Forty-Sixth Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 23-26, 2008

Large Deviations for Many Server Networks with Long Range Dependent and Batch Arrivals

Soummya Kar, José M. F. Moura and Kavita Ramanan

Abstract— **A sample path large deviations principle is established for the (scaled) vector of number of customers in a Jackson network of many-server queues, in the asymptotic regime where the arrival rate at each queue and the number of servers increase to infinity in a specified fashion. This result is obtained under the assumption of (possibly time-inhomogeneous) Markovian service and routing, and a condition on the scaled sequence of cumulative arrival processes, which holds for a wide class of long-range dependent and batch arrival processes. Moreover, the explicit form of the rate function is obtained in the single-queue setting, and the most likely way in which a large number of customers build up in the system is identified by solving a variational problem.**

I. Introduction

A. Background and Motivation

Large deviations in stochastic networks is a widely studied field with a vast literature. Prior work in this field has mostly emphasized the large buffer and the many sources regimes. Good accounts of this body of work have been presented in [1], [2], [3] and references therein. In this paper, we consider a different large deviations regime, which we refer to as the many-servers regime, where the number of servers in the system increases (in a suitable scale) as the arrival traffic increases. This regime is motivated by applications in decentralized data networks and call centers, where there are a large number of servers at each node. It is therefore natural to study the behavior of such systems in the asymptotic limit as the number of servers increases as the external traffic increases at each node. We consider a Jackson network of such multi-server queues or nodes, where the service and routing processes are assumed to be (possibly time-inhomogeneous) Markovian, and the arrival sequence can be general, but is required to satisfy the conditions 1 and 2 stated in Section I-C, which hold for a large class of long-range dependent and batch arrival processes. Our main result obtains a sample path large deviation principle (LDP) for the sequence of scaled number in system in the many server regime, with the rate function expressed in terms of the rate function associated with the scaled arrival sequence and a certain continuous map. In some cases, the obtained rate function is very well-characterized and the corresponding variational problems for estimating the exponential decay rate of probabilities of certain rare events are explicitly solvable, thus enabling the characterization of the most likely way in which the rare event occurs. It is worthwhile to point out that our conditions on the cumulative arrival processes preclude certain short-range dependent processes such as the Poisson process. In fact, the large deviations exhibits a phase transition in the sense that for Poisson arrivals (more generally, for processes for which condition 1 on the arrival sequence stated in Section I-C is satisfied by a sequence κ_N that remains bounded as $N \to \infty$), the form of the rate function is significantly different, as shown in a forthcoming paper.

We briefly summarize the organization of the rest of the paper. Section I-B introduces some common notation used throughout the paper, and the main results are summarized in Section I-C. The model description and assumptions are introduced in Section II. In Section III we rigorously state our main theorems in the Jackson network setting, and present the proofs of the theorems in Section IV. As an illustration of the usefulness of our main results, in Section V, we determine the LDP rate function in the specific setting of batch arrivals, for which the variational problem is solvable. The details of this solution will be presented in a subsequent paper. Concluding remarks are presented in Section VI.

B. Notation and Terminology

We denote the set of reals by $\mathbb{R}$, the set of non-negative reals by $\mathbb{R}_+$. Let $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ and $\overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{\infty\}$. A superscript M applied to any of these sets denotes the corresponding M dimensional Euclidean space. For example, the set $\mathbb{R}_+^M$ denotes the non-negative orthant in $\mathbb{R}^M$. For $a, b \in \mathbb{R}$, $a \wedge b$ denotes the minimum of a and b. When applied to vectors, $\|\cdot\|$ denotes the standard Euclidean 2-norm, while for matrices it corresponds to the induced 2-norm.

For $T > 0$, let $\mathcal{D}[0,T]$ be the space of real-valued, right-continuous functions on $[0,T]$ having left limits. Let $\mathcal{D}_+[0,T]$ and $\mathcal{AC}[0,T]$ be the respective subsets of non-negative and absolutely continuous functions in $\mathcal{D}[0,T]$. Let $\mathcal{AC}_+[0,T]$ be the subset of absolutely continuous functions in $\mathcal{D}_+[0,T]$. A superscript M applied to any of these sets denotes the corresponding M dimensional Euclidean space. For a set of functions, we abuse notation by denoting the subset of functions f with $f(0) = \mathbf{x}$ by putting a superscript $\mathbf{x}$. For example, the set $\mathcal{D}_+^{M,\mathbf{x}}[0,T]$ denotes the space of $\mathbb{R}_+^M$ valued, right-continuous functions on $[0,T]$ having left limits starting at $\mathbf{x}$.

The tuple $\left(\mathcal{D}^M[0,T], \mathcal{B}^M(T)\right)$ denotes the space $\mathcal{D}^M[0,T]$ with the Borel-algebra $\mathcal{B}^M(T)$ generated by

Names appear in alphabetical order.

Soummya Kar is with Department of Electrical and Computer Engineering, Carnegie Mellon University. soummyak@andrew.cmu.edu

José M. F. Moura is with the Department of Electrical and Computer Engineering, Carnegie Mellon University. moura@ece.cmu.edu

Kavita Ramanan is with Department of Mathematical Sciences, Carnegie Mellon University. kramanan@math.cmu.edu

978-1-4244-2926-4/08/$25.00 ©2008 IEEE

either the Skorokhod J_1 topology or the topology of uniform convergence. It will be clear from the context which topology we are using.

