Large Deviations of Queues under QoS Scheduling Algorithms

Alexander L. Stolyar
Bell Labs, Lucent Technologies
600 Mountain Ave, 2C-322
Murray Hill, NJ 07974, USA
stolyar@research.bell-labs.com

***Abstract*— We consider a model where multiple queues are served by a server whose capacity varies randomly and asynchronously with respect to different queues. The problem is to optimally control large deviations of the queues in the following sense: find a scheduling rule maximizing**

$$\min_i \left[\lim_{n\to\infty} \frac{-1}{n} \log P\left(a_i Q_i > n\right) \right] \ , \qquad (1)$$

where Q_i is the length of i-th queue in a stationary regime, and $a_i > 0$ are parameters. Thus, we seek to maximize the minimum of the exponential decay rates of the tails of distributions of weighted queue lengths $a_i Q_i$. We give a characterization of the upper bound on (1) under any scheduling rule, and of the lower bound on (1) under the *exponential* (EXP) rule. For the case of two queues, we prove that the two bounds match, thus proving optimality of EXP rule in this case.

The EXP rule is *not* asymptotically invariant with respect to scaling of the queues, which complicates its analysis in large deviations regime. To overcome this, we introduce and prove a refined sample path large deviations principle, or *refined Mogulsky theorem*, which is of independent interest.

I. Introduction

The model we consider in this paper is motivated primary by the problem of scheduling transmissions of multiple data users (flows) sharing the same wireless channel (server). As an example, one can think of the following scenario: a wireless access point, or base station, receives data traffic flows destined to several different mobile users, and needs to schedule data transmissions to the users over a shared wireless channel, so that the channel is used efficiently. (Cf. [3], [1], [15] for a more detailed discussion of this scenario.) The distinctive feature of this model, which separates it from more "conventional" queueing models, is the fact that the capacity (service rate) of the channel varies with time randomly and *asynchronously* with respect to different users.

A little more precisely (but still informally), the model is as follows. There are N exogenous input (traffic) flows, which are queued in separate (infinite capacity) buffers, before they can be served by a channel. Time is divided into slots. The channel can serve only one of the flows in one slot. The "aggregate state" of the channel varies randomly from slot to slot. If the channel state in a given slot is m and flow i is chosen for service in this slot, the service rate is $\mu_i^m \geq 0$, i.e., μ_i^m customers (bits of data) of flow i are served (transmitted) and leave the system. This and related models received a significant amount of attention in recent years (cf. [13] for an overview). It is well known that efficient scheduling rules cannot be "channel state oblivious." However, it is also known that large classes of rather "parsimonious" algorithms, making scheduling decisions based only on the current channel state and current queue lengths (and/or current head-of-the-line queueing delays) information can in fact achieve certain notions of efficiency. For example, MaxWeight-type algorithms (cf. [2] and references therein) and Exponential (EXP) algorithm [9] are *throughput optimal* in the sense that they ensure stochastic stability of the queues as long as such is feasible at all, under any rule. Also, both MaxWeight and EXP rules exhibit optimal behavior under heavy traffic conditions (see [13], [10]).

In this paper we would like to address the following issue. Suppose we want to find a scheduling algorithm (rule), or queueing discipline, under which the following Quality-of-Service condition is satisfied:

$$P\{Q_i > B_i\} \leq \delta_i \ , \ i = 1, \ldots, N \ , \qquad (2)$$

where Q_i is the steady state queue length for flow i, $B_i > 0$ is a predefined threshold, and δ_i is the maximum acceptable probability of queue length exceeding the threshold. (This problem appears in a variety of applications, cf. [4], [11], [12] for a further discussion and reviews.)

If thresholds B_i are "large," then conditions (2) can be "approximately" replaced by the following asymptotic - "tail" - conditions

$$\beta(Q_i) \geq a_i \ , \ i = 1, \ldots, N \ , \qquad (3)$$

where we use the notation

$$\beta(X) \doteq \lim_{n\to\infty} -\frac{1}{n} \log P(X > n) \qquad (4)$$

for the exponential decay rate of the tail of distribution of random variable X (assuming the above limit exists), and

$$a_i = -\log(\delta_i)/B_i.$$

This is precisely what we will do in this paper. *We consider the problem of finding a scheduling rule such that the tail conditions (3) are satisfied for some fixed set of positive parameters a_i.*

This problem in turn is equivalent to solving the following *optimization* problem

$$\text{maximize} \quad \min_{i=1,\ldots,N} a_i^{-1} \beta(Q_i) \ , \qquad (5)$$

where the maximization is over all scheduling disciplines. Indeed, a discipline satisfying (3) exists if and only if the

maximum in (5) is 1 or greater (and the maximum is attained, to be precise). Finally, if we denote by

$$Q_* \doteq \max_i a_i Q_i$$

the *maximal weighted queue length*, and observe that $\min_i a_i^{-1}\beta(Q_i) = \beta(Q_*)$, we see that the problem (5) is equivalent to

$$\text{maximize} \quad \beta(Q_*) \ . \tag{6}$$

To summarize, *we want to find a scheduling rule solving problem (6), i.e. a rule maximizing the exponential decay rate of the tail of the distribution of the maximal weighted queue length Q_*, with some fixed "weights" $a_i > 0$.*

In the case when channel is *not* time-varying, i.e., there is only one channel state and therefore the (potential) service rates μ_i are constant, our model essentially fits into the framework of [11], where, in particular, it is proved that an extremely simple rule always choosing for service the queue maximizing $a_i Q_i$ is an optimal solution to problem (5). (This result was extended in [12] to a queueing network setting.) However, for our model, where the channel *is* time-varying, the above simple rule *cannot possibly be optimal for problem (5)*, because it ignores the current state of the channel; moreover, except for degenerate cases, this rule is not even throughput-optimal - it can make queues unstable in cases when stability (under a different rule) is feasible. The main goal of this paper is to establish optimality of the EXP rule for the problem (5). The EXP rule is defined as follows: when channel is in state m,

$$\text{Serve flow } i \text{ maximizing} \quad \mu_i^m \exp\left(\frac{a_i Q_i}{1+\overline{Q}^\eta}\right), \tag{7}$$

where $\overline{Q} \doteq (1/N)\sum_i a_i Q_i$, and $\eta \in (0,1)$ is a fixed parameter.

Problems like (5) are naturally approached using Large Deviations (LD) theory techniques. It is well known in LD theory that, roughly speaking, the value of $\beta(Q_*)$ under a given scheduling rule is determined by a "most likely path" for the process $Q_*(t)$ to reach level n, starting from 0. (See the definition of $\beta(\cdot)$ in (4).) Or, equivalently, this is a most likely path for a "fluid-scaled" process $(1/n)Q_*(nt)$ to reach level 1. In turn, the likelihoods of such rescaled paths are determined by a sample path large deviations principle (Mogulsky theorem) for the sequence of fluid-scaled "driving processes" - namely, input flow and channel state processes, as $n \to \infty$. (If the value of the corresponding LD rate function of a path - or path "cost" - is c, then the "probability" of the path is "approximately" e^{-cn}, when n is large.)

One of the difficulties in the LD analysis of the EXP rule is that the "standard" sample path large deviations principle (SP-LDP) is not sufficient for "keeping track" of the path costs. The basic reason for this is that *EXP rule is not asymptotically invariant with respect to scaling of queue lengths*. An "asymptotically scaling-invariant" rule is roughly such that, when queue lengths are large, a scaling of all queue lengths by the same factor, at any given time, does not change the scheduling choice. An example of scaling-invariant rule is a MaxWeight-type algorithm, choosing for service a flow i maximizing $c_i Q_i^\gamma \mu_i^m$, where γ and all c_i are arbitrary positive parameters. A slightly more general rule, maximizing $[c_i Q_i^\gamma + d_i]\mu_i^m$, where d_i's are additional parameters, is not scaling-invariant, but is asymptotically scaling-invariant.

