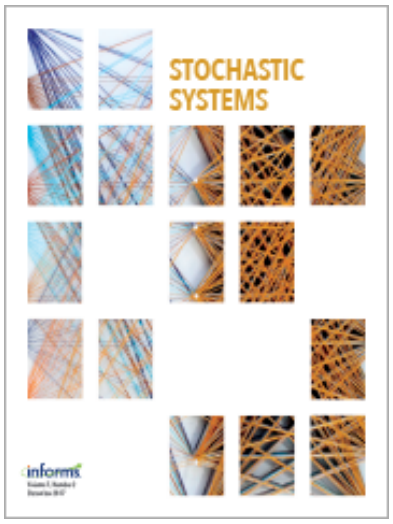
Scaling Limits for Infinite-server Systems in a Random Environment

Heemskerk, M.; van Leeuwaarden, J.; Mandjes, M.

DOI
10.1214/16-SSY214
Publication date
2017
Document Version
Final published version
Published in
Stochastic Systems



Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Scaling Limits for Infinite-server Systems in a Random Environment

Mariska Heemskerk, Johan van Leeuwaarden, Michel Mandjes

To cite this article:
Mariska Heemskerk, Johan van Leeuwaarden, Michel Mandjes (2017) Scaling Limits for Infinite-server Systems in a Random Environment. Stochastic Systems 7(1):1-31. https://doi.org/10.1287/16-SSY214

Stochastic Systems
2017, Vol. 7, No. 1, 1–31
DOI: 10.1214/16-SSY214

SCALING LIMITS FOR INFINITE-SERVER SYSTEMS IN A RANDOM ENVIRONMENT

By Mariska Heemskerk*, Johan van Leeuwaarden† and Michel Mandjes*

*University of Amsterdam** *and Eindhoven University of Technology*†

This paper studies the effect of an overdispersed arrival process on the performance of an infinite-server system. In our setup, a random environment is modeled by drawing an arrival rate Λ from a given distribution every Δ time units, yielding an i.i.d. sequence of arrival rates $\Lambda_1, \Lambda_2, \ldots$. Applying a martingale central limit theorem, we obtain a functional central limit theorem for the scaled queue length process. We proceed to large deviations and derive the logarithmic asymptotics of the queue length's tail probabilities. As it turns out, in a rapidly changing environment (i.e., Δ is small relative to Λ) the overdispersion of the arrival process hardly affects system behavior, whereas in a slowly changing random environment it is fundamentally different; this general finding applies to both the central limit and the large deviations regime. We extend our results to the setting where each arrival creates a job in multiple infinite-server queues.

1. Introduction. Empirical studies show that the number of arrivals in customer contact centers, hospital emergency departments and cloud computing systems typically varies strongly over time [8, 17]. This motivates modeling such arrival processes by a *non-homogeneous Poisson process* (NHPP) with time-dependent arrival rate $\lambda(t)$, see e.g. [9]. At the same time, various studies show that in a broad variety of real-life systems the intensity of the fluctuations in the arrival rate is so severe that the Poisson assumption ceases to hold [2, 8]. The observed level of *overdispersion* urges the need to develop stochastic models that can capture such persistent fluctuations.

Starting from the classical Poisson process, it is common practice to increase dispersion by using a *mixed* Poisson process [2, 13], to that end replacing a deterministic parameter λ by a random parameter Λ. This leads to the idea of modeling overdispersed arrival processes by a mixed version

Received January 2016.

AMS 2000 subject classifications: 60K25, 60F05, 60F10, 60F17, 60H20, 60K37, 97M40, 90B15

Keywords and phrases: Scaling limits, overdispersion, non-Poisson arrival processes, Cox processes, infinite-server queues, central limit theorem, large deviations.

of NHPPs, so-called *Cox processes* [5], where the time-dependent rate $\lambda(t)$ of the classical NHPP is replaced by a stochastic process $\Lambda(t)$. For instance, one could use Markov-modulated Poisson processes (MMPPs) in which the arrival rate $\Lambda(t) = \lambda_{J(t)}$ is a function of a continuous-time Markov chain $J(\cdot)$ on a finite state space S and non-negative rates λ_i for $i \in S$ (see e.g. [1, 3]). Although the MMPP is versatile and has various attractive properties, it has considerable drawbacks as well. First, while a substantial body of results for single-server queues with Markov modulation has been established, considerably less is known about their many-server and infinite-server counterparts; see e.g. an account of this issue for the infinite-server system in [4]. Second, due to the fact that the process $J(\cdot)$ is not observed, estimating the parameters of an MMPP from data is a non-trivial task [15].

The main objective of this paper is to develop an arrival process simpler than a Markov-modulated Poisson process – arguably the simplest in terms of analysis – that fits the overdispersed and time-dependent setting, and to assess the impact of these characteristics on a corresponding system's performance. The model we propose is a *mixed Poisson* arrival process in a *random environment*. It is defined as follows. Let Λ a non-negative random variable with finite first two moments and density $f_\Lambda(\cdot)$. Introduce a *sampling frequency* $\frac{1}{\Delta}$; then the arrival rate at time t is given by Λ_j when $t \in [j\Delta, (j+1)\Delta)$, where the Λ_j are independent and distributed as a non-negative random variable Λ, for $j \in \mathbb{Z}$. In other words, this arrival process is a special case of a stationary Cox process where the arrival rate at time t is given by

$$\Lambda(t) = \sum_j \Lambda_j 1_{[j\Delta,(j+1)\Delta)}(t). \tag{1.1}$$

To add nonstationarity in the arrivals, one could include a deterministic component $\bar{\lambda}(t)$ without intrinsically complicating the analysis; for ease of presentation we omit the extra component here. The resulting process can be viewed as an extension of the classical mixed Poisson setting, which is enriched by (independently) resampling the arrival rate after every time slot of length $\Delta > 0$. The intuition is that the arrival rate changes every Δ time units so that, when observed for a number of consecutive slots, the time between arrivals and hence the number of arrivals per time unit fluctuates more severely than one would expect in a standard Poisson setting. This can be made explicit via an elementary computation. Let the number of arrivals up to time t be given by $N_t \sim \text{Pois}(\int_0^t \Lambda_s \, \mathrm{d}s)$ and for simplicity, let t be an integer multiple of Δ. Then $\mathbb{E}N_t = t\mathbb{E}\Lambda$, whereas

$$\mathbb{V}\mathrm{ar}(N_t) = \sum_{j=1}^{t/\Delta} \mathbb{V}\mathrm{ar}(N_\Delta) = t\Delta^{-1}\big(\mathbb{E}[\mathbb{V}\mathrm{ar}(N_\Delta|\Lambda)] + \mathbb{V}\mathrm{ar}(\mathbb{E}[N_\Delta|\Lambda])\big)$$
$$= t\big(\mathbb{E}\Lambda + \Delta\mathbb{V}\mathrm{ar}(\Lambda)\big).$$

Conclude that, as desired, the variance-to-mean ratio is strictly larger than 1 for non-deterministic Λ, i.e.,

$$\frac{\mathbb{V}\mathrm{ar}(N_t)}{\mathbb{E}N_t} = 1 + \Delta\frac{\mathbb{V}\mathrm{ar}(\Lambda)}{\mathbb{E}\Lambda}. \tag{1.2}$$

Observe that the level of overdispersion is determined by the slot size Δ and the level of overdispersion in Λ (through its variance-to-mean ratio).

Given this model for the arrival process, various queueing models can be studied; in this paper we focus on single-class infinite-server systems with exponential service times. The proposed arrival process being overdispersed, the main objective of this paper is to reveal, in a compact manner, the impact of overdispersion on system performance. Infinite-server systems are a natural choice when the system at hand is designed to (almost) immediately serve all customers [16], but it may also serve as a tractable proxy for the more complicated multi-server systems, which is for instance exploited in the modified offered-load (MOL) and pointwise stationary approximation (PSA) methods for staffing large-scale service systems in a time-varying setting [10, 17].

Contributions. Infinite-server systems with overdispersed arrivals are, as described above, very tractable. As shown in Section 2, it is fairly straightforward to compute the probability generating function (PGF) of the stationary and time-dependent queue length processes (which in an infinite-server setting refers to the number of jobs in the system) in terms of transforms. This is due to the fact that customers are served immediately upon arrival, independently of each other; as a result, when analyzing the queue length at a given point in time, we can separately consider the individual (independent!) contributions that correspond to each of the preceding intervals of length Δ.

The queue length distribution can be characterized in terms of its PGF, which effectively means that evaluation of the accompanying performance measures requires numerical inversion. However, by imposing a scaling on both the time and scale parameters, Δ and Λ, we succeed in identifying an asymptotic regime in which the distribution *can* be explicitly given. We inflate the arrival rate and sampling frequency in the following way:

$$\Lambda \mapsto N\Lambda \qquad \Delta^{-1} \mapsto N^\alpha\Delta^{-1}, \tag{1.3}$$

where we let $N \to \infty$. The scaled counterpart of the variance-to-mean ratio in (1.2) is $1 + N^{1-\alpha}\Delta\mathbb{V}\mathrm{ar}(\Lambda)/\mathbb{E}\Lambda$. Due to the possibility of having different

growth rates for Λ and Δ^{-1} under scaling (1.3) this ratio will not always converge to 1. That is, the value of α determines the nature of the asymptotic behavior of the arrival process, giving rise to a *trichotomy*. In this paper we prove that the queue length process inherits this behavior from the input process. For $\alpha > 1$, in which case the arrival rate is resampled relatively frequently, we find that the system behaves as a standard infinite-server queue (no overdispersion), whereas for $\alpha < 1$ the overdispersion remains present in the asymptotic regime. The case $\alpha = 1$ essentially reflects a superposition of the two distinct types of behavior.

For preparatory purposes, we show in Section 2 that the centered and normalized stationary queue length is asymptotically normal under the scaling in (1.3). Next, in Section 3 we consider a multidimensional setting with correlated arrivals: an arrival triggers jobs in multiple queues. Hence, we work with a coupled system in which d parallel queues are fed by a single arrival process; cf. [11, 12]. With $\boldsymbol{U}^{(N)}(\cdot)$ denoting the vector of centered and normalized queue length processes, the asymptotic normality now translates to the corresponding limiting process $\boldsymbol{U}(\cdot)$ being Gaussian: $\boldsymbol{U}(\cdot)$ is a d-dimensional process of the Ornstein-Uhlenbeck type with parameters that depend on the scaling regime. Following the approach in [1], we show this by applying a lemma due to Kurtz and a martingale central limit theorem (MCLT) to a suitable stochastic integral equation.

Subsequently, in Section 4 we carry out a large deviations analysis to obtain the logarithmic tail asymptotics corresponding to the queue length distribution. The crucial observation in this analysis is that rare events can essentially be realized in two ways: (i) the random arrival rate attains an exceptionally high value, (ii) the Poisson process generates an unusually large number of arrivals given the (not so rare) value of the random parameter. Again, the value of α determines what type of tail behavior dominates: for $\alpha < 1$ this is effect (i), for $\alpha > 1$ effect (ii), and for $\alpha = 1$ a combination of effects (i) and (ii). These findings complement similar results that have been established for an infinite-server system with Markov-modulated input, where it is noted that the slow regime ($\alpha \in (0,1)$) was not covered in that setting [3, 6]. We conclude Section 4 by pointing out how the large deviations results can be extended to the multidimensional setting.

2. Overdispersion in an infinite-server context. In this section we present a stationary and transient analysis of the single-class Markovian infinite-server system in a random environment just introduced. A crucial role is played by $\Lambda(t)$, the arrival rate at time t given in (1.1). Remember that we assumed that the arrival rates are i.i.d. and distributed as a random variable $\Lambda \geqslant 0$ with finite first two moments and density $f_\Lambda(\cdot)$. The corre-

sponding service times are assumed i.i.d. (and in addition independent of the arrival process) exponentially distributed random variables with mean $1/\mu$.

First, in Section 2.1, we analyze the stationary system behavior, in terms of its PGF and the corresponding moments, which we then extend to the associated transient behavior. We then study the stationary behavior in a central limit regime under parameter scaling (1.3) in Section 2.2. This exposition serves as an illustration for the reader, and is intended to create intuition as for why the scaled stationary queue length is asymptotically normal and why the three different limiting regimes appear; in addition, in Section 4 we need a result that is proven along the same lines. We note that in Section 3 the normality is generalized in several directions: we establish a *functional* central limit theorem (FCLT) for the (scaled) transient process $M^{(N)}(\cdot)$ corresponding to the d-dimensional parallel system as defined in the introduction.

