Scheduling for Multiple Flows Sharing a Time-Varying Channel: The Exponential Rule

Sanjay Shakkottai*
Coordinated Science Laboratory
and Department of Electrical
and Computer Engineering
University of Illinois
at Urbana-Champaign
shakkott@uiuc.edu

Alexander L. Stolyar
Bell Labs, Lucent Technologies
600 Mountain Avenue
Murray Hill, NJ 07974
stolyar@research.bell-labs.com

Abstract

We consider the following queueing system which arises as a model of a wireless link shared by multiple users. Multiple flows must be served by a "channel" (server). The channel capacity (service rate) changes in time randomly and *asynchronously* with respect to different flows. In each time slot, a *scheduling discipline (rule)* picks a flow for service based on the current state of the channel and the queues.

We study a scheduling rule, which we call the *exponential rule*, and prove that this rule is *throughput-optimal*, i.e., it makes the queues stable if there exists any rule which can do so. In the proof we use the *fluid limit* technique, along with a *separation of time scales* argument. Namely, the proof of the desired property of a "conventional" fluid limit involves a study of a different fluid limit arising on a "finer" time scale.

In our companion paper [12] it is demonstrated that the exponential rule can be used to provide *Quality of Service* guarantees over a shared wireless link.

Key words and phrases: Queueing networks, scheduling, stability, fluid limit, separation of time scales, variable channel

AMS Subject Classification: 60K25, 90B22

*The work of Shakkottai was carried out in part when he was an intern at Bell Labs, Lucent Technologies, Murray Hill, NJ, and was partly supported by NSF Grant ITR 00-85929.

1 Introduction

The primary motivation of this work is the problem of scheduling transmissions of multiple data users (flows) sharing the same wireless channel (server). The unique feature of this problem is the fact that the capacity (service rate) of the channel varies with time randomly and asynchronously for different users. The variations of the channel capacity are due to different (and random) interference levels observed by different users, and also due to *fast fading* [3, 14], of a radio signal received by a moving user. At the very high level, the problem is: How to schedule transmission of different users in a rational way, so that the wireless channel is utilized efficiently? We will refer to this general problem and the corresponding queueing model as the *variable channel scheduling problem (model).*

The variable channel scheduling problem arises, for example, in the 3G CDMA High Data Rate (HDR) system [4], where multiple mobile users in a cell share the same CDMA wireless channel. On the downlink (the link from cell base station to users), time is divided into fixed size (1.67 msec) time slots. This slot size is short enough so that each user's channel quality stays approximately constant within one time slot. In each time slot, one user is scheduled for transmission. Each user constantly reports to the base station its "instantaneous" channel capacity, i.e., the rate at which data can be transmitted if this user is scheduled for transmission. In HDR system (and in the generic variable channel model as well) a "good" scheduling algorithm should take advantage of channel variations by giving some form of priority to users with instantaneously better channels.

In this paper, we study a scheduling algorithm which explicitly uses information on the state of the channel and the queues. We call it the *Exponential (EXP) rule.* Our main result is that
The EXP rule is throughput optimal*, i.e., it renders queues stable in any system for which stability is feasible at all, with any other rule.*

The specific variable channel scheduling model we study is the same as that in [2] (and its extended version [1]), where a scheduling rule called *Modified Largest Weighted Delay First (M-LWDF)* was proposed and proved to be throughput optimal. The Exponential rule was introduced in [1], but not studied analytically.

In a companion paper [12], we study the EXP rule using simulations, and show that this policy, in conjunction with a token queue mechanism allows us to support a mixture of real-time and non-real time data over HDR with high efficiency.

As in [2], our main tool for proving stability results is the *fluid limit* technique [11, 6, 5, 13, 7]. However, in this paper, the use of this technique is much less conventional. To prove the desired property of a "conventional" fluid limit process, we use a "separation of time scales" argument which leads to the analysis of another fluid limit, on a "finer" (space and time) scale. We believe, this approach and the constructions we use, are of independent interest.

In the next section we introduce the precise model and formulate the main result. At the end of that section, we describe the layout of the rest of the paper.