Fluid scaling is the "relevant" one to study the dynamics of the queue lengths under an asymptotically scaling-invariant rule in an (unscaled) time interval of the order of $O(n)$ (because rescaling of queue lengths by $1/n$, for any n, "preserves the information" on which scheduling choices are made), and a standard SP-LDP gives the likelihood of trajectories under this scaling. In contrast, the EXP rule is not asymptotically scaling-invariant, as seen from the expression in (7). Even if eventually we are interested in the dynamics of the queue lengths under EXP rule over an interval of the order $O(n)$, the "relevant" time and space scale which determines such dynamics is of the order $O(n^\eta)$. (The value of $\overline{Q}$ is "typically" $O(n)$. Therefore, the differences of the order $O(n^\eta)$ between weighted queue lengths $a_i Q_i$ result in the order $O(1)$ ratios of the exponent terms in (7) for different flows i. But, these ratios are what determines the scheduling choices.) Consequently, we need the likelihoods of (unscaled) trajectories over order $O(n^\eta)$ time intervals; fluid scaling, however, does not "preserve" this information. To resolve this difficulty, we introduce and prove what can be called a "refined" SP-LDP, or a *refined Mogulsky theorem* (RMT). Using RMT we introduce the notions of a *generalized fluid sample path* (GFSP) and its *refined cost*. (Roughly speaking, the refined cost of a GFSP "takes into account" the behavior of (unscaled) process trajectories on time scales that are "finer" than $O(n)$.) We show that the likelihood of building large value of Q_* under EXP rule can be given in terms of GFSP refined costs.

Our RMT result (Theorem 2) and the notions of GFSP and its refined cost are generic and are of independent interest. In particular, as the above discussion demonstrates, they are instrumental in LD analysis of scheduling rules that are not asymptotically scaling-invariant.

The **main results** of the paper are as follows. We prove the upper bound $\beta(Q_*) \le J_*$, which holds under any scheduling rule, where J_* is defined in terms of lowest cost "simple" (linear) paths to raise Q_*. The proof of this upper bound involves only a standard Mogulsky theorem for the sequence of fluid-scaled input flow and channel state processes. We introduce and prove a refined Mogulsky theorem, and introduce the related notion of GFSP. We then give the lower bound $\beta(Q_*) \ge J_{**}$, which holds for the EXP rule, where J_{**} is defined in terms of the lowest refined cost of a GFSP to raise Q_*. Finally, for the case of EXP rule and two flows, we show that the lower and upper bounds on $\beta(Q_*)$ match, that is $\beta(Q_*) = J_{**} = J_*$, thus proving that the EXP rule is indeed an optimal solution to problem (5) in this case. (In fact, the complete proof of the latter fact is given in [14]; in this paper, due to space limitation, we only give an informal description of the key points of the

proof.) Proving equality $J_{**} = J_*$ (and thus optimality of the EXP rule) for arbitrary number of flows is a subject of future work.

Previous work on the large deviations regime for queues served by a time-varying server includes [16], which contains results for a MaxWeight-type rule (maximizing $Q_i\mu_i^m$) in a symmetric model. ("Symmetric" means: all input flows have equal rate and are non-random; channel state $m = (m_1, \dots, m_N)$ is a direct product of N independent and identically distributed channel states m_i of the individual flows.) The optimality problem (5) is not addressed in [16], and the analysis relies in essential way on the symmetry assumptions.

The rest of the paper is organized as follows. In Section II we introduce basic notations, definitions, conventions used in the paper. The system model, formal definition of the EXP rule, and our main results (Theorem 1) regarding the estimates of $\beta(Q_*)$ under an arbitrary rule and EXP rule, including optimality of EXP in the case of two flows, are given in Section III. The necessary definitions of a sequence of scaled processes and a standard SP-LDP (Mogulsky theorem) are presented in Sections IV and V, respectively. In Section VI we prove the bound $\beta(Q_*) \le J_*$ (Theorem 1(i)) for any scheduling discipline. A refined Mogulsky theorem is formulated and proved in Section VII. Section VIII contains the definition of a GFSP and proof of the bound $\beta(Q_*) \ge J_{**}$ (Theorem 1(ii)) under EXP rule. Finally, Sections IX and X contain an informal description of the proof of optimality of EXP rule for two flows (Theorem 1(iii)). (The detailed proof in given in [14].)

II. Basic notation and definitions

We denote by $\mathbb{R}$ and $\mathbb{R}_+$ the sets of real and real non-negative numbers, respectively. The corresponding k-times product spaces are $\mathbb{R}^k$ and $\mathbb{R}_+^k$. Eucledian norm of vector $a \in \mathbb{R}^k$ is $\|a\|$.

The minimum of two real numbers ξ_1 and ξ_2 is $\xi_1 \wedge \xi_2$. and by $\lfloor \xi \rfloor$ and $\lceil \xi \rceil$ the integer part and the ceiling of a real number ξ, respectively.

Let $\mathcal{D}$ be the space of RCLL functions (i.e. right continuous functions with left limits) defined on $[0, \infty)$ and taking values in $\mathbb{R}$. Unless otherwise specified, we assume $\mathcal{D}$ is endowed with the topology of uniform convergence on compact sets (u.o.c.). As a measurable space, we always assume that $\mathcal{D}$ is endowed with the σ-algebra generated by the cylinder sets. By $\mathcal{A}$ we denote the subset of absolutely continuous functions in $\mathcal{D}$, and by $\mathcal{A}_0 \subset \mathcal{A}$ the subset of functions $h(\cdot)$ with $h(0) = 0$. For any function space S, and any $0 \le c < d$, $\zeta_c^d S$ denotes the space of functions in S with the domain "truncated" to $[c, d]$. The subspaces and spaces with truncated domains inherit the topology and σ-algebra of $\mathcal{D}$. Given any space S, we assume that the k times product space S^k has the product topology and product σ-algebra defined in the natural way.

For any $s \ge 0$ and $h = (h_1, \dots, h_k) \in \mathcal{D}^k$ [or $\zeta_c^d \mathcal{D}^k$]], we define the norm

$$||h||_s \doteq \max_{i=1,\dots,k} \sup_{t \le s} |h_i(s)|.$$

Thus the u.o.c. convergence in $\mathcal{D}^k$ [or $\zeta_c^d \mathcal{D}^k$] is equivalent to convergence in norm $||\cdot||_s$ for all $s > 0$.

We define the scaling operator $\Gamma^c,\ c > 0$, for $h \in \mathcal{D}^k$ as follows:

$$(\Gamma^c h)(t) \doteq \frac{1}{c} h(ct)\ . \tag{8}$$

For a function $h \in \mathcal{D}$, we define the domain truncation operator ζ_c^d, for $0 \le c < d$, in the natural way:

$$\zeta_c^d h \in \zeta_c^d \mathcal{D} \quad \text{and} \quad (\zeta_c^d h)(t) = h(t).$$

For $h \in \mathcal{D}$, and $0 \le c < d$, we also define operator $\bar{\zeta}_c^d$ (which is a simultaneous domain truncation and shift, as well as recentering) as follows:

$$\bar{\zeta}_c^d h \in \zeta_0^{d-c} \mathcal{D} \quad \text{and} \quad (\bar{\zeta}_c^d h)(t) = h(c+t) - h(c).$$

For a set of functions, operators ζ_c^d and $\bar{\zeta}_c^d$ are applied componentwise.

We use symbol $\Rightarrow$ for the *weak convergence of deterministic functions* in the space $\bar{\mathcal{D}}$, which is the space of RCLL functions taking values in the set $\bar{\mathbb{R}}$ of real numbers extended by including $+\infty$ and $-\infty$ (with the natural topology on $\bar{\mathbb{R}}$). If $h, g \in \bar{\mathcal{D}}$, then $h \Rightarrow g$ means $h(t) \to g(t)$ for every $t > 0$ where g is continuous. (Convergence at $t = 0$ is not required.) The weak convergence $\Rightarrow$ of functions in $\bar{\mathcal{D}}^k$ is understood component-wise.

Let $\mathbf{\Omega} \doteq (\Omega, \mathcal{F}, P)$ be a probability space. We assume that $\mathbf{\Omega}$ is large enough to support all the independent random processes that we use in the paper. Given any subset B of a topological space, we use $\bar{B}$ and B° to denote its closure and interior respectively.

Typically, we follow the convention of using bold font for stochastic processes and Roman font for deterministic functions, including realizations of the random processes.

The following is the standard definition of a large deviation principle [6, p.5].