2.1. *Pre-limit results.* This subsection presents 'pre-limit results'; later we study their counterparts in the limiting regime after imposing a parameter scaling.

Transform of stationary queue length. Let M be the random variable associated with the stationary number of jobs (also sometimes referred to as 'customers') in the system. Exploiting 'thinning' properties, we can identify the PGF $\phi(z) := \mathbb{E}z^M$ of M.

In the sequel we write $p_t := \mathrm{e}^{-\mu t}$ for the probability that a job present at kt is still present at $(k+1)t$ and $q_t := (1-\mathrm{e}^{-\mu t})/(\mu t)$ for the probability that a job arriving at a uniform epoch in $[kt,(k+1)t)$ is still present at $(k+1)t$. Denote $\bar{p}_t := 1 - p_t$.

Note that M can be written as the sum of $M_0, M_1, M_2, \ldots$, where M_k represents the number of jobs that arrived in $[-(k+1)\Delta, -k\Delta)$ and are still present at time 0. Furthermore, observe that these 'thinned' random variables M_k are independent. A job that arbitrarily arrived in $[-(k+1)\Delta, -k\Delta)$ (i.e., having arrived at a uniform epoch in this interval) is still in the system at time 0 with probability

$$\int_0^\Delta \frac{1}{\Delta}\mathrm{e}^{-\mu(k\Delta+s)}\mathrm{d}s = q_\Delta p_\Delta^k.$$

As a consequence, with $r_t := tq_t$,

$$\begin{aligned}\phi_k(z) &:= \mathbb{E}z^{M_k}\\ &= \sum_{\ell=0}^{\infty}\int_0^\infty f_\Lambda(\lambda)\mathrm{e}^{-\lambda\Delta}\frac{(\lambda\Delta)^\ell}{\ell!}\sum_{m=0}^{\ell} z^m\binom{\ell}{m}\left(q_\Delta p_\Delta^k\right)^m\left(1-q_\Delta p_\Delta^k\right)^{\ell-m}\mathrm{d}\lambda\end{aligned}$$

$$= \int_0^\infty \exp\big(-\lambda r_\Delta p_\Delta^k (1-z)\big) f_\Lambda(\lambda)\,\mathrm{d}\lambda$$

$$= \mathbb{E}\exp\big(-\Lambda_k r_\Delta p_\Delta^k (1-z)\big). \tag{2.1}$$

Observe that $\phi_k(z)$ is a PGF of 'mixed Poisson' type: conditional on $\Lambda_k = \lambda$ the PGF corresponds with that of a Poisson random variable with mean $\lambda r_\Delta p_\Delta^k$. We conclude that M_k is distributed as a mixed Poisson random variable with random parameter

$$\kappa_k(\Lambda_k) := \Lambda_k r_\Delta p_\Delta^k,$$

with Λ_k the value of the arrival rate in the interval $[-(k+1)\Delta, -k\Delta)$ (note that, in fact, we should have written $\Lambda_{-(k+1)}$ rather than Λ_k, but due to the i.i.d. assumption the processes $\{\Lambda(s)\}_{s\geqslant 0}$ and $\{\Lambda(-s)\}_{s\geqslant 0}$ have the same finite-dimensional distributions). Therefore, M is mixed Poisson as well and its random parameter is given by

$$\sum_{k=0}^\infty \kappa_k(\Lambda_k) = \int_0^\infty \Lambda(s)\mathrm{e}^{-\mu s}\,\mathrm{d}s =: \kappa(\Lambda). \tag{2.2}$$

(Note that $\kappa(\cdot)$ is defined as a functional; $\kappa(\Lambda)$ should be interpreted as $\kappa(\Lambda(\cdot))$.)

There is an alternative way to obtain this result. Indeed, since we observe the system in stationarity,

$$\phi(z) = \phi(zp_\Delta + \bar{p}_\Delta) g_{\Lambda,\Delta}(z), \tag{2.3}$$

where $g_{\Lambda,t}(z)$ is defined by

$$g_{\Lambda,t}(z) := \int_0^\infty \exp\big(-\lambda r_t(1-z)\big) f_\Lambda(\lambda)\,\mathrm{d}\lambda = \mathbb{E}\exp\big(-\Lambda r_t(1-z)\big).$$

Applying an iteration argument to (2.3) yields

$$\phi(z) = \prod_{k=0}^\infty g_{\Lambda,\Delta}\big(zp_\Delta^k + \bar{p}_\Delta \sum_{j=0}^{k-1} p_\Delta^j\big) = \prod_{k=0}^\infty g_{\Lambda,\Delta}\big(1-(1-z)p_\Delta^k\big). \tag{2.4}$$

In the factors $g_{\Lambda,\Delta}\big(1-(1-z)p_\Delta^k\big)$ we recognize the expression for $\phi_k(z)$ as in (2.1).

First two moments. We now evaluate the first two moments of M. This is an interesting computation in its own right, but it also provides useful results that can be exploited when considering this system under the central limit scaling (as is done in the next subsection).

Differentiating (2.3) and letting $z \uparrow 1$, we obtain a fixed-point equation,

$$\phi'(1) = \phi'(1)\mathrm{e}^{-\mu\Delta} + g'_{\Lambda,\Delta}(1) = \phi'(1)\mathrm{e}^{-\mu\Delta} + r_\Delta\, \mathbb{E}\Lambda.$$

Hence $\mathbb{E}M = \phi'(1) = r_\Delta\, \mathbb{E}\Lambda/(1-\mathrm{e}^{-\mu\Delta}) = \mathbb{E}\Lambda/\mu$. This quantity could have been computed more directly as well, using a standard identity for conditional means:

$$\mathbb{E}M = \sum_{k=0}^{\infty} \mathbb{E}M_k = \sum_{k=0}^{\infty} \mathbb{E}\big[\mathbb{E}[M_k \,|\, \Lambda_k]\big]. \tag{2.5}$$

Then observe that $(M_k \,|\, \Lambda_k)$ is Poisson, and hence its mean equals its parameter. As a result, (2.5) equals

$$\mathbb{E}M = \sum_{k=0}^{\infty} \mathbb{E}[\kappa_k(\Lambda_k)] = \mathbb{E}[\Lambda] r_\Delta \sum_{k=0}^{\infty} p_\Delta^k = \mathbb{E}\Lambda/\mu.$$

For the variance we use that

$$\phi''(1) = \phi''(1)p_\Delta^2 + 2\phi'(1)p_\Delta g'_{\Lambda,\Delta}(1) + g''_{\Lambda,\Delta}(1),$$

and hence

$$\phi''(1) = \frac{2\phi'(1)\, p_\Delta\, g'_{\Lambda,\Delta}(1)}{1-p_\Delta^2} + \frac{g''_{\Lambda,\Delta}(1)}{1-p_\Delta^2} = 2\frac{(\mathbb{E}\Lambda)^2}{\mu^2}\frac{p_\Delta}{1+p_\Delta} + \frac{\mathbb{E}\Lambda^2}{\mu^2}\frac{1-p_\Delta}{1+p_\Delta}.$$

It thus follows that, after some algebra,

$$\begin{aligned} \mathbb{V}\mathrm{ar}M &= \phi''(1) + \phi'(1) - (\phi'(1))^2 \\ &= \mathbb{E}\Lambda/\mu + C\, \mathbb{V}\mathrm{ar}\Lambda/\mu^2, \end{aligned} \tag{2.6}$$

where $C := (1-p_\Delta)/(1+p_\Delta)$.

Alternatively, one could use the 'law of total variance' to identify $\mathbb{V}\mathrm{ar}M$:

$$\mathbb{V}\mathrm{ar}M = \sum_{k=0}^{\infty} \mathbb{V}\mathrm{ar}M_k = \sum_{k=0}^{\infty} \mathbb{E}[\mathbb{V}\mathrm{ar}(M_k \,|\, \Lambda_k)] + \sum_{k=0}^{\infty} \mathbb{V}\mathrm{ar}(\mathbb{E}[M_k \,|\, \Lambda_k]). \tag{2.7}$$

Observe that, because of the 'mixed Poisson property', $\mathbb{E}[\mathbb{V}\mathrm{ar}(M_k \,|\, \Lambda_k)] = \mathbb{E}[\kappa_k(\Lambda_k)]$, and as a result the first term at the right-hand side of (2.7) equals $\mathbb{E}M$. The second term, which is inherently non-negative, gives rise to 'overdispersion', i.e., the effect that the variance of the stationary queue length *exceeds* the corresponding mean. This is a distinguishing feature compared to the analogous system in which the Poissonian arrival rate is deterministic: the stationary queue length in an M/M/∞ system is Poisson, and

cannot accommodate any overdispersion. In order to evaluate the second term in the right-hand side of (2.7), we note that

$$\mathbb{V}\mathrm{ar}(\mathbb{E}(M_k \mid \Lambda_k)) = \mathbb{V}\mathrm{ar}\big(\kappa_k(\Lambda_k)\big) = r_\Delta^2 p_\Delta^{2k} \cdot \mathbb{V}\mathrm{ar}\Lambda. \tag{2.8}$$

Substituting (2.8) in the second term in the right-hand side of (2.7), we find that $\mathbb{V}\mathrm{ar}\, M$ equals (2.6), as desired.

Formula (2.6) lends itself to a nice interpretation: the term $\mathbb{E}\Lambda/\mu$ is the contribution to the variance that one would have if the arrival rate would have had the *deterministic* value $\mathbb{E}\Lambda$, whereas the term $C\,\mathbb{V}\mathrm{ar}\Lambda/\mu^2$ needs to be added in order to deal with the non-Poisson variability due to the stochasticity of the arrival rate.

Transient behavior. As the analysis of the transient system behavior strongly resembles its stationary counterpart, we restrict ourselves to a short account of this. We let the system start empty (for ease of presentation; a non-empty initial condition can be analyzed without any additional difficulty). Denote by $M(t)$ the number of jobs present at time t. Then, for n the smallest integer such that $t - n\Delta < \Delta$,

$$M(t) = \sum_{j=0}^{n-1} \bar{M}_j + \bar{M}_{[n\Delta,t)},$$

where $\bar{M}_j$ ($\bar{M}_{[n\Delta,t)}$) represents the number of jobs that have arrived between in $[j\Delta,(j+1)\Delta)$ ($[n\Delta,t)$) and are still around at $n\Delta$ (t). As before, these have PGFs

$$\begin{aligned} \mathbb{E}z^{\bar{M}_j} &= \mathbb{E}\exp\big(-\Lambda r_\Delta p_\Delta^{n-(j+1)} \mathrm{e}^{-\mu(t-n\Delta)}(1-z)\big); \\ \mathbb{E}z^{\bar{M}_{[n\Delta,t)}} &= \mathbb{E}\exp\big(-\Lambda/\mu(1-\mathrm{e}^{-\mu(t-n\Delta)})(1-z)\big). \end{aligned}$$

As the individual random variables $\bar{M}_1, \bar{M}_2, \ldots$ and $\bar{M}_{[n\Delta,t)}$ are independent, $M(t)$ is mixed Poisson with random parameter

$$\kappa_t(\Lambda) := \int_0^t \Lambda(s)\mathrm{e}^{-\mu s}\,\mathrm{d}s. \tag{2.9}$$

2.2. *Limit results.* This section focuses on the central limit regime that results from simultaneously scaling, in a controlled way, both the arrival rate Λ and the sampling frequency $\frac{1}{\Delta}$ as in (1.3). Let the scaled counterpart of $\Lambda(t)$ be $N\Lambda^{(N)}(t)$, with

$$\Lambda^{(N)}(t) := \sum_{j=0}^{\infty} \Lambda_j 1_{[j\Delta N^{-\alpha},(j+1)\Delta N^{-\alpha})}(t). \tag{2.10}$$

That is, the sampling frequency and the arrival rates are both inflated as we let N tend to ∞, but, importantly, at rates that are not necessarily identical. As mentioned in the introduction, depending on the value of α, we obtain fundamentally different behavior.

We consider a sequence of systems indexed by N, where the N-scaled system uses a mixed Poisson arrival process with time-dependent random rate $N\Lambda^{(N)}(t)$. Let $M^{(N)}$ denote the stationary queue length in the N-scaled system, with parameter $N\kappa(\Lambda^{(N)})$ (cf. (2.2)). We start our exposition by a preliminary calculation, in which we compute the mean and variance of $M^{(N)}$ and study their behavior for large N, which indeed reveals the announced trichotomy. Then, after centering and normalizing $M^{(N)}$, we derive a central limit theorem.