2 Variable Channel Scheduling Model. Main Result

2.1 The Model and Notations

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

The channel operates in discrete time. A time interval $[t, t+1)$, with $t = 0, 1, 2, \dots$, we will call the *time slot* t. There is a finite set of channel states $M = \{1, \dots, M\}$, and the channel state is constant within each time slot. Associated with each state $m \in M$ is a fixed vector of data rates $(\mu_1^m, \dots, \mu_N^m)$, where all μ_i^m are strictly positive integers. The meaning of μ_i^m is as follows. If in a given time slot t the channel is in state m and all service (in this time slot) is allocated to queue i, then μ_i^m type i customers are served from those already present at time t (or the entire queue i content at t, whichever is less). Note that what we call a "channel state" here is actually a collection of channel states with respect to individual flows.

However, the *service in any time slot may be split* according to a (generally speaking random) stochastic vector $\sigma = (\sigma_1, \dots, \sigma_N)$, $\sigma_i \geq 0, \forall i$, $\sum_i \sigma_i = 1$. If in a given time slot t the channel is in state m and a "split" vector σ is chosen, then for each queue i, $\lfloor \sigma_i \mu_i^m \rfloor$ type i customers are served from those already present at time t (or the entire queue i content at t, whichever is less). Here and below $\lfloor \cdot \rfloor$ denotes the integer part, and $\lceil \cdot \rceil$ - the ceiling of a number.

The random channel state process $\mathbf{m}$ is assumed to be an irreducible discrete time Markov chain with the (finite) state space M. (See Feller [9] for the definitions of *irreducibility, aperiodicity*, and *ergodicity* of countable discrete time Markov chains.) The (unique) stationary distribution of this Markov chain we denote by $\pi = (\pi^1, \dots, \pi^M)$.

Denote by $A_i(t)$ the number of type i customers arrived in time slot t. We will adopt a convention that all arrivals in the time slot t actually happen at time t, but those arrivals are *not* available for service until time slot $t+1$. We assume that
The aggregate arrival process $A = \{(A_1(t), \dots, A_N(t)), t = 1, 2, \dots\}$ can be described by a finite number of regenerative processes with finite mean regeneration cycles.

Let us denote by λ_i, $i = 1, \dots, N$, the mean arrival rate for flow i, i.e., the mean number of type i customers arriving in one time slot.

In addition, we will assume that the average input rates converge to their mean rates exponentially fast. Namely, we make the following

Assumption 2.1 *For any $i \in N$ and any $\nu > 0$, there exists a constant $a = a(\nu) > 0$ such that the following estimate holds. Uniformly on the initial state of the input process, for all sufficiently large n,*

$$\Pr\{|\frac{1}{n}\sum_{1}^{n} A_i(t) - \lambda_i| \geq \nu\} < e^{-an} \ .$$

To simplify notation, let us assume that each input process A_i is an ergodic (discrete time) Markov chain with countable state space, and the input processes are mutually independent. (Ex-

tension of the proofs to more general input flows described above is straightforward.) In this case, Assumption 2.1 is satisfied, for example, if all $A_i(t)$ are i.i.d. random variables with a finite exponential moment, or if each $A_i(\cdot)$ is a *finite* state Markov chain.

The random process describing the behavior of the entire system is $S = (S(t),\ t = 0, 1, 2, \dots)$, where

$$S(t) = \{(U_{i1}(t), \dots, U_{iQ_i(t)}(t)),\ i = 1, \dots, N;\ m(t)\ \},$$

$Q_i(t)$ is the type i queue length at time t, and $U_{ik}(t)$ is the current *delay* of the k-th type i customer present in the system at time t. (Within each type, the customers are numbered in the order of their arrivals.) We will denote by $W_i(t) \doteq U_{i1}(t)$ the delay of flow i at time t (with $W_i(t) = 0$ if $Q_i(t) = 0$ by convention).