Definition 1: (LDP) Let $\mathcal{X}$ be a topological space and $\mathcal{B}$ a σ-algebra on $\mathcal{X}$ (which is not necessarily the Borel σ-algebra). A sequence of random variables $\{\mathbf{X}_n\}$ on Ω taking values in $\mathcal{X}$ is said to satisfy the LDP with good rate function I if for all $B \in \mathcal{B}$,

$$\limsup_{n\to\infty} \frac{1}{n} \log P(\mathbf{X}_n \in B) \le - \inf_{x \in \bar{B}} I(x),$$

and

$$\liminf_{n\to\infty} \frac{1}{n} \log P(\mathbf{X}_n \in B) \ge - \inf_{x \in B^\circ} I(x),$$

where $I : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ is a function with compact level sets.

III. The model and main results

A. The model.

The system has N input flows, consisting of discrete *customers*, which need to be served by a single *channel* (or server). We will denote by N both the set of flows $\{1,\ldots,N\}$ and its cardinality. Each flow has its own queue where customers wait for service. (Sometimes, we use terms "flow" and "queue" interchangeably.)

The system operates in discrete time. A time interval $[t, t+1)$, with $t = 0, 1, 2, \ldots$, we will call the *time slot* t. In each time slot the channel can be in one out of the finite set $M = \{1,\ldots,M\}$ of *channel states*, and it can pick one of the flows for service. If in a given time slot the channel is in state $m \in M$ and flow $i \in N$ is chosen for service, then the integer number $\mu_i^m \geq 0$ customers are served from the corresponding queue i (or the entire queue i content, if it is less than μ_i^m). Thus, associated with each channel state $m \in M$ is the fixed vector of service rates $(\mu_1^m,\ldots,\mu_N^m)$.

The channel state $m(t)$ in each time slot t is drawn independently according to some probability distribution $\pi = (\pi^1,\ldots,\pi^M)$. Without loss of generality, we can and will assume that $\pi_m > 0$ for all states m. (The i.i.d. assumption for the sequence of channel states is to simplify notation and exposition. All our results are easily generalized for example to the case when the random channel state process $m(t),\ t = 0, 1, 2, \ldots$, is an irreducible discrete time Markov chain (cf. [7]) with the finite state space M.)

Denote by $A_i(t)$ the number of type i customers arrived in time slot $t = 1, 2, \ldots$. We will adopt a convention that the customers arriving in slot t are immediately available for service in this slot. We will assume that all arrival processes are mutually independent, each sequence $A_i(t),\ t = 1, 2, \ldots$, is i.i.d., with finite exponential moments

$$Ee^{\theta A_i(1)} < \infty,\ \forall \theta \geq 0, \quad \forall i,$$

and, finally, that each $A_i(1)$ is unbounded. (The existence of exponential moments assumption is essential. The unboundedness of $A_i(1)$ is not essential at all, and assumed to simplify exposition. The flow independence and i.i.d. assumptions can be much relaxed, as long as the sequence of scaled joint arrival processes satisfies an LDP, and "has no memory" in the limit.)

Let us denote by $\bar{\lambda}_i \doteq EA_i(1),\ i = 1,\ldots,N$, the mean arrival rate for flow i, and assume that $\bar{\lambda}_i > 0$ for all i.

The random process describing the behavior of the system is

$$Q(t) = (Q_i(t),\ i = 1,\ldots,N), \quad t = 0, 1, 2, \ldots$$

where $Q_i(t)$ is the type i queue length at time t.

B. Scheduling Rules. Stability.

A *scheduling rule*, or a *queueing discipline*, picks one flow to be served in a given time slot t, depending on the current queue length vector $Q(t)$.

If we denote by $D_i(t) = \min\{Q_i(t-1), \mu_i^{m(t-1)}\}$, the number of type i customers served in the time slot $t-1$, then according to our conventions, for each $t = 1, 2, \ldots$,

$$Q_i(t) = Q_i(t-1) - D_i(t) + A_i(t),\ \forall i.$$

Note, that in any time slot t, $D_i(t)$ can be positive for at most one of the flows i, and it is zero for all other flows.

Obviously, under any scheduling rule, $Q(\cdot)$ is a Markov chain with countable state space. We say that the system under a given scheduling rule is *stable* if the Markov chain has a finite subset of states which is reachable from any other state with probability 1, and each state within the subset is positive recurrent. Stability implies existence of a stationary probability distribution. (If the Markov chain happens to be irreducible, stability is equivalent to ergodicity, and the stationary distribution is unique.)

Suppose a stochastic matrix $\phi = (\phi_{mi},\ m \in M,\ i = 1,\ldots,N)$ is fixed, which means that $\phi_{mi} \geq 0$ for all m and i, and $\sum_i \phi_{mi} = 1$ for every m. Given ϕ we define vector $v = (v_1,\ldots,v_N) = v(\phi)$ as follows:

$$v_i = \sum \pi^m \phi_{mi} \mu_i^m\ ,\ i \in N. \tag{9}$$

If each component ϕ_{mi} of matrix ϕ is interpreted as a "long-term" average fraction of time slots when flow i is chosen for service, out of those slots when the channel state is m, then $v(\phi)$ is simply the vector of average service rates which will be "given" to the flows. The set

$$V \doteq \{w \in R_+^N \mid w \leq v(\phi) \text{ for some } \phi\}$$

is called system *(service) rate region*.

It is well known (cf. [13] and references therein) that condition $\bar{\lambda} \in V$ is necessary for stability. Throughout this paper we assume a slightly stronger condition:

$$\bar{\lambda} < v^* \quad \text{for some} \quad v^* \in V. \tag{10}$$

(Under our arrival process assumptions, in particular the unboundedness of $A_i(1)$, it is not hard to show that condition (10) is also necessary for stability.)

C. Exponential Scheduling Rule.

Let a set of positive parameters $a_1,\ldots,a_N$ and $\eta \in (0,1)$ be fixed. The following scheduling rule is called Exponential [9], or EXP: it chooses for service in time slot t a single queue

$$i \in i(Q(t)) = \arg\max_i c_i\ \mu_i(t)\ \exp\left(\frac{a_i Q_i(t)}{c + [\overline{Q}(t)]^\eta}\right)\ , \tag{11}$$

where $\mu_i(t) \equiv \mu_i^{m(t)}$, $\overline{Q}(t) \doteq (1/N)\sum_i a_i Q_i(t)$, and $c, c_1, \ldots, c_N$, are some additional positive parameters. (Ties are broken in an arbitrary, but a priori fixed way, for example in favor of the smallest index within set $i(Q(t))$.)

Proposition 1: [9] If condition (10) holds, the system under the EXP rule is stable.

Proposition 1 says that the EXP rule is *throughput optimal* in the sense that it makes system stable as long as stability is feasible at all.

In the rest of the paper, to simplify exposition, we assume that parameters $c, c_1, \ldots, c_N$ are all equal to 1. (Setting these parameters to arbitrary values does not affect main results, and it does not affect the proofs in any essential way.)

D. Main results.

The function $Q_*(t) \doteq \max_i a_i Q_i(t)$ of the state $Q(t)$ will be called *maximal weighted queue length*. (The corresponding *random processes* are denoted by $\mathbf{Q}(t)$ and $\mathbf{Q}_*(t)$, $t = 0, 1, 2, \ldots$.) It will be convenient to extend the domain of $Q(\cdot)$ and $Q_*(\cdot)$ (as well as other functions introduced later in the paper), which are naturally defined in discrete time, to continuous time t by adopting the convention that the functions are constant within each time slot $[k, k+1)$, where k is integer. Now we are in position to formulate our main result.

Theorem 1: Suppose condition (10) is satisfied. Then, the following holds.

(i) There exists $T^0 \in (0, \infty)$ such that for any scheduling rule and any $t > T^0$, we have the following lower bound:

$$\liminf_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\mathbf{Q}_*(nt) > 1\right) \geq -J_* \ , \qquad (12)$$

where $J_* > 0$ is defined and explained later in Section VI.

(ii) Consider the system under the EXP scheduling rule and the Markov chain $\mathbf{Q}(\cdot)$ being in a stationary regime (which exists by Proposition 1). Then, we have the following upper bound:

$$\limsup_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\mathbf{Q}_*(0) > 1\right) \leq -J_{**} \ , \qquad (13)$$

where J_{**}, $0 \leq J_{**} \leq J_*$, is defined and explained later in Section VIII (see (28)).

