}\bigg(0,s_\alpha\mathcal{S}_\alpha(\cdot),\frac{\lambda^2}{2}t^2\bigg)\quad\text{in }(\mathcal{D}^3,M_1^{\mathrm{W}}). \qquad (41)$$

Note that [38, Theorem 13.2.3] does not hold in general in the finer J_1 topology. Since the three limit processes in (41) do not have common discontinuity points, it follows that addition is continuous in $(0,1/\mathbb{E}[S]^{(\alpha+1)/\alpha}\mathcal{S}_\alpha(\cdot),\lambda^2/2t^2)$ in the M_1 topology, so that

$$\boldsymbol{N}_n(t)\xrightarrow{\mathrm{D}}q_0+s_\alpha\mathcal{S}_\alpha(\cdot)-\frac{\lambda^2}{2}t^2\quad\text{in }(\mathcal{D},M_1).$$

https://doi.org/10.1017/jpr.2017.42 Published online by Cambridge University Press

The second claim (14) follows immediately from the continuous mapping theorem, since the reflection map is Lipschitz continuous in the M_1 topology by [38, Theorem 13.5.1]. □

Extension to general initial drift. Now assume that $c_n = 1 + \beta n^{-(\alpha-1)/(2\alpha-1)}\ell_2(n)^{-1}$, with $\beta \neq 0$. Write (28) as

$$\begin{aligned} N_n(t) &\overset{\mathrm{D}}{=} N_n(0) + \left(\Pi(t) - \frac{t}{\mathbb{E}[S]}\right) + \left(\frac{B_n(t)}{\mathbb{E}[S]} - \sigma(B_n(t))\right) - R_n(t) \\ &= N_n(0) + (\Pi(t) - \lambda c_n t) + \left(\frac{B_n(t)}{\mathbb{E}[S]} - \sigma(B_n(t))\right) - R_n(t) \\ &= N_n(0) + \lambda\beta n^{-(\alpha-1)/(2\alpha-1)}t + (\Pi(t) - \lambda t) + \left(\frac{B_n(t)}{\mathbb{E}[S]} - \sigma(B_n(t))\right) - R_n(t). \end{aligned}$$

In the first equality we have used assumption (3) and in the second $c_n = 1 - \beta n^{-(\alpha-1)/(2\alpha-1)} \times \ell_2(n)^{-1}$. By rescaling the process as in (29), we obtain

$$\boldsymbol{N}_n(t) = q_0 + \lambda\beta t + \boldsymbol{\Pi}_n(t) + \boldsymbol{\sigma}_n(\hat{\boldsymbol{B}}_n(t)) - \boldsymbol{R}_n(t).$$

Since $c_n \to 1$ as $n \to \infty$, the rescaled partial sums of the *double sequence* $(\bar{S}_i/c_n)_{i\geq 1}$ converge to the α-stable motion $\mathcal{S}_\alpha(\cdot)$ since c_n is deterministic and $c_n \to 1$ as $n \to \infty$, hence, Theorem 2 holds and $\boldsymbol{\sigma}_n(\cdot) \to s_\alpha \mathcal{S}_\alpha(\cdot)$. Moreover, $\boldsymbol{\Pi}_n(\cdot)$, $\hat{\boldsymbol{B}}_n(\cdot)$, and $\boldsymbol{R}_n(\cdot)$ can be shown to converge as before, and, thus, the vector of functions $(\boldsymbol{\Pi}_n(\cdot), \boldsymbol{\sigma}_n(\hat{\boldsymbol{B}}_n(\cdot)), \boldsymbol{R}_n(\cdot))$ converges as in (41). We again conclude that

$$\boldsymbol{N}_n(t) \xrightarrow{\mathrm{D}} q_0 + \lambda\beta t + s_\alpha \mathcal{S}_\alpha(\cdot) - \frac{\lambda^2}{2}t^2 \quad \text{in } (\mathcal{D}, M_1),$$

as desired. □

5. Discussion

We have considered a queueing model in which only a finite number of customers can potentially join the system, also referred to as the $\Delta_{(i)}/\mathrm{G}/1$ model [15]. For this model, we have defined a suitable heavy-traffic condition, in which the instantaneous arrival rate is assumed to be equal to the service rate. We have shown that, under the additional assumption that the service times obey a power-law with parameter $\alpha \in (1, 2)$, the queue-length process converges to an α-stable process with negative parabolic drift. To prove this, we have given a novel definition of the arrival process that enables us to obtain explicit bounds on the limiting drift. Using continuity arguments, we have proved that this implies that the length of the first busy period converges in distribution to the first excursion of the stable motion with negative drift. In this paper we have focused on heavy-tailed service times. Thus, Theorem 1 should be compared to the finite-variance case, where the rescaled queue-length process converges to a (reflected) Brownian motion with parabolic drift [4].