2.2 A Scheduling Rule. Throughput Optimality.

A mapping H which takes a system state $S(t)$ in a time slot into a fixed probability distribution $H(S(t))$ on the set of stochastic vectors σ, will be called a *scheduling rule*, or a *queueing discipline*. So, if we denote by $D_i(t)$ the number of type i customers served in the time slot t, then according to our conventions, for each time t,

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

where $D_i(t) = \min\{Q_i(t), \lfloor \sigma_i(t) \mu_i^{m(t)} \rfloor\}$ and $\sigma(t)$ is chosen randomly according to the distribution $H(S(t))$.

Our assumptions imply that with any scheduling rule, S is a discrete time countable Markov chain. To avoid trivial complications, we make an additional (not very restrictive) technical assumption that we will only consider scheduling rules H such that the Markov chain S is aperiodic and irreducible. By *stability* of the Markov chain S (and stability of the system) we understand its ergodicity, which (in the case of aperiodicity and irreducibility) is equivalent to the existence of a stationary distribution.

A scheduling rule H we will call *universally stable*, or *throughput optimal*, if it makes a system stable if the stability is feasible at all with any other rule. More precisely, a rule H is throughput-optimal if for *any fixed system*, the existence of a rule (possibly dependent on the system) which makes it stable, implies that the system is also stable under the rule H.

2.3 Exponential Rule. Main Result

Let an arbitrary set of positive constants $\gamma_1, \dots, \gamma_N$, $a_1, \dots, a_N$, and positive constants β and $\eta \in (0, 1)$ be fixed. The following two related rules we will call Exponential.
The Exponential (Queue length) rule (EXP-Q) chooses for service in time slot t a single queue

$$i \in i(S(t)) = \arg\max_i \gamma_i\ \mu_i(t)\ \exp\left(\frac{a_i Q_i(t)}{\beta + [\overline{Q}(t)]^\eta}\right)\ ,$$

where $\mu_i(t) \equiv \mu_i^{m(t)}$ and $\overline{Q}(t) \doteq (1/N)\sum_i a_i Q_i(t)$. Similarly, the Exponential (Waiting time) rule (EXP-W) chooses for service a queue

$$i \in i(S(t)) = \arg\max_i \gamma_i \ \mu_i(t) \ \exp\left(\frac{a_i W_i(t)}{\beta + [\overline{\overline{W}}(t)]^\eta}\right) \ ,$$

where $\overline{\overline{W}}(t) \doteq (1/N)\sum_i a_i W_i(t)$.

Remark 2.1 *1. Formally speaking, in the definition of the EXP rule, we also need to specify a "tie-breaking" convention. For example, we can assume the the queue $i = \max\{j \ : \ j \in i(S(t))\}$ is chosen.*
2. Note that without loss of generality we can assume $\gamma_1 = 1$. We will use this convention later in the paper.

The main result of this paper is the following

Theorem 2.1 *An EXP rule (either EXP-Q or EXP-W), with any fixed set of positive parameters β, $\eta \in (0,1)$, and $\gamma_i, a_i, \ i \in N$, is throughput optimal.*

2.4 Layout of the Rest of the Paper

In the next section we discuss the necessary and sufficient condition for a system to be stable. This condition is closely related to *Static Service Split (SSS)* scheduling rules. We study the properties of SSS rules which are needed for the proof of our main result. In Section 4 we first introduce preliminaries of the fluid limit technique, and then prove Theorem 2.1. The key element of the proof is a *local fluid limit* argument.

3 Necessary and Sufficient Stability Condition: Static Service Split Rule

From this point on, we consider a fixed system, with a fixed set of the parameters.

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. Consider a *Static Service Split* (SSS) scheduling rule [2], parameterized by the matrix ϕ. When the channel is in state m, the SSS rule chooses for service a (single) queue i with probability ϕ_{mi}. Clearly, the vector $v = (v_1, \ldots, v_N) = v(\phi)$, where

$$v_i = \sum \pi^m \phi_{mi} \mu_i^m \ ,$$

gives the long term average service rates allocated to different flows. This observation makes the following necessary and sufficient stability condition very intuitive. The proof of this result is available in [2].