(iii) Consider the system with two flows, $N = \{1, 2\}$, under the EXP scheduling rule, in a stationary regime. Then, $J_{**} = J_*$ and therefore

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\mathbf{Q}_*(0) > 1\right) = -J_* \ . \qquad (14)$$

Theorem 1(iii) shows that, in the case of two flows, the EXP rule is optimal in that it maximizes the exponential decay rate of the stationary distribution of the maximal weighted queue length $\mathbf{Q}_*(\cdot)$. Extending Theorem 1(iii) to arbitrary number of flows is a subject for future work.

IV. Extended description of the system process. Sequence of fluid-scaled processes

As formulation of Theorem 1 suggests (and is typical for this type of large deviations results for queueing systems), its proof involves considering a sequence of "fluid-scaled" versions of the queue length process $\mathbf{Q}$, namely the processes $\Gamma^n\mathbf{Q} = ((1/n)\mathbf{Q}(nt),\ t \geq 0)$, for $n = 1, 2, \ldots$. In this section we define this sequence formally. But first, we need to introduce additional functions associated with the system evolution.

For $t \geq 0$ let

$$F_i(t) \doteq \sum_{k=1}^{\lfloor t \rfloor} A_i(k) \quad \text{and} \quad \hat{F}_i \doteq \sum_{k=1}^{\lfloor t \rfloor} D_i(k)$$

denote the total number of flow i customers, respectively arrived to and departed from the system by (and including) time t, that is in the time slots $1 \leq k \leq \lfloor t \rfloor$. (Recall our convention, introduced in Section III-D, that we extend the domain of discrete time processes to continuous time $t \geq 0$.) Also, denote by $G_m(t)$ the total number of time slots $0 \leq k \leq \lfloor t-1 \rfloor$ when the channel was in state m; and by $\hat{G}_{mi}(t)$ the number of time slots $0 \leq k \leq \lfloor t-1 \rfloor$ when the server state was m and flow i was chosen for service.

The following set of functions describes the evolution of the system in time interval $[0, \infty)$:

$$(Q, Q_*, F, \hat{F}, G, \hat{G}),$$

where

$$Q = (Q(t) = (Q_1(t), \ldots, Q_N(t)),\ \ t \geq 0),$$
$$Q_* = (Q_*(t) \equiv \max_i Q_i(t),\ \ t \geq 0),$$
$$F = (F(t) = (F_1(t), \ldots, F_N(t)),\ \ t \geq 0),$$
$$\hat{F} = (\hat{F}(t) = (\hat{F}_1(t), \ldots, \hat{F}_N(t)),\ \ t \geq 0),$$
$$G = ((G_m(t),\ m \in M),\ t \geq 0) \ ,$$
$$\hat{G} = ((\hat{G}_{mi}(t),\ m \in M,\ i \in N),\ t \geq 0).$$

The set of functions $(Q, Q_*, F, \hat{F}, G, \hat{G})$ clearly has redundancies. The entire set is uniquely determined by the initial state $Q(0)$, the realizations F and G of the input flow and channel state processes, which "drive" the system, and the realization $\hat{G}$, which determines the scheduling choices.

In what follows, we will use bold font $(\mathbf{Q}, \mathbf{Q}_*, \mathbf{F}, \hat{\mathbf{F}}, \mathbf{G}, \hat{\mathbf{G}})$ when we view this set of functions as a random process, and use Roman font when we view it as a deterministic sample path.

For each index $n = 1, 2, \ldots$, consider a (stochastically equivalent) version of our system, and denote by $(\mathbf{Q}^{(n)}, \mathbf{Q}_*{}^{(n)}, \mathbf{F}^{(n)}, \hat{\mathbf{F}}^{(n)}, \mathbf{G}^{(n)}, \hat{\mathbf{G}}^{(n)})$ the corresponding process. The corresponding sequence of fluid-scaled processes is defined as

$$(\mathbf{q}^{(n)}, \mathbf{q}_*{}^{(n)}, \mathbf{f}^{(n)}, \hat{\mathbf{f}}^{(n)}, \mathbf{g}^{(n)}, \hat{\mathbf{g}}^{(n)}))$$
$$\doteq \Gamma^n(\mathbf{Q}^{(n)}, \mathbf{Q}_*{}^{(n)}, \mathbf{F}^{(n)}, \hat{\mathbf{F}}^{(n)}, \mathbf{G}^{(n)}, \hat{\mathbf{G}}^{(n)})),\ n = 1, 2, \ldots.$$

V. Sample path large deviations principle: Mogulsky Theorem

The sequence of processes $(\mathbf{f}^{(n)}, \mathbf{g}^{(n)})$ is known to satisfy a sample path LDP, described in this section.

Our assumptions on the input flows imply the following, for each flow i. The large deviations rate function, associated with the distribution of $A_i(1)$, is

$$L_i(\xi) \doteq \sup_{\theta \geq 0} [\theta\xi - \log E e^{\theta A_i(1)}],\ \xi \geq 0.$$

Function $L_i(\cdot)$ is a non-negative finite convex continuous function on $[0, \infty)$, attaining its unique minimum 0 at point λ_i, i.e.

$$L_i(\xi) = 0, \quad \text{and} \quad L_i(\xi) > 0 \text{ for } \xi \neq \lambda_i,$$

and it is superlinear on infinity, i.e.

$$L_i(\xi)/\xi \to \infty, \quad \xi \to \infty.$$

We adopt the convention that $L_i(\xi) = +\infty$ for $\xi < 0$.

For a vector $y \in R^N$ we will use notation

$$L_{(f)}(y) \doteq \sum_i L_i(y_i).$$

(Subscript (f) indicates that this is the rate function associated with input flows.)

The relative entropy of a probability distribution $\gamma = (\gamma_1, \ldots, \gamma_M)$ with respect to the distribution π we denote by

$$L_{(g)}(\gamma) \doteq \sum_{m \in M} \gamma_m \log \frac{\gamma_m}{\pi_m}.$$

According to Sanov theorem (cf. Theorem 2.1.10 in [6]), $L_{(g)}(\cdot)$ is the large deviations rate function for the sequence of empirical distributions of the channel state over n trials (with $n \to \infty$). Function $L_{(g)}(\cdot)$ is (finite) continuous and convex on the simplex of probability distributions γ; we adopt the convention that $L_{(g)}(\cdot)$ is defined on R^M and is $+\infty$ outside the above simplex.

For a pair (f, g) of vector-functions $f \in \mathcal{D}^N$ and $g \in \mathcal{D}^M$, its *cost* $J_t(f, g)$ in time interval $[0, t]$ is defined as

$$J_t(f, g) \doteq \int_0^t [L_{(f)}(f'(s)) + L_{(g)}(g'(s))]ds, \qquad (15)$$

if $\zeta_0^t(f, g) \in \zeta_0^t \mathcal{A}_0^{N+M}$, and as $+\infty$ otherwise. More generally, if functions f and g have a bounded domain $[0, d]$, that is $(f, g) \in \zeta_0^d \mathcal{D}^{N+M}$, the cost $J_t(f, g)$ is still defined as above, as long as $t \le d$.

The following is (a form of) Mogulsky theorem (cf. Theorem 5.1.2 in [6]).