Little is known about the (reflected) α-stable motion with negative quadratic drift. In particular, there are no explicit formulas for the maximum of the free process, and it is not known whether the excursions above past minima can be ordered. A striking property of the limiting process that we obtain is that $\sup_{t\geq 0} \phi(a + b\mathcal{S}_\alpha(t) - ct^2) = \infty$ almost surely. Indeed, it is well known that, for any Lévy process $X(\cdot)$ with unbounded Lévy measure,

$$\mathbb{P}(\text{for all } N \in \mathbb{N} \text{ for all } T > 0 \text{ there exists } t \geq T : \Delta X(t) \geq N) = 1,$$

https://doi.org/10.1017/jpr.2017.42 Published online by Cambridge University Press

where $\Delta X(t) := X(t) - \lim_{s \to t^-} X(s)$. However, due to the parabolic drift, the excursions of $\phi(\mathcal{N})(\cdot)$ containing a large jump become smaller as time passes. This suggests that the excursions of $\phi(\mathcal{N})(\cdot)$ can be ordered by their time duration and the largest one is finite. In particular, it should be possible to analytically prove that, for $q_0 > 0$ large enough, the probability that the first busy period is (one of) the largest ones is close to 1. This presents an interesting direction for future research.

Acknowledgements

This work was supported by the NWO Gravitation Networks grant 024.002.003. The work of RvdH was further supported by the NWO VICI grant 639.033.806. The work of JvL was further supported by an NWO TOP-GO grant and by an ERC Starting Grant.

References

[1] Aïdékon, E., van der Hofstad, R., Kliem, S. and van Leeuwaarden, J. S. H. (2016). Large deviations for power-law thinned Lévy processes. *Stoch. Process. Appl.* **126,** 1353–1384.

[2] Aldous, D. (1997). Brownian excursions, critical random graphs and the multiplicative coalescent. *Ann. Prob.* **25,** 812–854.

[3] Asmussen, S. (2003). *Applied Probability and Queues*, 2nd edn. Springer, New York.

[4] Bet, G., van der Hofstad, R. and van Leeuwaarden, J. S. H. (2015). Heavy-traffic analysis through uniform acceleration of queues with diminishing populations. Preprint. Available at https://arxiv.org/abs/1412.5329v2.

[5] Bet, G., van der Hofstad, R. and van Leeuwaarden, J. S. H. (2017). Big jobs arrive early: from critical queues to random graphs. Preprint. Available at https://arxiv.org/abs/1704.03406.

[6] Bhamidi, S., van der Hofstad, R. and van Leeuwaarden, J. S. H. (2012). Novel scaling limits for critical inhomogeneous random graphs. *Ann. Prob.* **40,** 2299–2361.

[7] Billingsley, P. (1999). *Convergence of Probability Measures*, 2nd edn. John Wiley, New York.

[8] Bingham, N. H., Goldie, C. M. and Teugels, J. L. (1987). *Regular Variation*. Cambridge University Press.

[9] Boxma, O. J. and Cohen, J. W. (1998). The M/G/1 queue with heavy-tailed service time distribution. *IEEE J. Selected Areas Commun.* **16,** 749–763.

[10] David, H. A. and Nagaraja, H. N. (2003). *Order Statistics*, 3rd edn. John Wiley, Hoboken, NJ.

[11] Dhara, S., van der Hofstad, R., van Leeuwaarden, J. S. H. and Sen, S. (2016). Heavy-tailed configuration models at criticality. Preprint. Available at https://arxiv.org/abs/1612.00650.

[12] Dhara, S., van der Hofstad, R., van Leeuwaarden, J. S. H. and Sen, S. (2017). Critical window for the configuration model: finite third moment degrees. *Electron. J. Prob.* **22,** 16.

[13] Groeneboom, P. (1989). Brownian motion with a parabolic drift and Airy functions. *Prob. Theory Relat. Fields* **81,** 79–109.