Qualitative behavior of first two moments: trichotomy in variance. First, we identify the steady-state mean and variance in our scaling regime, using (2.5), (2.7) and (2.8). We find that

$$\mathbb{E}M^{(N)} = N\mathbb{E}\Lambda/\mu; \tag{2.11}$$

$$\mathbb{V}\mathrm{ar}M^{(N)} = N\mathbb{E}\Lambda/\mu + N^2\frac{1-\mathrm{e}^{-\mu\Delta N^{-\alpha}}}{1+\mathrm{e}^{-\mu\Delta N^{-\alpha}}}\mathbb{V}\mathrm{ar}\Lambda/\mu^2, \tag{2.12}$$

where it is noted that for large N, (2.12) behaves approximately as

$$N\mathbb{E}\Lambda/\mu + N^{2-\alpha}\Delta\mathbb{V}\mathrm{ar}\Lambda/(2\mu)$$

(the ratio of the two converges to 1). We thus observe the trichotomy

$$\mathbb{V}\mathrm{ar}\, M^{(N)} \sim \begin{cases} N\mathbb{E}\Lambda/\mu & \text{if } \alpha > 1; \\ N^{2-\alpha}\Delta\mathbb{V}\mathrm{ar}\,\Lambda/(2\mu) & \text{if } \alpha < 1; \\ N\big(\mathbb{E}\Lambda/\mu + \Delta\mathbb{V}\mathrm{ar}\,\Lambda/(2\mu)\big) & \text{if } \alpha = 1. \end{cases} \tag{2.13}$$

For $\alpha > 1$, the sampling frequency dominates the variability of Λ. Consequently, the model behaves essentially as an M/M/∞ system, with the variance of $M^{(N)}$ being linear in N and equal to $\mathbb{E}M^{(N)}$, for large N. For $\alpha < 1$, we find a superlinear relation between N and $\mathbb{V}\mathrm{ar}\,\Lambda$, and both the sampling frequency (i.e., the reciprocal of the interval length Δ) and the variance of Λ play a role. Hence, the asymptotic variance indeed grows faster than the asymptotic mean for $\alpha < 1$; in this regime the system is overdispersed. For $\alpha = 1$, the variance is 'slightly larger' than for $\alpha > 1$, but it is still linear in N. In this case the sampling frequency and the variance of Λ grow at the same rate, so that the variance for $M^{(N)}$ combines the effects observed in the two former cases.

As observed from the above computation, the variance of $M^{(N)}$ is essentially proportional to N^γ with $\gamma := \max\{1, 2-\alpha\}$. As a consequence, one may expect that, under (1.3),

$$\check{M}^{(N)} := N^{-\gamma/2}(M^{(N)} - \mathbb{E}\, M^{(N)}) \tag{2.14}$$

converges to a (zero-mean) normally distributed random variable. It is this property that we verify now.

Asymptotic normality. We show how to establish asymptotic normality for the centered and normalized version of $M^{(N)}$ in (2.14) via evaluation of the corresponding Laplace transform. Appealing to Lévy's convergence theorem, we establish the desired convergence in distribution. For simplicity, the proof of Thm. 2.1 assumes that all moments of Λ are finite; however, as will appear from the proof of Thm. 3.2 only finiteness of the first two moments is necessary.

THEOREM 2.1 (CLT). *As $N \to \infty$, $\check{M}^{(N)}$ converges to a zero-mean normally distributed random variable with variance*

$$\sigma^2 := \frac{\mathbb{E}\Lambda}{\mu}\, 1_{\{\alpha \geqslant 1\}} + \frac{\Delta \mathbb{V}\mathrm{ar}\, \Lambda}{2\mu}\, 1_{\{\alpha \leq 1\}}.$$

PROOF. Let $\phi^{(N)}(z)$ be the counterpart of (2.4) under scaling as in (1.3); likewise $g^{(N)}_{\Lambda,\Delta}(z)$ is the counterpart of $g_{\Lambda,\Delta}(z)$. Then

$$\phi^{(N)}(z) = \prod_{k=0}^{\infty} g^{(N)}_{\Lambda,\Delta}\big(1 - (1-z)\mathrm{e}^{-\mu k N^{-\alpha}\Delta}\big).$$

We are interested in the behavior of $M^{(N)}$ in the central limit regime, hence we need to analyze the limiting distribution of $\check{M}^{(N)}$. To this end, we evaluate the logarithm of the corresponding Laplace transform:

$$\log \mathbb{E} \exp\left(-sN^{-\gamma/2}(M^{(N)} - \mathbb{E}M^{(N)})\right) = sN^{1-\gamma/2}\mathbb{E}\Lambda/\mu + \log \phi^{(N)}\big(\mathrm{e}^{-sN^{-\gamma/2}}\big). \tag{2.15}$$

We now use that $\log \phi^{(N)}(\mathrm{e}^{-sN^{-\gamma/2}})$ equals

$$\sum_{k=0}^{\infty} \log \mathbb{E}\mathrm{e}^{-N\Lambda/\mu\left(1-\mathrm{e}^{-\mu N^{-\alpha}\Delta}\right)\left(1-\mathrm{e}^{-sN^{-\gamma/2}}\right)\mathrm{e}^{-\mu k N^{-\alpha}\Delta}}. \tag{2.16}$$

Observe that (2.16) is the sum of cumulant generating functions (which is again a cumulant generating function), each of them related to the random

variable Λ but evaluated at different arguments. Let m_ℓ denote the ℓ-th cumulant of Λ (for $\ell \in \mathbb{N}$); in particular $m_1 = \mathbb{E}\Lambda$ and $m_2 = \mathbb{V}\mathrm{ar}\,\Lambda$. In addition, we define

$$\zeta_k^{(N)}(s) := -N/\mu\big(1 - \mathrm{e}^{-\mu N^{-\alpha}\Delta}\big)\big(1 - \mathrm{e}^{-sN^{-\gamma/2}}\big)\mathrm{e}^{-\mu k N^{-\alpha}\Delta}.$$

Then it follows that

$$\log \phi^{(N)}\big(\mathrm{e}^{-sN^{-\gamma/2}}\big) = \sum_{\ell=1}^{\infty} \frac{m_\ell}{\ell!} \sum_{k=0}^{\infty} \big(\zeta_k^{(N)}(s)\big)^\ell. \tag{2.17}$$

Let us first consider the contribution of the term corresponding to $\ell = 1$. Observe that, as $N \to \infty$,

$$\begin{aligned} m_1 \sum_{k=0}^{\infty} \zeta_k^{(N)}(s) &= -N\,\mathbb{E}\Lambda/\mu\big(1 - e^{-sN^{-\gamma/2}}\big) \\ &\sim -sN^{1-\gamma/2}\mathbb{E}\Lambda/\mu + \frac{1}{2}s^2 N^{1-\gamma}\mathbb{E}\Lambda/\mu. \end{aligned} \tag{2.18}$$

Note that the first term in (2.18) is canceled by the first term in the right-hand side of (2.15), so that we are left with the second term, i.e.,

$$\frac{1}{2}s^2 N^{1-\gamma}\mathbb{E}\Lambda/\mu. \tag{2.19}$$

The second term in (2.17) corresponding to $\ell = 2$ gives

$$\frac{1 - \mathrm{e}^{-\mu N^{-\alpha}\Delta}}{1 + \mathrm{e}^{-\mu N^{-\alpha}\Delta}} \frac{(1 - \mathrm{e}^{-sN^{-\gamma/2}})^2}{2\mu^2} N^2\,\mathbb{V}\mathrm{ar}\Lambda \sim \frac{1}{2}s^2 N^{2-\alpha-\gamma}\,\Delta\,\mathbb{V}\mathrm{ar}\Lambda/(2\mu). \tag{2.20}$$

Now compare the asymptotic expansion identified in (2.19) and (2.20). In case $\alpha > 1$, we have that $\gamma = 1$, so that (2.19) equals $\frac{1}{2}s^2\,\mathbb{E}\Lambda/\mu$, whereas (2.20) converges to zero. On the other hand, for $\alpha < 1$ we have $\gamma = 2-\alpha$, and hence (2.19) converges to zero, whereas (2.20) behaves as $\frac{1}{2}s^2\Delta\,\mathbb{V}\mathrm{ar}\Lambda/(2\mu)$. Finally, if $\alpha = 1$, we find that both terms converge to the expected finite positive limit.

We now check that the terms in (2.17) for $\ell \geqslant 3$ vanish as $N \to \infty$. For large N the terms can be approximated as follows,

$$\sum_{k=0}^{\infty} \big(\zeta_k^{(N)}(s)\big)^\ell \sim N^\ell \frac{\big(1 - e^{-\mu N^{-\alpha}\Delta}\big)^\ell}{1 - e^{-\ell\mu N^{-\alpha}\Delta}}\big(1 - e^{-sN^{-\gamma/2}}\big)^\ell \sim N^\ell \frac{\mu^\ell N^{-\alpha\ell}\Delta^\ell}{\ell\mu N^{-\alpha}\Delta} \frac{s^\ell}{N^{\gamma\ell/2}},$$

hence being of order N^δ with $\delta = \delta(\alpha) := \ell(1 - \alpha - \gamma/2) + \alpha$. In case $\alpha \geqslant 1$,

we get (bearing in mind that $\gamma = 1$ and $\ell \geqslant 1$)

$$\delta = \ell(\frac{1}{2} - \alpha) + \alpha = \frac{1}{2}\ell + \alpha(1-\ell) \leqslant \frac{1}{2}\ell + 1 - \ell = 1 - \frac{\ell}{2};$$

on the other hand, in case $\alpha < 1$ we get $\delta = -\ell\alpha/2 + \alpha = \alpha(1 - \ell/2)$ (with $\gamma = 2 - \alpha$). We conclude that $\delta < 0$ for $\ell \geqslant 3$ and the corresponding terms in (2.17) can indeed be neglected. We have therefore established that, as $N \to \infty$,

$$\log \mathbb{E} \exp\big(- sN^{-\gamma/2}(M^{(N)} - \mathbb{E}M^{(N)})\big) \to \frac{1}{2}\sigma^2 s^2,$$

as claimed. □

3. Functional central limit theorem. In this section we generalize the central limit result of Thm. 2.1 in two ways. First, we establish the functional version: the centered and normalized transient queue length process converges to a limiting process of Ornstein-Uhlenbeck type with parameters that depend on the value of α. Second, we extend this to a multidimensional setting with correlated arrivals: every arrival triggers jobs in multiple queues. The correlation structure of the resulting multidimensional Gaussian limiting process is explicitly identified.

Let us start by describing the mechanics of the generalized setting. We consider a parallel system in which d queues are fed by a *single* arrival process that was constructed in the same way as the one in the previous section: a Markovian process with arrival rate $\Lambda(t)$ as in (1.1). The service times in queue i are i.i.d. exponential random variables with mean μ_i^{-1}; the service processes of the individual queues are independent, and also independent of the arrival process. We perform the same scaling as before: the sampling frequency is sped up by a factor N^α, while the (random) arrival rate is blown up by a factor N. This results in a mixed Poisson arrival process with time-dependent rate $\Lambda^{(N)}(t)$ as in (2.10). Let

$$\boldsymbol{M}^{(N)}(t) = (M_1^{(N)}(t), \ldots, M_d^{(N)}(t))^{\mathrm{T}},$$

where $M_i^{(N)}(t)$ is the queue length at time t in the i-th queue of the N-scaled system, for $i \in \{1, \ldots, d\}$. Note that the $M_i^{(N)}(t)$ are mixed Poisson with time-dependent random parameter $N\kappa_{t,i}(\Lambda^{(N)})$, with $\kappa_{t,i}(\Lambda^{(N)})$ as defined in (2.9) but with μ replaced by μ_i.

We now present an alternative way of writing $M_i^{(N)}(t)$, which facilitates the use of a martingale central limit theorem. We introduce the functional

$$\Psi[X](t) := \int_0^t X(s)\,\mathrm{d}s,$$

mapping the stochastic process $\{X(s) : s \in [0,t]\}$ to a real number; then $\mu_i\,\Psi[M_i^{(N)}](t)$ is to be interpreted as the 'cumulative service capacity' in queue i over the interval $[0,t]$. In addition, for the 'cumulative arrival rate' we have $\Psi[\Lambda](t)$, with scaled counterpart $N\Psi[\Lambda^{(N)}](t)$. By the law of large numbers, $\Psi[\Lambda](t)/t$ converges a.s. to $\mathbb{E}\Lambda$ as $t \to \infty$ and for fixed t, $\Psi[\Lambda^{(N)}](t)$ converges a.s. to $t\,\mathbb{E}\Lambda$ as $N \to \infty$.