Proposition 2: Consider a sequence of scaled processes $(\mathbf{f}^{(n)},\ \mathbf{g}^{(n)})$, $n = 1, 2, \ldots$, as defined in Section IV. Then, for every $c \ge 0$ and $t \ge 0$, the sequence of processes $\bar{\zeta}_c^{c+t}(\mathbf{f}^{(n)},\ \mathbf{g}^{(n)})$ satisfies the LDP with good rate function $J_t(\cdot)$. In more detail, for any measurable $B \subseteq \zeta_0^t \mathcal{D}^{N+M}$, we have the following asymptotic (respectively lower and upper) bounds:

$$\liminf_{n\to\infty} \frac{1}{n} \log P(\bar{\zeta}_c^{c+t}(\mathbf{f}^{(n)},\ \mathbf{g}^{(n)}) \in B) \ge -\ \inf\{J_t(h) \mid h \in B^\circ\}\ , \qquad (16)$$

$$\limsup_{n\to\infty} \frac{1}{n} \log P(\bar{\zeta}_c^{c+t}(\mathbf{f}^{(n)},\ \mathbf{g}^{(n)}) \in B) \le -\ \inf\{J_t(h) \mid h \in \bar{B}\}\ . \qquad (17)$$

VI. Simple trajectories to raise maximal weighted queue length. LD lower bound under any scheduling rule

Let $\gamma = \{\gamma_m, m \in M\}$ be some ("twisted") probability distribution on the set of channel states, not necessarily equal to the distribution π. We denote by V_γ the corresponding "twisted" rate region, defined the same way as V but with π replaced by γ. (Thus $V = V_\pi$.) In addition, for every non-zero subset $N' \subseteq N$, we denote by $V_\gamma(N')$ the projection of V_γ onto the corresponding subspace $R^{|N'|}$, where $|N'|$ is the cardinality of set N'. We denote by $V_\gamma^*(N')$ the subset of maximal elements of $V_\gamma(N')$, that is

$$V_\gamma^*(N') \doteq \{v \in V_\gamma(N') \mid v \le w \in V_\gamma(N') \text{ implies } w = v\}.$$

For a fixed non-zero subset $N' \subseteq N$, consider pairs of a distribution γ and a vector $\lambda = \{\lambda_i,\ i \in N'\}$ such that there exists vector $\mu = \{\mu_i,\ i \in N'\} \in V_\gamma^*(N')$ for which the following condition holds:

$$a_i(\lambda_i - \mu_i) = \ell > 0, \quad \forall i \in N'.$$

(Note that if such vector $\hat{\mu}$ exists, it is unique, because this is the point where the ray emanating from point λ in the direction given by $\{-1/a_i,\ i \in N'\}$, hits region $V_\gamma(N')$.) Let us denote

$$J_*(N') \doteq \inf \frac{L_{(g)}(\gamma) + \sum_{i \in N'} L_i(\lambda_i)}{\ell},$$

where the $\inf$ is taken over all pairs of γ and λ, as specified above.

Finally, we define

$$J_* = \min_{N' \subseteq N,\ N' \ne \emptyset} J_*(N').$$

We now give the interpretation of the above definitions. Let $N' = N$ for simplicity. Consider the process with large index n on a (large) time interval $[0, nt]$ for some fixed t. Suppose the empirical distribution of the channel states in this interval is a "twisted" distribution γ, possibly different from π. Moreover, we assume that the fluid-scaled channel state process trajectory is "close to" linear: $g^{(n)}(s) \approx g(s) \equiv \gamma s,\ 0 \le s \le t$. Suppose also that the fluid-scaled input flow trajectory is "close to" linear: $f^{(n)}(s) \approx f(s) \equiv \lambda s,\ 0 \le s \le t$, for some vector λ not necessarily equal to the average rate vector $\bar{\lambda}$. The cost of this linear trajectory of the input and channel state processes is $J_t(f, g) = [L_{(f)}(\lambda) + L_{(g)}(\gamma)]t$. (In other words, the "probability" of $(\mathbf{f}^{(n)}, \mathbf{g}^{(n)})$ being close to (f, g) in the interval $[0, t]$ is roughly $\exp[-n(L_{(f)}(\lambda) + L_{(g)}(\gamma))t]$.) Suppose now that vectors γ and λ satisfy the conditions specified above, with the corresponding vector μ. Then, a scheduling rule can be chosen (at least in principle) such that the (scaled) service process trajectory is approximately linear: $\hat{f}^{(n)}(s) \approx \hat{f}(s) \equiv \mu s,\ 0 \le s \le t$. Then, if $q^{(n)}(0) = 0$, the queue length trajectory in $[0, t]$ is approximately linearly increasing as well, and moreover,

$$a_i q_i^{(n)}(s) \approx a_i q_i(s) = \ell s \ \text{ for each } i.$$

This means that for all flows, $a_i q_i^{(n)}(s)$ is approximately equal to their maximum $q_*^{(n)}(s)$ at any time s, and $q_*^{(n)}(s)$ reaches level ℓt at time t. Thus, the constructed linear *simple trajectory* (f, g, q), which is determined by the vectors λ, γ and μ, has the "unit cost of raising $q_*(s)$" equal to $[L_{(g)}(\gamma) + L_{(f)}(\lambda)]/\ell$. Therefore, the value J_* defined above in this section is the minimum possible unit cost of raising $q_*(s)$ along a simple trajectory.

The key property of the above construction of a simple trajectory (f,g,q) is as follows. Given vectors λ and γ, the corresponding vector of service rates μ is optimal in the sense that all $a_i q_i(s)$, and then $q_*(s)$, simultaneously reach level ℓt at time t. Using the condition that μ is a maximal element of $V_\gamma(N')$, it is easy to see that if (f,g) is the trajectory of input and channel state processes "offered" to the system, then *under any scheduling rule* and *for any initial condition* $q(0)$, at least one of the $a_i q_i(t)$, and then $q_*(t)$, is ℓt or greater. Thus, for any scheduling rule, J_* serves as an upper bound of the minimum possible cost of raising (scaled) maximal weighted queue length $q_*(\cdot)$ to level 1.

Our simple trajectory construction is in a sense analogous to, and serves the same purpose as, those in [11], [12]. It is however necessarily more involved, because in our case the rate region is more general convex, while in [11], [12] the outer boundary of the rate region is a hyperplane (which implies simple "work conservation" properties).

A. LD lower bound under any scheduling rule: Proof of Theorem 1(i).

The proof formalizes the argument presented above in this section, using the construction of a simple trajectory and Mogulsky theorem (Proposition 2). This formalization is done analogously to the proof of Theorem 3.2(ii) in [12] (or proof of Theorem 6.8(ii) in [11]). We omit details.

VII. Refined Mogulsky Theorem (upper bound)

From the standard large deviations principle for scalar random variables, we have the following bound, recorded here for future reference: for any interval $[\xi_1,\xi_2]$, where $0 \le \xi_1 < \xi_2 \le +\infty$, and any fixed $\delta > 0$, there exists a sufficiently large $\tau > 0$ such that, uniformly on non-negative $0 \le t_1 < t_2$ satisfying $t_2 - t_1 \ge \tau$:

$$\log P\{\frac{1}{t_2-t_1}[F_i(t_2)-F_i(t_1)] \in [\xi_1,\xi_2]\} \le -\left[\min_{[\xi_1,\xi_2]} L_i(\xi) - \delta\right](t_2-t_1). \quad (18)$$

Note that in the special case when $\xi_1 > \bar\lambda_i$, (18) takes the form

$$\log P\{\frac{1}{t_2-t_1}[F_i(t_2)-F_i(t_1)] \ge \xi_1\} \le -\left[L_i(\xi_1) - \delta\right](t_2-t_1). \quad (19)$$

If $B \subset R_+^M$ is compact, then according to Sanov theorem (cf. Theorem 2.1.10 in [6]), we can record the following property analogous to (18): for any fixed $\delta > 0$, there exists a sufficiently large $\tau > 0$ such that, uniformly on non-negative integers $0 \le t_1 < t_2$ satisfying $t_2 - t_1 \ge \tau$, we have

$$\log P\{\frac{1}{t_2-t_1}[G(t_2)-G(t_1)] \in B\} \le -\left[\min_{\gamma\in B} L_{(g)}(\gamma) - \delta\right](t_2-t_1). \quad (20)$$

We will need the following generalization of the definition of cost $J_t(f,g)$ (see (15)). For any constant $C > 0$ and a pair (f,g) of vector-functions $f \in \mathcal{D}^N$ and $g \in \mathcal{D}^M$, we define function $J_t^{(C)}(f,g),\ t \ge 0$, as follows:

$$J_t^{(C)}(f,g) \doteq \int_0^t [\sum_{i=1}^N L_i(f_i'(s)\wedge C) + L_{(g)}(g'(s))]ds, \quad (21)$$

if $\zeta_0^t(f,g) \in \zeta_0^t\mathcal{A}_0^{N+M}$, and is $+\infty$ otherwise.

Suppose we have an integer function $u(n) \uparrow \infty$ as $n \to \infty$, which is sublinear in n, i.e., $u(n)/n \downarrow 0$. (An example of such a function is $u(n) = \lceil n^\alpha \rceil$, with $0 < \alpha < 1$.) For any (non-decreasing) RCLL vector-function $h \in \mathcal{D}^{N+M}$, and each n, we denote by $U^n h$ the continuous piece-wise linear function obtained from h as follows: we divide the time interval $[0,\infty)$ into subintervals of equal length $u(n)/n$, that is $[0,u(n)/n],[u(n)/n,2u(n)/n],\ldots$ and linearize h within each subinterval.