In the sequel we adopt the convention that the supremum of an empty set is $-\infty$, $0\ln(0) = 0$, and $\ln(x) = -\infty$, $\forall x \leq 0$.

We assume there exists a complete probability space, $(\Omega, \mathcal{F}, \mathbb{P})$, where all processes of interest are defined. When the initial state of a process is fixed, say at the point $\mathbf{x}$, then we write the probability and expectation operators as $\mathbb{P}_{\mathbf{x}}(\cdot)$ and $\mathbb{E}_{\mathbf{x}}[\cdot]$, respectively.

C. Summary of Main Results

We first recall the definition of a large deviation principle (LDP). For a given $T > 0$, a sequence, $\{\overline{\mathbf{E}}^N\}_{N\in\mathbb{N}}$ of processes in $\left(\mathcal{D}^M[0,T], \mathcal{B}^M(T)\right)$ is said to satisfy an LDP at scale N^β ($\beta > 0$) with good rate function $I_T(\cdot)$ (see [4]), if $\forall B \in \mathcal{B}^M(T)$

$$\begin{aligned} -\inf_{\phi\in B^\circ} I_T(\phi) &\leq \liminf_{N\to\infty} \frac{1}{N^\beta}\log\mathbb{P}\left(\overline{\mathbf{E}}^N \in B\right) \\ \leq & \limsup_{N\to\infty}\frac{1}{N^\beta}\log\mathbb{P}\left(\overline{\mathbf{E}}^N \in B\right) \leq -\inf_{\phi\in\overline{B}} I_T(\phi) \quad (1) \end{aligned}$$

where B° and $\overline{B}$ denote the interior and closure of B respectively.

As mentioned above, we study a sequence of Jackson networks (see [5]), indexed by N, where each network consists of M queues and the service and routing disciplines are Markovian (possibly time-inhomogeneous.) Let $\mathbf{E}^N(t), \mathbf{X}^N(t) \in \mathcal{D}^M[0,T]$, denote the vectors of cumulative arrival streams over the interval $[0,t]$ and the number in system (including both those waiting and those being served) at time t, respectively. The main result of this paper states, that, if the following conditions are satisfied:

1: There exists $\beta > 0$ and a sequence, $\{\kappa_N\}_{N\in\mathbb{N}}$, such that $\kappa_N \to \infty$ as $N \to \infty$,

2: The sequence of scaled arrival processes, $\{\overline{\mathbf{E}}^N\}_{N\in\mathbb{N}}$, where
$$\overline{\mathbf{E}}^N(t) = \frac{1}{\kappa_N N^\beta}\mathbf{E}^N(t) \quad (2)$$
satisfies an LDP in $\left(\mathcal{D}^M[0,T], \mathcal{B}^M(T)\right)$ at scale N^β as given in eqn. (1) with good rate function $I_T(\cdot)$,

3: The number of servers at each of the M queues in the N-th network scales as $\kappa_N N^\beta$,

then the sequence of scaled processes, $\{\overline{\mathbf{X}}^N\}_{N\in\mathbb{N}}$, where

$$\overline{\mathbf{X}}^N(t) = \frac{1}{\kappa_N N^\beta}\mathbf{X}^N(t) \quad (3)$$

satisfies an LDP in $\left(\mathcal{D}^M[0,T], \mathcal{B}^M(T)\right)$ at scale N^β with a certain rate function $J_T(\cdot)$. As will be evident later, the scaling $\kappa_N N^\beta$ of the number of servers is the right scaling to satisfy a non-trivial LDP. Our results also characterize explicitly the resulting rate function $J_T(\cdot)$ in terms of $I_T(\cdot)$ and a continuous map. We note that the form of LDP in item 2 above is not very abstract and there are a wide class of arrival processes which fall under this category. As illustration, we provide two motivating examples.

Example 1) Arrival Processes with Long Range Dependence: Consider a sequence of single multi-server queues, indexed by N, where the arrival is a superposition of N sources with long range dependence. Let $\{E^N\}_{N\in\mathbb{N}}$ be the sequence of arrival streams (possibly centered.) In many cases, it can be shown that the scaled sequence, $\{\overline{E}^N\}_{N\in\mathbb{N}}$ satisfies an LDP at scale $N^{2(1-H)}$, where $H > 1/2$ is the Hurst parameter and

$$\overline{E}^N(t) = \frac{1}{N}E^N(t) \quad (4)$$

When $\beta = 2(1-H)$ and $\kappa_N = N^{2H-1}$, the sequence of arrival processes satisfy the conditions stated above. Hence, if the N-th queue contains N identical servers, each offering service at rate μ, the sequence of scaled processes, $\{\overline{\mathbf{X}}^N\}_{N\in\mathbb{N}}$, with the scaling in (3) satisfies an LDP at scale $N^{2(1-H)}$. Note, in this case, because of possible centering, the arrival sequence, $\{E^N\}_{N\in\mathbb{N}}$, may take negative values. This will not pose a problem, as the queueing functionals remain well-defined in this case, as will be shown later.