With $Y_0(\cdot), \ldots, Y_d(\cdot)$ denoting independent unit-rate Poisson processes,

$$M_i^{(N)}(t) \stackrel{\mathrm{d}}{=} M_i^{(N)}(0) + Y_0\big(N\Psi[\Lambda^{(N)}](t)\big) - Y_i\big(\mu_i\,\Psi[M_i^{(N)}](t)\big). \tag{3.1}$$

Our objective is to derive a d-dimensional FCLT for $\boldsymbol{M}^{(N)}(\cdot)$. This result characterizes the time-dependent queue length in the scaled system and makes explicit how the correlated arrivals lead to correlation between the individual queue length processes. It will be stated and proved in subsection 3.2; first we study the stationary behavior by presenting the corresponding first two moments of $\boldsymbol{M}^{(N)}$ (including the covariances).

3.1. *Qualitative behavior of first two moments in stationarity.* Note that the individual queue lengths are only coupled through the arrival process, so under (1.3) the mean and variance of $\boldsymbol{M}^{(N)}$ are, as in (2.11) and (2.12), given by

$$\begin{aligned}
\mathbb{E}M_i^{(N)} &= N\mathbb{E}\Lambda/\mu_i,\\
\mathbb{V}\mathrm{ar}M_i^{(N)} &= N\mathbb{E}\Lambda/\mu_i + N^2\mathbb{V}\mathrm{ar}\Lambda/\mu_i^2\frac{1-p_i(\Delta N^{-\alpha})}{1+p_i(\Delta N^{-\alpha})}\\
&\sim N\mathbb{E}\Lambda/\mu_i + N^{2-\alpha}\Delta\mathbb{V}\mathrm{ar}\Lambda/(2\mu_i),
\end{aligned}$$

for $i = 1, \ldots, d$. Hence, we find the same behavior as in (2.13). Interestingly, the same trichotomy is observed for the covariances between the individual queues, as stated in the next lemma.

LEMMA 3.1 (Covariance in $\boldsymbol{M}^{(N)}$). *For $i,k \in \{1,\ldots,d\}$ with $i \neq k$, and for large N,*

$$\mathbb{C}\mathrm{ov}(M_i^{(N)}, M_k^{(N)}) \sim \begin{cases} N\mathbb{E}\Lambda/(\mu_i+\mu_k) & \text{if } \alpha > 1;\\ N^{2-\alpha}\Delta\mathbb{V}\mathrm{ar}(\Lambda)/(\mu_i+\mu_k) & \text{if } \alpha < 1;\\ N\big(\mathbb{E}\Lambda/(\mu_i+\mu_k) + \Delta\mathbb{V}\mathrm{ar}(\Lambda)/(\mu_i+\mu_k)\big) & \text{if } \alpha = 1. \end{cases} \tag{3.2}$$

PROOF. Without loss of generality, we take $i = 1$ and $k = 2$. We first consider the non-scaled model, by studying the joint probability generating

function,

$$\mathbb{E}z_1^{M_1(n\Delta)}z_2^{M_2(n\Delta)} = \prod_{j=0}^{n-1}\xi_{jn}(z_1,z_2),$$

where $\xi_{jn}(z_1,z_2)$ is the contribution due to the slot between $j\Delta$ and $(j+1)\Delta$; as z_1 and z_2 are held fixed for the moment, we suppress them. Now we introduce functions (for $\ell = 1,2$)

$$f_\ell(r,n) := \mathrm{e}^{-\mu_\ell(n\Delta - r)}, \quad g_j(\mu,n) := \tfrac{1}{\mu\Delta}(1-\mathrm{e}^{-\mu\Delta})\,\mathrm{e}^{-\mu(n-j)\Delta},$$

where it is noted that $g_j(\mu,n)$ behaves as $\mathrm{e}^{-\mu(n-j)\Delta}$ for small Δ. In addition, we define the quantities

$$\begin{aligned}
\zeta_{jn}^{++} &:= \int_{j\Delta}^{(j+1)\Delta} \tfrac{1}{\Delta} f_1(r,n)f_2(r,n)\mathrm{d}r = g_j(\mu_1+\mu_2,n),\\
\zeta_{jn}^{+-} &:= \int_{j\Delta}^{(j+1)\Delta} \tfrac{1}{\Delta} f_1(r,n)(1-f_2(r,n))\mathrm{d}r = g_j(\mu_1,n) - g_j(\mu_1+\mu_2,n),\\
\zeta_{jn}^{-+} &:= \int_{j\Delta}^{(j+1)\Delta} \tfrac{1}{\Delta} (1-f_1(r,n))f_2(r,n)\mathrm{d}r = g_j(\mu_2,n) - g_j(\mu_1+\mu_2,n),\\
\zeta_{jn}^{--} &:= \int_{j\Delta}^{(j+1)\Delta} \tfrac{1}{\Delta} (1-f_1(r,n))(1-f_2(r,n))\mathrm{d}r\\
&\quad = 1 - g_j(\mu_1,n) - g_j(\mu_2,n) + g_j(\mu_1+\mu_2,n).
\end{aligned}$$

Using arguments similar to those we have used before,

$$\begin{aligned}
\xi_{jn} &:= \mathbb{E}\big(\sum_{m=0}^{\infty}\mathrm{e}^{-\lambda\Delta}\frac{(\lambda\Delta)^m}{m!}\big(\zeta_{jn}^{++}z_1z_2 + \zeta_{jn}^{+-}z_1 + \zeta_{jn}^{-+}z_2 + \zeta_{jn}^{--}\big)^m\big)\\
&= \mathbb{E}\exp\big(\Lambda\Delta\big(\zeta_{jn}^{++}z_1z_2 + \zeta_{jn}^{+-}z_1 + \zeta_{jn}^{-+}z_2 + \zeta_{jn}^{--} - 1\big)\big)\\
&\sim \mathbb{E}\exp\big(\Lambda\Delta\big(\prod_{i=1}^{2}\big((z_i-1)\mathrm{e}^{-\mu_i(n-j)\Delta}+1\big) - 1\big)\big) \text{ for small } \Delta.
\end{aligned}$$

Because the contributions to $M_1(n\Delta)$ and $M_2(n\Delta)$ resulting from different time intervals are independent, we obtain that

$$\begin{aligned}
&\mathbb{C}\mathrm{ov}\big(M_1(n\Delta), M_2(n\Delta)\big)\\
&\quad = \sum_{j=0}^{n-1}\big(\frac{\partial^2}{\partial z_1\partial z_2}\xi_{jn}(z_1,z_2) - \frac{\partial}{\partial z_1}\xi_{jn}(z_1,z_2)\frac{\partial}{\partial z_2}\xi_{jn}(z_1,z_2)\big)\Big|_{z_1\uparrow 1, z_2\uparrow 1}.
\end{aligned}$$

Now imposing scaling (1.3) and considering the stationary behavior by letting $n \to \infty$, it is readily derived that (for large N)

$$\mathbb{E}\, z_1^{M_1^{(N)}} z_2^{M_2^{(N)}} \sim \prod_{j=0}^{\infty} \mathbb{E} \exp\big(\Lambda \Delta N^{1-\alpha} \big(\prod_{\ell=1}^{2} \big((z_\ell - 1) e^{-\mu_\ell j \Delta / N^\alpha} + 1\big) - 1\big)\big);$$

observe that, for reasons of symmetry, it is allowed to replace $n - j$ by j in the definition of the $g_j(\mu, n)$. We thus arrive at

$$\begin{aligned}\mathbb{C}\mathrm{ov}\big(M_1^{(N)}, M_2^{(N)}\big) \sim \sum_{j=0}^{\infty} \big((\mathbb{E}\Lambda\, \Delta N^{1-\alpha} + \mathbb{E}[\Lambda^2] \Delta^2 N^{2-2\alpha}) \mathrm{e}^{-(\mu_1+\mu_2) j \Delta / N^\alpha} \\ - \prod_{\ell=1}^{2} \mathbb{E}\Lambda\, \Delta N^{1-\alpha} \mathrm{e}^{-\mu_\ell j \Delta / N^\alpha} \big) \\ = \frac{\mathbb{E}\Lambda\, \Delta N^{1-\alpha} + \mathbb{V}\mathrm{ar}(\Lambda)\, \Delta^2 N^{2-2\alpha}}{1 - \mathrm{e}^{-(\mu_1+\mu_2)\Delta N^{-\alpha}}},\end{aligned}$$

which behaves in accordance with (3.2) for N large. □

Recall that $\gamma = \max\{1, 2 - \alpha\}$; the above computation shows that the covariance matrix of $\boldsymbol{M}^{(N)}$ is essentially proportional to N^γ. Therefore, we expect that the centered and normalized version of the joint stationary queue length process converges to a (zero-mean) d-dimensional Gaussian random vector with covariance matrix C such that

$$C_{ik} = \begin{cases} 1_{\{\alpha \leqslant 1\}} \mathbb{E}\Lambda / \mu_i + 1_{\{\alpha \geqslant 1\}} \Delta \mathbb{V}\mathrm{ar}(\Lambda) / (2\mu_i) & \text{if } i = k, \\ 1_{\{\alpha \leqslant 1\}} \mathbb{E}\Lambda / (\mu_i + \mu_k) + 1_{\{\alpha \geqslant 1\}} \Delta \mathbb{V}\mathrm{ar}(\Lambda) / (\mu_i + \mu_k) & \text{if } i \neq k, \end{cases}$$

for $i, k \in \{1, \ldots, d\}$. This is verified in the next subsection.

3.2. *Proof of functional central limit theorem based on* MCLT. The main objective of this subsection is to derive a functional limit theorem for $\boldsymbol{M}^{(N)}(t)$, the vector describing the queue lengths of the scaled system at time t. To this end, we consider the process $\tilde{M}_i^{(N)}(\cdot) := M_i^{(N)}(\cdot)/N$, for which we have

(3.3)
$$\tilde{M}_i^{(N)}(t) = \tilde{M}_i^{(N)}(0) + N^{-1} Y_0\big(N \Psi[\Lambda^{(N)}](t)\big) - N^{-1} Y_i\big(N \mu_i\, \Psi[\tilde{M}^{(N)}](t)\big).$$

We will need the following lemma, which uses the law of large numbers for Poisson processes; see [1].

LEMMA 3.2. *Let Y be a unit-rate Poisson process. Then for any $U > 0$, almost surely*

$$\lim_{N\to\infty} \sup_{0\leqslant u\leqslant U} \left| \frac{Y(Nu)}{N} - u \right| = 0.$$

The uniform convergence in Lemma 3.2 entails that (3.3) converges almost surely to the solution of the functional equation

$$\varrho_i(t) = \varrho_{i,0} + t\,\mathbb{E}\Lambda - \mu_i\,\Psi[\varrho_i](t), \tag{3.4}$$

as $N \to \infty$, under the proviso that $\tilde{M}_i^{(N)}(0)$ converges a.s. to some value $\varrho_{i,0}$ for $i = 1, \ldots, d$. The solution is given by a convex mixture of the initial position $\varrho_i(0) = \varrho_{i,0}$ and the limiting value $\mathbb{E}\Lambda/\mu_i$:

$$\varrho_i(t) = \varrho_{i,0}\mathrm{e}^{-\mu_i t} + \frac{\mathbb{E}\Lambda}{\mu_i}\big(1 - \mathrm{e}^{-\mu_i t}\big). \tag{3.5}$$

Having identified this *fluid limit*, the next objective is to establish an FCLT for the centered and normalized process $\boldsymbol{U}^{(N)}(\cdot)$ given by

$$U_i^{(N)}(t) := N^{\frac{\beta}{2}}\big(\tilde{M}_i^{(N)}(t) - \varrho_i(t)\big), \tag{3.6}$$

with $\beta := 2 - \gamma = \min\{1, \alpha\}$. Here we closely follow the approach in [1], where the idea is to use an MCLT, so as to obtain weak convergence to a (generalized) Ornstein-Uhlenbeck process. The version of the MCLT that we need in our setting is stated below.