Theorem 2: Consider a sequence of scaled processes $(\mathbf{f}^{(n)},\ \mathbf{g}^{(n)})$, $n = 1,2,\ldots$, as defined in Section IV. Let $t > 0$ be fixed. Suppose, for each n there is a fixed measurable $B^{(n)} \subseteq \zeta_0^t\mathcal{D}^{N+M}$, which is a subset of the set of feasible realizations of $(\mathbf{f}^{(n)},\ \mathbf{g}^{(n)})$ in $[0,t]$. Then, for any fixed function $u(n)$ as defined above, we have the following asymptotic upper bound:

$$\limsup_{n\to\infty} \frac{1}{n}\log P(\zeta_0^t(\mathbf{f}^{(n)},\mathbf{g}^{(n)}) \in B^{(n)}) \le -\limsup_{C\to\infty}\ \liminf_{n\to\infty}\ \inf\{J_{t^{(n)}}^{(C)}(U^n h) \mid h \in B^{(n)}\}, \quad (22)$$

where $t^{(n)}$ is the largest multiple of $u(n)/n$ not greater than t, i.e.,

$$t^{(n)} \doteq \frac{u(n)}{n}\lfloor \frac{t}{u(n)/n}\rfloor. \quad (23)$$

Proof. To avoid clogging notation, assume that $t^{(n)} = t$ for each n, i.e., the time interval $[0,t]$ is divided into the integer number $tn/u(n)$ of $u(n)/n$-long subintervals.

Given the properties of rate functions L_i, the functional $J_t^{(C)}(\cdot)$ is non-decreasing in C when C is sufficiently large, namely for $C \ge \max_i \bar\lambda_i$. Therefore, it suffices to show that for a fixed $C > \max_i \bar\lambda_i$, we have

$$\limsup_{n\to\infty} \frac{1}{n}\log P(\zeta_0^t(\mathbf{f}^{(n)},\mathbf{g}^{(n)}) \in B^{(n)}) \le -\liminf_{n\to\infty}\ \inf\{J_t^{(C)}(U^n h) \mid h \in B^{(n)}\}. \quad (24)$$

The rest of the proof is a fairly straightforward combinatorial estimate.

Let us fix a small $\delta > 0$. We choose a large integer $K > 0$ and divide interval $[0,C)$ into K subintervals, each $\epsilon = C/K$-long, namely $[k\epsilon,(k+1)\epsilon)$ with $k = 0,1,\ldots,K-1$. The k-th interval defined above, with $k = 0,1,\ldots,K-1$, we will call k-th "bin". In addition, the interval $[C,\infty)$ we also call a bin and give it the index $k = K$. We will choose K to be large enough so that the total variation of each function L_i in each of the bins $k = 0,1,\ldots,K-1$ is less than $\delta/(4N)$. We will choose $\tau > 0$ such that the estimates (18)-(19) hold for all i and for the intervals $[\xi_1,\xi_2]$ in (18) being closures of all bins $k = 0,1,\ldots,K-1$ and with ξ_1

in (19) replaced by C; in addition, we require that (18)-(19) hold with δ replaced by $\delta/(4N)$.

Let us divide the simplex of all vectors representing probability distributions γ on the set of channel states M into $K+1$ non-intersecting subsets ("bins"), such that the total variation of $L_{(g)}(\gamma)$ within the closure of each bin is at most $\delta/4$. (The latter can always be achieved by making K larger, if necessary.) We also will increase τ, if necessary, to make sure that the estimate (20) holds for all the bins, with δ replaced by $\delta/4$.

Let $\hat{J}$ denote the $\liminf$ in the RHS of (24). From now on in this proof we will only consider sufficiently large n such that $\inf\{J_t^{(C)}(U^n h) \mid h \in B^{(n)}\} > \hat{J} - \delta$, and $u(n) > \tau$, where τ is the one chosen above.

Consider a fixed n, a vector-function $h = (f_i,\ i = 1,\dots,N; g) \in B^{(n)}$, and its piece-wise linearization $U^n h = (U^n f_i,\ i = 1,\dots,N; U^n g)$. Recall that each component of $U^n h$ has a constant non-negative derivative in each of the $tn/u(n)$-long time-subintervals of $[0,t]$. Thus, vector-function h can belong to one of the finite number $[(K+1)^{(N+1)}]^{tn/u(n)}$ of "aggregate bins," according to which bins the (constant) slopes of the components $U^n f_i$ and $U^n g$ belong to, in each of the time-subintervals.

Now, consider any fixed aggregate bin, let us call it B_{ab}, containing at least one function belonging to $B^{(n)}$, and let us pick some fixed $h \in B^{(n)}$. (Recall that $J_t^{(C)}(U^n h) > \hat{J} - \delta$.) Then, using estimates (18),(19) and (20), for each of the time subintervals (more precisely - for each of the corresponding unscaled $u(n)$-long intervals), we obtain the following upper bound:

$$\log P\{\zeta_0^t(\mathbf{f}^{(n)}, \mathbf{g}^{(n)}) \in B_{ab}\}$$

$$\leq -[J_T^{(C)}(U^n h)n - u(n)\delta\frac{tn}{u(n)}] \leq -\hat{J}n + \delta n + \delta t n\ .$$

Since the total number of aggregate bins is $\exp\{\frac{(N+1)\log(K+1)t}{u(n)}n\}$ with $[(N+1)\log(K+1)t]/u(n) \to 0$ as $n \to \infty$, we have

$$\limsup_{n\to\infty} \frac{1}{n}\log P(\zeta_0^t(\mathbf{f}^{(n)}, \mathbf{g}^{(n)}) \in B^{(n)}) \leq -\ \hat{J} + \delta(1+t)\ .$$

Since δ can be chosen arbitrarily small, the proof is complete. ∎

VIII. Large deviations upper bound via Refined Mogulsky Theorem: Proof of Theorem 1(ii)

From now on we specify the function $u(n)$, defined in Section VII, to be

$$u(n) = \lceil n^\alpha \rceil,\ n = 1, 2, \dots,$$

for some fixed $\alpha \in (0, \eta)$. Also, consider some fixed sequence $C_n,\ n = 1, 2, \dots$, such that $C_n > 0$ and $C_n \uparrow \infty$.

Definition 2: Suppose an increasing subsequence $\mathcal{N}$ of the sequence of positive integers is fixed, and the following conditions (i) and (ii) hold.

(i) For each $n \in \mathcal{N}$, there is a fixed (feasible) realization $(q^{(n)}, q_*^{(n)}, f^{(n)}, \hat{f}^{(n)}, g^{(n)}, \hat{g}^{(n)})$ of the process $(\mathbf{q}^{(n)}, \mathbf{q_*}^{(n)}, \mathbf{f}^{(n)}, \hat{\mathbf{f}}^{(n)}, \mathbf{g}^{(n)}, \hat{\mathbf{g}}^{(n)})$.

(ii) As $n \to \infty$, we have the weak convergence

$$(q^{(n)}, q_*^{(n)}, f^{(n)}, \hat{f}^{(n)}, g^{(n)}, \hat{g}^{(n)}) \Rightarrow (q, q_*, f, \hat{f}, g, \hat{g}) \quad (25)$$

for some set of functions $(q, q_*, f, \hat{f}, g, \hat{g})$, and the weak convergence

$$\bar{J}^{(C_n)} \doteq (J_{t^{(n)}}^{(C_n)}[U^n(f^{(n)}, g^{(n)})],\ t \geq 0) \Rightarrow \bar{J} = (\bar{J}_t,\ t \geq 0) \quad (26)$$

for some non-negative non-decreasing function $\bar{J}$.

Then, the entire construction

$$\psi = [\mathcal{N};\ \ (q^{(n)}, q_*^{(n)}, f^{(n)}, \hat{f}^{(n)}, g^{(n)}, \hat{g}^{(n)}),\ \bar{J}^{(C_n)},\ n \in \mathcal{N};$$
$$(q, q_*, f, \hat{f}, g, \hat{g});\ \ \bar{J}]$$

will be called a *generalized fluid sample path* (GFSP). The non-decreasing function $\bar{J}$ will be called the *refined cost function* of the GFSP.