Example 2) Batch Arrival Processes: Consider a sequence of single multi-server queues, indexed by N, with external batch arrivals. As N increases, both the arrival rate and batch size increase. This form of arrival process would, for example, model data networks in which a central hub (the queue in our case) receives requests from a large number of local stations. Then, as the number of users increases, the traffic from each station, and at the same time the number of stations, would also increase. We model the overall arrival process to the N-th queue (the central hub) by a process, $E^N \in \mathcal{D}[0,T]$, where

$$E^N(t) = \kappa_N A(\lambda N t) \quad (5)$$

where κ_N is the fixed batch size and λN is the arrival rate. Here A is a unit rate Poisson process. We assume that $\kappa_N \to \infty$ as $N \to \infty$ and $\beta = 1$ in this case. It is easy to see that, with the scaling in item 2, the sequence $\{\overline{\mathbf{E}}^N\}_{N\in\mathbb{N}}$ satisfies an LDP at scale N. Since the mean number of arrivals per unit time is $\kappa_N N$, it is not hard to see that a fair policy should scale the number of servers in the N-th queue as $\kappa_N N$, which is the scaling in item 3, each server having a constant service rate μ. We can show that the sequence of scaled processes, $\{\overline{\mathbf{X}}^N\}_{N\in\mathbb{N}}$, with the scaling in (3), satisfies an LDP at scale N.

II. Model Description and Assumptions

In this section, we precisely formulate our model of a Jackson network of many-server queues. Specifically, we consider a sequence of such networks, indexed by $N \in \mathbb{N}$, where the N-th system is a network of M many-server queues (nodes) with Markovian (possibly time inhomogeneous) service and routing. In the following we summarize the dynamics of the N-th network.

A.1) External Arrival Processes: The cumulative arrival process E_i^N at the i-th node, $1 \leq i \leq M$, in the N-th system is a stochastic process with sample paths that are non-decreasing and lie in $\mathcal{D}[0,T]$[1] We denote by $\mathbf{E}^N(t) = [E_1^N(t) \cdots E_M^N(t)]$ the vector of arrival processes for the N-th system. We assume that there exists a sequence $\{\kappa_N\}_{N\in\mathbb{N}}$ with $\kappa_N \to \infty$ as $N \to \infty$ and $\beta > 0$, such that the scaled arrival sequence $\{\overline{\mathbf{E}}^N\}_{N\in\mathbb{N}}$ satisfies an LDP in $\left(\mathcal{D}^M[0,T], \mathcal{B}^M(T)\right)$ at scale N^β as given in eqn. (1) with good rate function $I_T(\cdot)$, where

$$\overline{\mathbf{E}}^N(t) = \frac{1}{\kappa_N N^\beta}\mathbf{E}^N(t) \tag{6}$$

No restrictions, like Markovian etc., are imposed on the arrival processes.

A.2) Time-Inhomogeneous Markovian Service: We assume that the i-th node in the N-th system is equipped with $\kappa_N N^\beta$ identical Markovian servers. We model such generic Markovian service by assigning a rate function $\mu_i(t) : \mathbb{R}_+ \longmapsto \mathbb{R}_+$ to each server in the i-th node. We assume that the service rate functions, $\mu_i(t)$, are measurable and locally bounded such that for every $T > 0$

$$\mu_i(t) \le k_T < \infty, \quad \forall 1 \le i \le M \tag{7}$$

We define

$$\Upsilon(t) = \text{diag}\left(\mu_1(t), \cdots, \mu_M(t)\right), \quad \forall t \ge 0 \tag{8}$$

As a matter of fact, all our results will hold if the assumption of local boundedness of the rates is replaced by local integrability. For clarity of presentation, in this manuscript we assume local boundedness.

A.3) Time-Inhomogeneous Routing and Exit Probabilities: The probability of getting transferred to node j after service completion at node i is given by the measurable function $p_{ij}(t) : \mathbb{R}_+ \longmapsto [0,1]$. Similarly, the probability of departure upon service completion at node i is given by the measurable function $p_i(t) : \mathbb{R}_+ \longmapsto [0,1]$. We assume the following holds:

$$\sum_{j=1}^{M} p_{ij}(t) + p_i(t) = 1, \quad \forall t, i \tag{9}$$

We define the following $M \times M$ matrices $\forall t \ge 0$:

$$[P(t)]_{i,j} = p_{ij}(t), \quad 1 \le i,j \le M \tag{10}$$

A.4) Independence Assumptions: We assume that the service/routing processes are independent of the arrival processes and the initial number in the system.

For the N-th system, we denote by $X_i^N(t)$ the actual number of customers (both waiting and being served) in the i-th node at time t. From the construction below, it is easy to deduce that X_i^N is a process with sample paths in $\mathcal{D}[0,\infty)$. Let $\mathbf{X}^N = [X_1^N \cdots X_M^N]$. We now provide a pathwise construction (based on the pathwise construction used in [6], [7]) of the process $\mathbf{X}_i^N$ under Assumptions **A.1)-A.4)**. To this end, define $\{T_{ij}\}_{1\le i,j\le M}$ and $\{D_i\}_{1\le i\le M}$ to be independent rate 1 Poisson processes, independent of the arrival processes and the initial number $\mathbf{X}^N(0)$ in the system. We then have the following representation (in the sense of the distribution induced in path space) for the process $\mathbf{X}^N$: for $1 \le i \le M$,

$$\begin{aligned} X_i^N(t) &= X_i^N(0) + E_i^N(t) \\ &+ \sum_{j=1}^{M} T_{ji}\left(\int_0^t \mu_j(s)p_{ji}(s)\left(X_j^N(s) \wedge \kappa_N N^\beta\right)^+ ds\right) \\ &- \sum_{j=1}^{M} T_{ij}\left(\int_0^t \mu_i(s)p_{ij}(s)\left(X_i^N(s) \wedge \kappa_N N^\beta\right)^+ ds\right) \\ &- D_i\left(\int_0^t \mu_i(s)p_i(s)\left(X_i^N(s) \wedge \kappa_N N^\beta\right)^+ ds\right) \end{aligned} \tag{11}$$

The above pathwise construction was used in [6], [7] for the many-server queue with Poisson arrival, where fluid and diffusion limits of many-server models were analyzed, though under a different scaling. We note, that, in our case, the cumulative arrivals may become negative occasionally (see, **Example 1** in Subsection I-C), and which is taken into account by considering the non-negative part of $X_j^N(s) \wedge \kappa_N N^\beta$ in the time-change expression in eqn. (11).