Theorem 3.1 *For a given set of system parameters, a scheduling rule H under which the system is stable exists if and only if there exists a stochastic matrix ϕ such that*

$$\lambda_i < v_i(\phi), \ \forall i \ . \tag{1}$$

In the rest of this paper, we assume that the system parameters are such that stability of the system is feasible, i.e., condition (1) holds for some ϕ and, consequently, the SSS rule associated with this ϕ makes system stable.

An SSS rule associated with a stochastic matrix ϕ^* we will call *maximal* if the vector $v(\phi^*)$ is not dominated by $v(\phi)$ for any other stochastic matrix ϕ. (We say that vector $v^{(1)}$ is dominated by vector $v^{(2)}$ if $v_i^{(1)} \leq v_i^{(2)}$ for all i, and the strict inequality $v_i^{(1)} < v_i^{(2)}$ holds for at least one i.)

Remark 3.1 *We note that any* fixed *maximal SSS rule can* not *be throughput optimal, because throughput-optimality requires that a rule makes stable* any *system for which stability is feasible. (And the necessary and sufficient condition for such feasibility is given by Theorem 3.1.)*

We next present a very useful characterization of a maximal SSS rule. The following result is proved in [2].

Theorem 3.2 *Consider a maximal SSS rule associated with a stochastic matrix ϕ^*. Suppose in addition that all components of $v^* = v(\phi^*)$ are strictly positive. Then there exists a set of strictly positive constants $\alpha_i, \ i = 1, 2, \ldots, N$ such that*

$$\phi^*_{mi} > 0 \ \textit{implies} \ i \in \arg\max_j \alpha_j \mu_j^m \ . \tag{2}$$

The theorem says that basically a maximal SSS rule simply chooses for service at any time t the queue i for which $\alpha_i \mu_i^{m(t)}$ is maximal. It does not specify what to do in case of a tie (when $\alpha_j \mu_j^m$ is same for multiple queues); as a result the same set of $\{\alpha_i\}$ may (and typically will) correspond to different maximal SSS rules.

3.1 A "Diagonal Drift" Maximal SSS Rule

In this section, we show that there exists a maximal SSS rule which keeps the (weighted) difference of the arrival rates and service rates equal. Let us define

$$\begin{aligned} \mathcal{G} &= \left\{ y = (y_1, \ldots, y_N) \mid 0 \leq y_i \leq \max_{i,m} \mu_i^m, \ a_i(\lambda_i - y_i) = a_j(\lambda_j - y_j) \ \forall i, j \in N \right\} \\ \widehat{\mathcal{G}} &= \{ y \in \mathcal{G} \mid \exists \text{ SSS rule parameterized by } \phi \text{ s.t. } v(\phi) \geq y \} \end{aligned}$$

where $\{a_i\}$ are arbitrary positive constants. Then, we have the following simple result.

Lemma 3.1 *There exists a unique maximal element y^* in $\widehat{\mathcal{G}}$, and a maximal SSS rule ϕ^* such that $y^* = v(\phi^*)$.*

Roughly speaking, each maximal SSS rule ϕ^* described in the lemma, provides the maximal absolute value negative drift along the diagonal $a_i Q_i = a_j Q_j, \ \forall i, j$.

Proof. The existence of a matrix ϕ such that condition (1) holds, immediately implies that the set $\widehat{\mathcal{G}}$ is non-empty. Since $\widehat{\mathcal{G}}$ is a completely ordered bounded set, we can consider $y^* = \sup \widehat{\mathcal{G}}$, where the supremum is component-wise. Using compactness of the set of possible matrices ϕ and continuity of the mapping $v(\phi)$, we easily establish the existence of ϕ^* such that $y^* \leq v(\phi^*)$, and therefore $y^* \in \widehat{\mathcal{G}}$. The equality $y^* = v(\phi^*)$ must hold. Indeed, if the strict inequality $y_i^* < v_i(\phi^*)$ would hold for at least one i, we could always "adjust" elements of the matrix ϕ^* to produce another matrix ϕ^{**} such that $y_i^* < v_i(\phi^{**})$ for all i, which would contradict the maximality of y^*. The same argument shows the maximality of the SSS rule associated with ϕ^*. ∎

4 Throughput Optimality of the Exponential Rule

In this section, we prove that the exponential rule is throughput-optimal. The detailed proof will be for the EXP-Q only. The proof for the EXP-W is virtually same, but requires a slight adjustment which we will sketch at the end of this section.