Remark. A set of functions $(q, q_*, f, \hat{f}, g, \hat{g})$, defined as a limit of a sequence of fluid scaled trajectories of a process, is sometimes called a *fluid sample path* (FSP), cf. [12]. Therefore, the term "generalized" in the above definition of a GFSP refers to the fact that GFSP contains not only the "fluid limit" of a (pre-limit) sequence, but the sequence itself. Moreover, the pre-limit sequence is required to satisfy condition (26).

Given that (unscaled) functions $\hat{F}_i^{(n)}$, $G_m^{(n)}$, $\hat{G}_{mi}^{(n)}$ obviously have uniformly bounded increments within one time slot, it is easy to observe that GFSP components $\hat{f}, g, \hat{g}$ are non-decreasing Lipschitz continuous (and then absolutely continuous) functions with $\hat{f}(0) = 0, g(0) = 0, \hat{g} = 0$.

For a givin GFSP, if we denote $T^* = \sup\{t \mid \|f(t)\| < \infty,\ \bar{J}_t < +\infty\}$, the following is easy to verify: $\|q(t)\| < \infty$ for all $t < T^*$; if t_1 and t_2 are points of continuity of f (or, equivalently, of q), and $0 \leq t_1 < t_2 < T^*$, then

$$\bar{J}_{t_2} \geq J_{t_2}(f, g), \quad \bar{J}_{t_2} - \bar{J}_{t_1} \geq J_{t_2}(f, g) - J_{t_1}(f, g). \quad (27)$$

Consequently both f and q are absolutely continuous in the interval $[0, T^*)$, with $f(0) = 0$, and therefore the convergence in (25) is in fact uniform on compact subsets of $[0, T^*)$. (It also follows that (27) holds for any $0 \leq t_1 < t_2 < T^*$.)

The following simple facts (in Lemmas 1 and 2 below) have straightforward proofs and recorded for future reference.

Lemma 1: Suppose, there exists a sequence $\{(f^{(n)}, g^{(n)}),\ n \in \mathcal{N}\}$ of feasible realizations of the (scaled) input and channel state processes, such that

$$(f^{(n)}, g^{(n)}) \Rightarrow (f, g).$$

Then, there exists a GFSP, having (f, g) as its components and such that its refined cost function $\bar{J}$ is equal to the cost function $(J_t(f, g),\ t \geq 0)$.

Lemma 2: [Compactness.] Suppose, a sequence of GFSP ${}^k\psi,\ k = 1, 2, \dots$, is such that the values of $\|{}^k q(0)\|$ are uniformly bounded. Then, there exists a GFSP ψ such that, along some subsequence of k,

$$[({}^k q,\ {}^k q_*,\ {}^k f,\ {}^k \hat{f},\ {}^k g,\ {}^k \hat{g});\ \ {}^k \bar{J}] \Rightarrow [(q, q_*, f, \hat{f}, g, \hat{g});\ \ \bar{J}].$$

Let J_{**} denote the lowest refined cost of a GFSP which "brings" $q_*(t)$ to level 1 from the zero initial state $q(0) = 0$. Namely,

$$J_{**} \doteq \inf_{t\geq 0} J_{**,t}, \tag{28}$$

where

$$J_{**,t} \doteq \inf\{\bar{J}_t \mid \psi:\ q(0) = 0,\ q_*(t) \geq 1\}. \tag{29}$$

From the definition of J_* in Section VI (via the construction of simple trajectories) and Lemma 1, we conclude that

$$J_{**} \leq J_*, \tag{30}$$

because we can always construct a GFSP for which (f, g) is a simple trajectory with $q_*(t) \geq 1$ for some $t > 0$, and with $J_t(f, g)$ being arbitrarily close to J_*.

The goal of this section is to establish the following fact, which is Theorem 1(ii) rephrased in terms of the sequence of fluid-scaled processes.

Theorem 3: For each parameter $n = 1, 2, \ldots$, consider a version of the system in a stationary regime. Then, the corresponding sequence of fluid-scaled processes is such that

$$\limsup_{n\to\infty} \frac{1}{n} \log P\left(\mathbf{q}_*{}^{(n)}(0) > 1\right) \leq -J_{**}\ . \tag{31}$$

We will not give a complete proof of Theorem 3 here. We only provide a key step of such proof, Theorem 4 below, which demonstrates how Theorem 2 is applied to obtain LD upper bounds for the sequence of fluid-scaled processes *on a finite time interval*, in terms of GFSPs and their refined costs. The rest of the proof of Theorem 3 is carried out analogously to the proof of a similar result in Section 10 of [12], which uses classical Wentzel-Freidlin constructions to "translate" LD results on a finite time interval into the LD results in a stationary regime. (In addition, the proof of Theorem 3 will utilize some of the properties of local fluid sample paths, defined and analyzed later in this paper.)

Theorem 4: For a fixed $T \geq 0$, let us denote

$$J_{**,\leq T} \doteq \inf\{\bar{J}_t \mid \psi:\ q(0) = 0,\ q_*(t) \geq 1 \text{ for some } t \leq T\}$$
$$\equiv \inf_{t\leq T} J_{**,t}.$$

Then, we have:

$$\lim_{c\downarrow 0} \limsup_{n\to\infty} \frac{1}{n} \log \sup_{\|q^{(n)}(0)\|\leq c} P\left(\sup_{t\in[0,T]} \mathbf{q}_*{}^{(n)}(t) > 1\right)$$
$$\leq -J_{**,\leq T}\ . \tag{32}$$

(The sup over $\|q^{(n)}(0)\| \leq c$ is supremum over all processes with non-random initial state satisfying this condition.)

Proof. We have

$$\sup_{\|q^{(n)}(0)\|\leq c} P\left(\sup_{t\in[0,T]} \mathbf{q}_*{}^{(n)}(t) > 1\right)$$
$$\leq P\left((\mathbf{f}^{(n)}, \mathbf{g}^{(n)}) \in B^{(n,c)}\right),$$

where

$$B^{(n,c)} \doteq \{(f^{(n)}, g^{(n)}) \mid \exists q^{(n)}(0),\ t \leq T,$$
$$\text{such that } \|q^{(n)}(0)\| \leq c \text{ and } q_*^{(n)}(t) > 1\}.$$

By Theorem 2,

$$\limsup_{n\to\infty} \frac{1}{n} \log P\left((\mathbf{f}^{(n)}, \mathbf{g}^{(n)}) \in B^{(n,c)}\right)$$
$$\leq -\limsup_{C\to\infty} \liminf_{n\to\infty} \inf\{J^{(C)}_{t^{(n)}}(U^n h) \mid h \in B^{(n,c)}\}\ , \tag{33}$$

where $t^{(n)}$ is as in (23), but with t replaced by T in the RHS. Using GFSP definition (and the fact that for large C cost $J_t^{(C)}$ is non-decreasing in C), it is easy to see that the $\limsup$ in the RHS of (33), let us denote it by $J_{**,\leq T,c}$, is exactly the lowest refined cost of a GFSP such that $\|q(0)\| \leq c$ and $q_*(t) \geq 1$ for some $t \leq T$. (We choose a subsequence with infinitely increasing C to construct a GFSP with refined cost $J_{**,\leq T,c}$ and satisfying the above conditions; at the same time all such GFSP have the refined cost at least $J_{**,\leq T,c}$.) Since $J_{**,\leq T,c}$ is non-increasing in c, $J_{**,\leq T,c} \downarrow J_{**,\leq T,0} \equiv J_{**,\leq T}$. (The fact that the limit of $J_{**,\leq T,c}$ cannot be strictly greater than $J_{**,\leq T,0}$ easily follows from Lemma 2.) This completes the proof. ∎

IX. Local fluid sample paths

In view of (30), to prove optimality of the EXP rule (or any given rule) in the sense of (6), it suffices to show that under this rule the equality $J_{**} = J_*$ in fact holds. One way to approach this is to show that assumption $J_{**} < J_*$ leads to a construction of a simple trajectory (as defined in Section VI) with *unit cost* (of raising $q_*(\cdot)$ from 0 to 1) being strictly less than J_* - a contradiction.

Using this approach for the EXP rule naturally leads to the notion of a local fluid sample path (LFSP), and its cost function. The definition of the bound J_{**} in terms of GFSPs and their refined costs, allows us to "properly" define LFSP cost function.