III. Statement of Main Theorems

We briefly summarize the main results in this manuscript. For a suitably scaled sequence of the processes, $\{\mathbf{X}^N(t)\}_{N\in\mathbb{N}}$, we establish a sample path LDP. To this end, define the scaled sequence of processes:

$$\overline{\mathbf{X}}_i^N(t) = \frac{1}{\kappa_N N^\beta}\mathbf{X}_i^N(t) \tag{12}$$

For a given $T > 0$, our results will be of the following general form: $\forall B \in \mathcal{B}^M(T)$

$$\begin{aligned} -\inf_{\phi\in B^\circ} J_T(\phi) &\le \liminf_{N\to\infty}\frac{1}{N}\log\mathbb{P}\left(\overline{\mathbf{X}}^N \in B\right) \\ \le \quad &\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{P}\left(\overline{\mathbf{X}}^N \in B\right) \le -\inf_{\phi\in\overline{B}} J_T(\phi) \end{aligned} \tag{13}$$

We characterize the rate function $J_T(\phi) : \mathcal{D}^M[0,T] \longmapsto \overline{\mathbb{R}}_+$ explicitly, so that, evaluation of probabilities reduce to solving variational problems, as given in eqn. (13). As mentioned earlier, characterizing the rate function explicitly not only gives the asymptotic decay rate of the probabilities, but also identifies the most likely path in which a rare event occurs, providing much more insight on the system behavior.

Let us consider the following scaled sequence:

$$\overline{\mathbf{X}}^N(0) = \frac{1}{\kappa_N N^\beta}\mathbf{X}^N(0) \tag{14}$$

We now state the main result regarding the LDP for the sequence of scaled processes, $\{\overline{\mathbf{X}}^N\}_{N\in\mathbb{N}}$, corresponding to the number in system for the sequence of Jackson networks indexed by N.

Theorem 1 Consider the sequence of Jackson networks, indexed by $N \in \mathbb{N}$, under the Assumptions **A.1)-A.4)**. For

[1] We assume that the number of jumps of $E_i^N(t)$ is a.s. at most finite in a finite time interval.

any $T > 0$, let the space $\mathcal{D}^M[0,T]$ be equipped with the Skorokhod J_1 topology (or the topology of uniform convergence.) Let the space $R^M \times \mathcal{D}^M[0,T]$ be equipped with the product topology and the sequence $\{\overline{\mathbf{X}}^N(0), \overline{\mathbf{E}}^N\}_{N\in\mathbb{N}}$ satisfy an LDP at scale N^β in $R^M \times \mathcal{D}^M[0,T]$ with good rate function. Then, the sequence, $\{\overline{\mathbf{X}}^N\}_{N\in\mathbb{N}}$, satisfies an LDP at scale N^β in $\mathcal{D}^M[0,T]$ under the Skorokhod J_1 topology (or the topology of uniform convergence) with good rate function.

Given the LDP rate function for the sequence $\{\overline{\mathbf{X}}^N(0), \overline{\mathbf{E}}^N\}_{N\in\mathbb{N}}$ we can explicitly characterize the LDP rate function for the sequence $\{\overline{\mathbf{X}}^N\}_{N\in\mathbb{N}}$. To this end, define the function $\Lambda : \mathbb{R}^M \times \mathcal{D}^M[0,T] \longmapsto \mathcal{D}^M[0,T]$, which maps the pair $(\mathbf{a}, \widetilde{\phi}) \in \mathbb{R}^M \times \mathcal{D}^M[0,T]$ to $\phi \in \mathcal{D}^M[0,T]$ by

$$\phi(t) = \mathbf{a} + \widetilde{\phi}(t) - \int_0^t \left(I - P^T(s)\right) \Upsilon(s) \left(\phi(s) \wedge \mathbf{1}\right)^+ ds \tag{15}$$

Note, under Assumption **A.2)**, it can be shown eqn. (15) constitutes a well-defined map from the space $R^M \times \mathcal{D}^M[0,T]$ to $\mathcal{D}^M[0,T]$. In fact, it can be shown that the function $\Lambda(\cdot)$ is continuous both in the topology of uniform convergence and the Skorokhod J_1 topology (see [6], [7].)

We now state the following theorem which characterizes the LDP rate function of the sequence $\{\overline{\mathbf{X}}^N\}_{N\in\mathbb{N}}$. As conveniently assumed in most LDP results, we assume that the scaled initial number $\overline{\mathbf{X}}^N(0)$ is equal to a constant $\mathbf{x} \in \mathbb{R}^M$ for all $N \in \mathbb{N}$.