Consider the system we fixed earlier in this paper. We remind that for this system the necessary and sufficient condition described in Theorem 3.1 holds. Suppose now that this system operates under an EXP-Q scheduling rule with a fixed set of parameters $\beta > 0$, $\eta \in (0, 1)$, and $a_i > 0$, $\gamma_i > 0$ for $i \in N$. Without loss of generality, we assume that $\gamma_1 = 1$. Also, just to make the proof more readable, we put $\eta = 1/2$. (The proof for any $0 < \eta < 1$ is obtained by trivial modifications.) Then the EXP-Q rule is given by

$$i \in i(S(t)) = \arg\max_i \gamma_i \ \mu_i(t) \ \exp\left(\frac{a_i Q_i(t)}{\beta + \sqrt{\overline{Q}(t)}}\right) \ .$$

From this point on, let us fix a matrix ϕ^* (and the corresponding maximal SSS), as in Lemma 3.1, with the constants a_i being the parameters of the EXP-Q rule. Let us denote

$$\epsilon^* \doteq a_1(y_1^* - \lambda_1) = a_i(v_i(\phi^*) - \lambda_i), \ \forall i \in N \ . \tag{3}$$

For this matrix ϕ^*, let us fix a corresponding set of positive constants $\{\alpha_i\}$, as in Theorem 3.2. Without loss of generality we assume $\alpha_1 = 1$.

Let us define b_i by

$$\gamma_i e^{b_i} \quad = \quad \alpha_i \ , \ i \in N \ .$$

Note that $b_1 = 0$ (since $\alpha_1 = 1$ by the convention adopted earlier).

To prove Theorem 2.1, it will suffice to prove the following "narrower" statement.

Lemma 4.1 *The fixed system described above is stable.*

The proof will use the *fluid limit* technique. In the next subsection we describe preliminaries needed to use the technique. In the following subsection, we apply the technique to prove Lemma 4.1. Our application of the fluid limit technique is not straight forward. It involves a *separation of time scales* argument: namely, it requires the analysis of fluid limit processes on two different time scales.

4.1 Fluid Limit Technique Preliminaries

Let us define the norm of the state $S(t)$ as follows:

$$\|S\| \equiv \sum_{i}^{N} Q_i(t) \ .$$

Let $S^{(n)}$ denote a process S with an initial condition such that $\|S^{(n)}(0)\| = n$. In the analysis to follow, all variables associated with a process $S^{(n)}$ will be supplied with the upper index (n).

The following theorem is a corollary of a more general result of Malyshev and Menshikov [10].

Theorem 4.1 *Suppose there exist $\epsilon > 0$ and an integer $T > 0$ such that for any sequence of processes $S^{(n)}, n = 1, 2, \ldots,$ we have*

$$\limsup_{n\to\infty} E[\frac{1}{n}\|S^{(n)}(nT)\|] \leq 1 - \epsilon \ . \tag{4}$$

Then S is ergodic.

It was shown by Rybko and Stolyar [11] that an ergodicity condition of the type (4) naturally leads to a fluid-limit approach to the stability problem of queueing systems. This approach was further developed by Dai [6], Chen [5], Stolyar [13], and Dai and Meyn [7]. As the form of (4) suggests, the approach studies a fluid process $s(t)$ obtained as a limit of the sequence of scaled processes $\frac{1}{n}S^{(n)}(nt), t \geq 0$. At the heart of the approach (in its standard form) is a proof that $s(t)$ starting from any initial state with norm $\|s(0)\| = 1$ reaches 0 in finite time T and stays there.

The first thing we need to do is to define what the scaling $\frac{1}{n}S^{(n)}(nt)$ means in our setting. In order for this scaling to make sense, we will need an alternative definition of the process.