THEOREM 3.1 (MCLT, [1]). *Let $\{\boldsymbol{R}^{(N)}\}_{N\in\mathbb{N}}$ be a sequence of $\mathbb{R}^d$-valued martingales. Assume that the following condition on the jump sizes is met:*

$$\lim_{N\to\infty} \mathbb{E}\big[\sup_{s\leqslant t} \big|\boldsymbol{R}^{(N)}(s) - \boldsymbol{R}^{(N)}(s^-)\big|\big] = 0; \tag{3.7}$$

in addition, assume that, as $N \to \infty$,

$$\big[R_i^{(N)}, R_k^{(N)}\big]_t \to C_{ik}(t)$$

for a deterministic function $C_{ik}(t)$, continuous in t for all $t > 0$ and for $i, k = 1, \ldots, d$. Then the process $\boldsymbol{R}^{(N)}$ converges weakly to a centered Gaussian process $\boldsymbol{W}$ with independent increments whose covariance matrix is characterized by

$$\mathbb{E}\big[W_i(t) \cdot W_k(t)^{\mathrm{T}}\big] = C_{ik}(t).$$

Introducing compensated unit-rate Poisson processes $\tilde{Y}_i(t) := Y_i(t) - t$, we define

$$\check{\boldsymbol{Y}}_0^{(N)}(t) := N^{\frac{\beta}{2}-1} \begin{pmatrix} \tilde{Y}_0\big(N\Psi[\Lambda^{(N)}](t)\big) \\ \vdots \\ \tilde{Y}_0\big(N\Psi[\Lambda^{(N)}](t)\big) \end{pmatrix}, \tag{3.8}$$

$$\check{\boldsymbol{Y}}^{(N)}(t) := N^{\frac{\beta}{2}-1} \begin{pmatrix} \tilde{Y}_1\big(N\mu_1\,\Psi[\tilde{M}_1^{(N)}](t)\big) \\ \vdots \\ \tilde{Y}_d\big(N\mu_d\,\Psi[\tilde{M}_d^{(N)}](t)\big) \end{pmatrix}. \tag{3.9}$$

LEMMA 3.3. *Consider the d-dimensional processes $\check{\boldsymbol{Y}}_0^{(N)}(\cdot)$ and $\check{\boldsymbol{Y}}^{(N)}(\cdot)$. If $\alpha \geqslant 1$, then as $N \to \infty$ these processes converge weakly to d-dimensional zero-mean Brownian motions with covariance matrices $K_0(t) := (t\,\mathbb{E}\Lambda)\mathbf{1}\mathbf{1}^{\mathrm{T}}$ and $K(t) := \operatorname{diag}\{\mu_1\,\Psi[\varrho_1](t), \ldots, \mu_d\,\Psi[\varrho_d](t)\}$, respectively; if $\alpha < 1$ the limiting covariance matrices equal $\mathbf{0}$.*

PROOF. We start by checking the conditions of Thm. 3.1. First, observe that for each N, $\check{\boldsymbol{Y}}_0^{(N)}(\cdot)$ and $\check{\boldsymbol{Y}}^{(N)}(\cdot)$ are d-dimensional real-valued martingales. Also, condition (3.7) is met, as both for $\boldsymbol{R}^{(N)} = \check{\boldsymbol{Y}}_0^{(N)}$ and $\boldsymbol{R}^{(N)} = \check{\boldsymbol{Y}}^{(N)}$,

$$\lim_{N\to\infty} \mathbb{E}\big[\sup_{s\leqslant t} \big|\boldsymbol{R}^{(N)} - \boldsymbol{R}^{(N)}(s^-)\big|\,\big] < \infty,$$

whereas $N^{\frac{\beta}{2}-1} \leqslant 1/\sqrt{N}$ converges to zero.

Note that $\beta = \min\{1, \alpha\}$, so that $\beta - 2 = \min\{-1, \alpha - 2\}$. Now observe that for $\alpha \geqslant 1$ (and hence $\beta - 2 = -1$) the quadratic covariation of $\check{\boldsymbol{Y}}_0^{(N)}(\cdot)$,

$$\big[N^{\frac{\beta}{2}-1}\tilde{Y}_0\big(N\Psi[\Lambda^{(N)}](t)\big)\big]_t = N^{\beta-2}Y_0\big(N\Psi[\Lambda^{(N)}](t)\big),$$

converges to $t\,\mathbb{E}\Lambda$ as $N \to \infty$ (0 for $\alpha < 1$), by virtue of Lemma 3.2. The covariance matrix for $\check{\boldsymbol{Y}}^{(N)}(\cdot)$ is determined in the same way; for $\alpha \geqslant 1$ the diagonal entries are given by

$$\begin{aligned} \lim_{N\to\infty} \big[N^{\frac{\beta}{2}-1}\tilde{Y}_i\big(N\mu_i\,\Psi[\tilde{M}_i^{(N)}](t)\big)\big]_t &= \lim_{N\to\infty} N^{\beta-2}Y_i\big(N\mu_i\,\Psi[\tilde{M}^{(N)}](t)\big) \\ &= \mu_i\,\Psi[\varrho_i](t) \end{aligned}$$

(which would equal 0 for $\alpha < 1$), whereas for $i \neq k$ (then $\tilde{Y}_i(\cdot)$ and $\tilde{Y}_k(\cdot)$ are independent)

$$\lim_{N\to\infty} \big[N^{\frac{\beta}{2}-1}\tilde{Y}_i\big(N\mu_i\,\Psi[\tilde{M}_i^{(N)}](t)\big), N^{\frac{\beta}{2}-1}\tilde{Y}_k\big(N\mu_k\,\Psi[\tilde{M}_k^{(N)}](t)\big)\big]_t = 0,$$

with $i,k \in \{1,\ldots,d\}$. For $\alpha \geqslant 1$, Thm. 3.1 yields that the processes converge weakly to d-dimensional Brownian motions with covariance matrices $K_0(t)$ and $K(t)$. On the other hand, for $\alpha < 1$ the entries of the covariance matrices all vanish as $N \to \infty$. As a result, both $\check{\boldsymbol{Y}}_0^{(N)}(\cdot)$ and $\check{\boldsymbol{Y}}^{(N)}(\cdot)$ converge to a process identical to $\mathbf{0}$. □

Stated below is the main theorem of this section: an FCLT for $\boldsymbol{U}^{(N)}(\cdot)$, the process defined via (3.6). In line with earlier findings, three regimes need to be distinguished: $\alpha > 1$ (the fast regime), $\alpha < 1$ (the slow regime) and $\alpha = 1$ (the intermediate regime).

THEOREM 3.2 (FCLT). *As $N \to \infty$, $\boldsymbol{U}^{(N)}(\cdot)$ converges weakly to a zero-mean d-dimensional Gaussian process with covariance matrix given by*

(3.10)
$$C_{ii}(t) := 1_{\{\alpha\geqslant 1\}}\big(\mathbb{E}\Lambda/\mu_i + \varrho_{i,0}\mathrm{e}^{-\mu_i t}\big)(1-\mathrm{e}^{-\mu_i t}) + 1_{\{\alpha\leqslant 1\}}\Delta\mathbb{V}\mathrm{ar}\,\Lambda/(2\mu_i)(1-\mathrm{e}^{-2\mu_i t}),$$

(3.11)
$$C_{ik}(t) := \big(1_{\{\alpha\geqslant 1\}}\mathbb{E}\Lambda/(\mu_i+\mu_k) + 1_{\{\alpha\leqslant 1\}}\Delta\mathbb{V}\mathrm{ar}\,\Lambda/(\mu_i+\mu_k)\big)\cdot(1-\mathrm{e}^{-(\mu_i+\mu_k)t}),$$

for $i \neq k$ $(i,k \in \{1,\ldots,d\})$.

PROOF. Using (3.3), we write

(3.12)
$$U_i^{(N)}(t) = N^{\frac{\beta}{2}}\big(\tilde{M}_i^{(N)}(0) + N^{-1}Y_0\big(N\Psi[\Lambda^{(N)}](t)\big) - N^{-1}Y_i\big(N\mu_i\,\Psi[\tilde{M}_i^{(N)}](t)\big) - \varrho_i(t)\big),$$

for $i = 1,\ldots,d$. Adding and subtracting $\varrho_{i,0}$, (3.12) is equivalent to

$$\begin{aligned} U_i^{(N)}(t) =& N^{\frac{\beta}{2}}\big(\tilde{M}_i^{(N)}(0) - \varrho_{i,0}\big) - N^{\frac{\beta}{2}}\big(\varrho_i(t) - \varrho_{i,0}\big) \\ &+ N^{\frac{\beta}{2}-1}\big(Y_0\big(N\Psi[\Lambda^{(N)}](t)\big) - Y_i\big(N\mu_i\,\Psi[\tilde{M}_i^{(N)}](t)\big)\big), \end{aligned}$$

which, by filling out the implicit form of $\varrho_i(t)$ as in (3.4), simplifies to

$$U_i^{(N)}(t) = U_{0,i}^{(N)}(t) + U_1^{(N)}(t) + U_{2,i}^{(N)}(t),$$

with

$$U_{0,i}^{(N)}(t) := U_i^{(N)}(0) - \mu_i\,\Psi[U_i^{(N)}](t),$$
$$U_1^{(N)}(t) := N^{\frac{\beta}{2}}\big(\Psi[\Lambda^{(N)}](t) - t\,\mathbb{E}\Lambda\big),$$
$$U_{2,i}^{(N)}(t) := N^{\frac{\beta}{2}-1}\big(\tilde{Y}_0\big(N\Psi[\Lambda^{(N)}](t)\big) - \tilde{Y}_i\big(N\mu_i\,\Psi[\tilde{M}_i^{(N)}](t)\big)\big).$$

We consider the three individual components separately.

(i) Component $U_{0,i}^{(N)}(t)$ consists of the starting value of the process, which is assumed to converge to some value $U_i(0)$, minus a reverting term. It is now straightforward that, as $N \to \infty$, $U_{0,i}^{(N)}(\cdot)$ converges to $U_{0,i}(t) = U_i(0) - \mu_i\,\Psi[U_i](t)$.

(ii) Then consider $U_1^{(N)}(\cdot)$. For $\alpha \leq 1$ (and hence $\frac{\beta}{2} = \frac{\alpha}{2}$), the standard functional central limit theorem for partial sums of i.i.d. random variables entails that, as $N \to \infty$,

$$U_1^{(N)}(\cdot) \to \sqrt{\Delta\,\mathbb{V}\mathrm{ar}\,\Lambda}\cdot V(\cdot),$$

with $V(\cdot)$ a standard Brownian motion. On the other hand, for $\alpha > 1$ the limiting process is identical to $\mathbf{0}$, as a consequence of $\frac{\beta}{2} = \frac{1}{2} < \frac{\alpha}{2}$.

(iii) Finally, from Lemma 3.3, we conclude that $\boldsymbol{U}_2^{(N)}(\cdot)$ converges weakly to a d-dimensional zero-mean Brownian motion with covariance matrix $K_0(t) + K(t)$ for $\alpha \geqslant 1$, and to $\mathbf{0}$ else.

Using the above observations, we can now complete the proof. Each of the three regimes will be considered separately.

1. Fast regime ($\alpha > 1$). We have obtained above that $\boldsymbol{U}^{(N)}(\cdot)$ converges weakly to the solution $\boldsymbol{U}(\cdot)$ of the d-dimensional stochastic integral equation given by

$$U_i(t) = U_i(0) - \mu_i\,\Psi[U_i](t) + W_i(t\,\mathbb{E}\Lambda + \mu_i\,\Psi[\varrho_i](t)) \text{ for } i = 1,\ldots,d$$

with $W_1(\cdot),\ldots,W_d(\cdot)$ standard Brownian motions (but not independent), or equivalently

$$U_i(t) = U_i(0) - \mu_i\,\Psi[U_i](t) + \tilde{W}_0(t\,\mathbb{E}\Lambda) + \tilde{W}_i\big(\mu_i\,\Psi[\varrho_i](t)\big)$$

with $\tilde{W}_0(\cdot), \tilde{W}_1(\cdot),\ldots,\tilde{W}_d(\cdot)$ independent standard Brownian motions. It takes a routine calculation to derive that

$$U_i(t) = \mathrm{e}^{-\mu_i t}\big(U_i(0) + \int_0^t \sqrt{\mathbb{E}\Lambda + \mu_i\varrho_i(s)}\,\mathrm{e}^{\mu_i s}\,\mathrm{d}W_i(s)\big).$$

All linear combinations of the $U_i(\cdot)$ are Gaussian processes, so we conclude that this d-dimensional process is Gaussian. It is readily seen that $\mathbb{E}\,U_i(t) = U_i(0)\mathrm{e}^{-\mu_i t}$. For the variance, an elementary computation gives

$$\mathbb{V}\mathrm{ar}\,U_i(t) = \mathrm{e}^{-2\mu_i t}\Big(\int_0^t \big(\mathbb{E}\Lambda + \mu_i \varrho_i(s)\big)\mathrm{e}^{2\mu_i s}\,\mathrm{d}s\Big) = \big(\mathbb{E}\Lambda/\mu_i + \varrho_{i,0}\mathrm{e}^{-\mu_i t}\big)(1-\mathrm{e}^{-\mu_i t}).$$

Likewise, for the covariance, with

$$\mathcal{U}_i(t) := \sqrt{\mathbb{E}\Lambda}\int_0^t \mathrm{e}^{\mu_i s}\,\mathrm{d}\tilde{W}_0(s) + \int_0^t \sqrt{\mu_i \varrho_i(s)}\mathrm{e}^{\mu_i s}\,\mathrm{d}\tilde{W}_i(s),$$

it follows that, for $i \neq k$,

$$\begin{aligned}
\mathbb{C}\mathrm{ov}(U_i(t), U_k(t)) &= \mathrm{e}^{-(\mu_i+\mu_k)t}\,\mathbb{E}\left[\mathcal{U}_i(t)\,\mathcal{U}_k(t)\right] \\
&= \mathrm{e}^{-(\mu_i+\mu_k)t}\,\mathbb{E}\Lambda\cdot\mathbb{E}\Big[\int_0^t \mathrm{e}^{-\mu_i s}\,\mathrm{d}\tilde{W}_0(s)\cdot\int_0^t \mathrm{e}^{-\mu_k s}\,\mathrm{d}\tilde{W}_0(s)\Big] \\
&= \mathrm{e}^{-(\mu_i+\mu_k)t}\,\mathbb{E}\Lambda\int_0^t \mathrm{e}^{(\mu_i+\mu_k)s}\,\mathrm{d}s \\
&= \mathbb{E}\Lambda/(\mu_i+\mu_k)(1-\mathrm{e}^{-(\mu_i+\mu_k)t}).
\end{aligned}$$

This shows (3.10) for $\alpha > 1$.