Theorem 2 Let the sequence $\{\overline{\mathbf{E}}^N\}_{N\in\mathbb{N}}$ satisfy an LDP at scale N^β in $\mathcal{D}^M[0,T]$ under the Skorokhod J_1 topology (or the topology of uniform convergence) with good rate function $I_T(\cdot)$. Then, the sequence, $\{\overline{\mathbf{X}}^N\}_{N\in\mathbb{N}}$, satisfies an LDP at scale N^β in $\mathcal{D}^{\mathbf{x},M}[0,T]$ under the Skorokhod J_1 topology (or the topology of uniform convergence) with good rate function $J_T^{\mathbf{x}}$, where

$$J_T^{\mathbf{x}}(\phi) = \inf_{\widetilde{\phi}\in\mathcal{D}^M[0,T] \quad \phi=\Lambda(\mathbf{x},\widetilde{\phi})} I_T(\widetilde{\phi}) \tag{16}$$

where $\mathbf{x} \in \mathbb{R}^M$ is the scaled initial number at each system and $\Lambda(\cdot)$ is the function defined in eqn. (15). In particular, for every $\mathbf{x} \in \mathbb{R}^M$, we have $\forall B \in \mathcal{B}^{\mathbf{x},M}(T)$

$$\begin{aligned} &- \inf_{\phi\in B^\circ} J_T^{\mathbf{x}}(\phi) \leq \liminf_{N\to\infty} \frac{1}{N^\beta} \log \mathbb{P}_{\mathbf{x}}\left(\overline{\mathbf{X}}^N \in B\right) \\ \leq \quad &\limsup_{N\to\infty} \frac{1}{N^\beta} \log \mathbb{P}_{\mathbf{x}}\left(\overline{\mathbf{X}}^N \in B\right) \leq - \inf_{\phi\in\overline{B}} J_T^{\mathbf{x}}(\phi) \end{aligned} \tag{17}$$

where $\mathcal{B}^{\mathbf{x},M}(T)$ is the Borel algebra in $\mathcal{D}^{\mathbf{x},M}[0,T]$ generated either by the Skorokhod J_1 topology (or the topology of uniform convergence.)

IV. Proof of Theorems 1 and 2

This section is devoted to the proofs of Theorems 1 and 2. We start by briefly sketching the main steps of the proof. Define the sequence of auxiliary processes, $\{\overline{\mathbf{Z}}^N\}_{N\in\mathbb{N}}$ with sample paths in $\mathcal{D}^M[0,T]$ by

$$\begin{aligned} \overline{\mathbf{Z}}^N(t) &= \overline{\mathbf{X}}^N(0) + \overline{\mathbf{E}}^N(t) \\ &- \int_0^t \left(I - P^T(s)\right) \Upsilon(s) \left(\overline{\mathbf{Z}}^N(s) \wedge \mathbf{1}\right)^+ ds \end{aligned} \tag{18}$$

Under the assumptions of Theorems 1,2 we show that the sequence $\{\overline{\mathbf{Z}}^N\}_{N\in\mathbb{N}}$ satisfies an LDP at scale N^β in $\mathcal{D}^M[0,T]$ with good rate function and we explicitly characterize this rate function. We then show that the sequences $\{\overline{\mathbf{X}}^N\}_{N\in\mathbb{N}}$ and $\{\overline{\mathbf{Z}}^N\}_{N\in\mathbb{N}}$ are exponentially equivalent at scale N^β, thus leading to Theorems 1 and 2.

We start by developing a representation for the sequence of scaled processes, $\{\overline{\mathbf{X}}^N\}_{N\in\mathbb{N}}$, as indicated in eqn. (12.)

A Representation: For $1 \leq i, j \leq M$, define the process

$$\begin{aligned} \overline{M}_{ij}^N(t) &= \frac{1}{\kappa_N N^\beta} T_{ij}\left(\kappa_N N^\beta \int_0^t \mu_i(s) p_{ij}(s)\right. \\ &\left.(\overline{X}_i^N(s) \wedge 1)^+ ds\right) - \int_0^t \mu_i(s) p_{ij}(s) (\overline{X}_i^N(s) \wedge 1)^+ ds \end{aligned} \tag{19}$$

Similarly, define the processes

$$\begin{aligned} \overline{M}_{i}^N(t) &= \frac{1}{\kappa_N N^\beta} D_{i}\left(\kappa_N N^\beta \int_0^t \mu_i(s) p_{i}(s)\right. \\ &\left.(\overline{X}_i^N(s) \wedge 1)^+ ds\right) - \int_0^t \mu_i(s) p_{i}(s) (\overline{X}_i^N(s) \wedge 1)^+ ds \end{aligned} \tag{20}$$

Then, for each i, the scaled process, $\overline{X}_i^N(t)$, in eqn. (14) can be written as

$$\begin{aligned} \overline{X}_i^N(t) &= \overline{X}_i^N(0) + \overline{E}_i^N(t) + \sum_{j=1}^M \left(\int_0^t \mu_j(s) p_{ji}(s)\right. \\ &\left.(\overline{X}_j^N(s) \wedge 1)^+ ds\right) \\ &- \sum_{j=1}^M \left(\int_0^t \mu_i(s) p_{ij}(s) (\overline{X}_i^N(s) \wedge 1)^+ ds\right) \\ &- \int_0^t \mu_i(s) p_i(s) (\overline{X}_i^N(s) \wedge 1)^+ ds \\ &+ \sum_{j=1}^M \overline{M}_{ji}^N(t) - \sum_{j=1}^M \overline{M}_{ij}^N(t) - \overline{M}_i^N(t) \end{aligned} \tag{21}$$