To this end, let us define the following random functions associated with the process $S^{(n)}(t)$. Let $F_i^{(n)}(t)$ be the total number of type-i customers that arrived by time $t \geq 0$, including the customers present at time 0; and $\hat{F}_i^{(n)}(t)$ be the number of type-i customers that were served by time $t \geq 0$. Obviously, $\hat{F}_i^{(n)}(0) = 0$ for all i. As in [11] and [13], we "encode" the initial state of the system; in particular, we extend the definition of $F_i^{(n)}(t)$ to the negative interval $t \in [-n, 0)$ by assuming that the customers present in the system in its initial state $S^{(n)}(0)$ arrived in the past at some of the time instants $-(n-1), -(n-2), \ldots, 0$, according to their delays in the state $S(0)$. By this convention $F_i^{(n)}(-n) = 0$ for all i and n, and $\sum_{i=1}^{N} F_i^{(n)}(0) = n$. Also, denote by $G_m^{(n)}(t)$ the total number of time slots before time t (i.e., among the slots $0, 1, \ldots, t-1$), when the channel was in state m; and by $\hat{G}_{mi}^{(n)}(t)$ the number of time slots before time t when the channel state was m and the channel was allocated to serve queue i. Then the following relations obviously hold:

$$Q_i^{(n)}(t) \equiv F_i^{(n)}(t) - \hat{F}_i^{(n)}(t), \quad t \geq 0, \quad i = 1, 2, \ldots, N, \tag{5}$$

Finally, let

$$\widetilde{Q}_i^{(n)}(t) \quad \equiv \quad a_i Q_i^{(n)}(t) - b_i \sqrt{\frac{1}{N}\sum_j a_j Q_j^{(n)}(t)}$$

It is clear that the process $S^{(n)} = (S^{(n)}(t), t \geq 0)$ is a projection of the process $X^{(n)} = (F^{(n)}, \hat{F}^{(n)}, G^{(n)}, \hat{G}^{(n)}, Q^{(n)})$, where

$$F^{(n)} = (F_i^{(n)}(t), \quad t \geq -n, \quad i = 1, 2, \ldots, N) \ ,$$

$$\hat{F}^{(n)} = (\hat{F}_i^{(n)}(t), \quad t \geq 0, \quad i = 1, 2, \ldots, N),$$

$$G^{(n)} = (G_m^{(n)}(t), \quad t \geq 0, \quad m \in M) \ ,$$

$$\hat{G}^{(n)} = (\hat{G}_{mi}^{(n)}(t), \quad t \geq 0, \ m \in M, \ i = 1, 2, \ldots, N),$$

$$Q^{(n)} = (Q_i^{(n)}(t), \quad t \geq 0, \quad i = 1, 2, \ldots, N),$$

In other words, a sample path of $X^{(n)}$ uniquely defines the sample path of $S^{(n)}$.

Let us also adopt the convention

$$Y(t) = Y(\lfloor t \rfloor), \ \text{for } Y = S^{(n)}, F_i^{(n)}, \hat{F}_i^{(n)}, G_m^{(n)}, \hat{G}_{mi}^{(n)}, Q_i^{(n)}$$

with $t \geq -n$ for $Y = F_i^{(n)}$ and $t \geq 0$ for all other functions. This convention allows us to view the above functions as continuous-time processes defined for all $t \geq 0$ (or $t \geq -n$), but having constant values in each interval $[t, t+1)$.

Now consider the scaled process $x^{(n)} = (f^{(n)}, \hat{f}^{(n)}, g^{(n)}, \hat{g}^{(n)}, q^{(n)})$, where

$$f^{(n)} = (f_i^{(n)}(t), \quad t \geq -1, \quad i = 1, 2, \ldots, N) \ ,$$

$$\hat{f}^{(n)} = (\hat{f}_i^{(n)}(t), \quad t \geq 0, \quad i = 1, 2, \ldots, N),$$

$$g^{(n)} = (g_m^{(n)}(t), \quad t \geq 0, \quad m \in M) \ ,$$

$$\hat{g}^{(n)} = (\hat{g}_{mi}^{(n)}(t), \quad t \geq 0, \ m \in M, \ i = 1, 2, \ldots, N),$$