2. Slow regime ($\alpha < 1$). In the slow regime, $\boldsymbol{U}^{(N)}(\cdot)$ converges to the solution of

$$U_i(t) = U_i(0) - \mu_i\,\Psi[U_i](t) + \big(\sqrt{\Delta\,\mathbb{V}\mathrm{ar}\,\Lambda}\big)V(t) \;\text{ for } i = 1,\ldots,d,$$

which can be written as

$$\mathrm{d}U_i(t) = -\mu_i U_i(t)\,\mathrm{d}t + \big(\sqrt{\Delta\,\mathbb{V}\mathrm{ar}\,\Lambda}\big)\,\mathrm{d}V(t).$$

Therefore the $U_i(\cdot)$ are Ornstein-Uhlenbeck processes given by:

$$U_i(t) = \mathrm{e}^{-\mu_i t}\big(U_i(0) + \int_0^t \sqrt{\Delta\mathbb{V}\mathrm{ar}(\Lambda)}\,\mathrm{e}^{\mu_i s}\,\mathrm{d}V(s)\big).$$

As before, we can conclude that this d-dimensional process is Gaussian with expectation vector given by $U_i(0)\mathrm{e}^{-\mu_i t}$. Computations as above reveal that for $\alpha < 1$, as claimed in (3.10),

$$\mathbb{C}\mathrm{ov}(U_i(t), U_k(t)) = \Delta\mathbb{V}\mathrm{ar}(\Lambda)/(\mu_i+\mu_k)(1-\mathrm{e}^{-(\mu_i+\mu_k)t}).$$

3. Intermediate regime ($\alpha = 1$). In this regime, a combination of the processes from the other cases appears:

$$\mathrm{d}U_i(t) = -\mu_i U_i(t)\,\mathrm{d}t + \sqrt{\mathbb{E}\Lambda}\,\mathrm{d}\tilde{W}_0(t) + \sqrt{\mu_i\varrho_i(t)}\,\mathrm{d}\tilde{W}_i(t) + \sqrt{\Delta\mathbb{V}\mathrm{ar}(\Lambda)}\,\mathrm{d}V(t).$$

The marginal solutions are, for $i = 1, \ldots, d$, equal to

$$U_i(t) = \mathrm{e}^{-\mu_i t}\big(U_i(0) + \int_0^t \sqrt{\mathbb{E}\Lambda}\,\mathrm{e}^{\mu_i s}\,\mathrm{d}\tilde{W}_0(s) + \int_0^t \sqrt{\mu_i \varrho_i(s)}\,\mathrm{e}^{\mu_i s}\,\mathrm{d}\tilde{W}_i(s) + \int_0^t \sqrt{\Delta\mathbb{V}\mathrm{ar}(\Lambda)}\,e^{\mu_i s}\,\mathrm{d}V(s)\big).$$

Again, we conclude that this d-dimensional process is Gaussian with expectation vector given by $U_i(0)e^{-\mu_i t}$; routine computations yield the desired covariance matrix, as given in (3.10) and (3.11). This completes the proof. □

It is interesting to study the impact of the scaling parameter α on the correlation between the individual queue lengths. Remarkably, it turns out that for $\alpha \neq 1$ this correlation depends on the service rates only, whereas for $\alpha = 1$ also the first and second moment of Λ play a role; see the following corollary for a result on the stationary regime.

COROLLARY 3.1 (Correlation coefficients). *In stationarity, the correlation coefficient for $i \neq k$ satisfies*

$$\lim_{N\to\infty} \mathbb{C}\mathrm{orr}(M_i^{(N)}, M_k^{(N)}) = c_{ik}(\alpha) \cdot \frac{\sqrt{\mu_i \mu_k}}{\mu_i + \mu_k}, \tag{3.13}$$

for some constant $c_{ik}(\alpha) \in [1, 2]$. The constant $c_{ik}(\alpha)$ equals 1 for $\alpha > 1$ and 2 for $\alpha < 1$.

PROOF. From Thm. 3.2, as $t \to \infty$,

$$C_{ik}(t) \to 1_{\{\alpha \geqslant 1\}} \frac{\mathbb{E}\Lambda}{\mu_i + \mu_k 1_{\{i\neq k\}}} + 1_{\{\alpha \leqslant 1\}} \frac{\Delta\mathbb{V}\mathrm{ar}(\Lambda)}{\mu_i + \mu_k}.$$

We observe that (3.13) holds, with

$$c_{ik}(\alpha) = \frac{\mathbb{E}\Lambda\, 1_{\{\alpha\geqslant 1\}} + \Delta\mathbb{V}\mathrm{ar}(\Lambda)\, 1_{\{\alpha\leqslant 1\}}}{\mathbb{E}\Lambda\, 1_{\{\alpha\geqslant 1\}} + \frac{1}{2}\Delta\mathbb{V}\mathrm{ar}(\Lambda)\, 1_{\{\alpha\leqslant 1\}}}.$$

□

4. Large deviations. Where the previous section studied the random-environment infinite-server system under the *central limit scaling*, we now focus on the *large deviations domain*. As it turns out, the previously observed trichotomy remains valid. The results again translate to the setting with d coupled queues; for ease we first present (and prove) the results for $d = 1$, to return to the coupled model at the end of the section.

4.1. *Univariate large deviations.* Let the arrival rate of the N-scaled model again be given by $N\Lambda^{(N)}(t)$ (see (2.10)). An important quantity in our analysis is

$$\begin{aligned}\kappa_t\big(\Lambda^{(N)}\big) &= \int_0^t \Lambda^{(N)}(s)\mathrm{e}^{-\mu s}\,\mathrm{d}s \\ &= \frac{1}{\mu}(1-\mathrm{e}^{-\mu\Delta N^{-\alpha}}) \sum_{j=0}^{\lfloor t/(\Delta N^{-\alpha})\rfloor-1} \Lambda_j \mathrm{e}^{-\mu j\Delta N^{-\alpha}} + o_p(1),\end{aligned}$$

as $N\to\infty$. As observed earlier, $M^{(N)}(t)$ is a mixed Poisson random variable, with random parameter distributed as $N\kappa_t(\Lambda^{(N)})$. In the large deviations setting we are interested in the tail probabilities of $M^{(N)}(t)$ for given t and N large. More specifically, our objective is to evaluate the decay rate

$$\lim_{N\to\infty} N^{-\beta}\log\mathbb{P}\big(M^{(N)}(t)/N\geqslant a\big),$$

for any $a>\rho(t)=\rho(1-\mathrm{e}^{-\mu t})$ (where $\rho:=\lambda/\mu$) and some specific $\beta>0$. Given the results obtained in the central limit regime, we expect that $\beta=\min\{1,\alpha\}$.

The main idea behind our analysis is to condition on the value of the random Poisson parameter. In self-evident notation,

$$\begin{aligned}\mathbb{P}\big(M^{(N)}(t)/N\geqslant a\big) &= \mathbb{P}(\mathrm{Pois}(N\kappa_t(\Lambda^{(N)})\geqslant Na) \\ &= \int_0^\infty \mathbb{P}(\mathrm{Pois}(Nx)\geqslant Na)\mathbb{P}(\kappa_t(\Lambda^{(N)})\in\mathrm{d}x). \qquad (4.1)\end{aligned}$$

In some parts of the analysis we rely on the following lemma, in which we establish a large deviation result for $\mathbb{P}(\kappa_t(\Lambda^{(N)})\geqslant a)$.

LEMMA 4.1. *Let $a>\rho(t)$. Then, with $M(\theta):=\mathbb{E}\,\mathrm{e}^{\theta\Lambda}$,*

$$\lim_{N\to\infty}\Delta N^{-\alpha}\log\mathbb{P}(\kappa_t(\Lambda^{(N)})\geqslant a) = -\sup_{\theta>0}\big(\theta a-\int_0^t \log M(\theta\,\mathrm{e}^{-\mu s})\,\mathrm{d}s\big) \qquad (4.2)$$

PROOF. As a first step, we define a proxy for $\kappa_t(\Lambda^{(N)})$ that is easier to work with:

$$k_t\big(\Lambda^{(N)}\big) := \Delta N^{-\alpha}\sum_{j=0}^{\lfloor t/(\Delta N^{-\alpha})\rfloor-1}\Lambda_j\mathrm{e}^{-\mu j\Delta N^{-\alpha}}; \qquad (4.3)$$

later we show that $\kappa_t(\Lambda^{(N)})$ and $k_t(\Lambda^{(N)})$ are 'close enough'. Let $P_N(a):=\mathbb{P}(k_t(\Lambda^{(N)})\geqslant a)$. Writing, for arbitrary $\theta>0$,

$$P_N(a) = \mathbb{P}\big(\mathrm{e}^{\theta k_t(\Lambda^{(N)})/(\Delta N^{-\alpha})} \geqslant \mathrm{e}^{\theta a/(\Delta N^{-\alpha})}\big)$$
$$= \mathbb{P}\Big(\prod_{j=0}^{\lfloor t/(\Delta N^{-\alpha})\rfloor -1} \mathrm{e}^{\theta\,\Lambda_j \mathrm{e}^{-\mu j\,\Delta N^{-\alpha}}} \geqslant \mathrm{e}^{\theta a/(\Delta N^{-\alpha})}\Big),$$

Markov's inequality immediately yields the upper bound

$$P_N(a) \leqslant \mathrm{e}^{-\theta a/(\Delta N^{-\alpha})} \prod_{j=0}^{\lfloor t/(\Delta N^{-\alpha})\rfloor -1} M(\theta \mathrm{e}^{-\mu j\,\Delta N^{-\alpha}}).$$

Recognizing a Riemann sum, we thus obtain

$$\limsup_{N\to\infty} \Delta N^{-\alpha} \log P_N(a) \leqslant \limsup_{N\to\infty} \Delta N^{-\alpha} \sum_{j=0}^{\lfloor t/(\Delta N^{-\alpha})\rfloor -1} \log M(\theta \mathrm{e}^{-\mu j\,\Delta N^{-\alpha}}) - \theta a$$
$$= \int_0^t \log M(\theta \mathrm{e}^{-\mu s})\,\mathrm{d}s - \theta a.$$

As the established upper bound holds for any $\theta > 0$,

$$\limsup_{N\to\infty}\ \Delta N^{-\alpha} \log P_N(a) \leqslant \inf_{\theta>0} \Big(\int_0^t \log M(\theta \mathrm{e}^{-\mu s})\,\mathrm{d}s - \theta a\Big). \tag{4.4}$$

The next goal is to prove that the above upper bound is tight. We do so by first noting that, due to the convexity of the function involved, the infimum in the right-hand side of (4.4) is attained by $\theta^\star$, being the unique solution to

$$\frac{\partial}{\partial\theta} \int_0^t \log M(\theta \mathrm{e}^{-\mu s})\,\mathrm{d}s \bigg|_{\theta=\theta^\star} = a.$$