[14] Groeneboom, P. (2010). The maximum of Brownian motion minus a parabola. *Electron. J. Prob.* **15,** 1930–1937.

[15] Honnappa, H., Jain, R. and Ward, A. R. (2014). On transitory queueing. Preprint. Available at https://arxiv.org/abs/1412.2321.

[16] Honnappa, H., Jain, R. and Ward, A. R. (2015). A queueing model with independent arrivals, and its fluid and diffusion limits. *Queueing Systems* **80,** 71–103.

[17] Iglehart, D. L. and Whitt, W. (1970). Multiple channel queues in heavy traffic. I. *Adv. Appl. Prob.* **2,** 150–177.

[18] Iglehart, D. L. and Whitt, W. (1970). Multiple channel queues in heavy traffic. II. Sequences, networks, and batches. *Adv. Appl. Prob.* **2,** 355–369.

[19] Jacod, J. and Shiryaev, A. N. (2003). *Limit Theorems for Stochastic Processes*, 2nd edn. Springer, Berlin.

[20] Janson, S., Louchard, G. and Martin-Löf, A. (2010). The maximum of Brownian motion with parabolic drift. *Electron. J. Prob.* **15,** 1893–1929.

[21] Joseph, A. (2014). The component sizes of a critical random graph with given degree sequence. *Ann. Appl. Prob.* **24,** 2560–2594.

[22] Klenke, A. (2008). *Probability Theory: A Comprehensive Course*. Springer, London.

[23] Louchard, G. (1994). Large finite population queueing systems. The single-server model. *Stoch. Process. Appl.* **53,** 117–145.

[24] Mandelbaum, A. and Massey, W. A. (1995). Strong approximations for time-dependent queues. *Math. Operat. Res.* **20,** 33–64.

[25] Massey, W. A. (1981). Non-stationary queues. Doctoral Thesis, Stanford University.

[26] Massey, W. A. (1985). Asymptotic analysis of the time dependent $M/M/1$ queue. *Math. Operat. Res.* **10,** 305–327.

https://doi.org/10.1017/jpr.2017.42 Published online by Cambridge University Press

[27] Newell, G. F. (1968). Queues with time-dependent arrival rates. III. A mild rush hour. *J. Appl. Prob.* **5,** 591–606.

[28] Pittel, B. (2001). On the largest component of the random graph at a nearcritical stage. *J. Combinatorial Theory B* **82,** 237–269.

[29] Roberts, M. I. (2016). The probability of unusually large components in the near-critical Erdős-Rényi graph. Preprint. Available at https://arxiv.org/abs/1610.05485.

[30] Roberts, M. I. and Sengul, B. (2017). Exceptional times of the critical dynamical Erdős-Rényi graph. Preprint. Available at https://arxiv.org/abs/1610.06000.

[31] Samorodnitsky, G. and Taqqu, M. S. (1994). *Stable Non-Gaussian Random Processes*. Chapman & Hall, New York.

[32] Sato, K.-I. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press.

[33] Simon, T. (2011). Hitting densities for spectrally positive stable processes. *Stochastics* **83,** 203–214.

[34] Skorokhod, A. V. (1956). Limit theorems for stochastic processes. *Theory Prob. Appli.* **1,** 261–290.

[35] Van der Hofstad, R., Janssen, A. J. E. M. and van Leeuwaarden, J. S. H. (2010). Critical epidemics, random graphs, and Brownian motion with a parabolic drift. *Adv. Appl. Prob.* **42,** 1187–1206.

[36] Van der Hofstad, R., Kliem, S. and van Leeuwaarden, J. S. H. (2014). Cluster tails for critical power-law inhomogeneous random graphs. Preprint. Available at https://arxiv.org/abs/1404.1727.

[37] Van der Hofstad, R., van Leeuwaarden, J. S. H. and Stegehuis, C. (2016). Mesoscopic scales in hierarchical configuration models. Preprint. Available at https://arxiv.org/abs/1612.02668.

[38] Whitt, W. (2002). *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer, New York.

[39] Whitt, W. (2016). Heavy-traffic limits for a single-server queue leading up to a critical point. *Operat. Res. Lett.* **44,** 796–800.

[40] Whitt, W. and You, W. (2017). Time-varying robust queueing. Submitted.

https://doi.org/10.1017/jpr.2017.42 Published online by Cambridge University Press