$$q^{(n)} = (q_i^{(n)}(t), \quad t \geq 0, \quad i = 1, 2, \ldots, N) \ ,$$

$$\tilde{q}^{(n)} = (\tilde{q}_i^{(n)}(t), \quad t \geq 0, \quad i = 1, 2, \ldots, N) \ ,$$

and the scaling is defined as

$$z^{(n)}(t) = \frac{1}{n} Z^{(n)}(nt) \ .$$

From (5) we get:

$$q_i^{(n)}(t) \equiv f_i^{(n)}(t) - \hat{f}_i^{(n)}(t), \quad t \geq 0, \quad i = 1, 2, \ldots, N. \tag{6}$$

The following lemma states the convergence to and basic properties of a "fluid limit" process, and is a variant of Theorem 4.1 in [6] or Theorem 7.1 in [13].

Lemma 4.2 *The following statements hold with probability* 1. *For any sequence of processes* $X^{(n)}$, *there exists a subsequence* $X^{(k)}$, $\{k\} \subseteq \{n\}$, *such that for each* i, $1 \leq i \leq N$ *and* $m \in M$,

$$(f_i^{(k)}(t), t \geq -1) \Rightarrow (f_i(t), t \geq -1) \tag{7}$$

$$(f_i^{(k)}(t), t \geq 0) \to (f_i(t), t \geq 0) \quad u.o.c. \tag{8}$$

$$(\hat{f}_i^{(k)}(t), t \geq 0) \to (\hat{f}_i(t), t \geq 0) \quad u.o.c. \tag{9}$$

$$(q_i^{(k)}(t), t \geq 0) \to (q_i(t), t \geq 0) \quad u.o.c. \tag{10}$$

$$(\tilde{q}_i^{(k)}(t), t \geq 0) \to (a_i q_i(t), t \geq 0) \quad u.o.c. \tag{11}$$

$$(g_m^{(k)}(t), t \geq 0) \to (g_m(t), t \geq 0) \quad u.o.c. \tag{12}$$

$$(\hat{g}_{mi}^{(k)}(t), t \geq 0) \to (\hat{g}_{mi}(t), t \geq 0) \quad u.o.c. \tag{13}$$

where the functions f_i are non-negative non-decreasing right continuous with left limits (RCLL) in $[-1, \infty)$, the functions $f_i, \hat{f}_i, g_m, \hat{g}_{mi}$ are non-negative non-decreasing Lipschitz-continuous in $[0, \infty)$, functions q_i are continuous in $[0, \infty)$, "$\Rightarrow$" signifies convergence at continuity points of the limit, and "u.o.c." means uniform convergence on compact sets, as $k \to \infty$. The limiting set of functions

$$x = (f, \hat{f}, g, \hat{g}, q)$$

also satisfies the following properties:

$$\sum_{i=1}^{N} f_i(0) \leq 1, \tag{14}$$

and for all $i,\ 1 \leq i \leq N$ and $m \in M$,

$$f_i(t) - f_i(0) = \lambda_i t, \quad t \geq 0, \tag{15}$$

$$\hat{f}_i(0) = 0, \tag{16}$$

$$\hat{f}_i(t) \leq f_i(t), \quad t \geq 0\ , \tag{17}$$

$$g_m(t) = \pi^m t, \quad t \geq 0, \tag{18}$$

$$q_i(t) = f_i(t) - \hat{f}_i(t), \quad t \geq 0\ , \tag{19}$$

$$\hat{g}_{mi}(0) = 0, \tag{20}$$

$$\sum_{i=1}^{N} \hat{g}_{mi}(t) = g_m(t), \tag{21}$$

for any interval $[t_1, t_2] \subset [0, \infty)$,

$$\hat{f}_i(t_2) - \hat{f}_i(t_1) \leq \sum_{m \in M} \mu_i^m (\hat{g}_{mi}(t_2) - \hat{g}_{mi}(t_1))\ , \tag{22}$$

if $q_i(t) > 0$ for $t \in [t_1, t_2] \subset [0, \infty)$, then

$$\hat{f}_i(t_2) - \hat{f}_i(t_1) = \sum_{m \in M} \mu_i^m (\hat{g}_{mi}(t_2) - \hat{g}_{mi}(t_1))\ , \tag{23}$$