Now we apply a change-of-measure technique. Define a measure $\mathbb{Q}$ by exponential twisting; the density of the Λ_j is changed into

$$\mathbb{Q}(\Lambda_j \in \mathrm{d}x) := \frac{\mathrm{e}^{\theta^\star \mathrm{e}^{-\mu j\,\Delta N^{-\alpha}} x}}{M(\theta^\star \mathrm{e}^{-\mu j\,\Delta N^{-\alpha}})} \mathbb{P}(\Lambda_j \in \mathrm{d}x).$$

Fix an $\varepsilon > 0$, and let the event $\mathcal{E}_a^{(N)} := \{k_t(\Lambda^{(N)}) \in [a, a+\varepsilon)\}$. Then

$$P_N(a) \;=\; \mathbb{E}_{\mathbb{Q}}\big[1_{\mathcal{E}_a^{(N)}} \prod_{j=0}^{\lfloor t/(\Delta N^{-\alpha})\rfloor -1} M(\theta^\star \mathrm{e}^{-\mu j\,\Delta N^{-\alpha}})\, \mathrm{e}^{-\theta^\star \Lambda_j \mathrm{e}^{-\mu j\,\Delta N^{-\alpha}}}\big]$$
$$\geqslant\; \mathbb{Q}\big(k_t(\Lambda^{(N)}) \in [a, a+\varepsilon)\big)\, \mathrm{e}^{-\theta^\star (a+\varepsilon)/(\Delta N^{-\alpha})} \prod_{j=0}^{\lfloor t/(\Delta N^{-\alpha})\rfloor -1} M(\theta^\star \mathrm{e}^{-\mu j\,\Delta N^{-\alpha}}).$$

To obtain that $\mathbb{Q}(k_t(\Lambda^{(N)}) \in [a, a+\varepsilon)) \to \frac{1}{2}$ as $N \to \infty$, we now show that $k_t(\Lambda^{(N)})$ is asymptotically normal. It is verified that $\mathbb{E}_{\mathbb{Q}} k_t(\Lambda^{(N)}) \to a$ as $N \to \infty$, due to the specific construction of the measure $\mathbb{Q}$. Also,

$$\begin{aligned}\lim_{N\to\infty} \mathbb{V}\mathrm{ar}_{\mathbb{Q}}(k_t(\Lambda^{(N)})) &= \lim_{N\to\infty} \mathbb{V}\mathrm{ar}_{\mathbb{Q}}\big(\Delta N^{-\alpha} \sum_{j=0}^{\lfloor t/(\Delta N^{-\alpha})\rfloor -1} \Lambda_j \, \mathrm{e}^{-\mu j \Delta N^{-\alpha}}\big) \\ &= \lim_{N\to\infty} \Delta N^{-\alpha} \big(\Delta N^{-\alpha} \sum_{j=0}^{\lfloor t/(\Delta N^{-\alpha})\rfloor -1} \mathrm{e}^{-2\mu j \Delta N^{-\alpha}}\big) \mathbb{V}\mathrm{ar}_{\mathbb{Q}}(\Lambda) \\ &= \lim_{N\to\infty} \Delta N^{-\alpha} \int_0^t \mathrm{e}^{-2\mu s} \mathrm{d}s \, \mathbb{V}\mathrm{ar}_{\mathbb{Q}}(\Lambda) = 0.\end{aligned}$$

Copying the approach – using cumulant generating functions – underlying the proof of Theorem 2.1, it is readily derived that

$$N^{\frac{\alpha}{2}}(k_t(\Lambda^{(N)}) - a) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2), \quad \text{where} \quad \sigma^2 := \Delta/(2\mu)(1 - \mathrm{e}^{-2\mu t}) \mathbb{V}\mathrm{ar}_{\mathbb{Q}}(\Lambda).$$

Hence,

$$\begin{aligned}\liminf_{N\to\infty} \Delta N^{-\alpha} \log \mathbb{P}_N(a) &\geqslant \liminf_{N\to\infty} \Delta N^{-\alpha} \log \mathbb{Q}\big(k_t(\Lambda^{(N)}) \in [a, a+\varepsilon)\big) \\ &\qquad - \theta^\star(a+\varepsilon) + \int_0^t \log M(\theta^\star \mathrm{e}^{-\mu s}) \, \mathrm{d}s \\ &\geqslant \int_0^t \log M(\theta^\star \mathrm{e}^{-\mu s}) \, \mathrm{d}s - \theta^\star(a+\varepsilon).\end{aligned}$$

By letting $\varepsilon \downarrow 0$, together with the upper bound this leads to

$$\lim_{N\to\infty} \Delta N^{-\alpha} \log P_N(a) = -\sup_{\theta>0} \big(\theta a - \int_0^t \log M(\theta \mathrm{e}^{-\mu s}) \, \mathrm{d}s\big). \tag{4.5}$$

Now it remains to show that $k_t(\Lambda^{(N)})$ can again be replaced by $\kappa_t(\Lambda^{(N)})$ (which we abbreviate for compactness to k_t and κ_t). Note that, as $N \to \infty$,

$$|\kappa_t - k_t| = |\big(\frac{1 - \mathrm{e}^{-\mu \Delta N^{-\alpha}}}{\mu \Delta N^{-\alpha}} - 1\big) \Delta N^{-\alpha} \sum_{j=0}^{\lfloor t/(\Delta N^{-\alpha})\rfloor -1} \Lambda_j \mathrm{e}^{-\mu j \Delta N^{-\alpha}} + o_p(1)| = o_p(1).$$

Let $\eta > 0$ small enough to guarantee $a - \eta > \rho(t)$. Then $\mathbb{P}(\kappa_t \in (k_t - \eta, k_t + \eta)) \to 1$ as $N \to \infty$, hence

$$\begin{aligned}\lim_{N\to\infty} \Delta N^{-\alpha} \log \mathbb{P}_N(a+\eta) &\leqslant \lim_{N\to\infty} \Delta N^{-\alpha} \log \mathbb{P}(\kappa_t \geqslant a) \\ &\leqslant \lim_{N\to\infty} \Delta N^{-\alpha} \log \mathbb{P}_N(a-\eta),\end{aligned}$$

which provides bounds for the decay rate of interest of the form

$$-\sup_{\theta>0}\Big(\int_0^t \log M\mathrm{e}^{-\mu s})\,\mathrm{d}s-\theta(a\pm\eta)\Big). \tag{4.6}$$

The rate function in (4.6) is continuous in η, so now letting $\eta\downarrow 0$ yields (4.2). □

As in the central limit regime, we distinguish between the cases $\alpha>1$, $\alpha=1$, and $\alpha<1$. For all three cases we derive the logarithmic asymptotics.

1. Fast regime ($\alpha>1$). We can bound (4.1) from below by

$$\mathbb{P}(\mathrm{Pois}(N(\rho(t)-\varepsilon))\geqslant Na)\cdot\mathbb{P}(\kappa_t(\Lambda^{(N)})\geqslant\rho(t)-\varepsilon), \tag{4.7}$$

for some $\varepsilon\in(0,a-\rho(t))$. As N tends to infinity, it is directly shown that the second factor in (4.7) converges to 1, and hence has exponential decay rate 0. Now an application of Cramér's theorem [7] yields

$$\begin{aligned}\liminf_{N\to\infty} N^{-1}\log\mathbb{P}(\mathrm{Pois}(N\,\kappa_t(\Lambda^{(N)}))\geqslant Na)&\\ \geqslant\lim_{N\to\infty}N^{-1}\log\mathbb{P}(\mathrm{Pois}(N(\rho(t)-\varepsilon))\geqslant Na)&\\ =-\sup_\theta\big(\theta a-(\rho(t)-\varepsilon)(\mathrm{e}^\theta-1)\big)&\\ =a\log\big(\frac{\rho(t)-\varepsilon}{a}\big)-(\rho(t)-\varepsilon)+a.&\end{aligned}$$

On the other hand, (4.1) is majorized by

$$\mathbb{P}(\mathrm{Pois}(N(\rho(t)+\varepsilon))\geqslant Na)+\mathbb{P}(\kappa_t(\Lambda^{(N)})\geqslant\rho(t)+\varepsilon). \tag{4.8}$$

By Cramér's theorem, the first term in (4.8) decays exponentially in N. As a consequence of Lemma 4.1, the second term decays exponentially in N^α, i.e., superexponentially in N. This yields

$$\begin{aligned}\limsup_{N\to\infty} N^{-1}\log\mathbb{P}(\mathrm{Pois}(N\kappa_t(\Lambda^{(N)})\geqslant Na)&\\ \leqslant\lim_{N\to\infty}N^{-1}\log\mathbb{P}(\mathrm{Pois}(N(\rho(t)+\varepsilon))\geqslant Na)&\\ =a\log\big(\frac{\rho(t)+\varepsilon}{a}\big)-(\rho(t)+\varepsilon)+a.&\end{aligned}$$

As this holds for all $\varepsilon>0$, we conclude that

$$\lim_{N\to\infty}N^{-1}\log\mathbb{P}\big(M_N(t)/N\geqslant a\big)=a\log\big(\frac{\rho(t)}{a}\big)-\rho(t)+a.$$

Recognizing the decay rate of a Poisson distribution with mean $\rho(t)$, we observe that the essential behavior in the fast regime is again of M/M/∞ type.

2. Slow regime ($\alpha < 1$). In this regime we need to distinguish between the situation in which the random variable Λ almost surely results in a $\kappa_t(\Lambda^{(N)})$ below a, and the situation in which this is not the case. The proof of the following lemma is straightforward hence omitted.

LEMMA 4.2. *Given* Λ, *let* $y = \inf\{x > 0 : \mathbb{P}(\Lambda \leqslant x) = 1\}$ *and* $u(t) = y/\mu\,(1 - e^{-\mu t})$. *Then, as* $N \to \infty$,

$$\mathbb{P}(\kappa_t(\Lambda^{(N)}) \leqslant u(t)) \to 1.$$

The cases $u(t) \geqslant a$ and $u(t) < a$ should be treated differently, as follows from the following intuitive explanation that is based on the decomposition (4.1). If $u(t) \geqslant a$, then the random variable Λ can be 'large' with respect to a, which enables $M^{(N)}(t)$ to reach Na without the Poisson random variable attaining an unlikely value. If on the contrary $u(t) < a$, then Λ is 'small' with respect to a; $M^{(N)}(t)$ can only exceed level Na by the Poisson random variable attaining an extraordinarily large value.

We first consider the case $u(t) < a$. For ease we assume that Λ attains values in a discrete set of positive values, of which y is the largest (occurs with probability $p \in (0,1)$) and $y' < y$ the one-but-largest. It is directly seen that, for $\theta > 0$,

$$\mathbb{E}\mathrm{e}^{\theta M^{(N)}(t)} \geqslant p^{\lceil t/(\Delta N^{-\alpha})\rceil}\,\mathbb{E}\exp\big(\theta\,\mathrm{Pois}\big(Nu(t)\big)\big).$$

As $\alpha < 1$, this leads to

$$\lim_{N\to\infty} N^{-1}\log\mathbb{E}\mathrm{e}^{\theta M^{(N)}(t)} \geqslant u(t)(\mathrm{e}^{\theta} - 1). \tag{4.9}$$

In addition, $\mathbb{E}\mathrm{e}^{\theta M^{(N)}(t)}$ is majorized by

$$\begin{aligned} &p\,\mathbb{E}\exp\big(\theta\,\mathrm{Pois}(Nu(t))\big) + (1-p)\mathbb{E}\exp\big(\theta\,\mathrm{Pois}(N(y'/\mu)(1-\mathrm{e}^{-\mu t}))\big)\\ &= p\exp\big(Nu(t)(\mathrm{e}^{\theta}-1)\big) + (1-p)\exp\big(N\,(y'/\mu)(1-\mathrm{e}^{-\mu t})(\mathrm{e}^{\theta}-1)\big), \end{aligned}$$

which converges to the right-hand side of (4.9) on an exponential scale (use $y > y'$). Applying 'Cramér', we thus find that the probability of interest decays exponentially:

$$\begin{aligned} \lim_{N\to\infty} N^{-1}\log\mathbb{P}\big(M^{(N)}(t)/N \geqslant a\big) &= -\sup_{\theta>0}\big(\theta a - u(t)(\mathrm{e}^{\theta}-1)\big)\\ &= a\log\big(\frac{u(t)}{a}\big) + a - u(t). \end{aligned}$$

Now we focus on $u(t) \geqslant a$; in this case

$$\mathbb{P}(\mathrm{Pois}(Na) \geqslant Na)\mathbb{P}(\kappa_t(\Lambda^{(N)}) \geqslant a) \tag{4.10}$$

gives an asymptotically non-trivial lower bound for (4.1). Note that for every $\delta > 0$, there is an N large enough such that

$$\mathbb{P}(\mathrm{Pois}(Na) \geqslant Na) \geqslant \big(\frac{1}{2} - \delta\big),$$

so the first factor in (4.10) will not contribute to the decay rate. The tail behavior of the second factor follows from Lemma 4.1. On the other hand, (4.1) is majorized by

$$\mathbb{P}(\mathrm{Pois}(N(a-\varepsilon)) \geqslant Na) + \mathbb{P}(\kappa_t(\Lambda^{(N)}) \geqslant a - \varepsilon). \tag{4.11}$$

Again it is observed that only the second term in (4.11) contributes to the decay rate: by 'Cramér' the first term in (4.11) decays exponentially, whereas the decay of the second term is subexponential (by Lemma 4.1) for $\varepsilon > 0$ small enough (we need $a - \varepsilon > \rho(t)$). Letting $\varepsilon \downarrow 0$ while using that the rate function in (4.2) is continuous in a, we arrive at

$$\lim_{N\to\infty} \Delta N^{-\alpha} \log \mathbb{P}\big(M^{(N)}(t)/N \geqslant a\big) = -\sup_{\theta>0} \big(\theta a - \int_0^t \log M(\theta e^{-\mu s})\,\mathrm{d}s\big).$$

Note that the decay rate in this fast regime depends on more detailed information on the distribution of Λ than just the mean.