Due to space limitation, in the rest of this section we give an informal intuitive description of LFSPs. (Rigorous definitions and results are given in [14].) For the purposes of this discussion, let assume that EXP parameter $\eta = 1/2$, and parameter $0 < \alpha < 1/2$ is arbitrary.

Let us fix arbitrary number $S > 0$. Suppose, we have a GFSP ψ such that $q_*(0) = 0$ (which is equivalent to $q(0) = 0$) and, for some finite $T > 0$, $q_*(T) = 1$ and $\bar{J}_T = J_{**} + \epsilon < J_*$, for a small $\epsilon > 0$. Then, for each n, breaking down the interval $[0, T]$ into subintervals of the length of the order $n^{1/2}$, and working with the functions $q^{(n)}$ and other components of GFSP ψ, the following can be shown. For every sufficiently large n, we can find an interval $[t_1^{(n)}, t_2^{(n)}]$ within $[0, T]$, satisfying the following conditions:

$$t_2^{(n)} - t_1^{(n)} = S\sigma_n,\quad \sigma_n \doteq [\bar{q}^{(n)}(t_1^{(n)})]^{1/2} n^{1/2}/n,$$
$$q_*^{(n)}(t_2^{(n)}) - q_*^{(n)}(t_1^{(n)}) > 0,$$
$$\frac{\bar{J}^{(C_n)}_{t_2^{(n)}} - \bar{J}^{(C_n)}_{t_1^{(n)}}}{q_*^{(n)}(t_2^{(n)}) - q_*^{(n)}(t_1^{(n)})} < J_*. \tag{34}$$

We introduce rescaled functions as follows (the actual definition in [14] is somewhat different), each defined for $s \in [0, S]$:

$$ {}_\diamond q_i^{(n)}(s) \doteq \frac{1}{\sigma_n}[q_i^{(n)}(t_1^{(n)} + \sigma_n s) - q_*^{(n)}(t_1^{(n)})],\ i \in N, $$

$$ {}_\diamond q_*^{(n)}(s) \doteq \max_i {}_\diamond q_i^{(n)}(s), $$

$$ {}_\diamond f_i^{(n)}(s) \doteq \frac{1}{\sigma_n}[f_i^{(n)}(t_1^{(n)} + \sigma_n s) - f_i^{(n)}(t_1^{(n)})],\ i \in N, $$

$$ {}_\diamond g_m^{(n)}(s) \doteq \frac{1}{\sigma_n}[g_m^{(n)}(t_1^{(n)} + \sigma_n s) - g_m^{(n)}(t_1^{(n)})],\ m \in M. $$

Then we show that we can find a subsequence of $\{n\}$ such that

$$ ({}_\diamond q^{(n)}, {}_\diamond q_*^{(n)}, {}_\diamond f^{(n)}, {}_\diamond g^{(n)}) \to ({}_\diamond q, {}_\diamond q_*, {}_\diamond f, {}_\diamond g), \tag{35} $$

where all functions are defined on the interval $[0, S]$, all functions in the RHS are absolutely continuous, and the convergence is uniform. The 4-tuple $({}_\diamond q, {}_\diamond q_*, {}_\diamond f, {}_\diamond g)$ is what we will call an LFSP. It can be shown (using convexity of the rate functions $L_{(f)}(\cdot)$ and $L_{(g)}(\cdot)$, along with the fact that $n^\alpha/n = o(n^{1/2}/n)$ as $n \to \infty$), that

$$ \liminf_{n\to\infty} \sigma_n^{-1}[\bar{J}_{t_2^{(n)}}^{(C_n)} - \bar{J}_{t_1^{(n)}}^{(C_n)}] \geq J_S({}_\diamond f, {}_\diamond g). \tag{36} $$

We then show that the constructed LFSP satisfies the following conditions:

$$ {}_\diamond q_*(S) - {}_\diamond q_*(0) > 0, \tag{37} $$

$$ \frac{J_S({}_\diamond f, {}_\diamond g) - J_0({}_\diamond f, {}_\diamond g)}{{}_\diamond q_*(S) - {}_\diamond q_*(0)} < J_*, \tag{38} $$

where (38) is obtained using bounds (34) and (36).

We also show (see [14]) that an LFSP satisfies a certain set of differential inclusions, some of which a generic for fluid limits (and LFSP is a fluid limit, although defined on a "local" scale) and some are specific to EXP rule.

X. EXP RULE OPTIMALITY IN THE TWO FLOWS CASE

As described in the previous section, the assumption $J_{**} < J_*$ leads to the existence of LFSPs with (roughly speaking) the unit cost of raising ${}_\diamond q_*(\cdot)$ being strictly less than J_* (see (38)). It is shown in [14], that, in the case of EXP scheduling rule and two flows, this fact (along with the dynamic properties of LFSPs under EXP rule) allows us to construct a simple trajectory (see Section VI) with unit cost strictly less than J_*. This is a contradiction proving Theorem 1(iii).

REFERENCES

[1] M. Andrews, K. Kumaran, K. Ramanan, A. L. Stolyar, R. Vijayakumar, P. Whiting. Providing Quality of Service over a Shared Wireless Link. *IEEE Communications Magazine*, 2001, Vol.39, No.2, pp.150-154.

[2] M. Andrews, K. Kumaran, K. Ramanan, A. L. Stolyar, R. Vijayakumar, P. Whiting. Scheduling in a Queueing System with Asynchronously Varying Service Rates. *Probability in Engineering and Informational Sciences*, 2004, Vol. 18, pp. 191-217.

[3] P.Bender, P.Black, M.Grob, R.Padovani, N.Sindhushayana, A.Viterbi, "CDMA/HDR: A Bandwidth Efficient High Speed Wireless Data Service for Nomadic Users," *IEEE Communications Magazine*, July 2000.

[4] D. Bertsimas, I. C. Paschalidis, J. N. Tsitsiklis. Asymptotic Buffer Overflow Probabilities in Multiclass Multiplexers: An Optimal Control Approach. *IEEE Trans. Automat. Control*, 43:315-335, 1998.

[5] P. Billingsley. Convergence of Probability Measures. Wiley, 1968.

[6] A.Dembo, O.Zeitouni. Large Deviations Techniques and Applications. Springer, 1998. (2nd edition)

[7] W. Feller. *An Introduction to Probability Theory and its Applications*, Wiley, 1950.

[8] M.I.Freidlin, A.D.Wentzell. Random Perturbations of Dynamical Systems. Springer, 1998. (2nd edition)

[9] S. Shakkottai and A. L. Stolyar. Scheduling for Multiple Flows Sharing a Time-Varying Channel: The Exponential Rule. *Analytic Methods in Applied Probability. In Memory of Fridrih Karpelevich. Yu. M. Suhov, Editor.* American Mathematical Society Translations, Series 2, Volume 207, pp. 185-202. American Mathematical Society, Providence, RI, 2002.

[10] S. Shakkottai, R. Srikant, and A. L. Stolyar. Pathwise Optimality of the Exponential Scheduling Rule for Wireless Channels. *Advances in Applied Probability*, 2004, Vol. 36, No. 4, pp. 1021-1045.

[11] A. L. Stolyar and K. Ramanan. Largest Weighted Delay First Scheduling: Large Deviations and Optimality. *Annals of Applied Probability*, 2001, Vol. 11, pp. 1-48.

[12] A.L. Stolyar. Control of End-to-End Delay Tails in a Multiclass Network: LWDF Discipline Optimality. *Annals of Applied Probability*, 2003, Vol.13, No.3, pp.1151-1206.

[13] A. L. Stolyar. MaxWeight Scheduling in a Generalized Switch: State Space Collapse and Workload Minimization in Heavy Traffic. *Annals of Applied Probability*, 2004, Vol.14, No.1, pp.1-53.

[14] A. L. Stolyar. Large Deviations of Queues Sharing a Randomly Time-varying Server. Bell Labs Technical Memo, July 2006. Submitted for publication.

[15] P. Viswanath, D. Tse, and R. Laroia. Opportunistic Beamforming using Dumb Antennas. *IEEE Transactions on Information Theory*, 2002, Vol. 48(6), pp. 1277-1294.

[16] L. Ying, R. Srikant, A. Eryilmaz, and G. E. Dullerud. A Large Deviations Analysis of Scheduling in Wireless Networks. Preprint, 2005.