In a compact form, eqn. (21) becomes

$$\begin{aligned} &\overline{\mathbf{X}}^N(t) \\ = \quad &\overline{\mathbf{X}}^N(0) + \overline{\mathbf{E}}^N(t) - \int_0^t \left[\left(I - P^T(s)\right) \Upsilon(s)\right. \\ &\left.\left(\overline{\mathbf{X}}^N(s) \wedge \mathbf{1}\right)^+ ds + \overline{\mathbf{M}}^N(t)\right] \end{aligned} \tag{22}$$

where:

$$\left[\overline{\mathbf{M}}^N(t)\right]_i = \sum_{j=1}^M \overline{M}_{ji}^N(t) - \sum_{j=1}^M \overline{M}_{ij}^N(t) - \overline{M}_i^N(t), \ \ 1 \leq i \leq M \tag{23}$$

Using the representation in eqn. (22), we now prove the exponential equivalence of $\{\overline{\mathbf{X}}^N\}_{N\in\mathbb{N}}$ and $\{\overline{\mathbf{Z}}^N\}_{N\in\mathbb{N}}$ at scale N^{β}.

Lemma 3 Under the assumptions of Theorem 1, the sequences $\{\overline{\mathbf{X}}^N\}_{N\in\mathbb{N}}$ and $\{\overline{\mathbf{Z}}^N\}_{N\in\mathbb{N}}$ are exponentially equivalent at scale N^{β} in both the Skorokhod J_1 topology or the topology of uniform convergence.

Proof: In the following we work with the stronger topology of uniform convergence. Clearly, the results follow for the weaker J_1 topology. We start by showing that the sequence $\{\overline{\mathbf{M}}^N\}_{N\geq 0}$ converges to $\mathbf{0}$ in probability at a scale $\kappa_N N^{\beta}$. For any $\varepsilon > 0$

$$\begin{aligned}&\mathbb{P}\left\{\sup_{0\le t\le T}\left\|\overline{\mathbf{M}}^N(t)\right\|>\varepsilon\right\}\\ \le\;& \sum_{i=1}^{M}\mathbb{P}\left\{\sup_{0\le t\le T}\left|\left[\overline{\mathbf{M}}^N(t)\right]_i\right|>\frac{\varepsilon}{\sqrt{M}}\right\}\end{aligned}\tag{24}$$

From eqn. (22) we have

$$\begin{aligned}&\mathbb{P}\left\{\sup_{0\le t\le T}\left|\left[\overline{\mathbf{M}}^N(t)\right]_i\right|>\frac{\varepsilon}{\sqrt{M}}\right\}\\ \le\;& \sum_{j=1}^{M}\mathbb{P}\left\{\sup_{0\le t\le T}\left|\overline{M}_{ji}^N(t)\right|>\frac{\varepsilon}{\sqrt{M}(2M+1)}\right\}\\ &+\sum_{j=1}^{M}\mathbb{P}\left\{\sup_{0\le t\le T}\left|\overline{M}_{ij}^N(t)\right|>\frac{\varepsilon}{\sqrt{M}(2M+1)}\right\}\\ &+\mathbb{P}\left\{\sup_{0\le t\le T}\left|\overline{M}_{i}^N(t)\right|>\frac{\varepsilon}{\sqrt{M}(2M+1)}\right\}\end{aligned}\tag{25}$$

Now, using the fact, that,

$$\int_0^T \mu_i(s)p_{ij}(s)\left(\overline{X}_i^N(s)\wedge 1\right)^+ ds \le k_T T \tag{26}$$

we have

$$\begin{aligned}&\mathbb{P}\left\{\sup_{0\le t\le T}\left|\overline{M}_{ij}^N(t)\right|>\frac{\varepsilon}{\sqrt{M}(2M+1)}\right\}\\ =\;& \mathbb{P}\left\{\sup_{0\le t\le T}\left\|\frac{1}{\kappa_N N^{\beta}}T_{ij}\left(\kappa_N N^{\beta}\int_0^t \mu_i(s)p_{ij}(s)\right.\right.\right.\\ &\left.(\overline{X}_i^N(s)\wedge 1)^+ds\right)-\int_0^t \mu_i(s)p_{ij}(s)(\overline{X}_i^N(s)\wedge 1)^+ds\Bigg\|\\ &\left.>\frac{\varepsilon}{\sqrt{M}(2M+1)}\right\}\\ \le\;& \mathbb{P}\left\{\sup_{0\le t\le k_T T}\left\|\frac{1}{\kappa_N N^{\beta}}T_{ij}(\kappa_N N^{\beta}t)-t\right\|\right.\\ &\left.>\frac{\varepsilon}{\sqrt{M}(2M+1)}\right\}\\ \le\;& C_1(\varepsilon,T)e^{-\kappa_N N^{\beta}C_2(\varepsilon,T)}\end{aligned}\tag{27}$$

for $C_1(\varepsilon,T), C_2(\varepsilon,T) > 0$, and the last step follows from Kurtz' Theorem for scaled Poisson processes (c.f. Theorem 5.3 in [3].) Similarly,

$$\begin{aligned}&\mathbb{P}\left\{\sup_{0\le t\le T}\left|\overline{M}_{i}^N(t)\right|>\frac{\varepsilon}{\sqrt{M}(2M+1)}\right\}\\ \le\;& C_1(\varepsilon,T)e^{-\kappa_N N^{\beta}C_2(\varepsilon,T)}\end{aligned}\tag{28}$$