3. Intermediate regime ($\alpha = 1$). In this regime we expect exponential decay. Indeed, it is directly derived that

$$\lim_{N\to\infty} \Delta N^{-1} \log \mathbb{E}\mathrm{e}^{\theta M^{(N)}(t)} = \int_0^t \log M(\Delta(\mathrm{e}^{\theta} - 1)\mathrm{e}^{-\mu s})\,\mathrm{d}s,$$

and hence 'Gärtner-Ellis' [7] gives

$$\lim_{N\to\infty} \Delta N^{-1} \log \mathbb{P}\big(M^{(N)}(t)/N \geqslant a\big) = -\sup_{\theta>0} \big(\theta a - \int_0^t \log M(\Delta(\mathrm{e}^{\theta/\Delta} - 1)\mathrm{e}^{-\mu s})\,\mathrm{d}s\big).$$

For deterministic Λ the above result would equal that of the fast regime; the resemblance with the slow regime on the other hand becomes more pronounced for larger values of Δ.

4.2. *Large deviations for the coupled model.* We conclude this section by pointing out how the large devations for the coupled model (where each arrival generates work in d queues) can be dealt with. For $\alpha \geqslant 1$ we are in the regime of exponential decay. The multivariate version of the Gärtner-Ellis theorem entails that, modulo the validity of mild regularity conditions to be imposed on the set $A \subset \mathbb{R}^d_+$,

$$\lim_{N\to\infty} N^{-1} \log \mathbb{P}\big(\boldsymbol{M}^{(N)}(t)/N \in A\big) \\ = -\inf_{\boldsymbol{a}\in A} \sup_{\boldsymbol{\theta}} \big(\sum_{i=1}^d \theta_i a_i - \lim_{N\to\infty} \frac{1}{N} \log \mathbb{E} \exp\big[\sum_{i=1}^d \theta_i M_i^{(N)}(t)\big]\big).$$

The problem therefore reduces to characterizing the limiting log moment generating function. It takes standard computations to verify that for $\alpha > 1$, with an argumentation borrowed from specific intermediate results in the proof of Lemma 3.1,

$$\lim_{N\to\infty} N^{-1} \log \mathbb{E} \exp\big[\sum_{i=1}^d \theta_i M_i^{(N)}(t)\big] \\ = t\, \mathbb{E}\Lambda\Big(\int_0^t \frac{1}{t} \prod_{i=1}^d \big(e^{-\mu_i s}(e^{\theta_i}-1)+1\big)\mathrm{d}s - 1\Big),$$

whereas for $\alpha = 1$ it turns out to equal

$$\frac{1}{\Delta} \int_0^t \log \mathbb{E} \exp\big[\Lambda\Delta\big(\big(\prod_{i=1}^d e^{-\mu_i s}(e^{\theta_i}-1)+1\big)-1\big)\big]\mathrm{d}s.$$

For $\alpha < 1$, as before, the decay is either exponential in N (if the the multi-dimensional random Poisson parameter cannot attain values that are contained in A), or exponential in N^α. The latter regime being the more complicated one, we here include the corresponding decay rate. The probability of our interest can be rewritten as

$$\int_{x_1=0}^\infty \cdots \int_{x_d=0}^\infty F_A(\boldsymbol{x}) \cdot \pi(\mathrm{d}x_1, \ldots, \mathrm{d}x_d), \tag{4.12}$$

where

$$F_A(\boldsymbol{x}) := \mathbb{P}\big(\big(\mathrm{Pois}_1(N\,\kappa_t(\Lambda^{(N)})), \ldots, \mathrm{Pois}_d(N\,\kappa_t(\Lambda^{(N)}))/N\big) \in A\big),$$
$$\pi(\mathrm{d}x_1, \ldots, \mathrm{d}x_d) := \mathbb{P}(\kappa_{t,1}(\Lambda^{(N)}) \in \mathrm{d}x_1, \ldots, \kappa_{t,d}(\Lambda^{(N)}) \in \mathrm{d}x_d);$$

here the d Poisson random variables are independent. Using the same ideas as above, it can be shown that (4.12) decays exponentially in N^α, where the decay rate is now given by

$$\lim_{N\to\infty} \Delta N^{-\alpha} \log \mathbb{P}\big((\kappa_{t,1},\ldots,\kappa_{t,d})(\Lambda^{(N)}) \in A\big) = -\inf_{\boldsymbol{a}\in A} \sup_{\boldsymbol{\theta}} \Big(\sum_{i=1}^{d} \theta_i a_i - \int_0^t \log M\Big(\sum_{i=1}^{d} \theta_i e^{-\mu_i s}\Big)\mathrm{d}s\Big).$$

5. Discussion and future research. In this paper we propose to model an overdispersed arrival process by a mixed Poisson process in a random environment. We assess the impact of overdispersion on system performance when feeding such an arrival process into an infinite-server system. Under a specific scaling, we derive (functional) central limit results and large deviations asymptotics.

Various extensions can be explored before using the model in an operational context; a few of them are mentioned here. To start with, many results seem to carry over to the setting with generally distributed service times. In addition, systematically studying the effect of adding a deterministic trend $\bar{\lambda}(\cdot)$ (e.g. to account for diurnal patterns) to the random environment $\Lambda(\cdot)$, the results could be generalized to a setting with nonstationary Cox arrival processes. Moreover, we could pursue to upgrade our model to not only allow for a deterministic trend, but also for dependence between arrival rates corresponding to subsequent time slots; such properties have been observed in (overdispersed) datasets, and are therefore desirable to incorporate [14]. Another challenge lies in refining the logarithmic asymptotics, as obtained in Section 4, to exact asymptotics.

In all of the results obtained, we revealed a trichotomy in system behavior depending on the imposed scaling on system size and sampling frequency. Here the scaling primarily serves to tweak the level of overdispersion in the system. The combination of tunable sampling and tunable overdispersion provides a rich framework for modeling real-world arrival processes. One could imagine that in a rapidly changing environment, the inherent overdispersion of the arrival process hardly plays a role, whereas in a slowly changing random environment, overdispersion is expected to be more dominant. This interplay between sampling frequency and overdispersion is a convenient feature of our model, which could be used to calibrate the model to real-world data. The latter could be a promising direction for future research, involving challenging statistical issues related to determining which of the three asymptotic regimes that arise in the revealed trichotomy suits best with a given data set for a finite N sys-

tem, and estimating the model parameters conditioned on the asymptotic regime.

An application of our model would be in the area of dimensioning service systems or staffing, and in particular square-root staffing in many-server systems. The general idea behind square-root staffing is as follows: a finite-server system is modeled as a system in heavy traffic, where the number of servers s is large and at the same time, the system is critically loaded. Under Markovian assumptions this can be achieved by setting $s = \lambda + \beta\sqrt{\lambda}$ (denoting the load on the system by λ) and letting $\lambda \to \infty$ while keeping $\beta > 0$ fixed. The system then reaches the desirable Quality-and-Efficiency-Driven (QED) regime, in which the system load approaches 100% while the delays experienced by customers remain limited. In such large-scale service operations, it is natural to use an infinite-server system as a proxy to the many-server system. Infinite-server models are extremely useful because of their tractability; this can be exploited by translating detailed knowledge of the infinite-server system state to the finite-server setting. This returns rather good estimates of future arrivals, even in situations of time-varying arrival processes [16, 17]. The model developed in this paper provides a new way of modeling such large-scale service systems, with the additional feature of a tunable level of overdispersion, essentially replacing a deterministic λ by a stochastically fluctuating Λ. The possibility of using this model, including the above mentioned extensions, for designing dimensioning schemes is currently being investigated by the authors.

REFERENCES

[1] D. Anderson, J. Blom, M. Mandjes, H. Thorsdottir, and K. de Turck. A functional central limit theorem for a Markov-modulated infinite-server queue. *Methodology and Computing in Applied Probability*, 2016. MR3465473

[2] A. Bassamboo, R.S. Randhawa, and A. Zeevi. Capacity sizing under parameter uncertainty: safety staffing principles revisited. *Management Science*, 56(10):1668–1686, 2010.

[3] J. Blom, K. de Turck, O. Kella, and M. Mandjes. Tail asymptotics of a Markov-modulated infinite-server queue. *Queueing Systems*, 78(4):337–357, 2014. MR3269262

[4] J. Blom, K. de Turck, and M. Mandjes. Analysis of Markov-modulated infinite-server queues in the central-limit regime. *Probability in the Engineering and Informational Sciences*, 29:433–459, 2015. MR3355613

[5] D.R. Cox. Some statistical methods connected with series of events journal of the royal statistical society. *Journal of the Royal Statistical Society, Series B (Methodological)*, 17(2):129–164, 1955. MR0092301

[6] K. de Turck and M. Mandjes. Large deviations of an infinite-server system with a linearly scaled background process. *Performance Evaluation*, 75–76:36–49, 2014.

[7] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer, New York, 1998. MR1619036

[8] S. Kim and W. Whitt. Are call center and hospital arrivals well modeled by nonhomogeneous poisson processes? *Manufacturing & Service Operations Management*, 16(3):464–480, 2014.

[9] S. Kim and W. Whitt. Choosing arrival process models for service systems: Tests of a nonhomogeneous Poisson process. *Naval Research Logistics*, 61(1):66–90, 2014. MR3162951

[10] Y. Liu and W. Whitt. Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Operations Research*, 60(6):1551–1564, 2012. MR3009182

[11] H. Lu and G. Pang. Gaussian limits for a fork-join network with non-exchangeable synchronization in heavy traffic. *Mathematics of Operations Research*, 2015.

[12] H. Lu and G. Pang. Heavy-traffic limits for a fork-join network in a renewal random environment. *Submitted*, 2015. MR3486808

[13] S. Maman. Uncertainty in the demand of service: the case of call centers and emergency departments. *M. Sc. Thesis, Technion – Israel Institute of Technology, Haifa, Israel*, 2009.

[14] Ibrahim. R. H. Ye, P. L'Ecuyer, and H. Shen. Modeling and forecasting call center arrivals: a literature survey. *International Journal of Forecasting*, 32:865–874, 2016.

[15] T. Rydén. An EM algorithm for estimation in Markov-modulated Poisson processes. *Computational Statistics & Data Analysis*, 21:431–447, 1996. MR1394060

[16] W. Whitt. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters*, 24(5):205–212, 1999. MR1719916

[17] W. Whitt, L. V. Green, and P. J. Kolesar. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, 16(1):13–39, 2007.

Mariska Heemskerk
Michel Mandjes
University of Amsterdam
Korteweg-de Vries Institute for Mathematics
Science Park 107
1098 XG Amsterdam
The Netherlands
E-mail: j.m.a.heemskerk@uva.nl
m.r.h.mandjes@uva.nl

Johan van Leeuwaarden
Eindhoven University of Technology
Department of Mathematics
and Computer Science
PO Box 513
5600 MB Eindhoven
The Netherlands
E-mail: j.s.h.v.leeuwaarden@tue.nl.