It then follows from eqns. (24,25)

$$\begin{aligned}&\mathbb{P}\left\{\sup_{0\le t\le T}\left\|\overline{\mathbf{M}}^N(t)\right\|>\varepsilon\right\}\\ \le\;& \left(M(2M+1)C_1(\varepsilon,T)\right)e^{-\kappa_N N^{\beta}C_2(\epsilon,T)}\end{aligned}\tag{29}$$

From eqns. (22,18) we have

$$\begin{aligned}&\overline{\mathbf{X}}^N(t)-\overline{\mathbf{Z}}^N(t)\\ =\;& \int_0^t\left[I-P^T(s)\right]\Upsilon(s)\left(\left(\overline{\mathbf{Z}}^N(s)\wedge 1\right)^+\right.\\ &\left.-\left(\overline{\mathbf{X}}^N(s)\wedge 1\right)^+\right)ds+\overline{\mathbf{M}}^N(t)\quad 0\le t\le T\end{aligned}\tag{30}$$

Now choose $\overline{k}_T > 0$ sufficiently large, such that,

$$\left\|\left(I-P^T(s)\right)\Upsilon(s)\right\|\le \overline{k}_T,\quad 0\le s\le T \tag{31}$$

Using the fact that

$$\left\|\left(\overline{\mathbf{X}}^N(s)\wedge 1\right)^+-\left(\overline{\mathbf{Z}}^N(s)\wedge 1\right)^+\right\|\le\left\|\overline{\mathbf{X}}^N(s)-\overline{\mathbf{Z}}^N(s)\right\| \tag{32}$$

we have from eqn. (30)

$$\begin{aligned}&\left\|\overline{\mathbf{X}}^N(t)-\overline{\mathbf{Z}}^N(t)\right\|\\ \le\;& \overline{k}_T\int_0^t\left\|\overline{\mathbf{X}}^N(s)-\overline{\mathbf{Z}}^N(s)\right\|ds\\ &+\left\|\overline{\mathbf{M}}^N(t)\right\|\quad 0\le t\le T\end{aligned}\tag{33}$$

To show exponential equivalence, consider the set $\Omega_{\varepsilon}\subset\Omega$, where $\sup_{0\le t\le T}\left\|\overline{\mathbf{M}}^N(t)\right\|\le\varepsilon$. Then from eqn. (33) we have on Ω_{ε}

$$\begin{aligned}&\left\|\overline{\mathbf{X}}^N(t)-\overline{\mathbf{Z}}^N(t)\right\|\\ \le\;& \overline{k}_T\int_0^t\left\|\overline{\mathbf{X}}^N(s)-\overline{\mathbf{Z}}^N(s)\right\|ds\\ &+\varepsilon\quad 0\le t\le T\end{aligned}\tag{34}$$

Then by Gronwall's inequality, on Ω_{ε}

$$\left\|\overline{\mathbf{X}}^N(t)-\overline{\mathbf{Z}}^N(t)\right\|\le\varepsilon e^{\overline{k}_T t}\quad 0\le t\le T \tag{35}$$

Thus, on Ω_{ε}, we have

$$\sup_{0\le t\le T}\left\|\overline{\mathbf{X}}^N(t)-\overline{\mathbf{Z}}^N(t)\right\|\le\varepsilon e^{\overline{k}_T T} \tag{36}$$

Therefore, we obtain

$$\mathbb{P}\left\{\sup_{0\le t\le T}\left\|\overline{\mathbf{X}}^N(t)-\overline{\mathbf{Z}}^N(t)\right\|\le\varepsilon e^{\overline{k}_T T}\right\}\ge\mathbb{P}\left\{\Omega_{\varepsilon}\right\} \tag{37}$$

From eqns. (29,37)

$$\begin{aligned}
& \mathbb{P}\left\{\sup_{0\le t\le T}\left\|\overline{\mathbf{X}}^N(t)-\overline{\mathbf{Z}}^N(t)\right\|>\varepsilon e^{\overline{k}_T T}\right\} \\
\le\ & 1-\mathbb{P}\left\{\Omega_\varepsilon\right\} \\
=\ & 1-\mathbb{P}\left\{\sup_{0\le t\le T}\left\|\overline{\mathbf{M}}^N(t)\right\|\le\varepsilon\right\} \\
=\ & \mathbb{P}\left\{\sup_{0\le t\le T}\left\|\overline{\mathbf{M}}^N(t)\right\|>\varepsilon\right\} \\
\le\ & \left(M(2M+1)C_1(\varepsilon,T)\right)e^{-\kappa_N N^\beta C_2(\varepsilon,T)} \qquad (38)
\end{aligned}$$

Now consider any $\delta > 0$. Then taking $\varepsilon = \delta e^{-\overline{k}_T T}$ in eqn. (38) we have

$$\mathbb{P}\left\{\sup_{0\le t\le T}\left\|\overline{\mathbf{X}}^N(t)-\overline{\mathbf{Z}}^N(t)\right\|>\delta\right\}\le C_3(\delta,T)e^{-\kappa_N N^\beta C_4(\delta,T)} \qquad (39)$$

where

$$C_3(\delta,T)=(M(2M+1)C_1(\delta e^{-\overline{k}_T T},T)>0 \qquad (40)$$

$$C_4(\delta,T)=C_2(\delta e^{-\overline{k}_T T},T)>0 \qquad (41)$$

Since, $\kappa_N\to\infty$, it follows from eqn. (39) that

$$\limsup_{N\to\infty}\frac{1}{N^\beta}\log\mathbb{P}\left\{\sup_{0\le t\le T}\left\|\overline{\mathbf{X}}^N(t)-\overline{\mathbf{Z}}^N(t)\right\|>\delta\right\}=-\infty \qquad (42)$$

thus establishing exponential equivalence at scale N^β.

We now complete the last step in the proof of Theorems 1 and 2, which is establishing the LDP for the sequence $\{\overline{\mathbf{Z}}^N\}$.

Lemma 4 Under the assumptions of Theorem 1, the sequence $\{\overline{\mathbf{Z}}^N\}$ satisfies an LDP at scale N^β with good rate function in $\mathcal{D}^M[0,T]$ equipped with either the Skorokhod J_1 topology or the topology of uniform convergence. In particular, if $I(\cdot)$ be the rate function of $\{\overline{\mathbf{E}}^N\}_{N\in\mathbb{N}}$ and the sequence of the scaled initial states are fixed at $\mathbf{x}\in\mathbb{R}_+^M$, the rate function $J_T^{\mathbf{x}}(\cdot)$ for $\{\overline{\mathbf{Z}}^N\}$ is given by eqn. (17.)

Proof: We note that

$$\overline{\mathbf{Z}}^N=\Lambda\left(\overline{\mathbf{X}}^N(0),\overline{\mathbf{E}}^N\right),\quad\forall N\in\mathbb{N} \qquad (43)$$

As established earlier, the function $\Lambda(\cdot)$ is continuous in both the Skorokhod J_1 topology and the topology of uniform convergence. Hence the claim follows from the contraction principle (c.f. Lemma 3.13 in [4].)

Theorems 1 and 2 now follow as immediate consequences of Lemmas 3 and 4.

V. An Example of Batch Arrivals

Consider the batch arrival process described in Example 2 of Section I-C in heavy traffic. Without loss of generality assume $\lambda=\mu=1$. Let the "local rate function" $\ell(\cdot,\cdot)$ be defined by

$$\ell(x,y)=(y+(x\wedge 1))\ln(y+\ln(x\wedge 1))-(y+(x\wedge 1))+1.$$

Then, it can be shown from Theorems 1 and 2 that the sequence of the scaled number in system, $\{\overline{X}^N\}_{N\ge 0}$ satisfies an LDP at scale N^β in $\mathcal{D}^x[0,T]$ (equipped with the Skorokhod J_1 topology) with good rate function, $J_T^x(\cdot)$, where

$$J_T^x(\phi)=\begin{cases}\displaystyle\int_0^T\ell(\phi(s),\dot{\phi}(s))\,ds & \text{if }\phi\in\mathcal{AC}_+^x[0,T]\\ \infty & \text{otherwise}\end{cases} \qquad (44)$$

and $\mathcal{AC}_+^x[0,T]$ is the space of non-negative absolutely continuous functions starting at x.

The variational problems leading to the decay rate of the probabilities of rare events are solvable in this case and will be presented in a forthcoming paper.

VI. Conclusion

The paper presents a sample path large deviations analysis in the many-server regime, which is very relevant for applications but appears not to have received much attention in the literature thus far. Specifically, a sample path large deviations is established in the setting of a Jackson network of many-server queues for a certain class of arrival processes that includes long range dependent and batch arrivals. We show that the LDP rate function of the sequence of scaled number in system is related to the LDP rate function of the arrival sequence through a continuous map. This reduces the problem of estimating the decay rate of probabilities of rare events into solving variational problems, which also shed light into the way rare events occur. As an illustration of the usefulness of our result, we present an example, namely the batch arrival case, for which the variational problem is explicitly solvable. The analysis presented here complements forthcoming work on sample path large deviations in the many-server regime with different assumptions on the cumulative arrival processes, that allows for the consideration of Poisson arrival processes.

VII. Acknowledgment

The work of Soummya Kar and José M. F. Moura was supported by the DARPA DSO Advanced Computing and Mathematics Program Integrated Sensing and Processing (ISP) Initiative under ARO grant # DAAD19-02-1-0180, by NSF under grants # ECS-0225449 and # CNS-0428404, and by an IBM Faculty Award. The work of Kavita Ramanan was supported by the NSF under grants DMS 0405343 and CMMI 0728064.

References

[1] A. Ganesh, N. O'Connell, and D. Wischik, *Big Queues*. Springer-Verlag Berlin Heidelberg, 2004.

[2] M. Mandjes, *Large Deviations for Gaussian Queues*. Wiley, Chichester, UK, 2007.

[3] A. Shwartz and A. Weiss, *Large Deviations for Perfomance Analysis*. Chapman and Hall, 1994.

[4] J. Feng and T. G. Kurtz, *Large Deviations for Stochastic Processes*. American Mathematical Society, 2006.

[5] H. Chen and D. D. Yao, *Fundamentals of Queueing Networks*. Springer-Verlag New York, Inc., 2001.

[6] A. Mandelbaum, W. Massey, and M. I. Reiman, "Strong approximations for Markovian service networks," *Queueing Systems*, vol. 30, pp. 149–201, 1998.

[7] G. Pang, R. Talreja, and W. Whitt, "Martingale proofs of many-server heavy-traffic limits for Markovian queues," *Probability Surveys*, vol. 4, pp. 193–267, 2007.