Proof. It follows from the strong law of large numbers that, with probability 1 for every i,

$$(f_i^{(n)}(t) - f_i^{(n)}(0),\ t \geq 0) \to (\lambda_i t,\quad t \geq 0) \quad u.o.c.$$

So, to prove (8), (14), and (15) it suffices to choose a subsequence $\{k\} \subseteq \{n\}$ such that for every i, $\lim f_i^{(k)}(0)$ exists, and denote the limit by $f_i(0)$. Since all $f_i^{(k)}$ are non-decreasing, we can always choose a further subsequence such that (7) holds.

The properties (12) and (18) follow from the ergodicity of the channel state process.

Also, for any fixed $0 \leq t_1 \leq t_2$, for every i, m, and any n, we have (using the notation $\mu^* \equiv \max_{m,j} \mu_j^m$):

$$\hat{f}_i^{(n)}(t_2) - \hat{f}_i^{(n)}(t_1) \leq \sum_{m \in M} \mu_i^m (\hat{g}_{mi}^{(n)}(t_2) - \hat{g}_{mi}^{(n)}(t_1) + 1/n) \leq \mu^*(t_2 - t_1 + 1/n)\ .$$

From this inequality we deduce the existence of a subsequence (of the subsequence already chosen) such that the convergences (9) and (13) take place, and (22) holds.

The relations (16), (17), (20), and (21), follow from the corresponding relations which trivially hold for the prelimit functions (with index (n)). The convergence (10) and identity (19) trivially follow from identity (6). Convergence (11) follows from (10) as the (scaled by $1/n$) $\sqrt{\cdot}$ term goes to zero u.o.c., as $n \to \infty$.

Suppose, $q_i(t) > 0$ for $t \in [t_1, t_2] \subset [0, \infty)$. Let us fix $\delta \in (0, \min_{t \in [t_1, t_2]} q_i(t))$. The Lipschitz continuity of $q_i(\cdot)$, along with u.o.c. convergence of $q_i^{(k)}$ to q_i, implies that (with probability 1) the sequence $X^{(k)}$ is such that for all sufficiently large k, the following inequalities hold:

$$\min_{t \in [\lfloor t_1 k \rfloor, t_2 k + 1]} Q^{(k)}(t) > \delta k > \max_m \mu_i^m\ .$$

The latter property implies that if the queue i was chosen for service anywhere in the interval $[\lfloor t_1 k \rfloor, t_2 k + 1]$ when the channel state was m, then exactly μ_i^m type i customers were served. So, we must have

$$|\hat{F}_i^{(k)}(kt_2) - \hat{F}_i^{(k)}(kt_1) - \sum_{m \in M} \mu_i^m (\hat{G}_{mi}^{(k)}(kt_2) - \hat{G}_{mi}^{(k)}(kt_1))| \leq 2 \max_m \mu_i^m\ .$$

Scaling the last inequality by k and taking the limit $k \to \infty$ we get (23).

∎

Since some of the component functions included in x, namely $f_i(\cdot)$, $\hat{f}_i(\cdot)$, $g_m(\cdot)$, $\hat{g}_{mi}(\cdot)$, $q_i(\cdot)$, are Lipschitz in $[0, \infty)$, they are absolutely continuous. Therefore, at almost all points $t \in [0, \infty)$ (with respect to Lebesgue measure), the derivatives of all those functions exist. We will call such points *regular*.

4.2 More Preliminaries

Let arbitrary $\nu > 0$ and $T > 0$ be fixed. For each n, let us cover the interval $[0, nT]$ with $P_T^n \doteq \lfloor nT/n^{\frac{1}{4}} \rfloor + 1$ equal non-overlapping $n^{\frac{1}{4}}$-long intervals $[(i-1)n^{\frac{1}{4}}, in^{\frac{1}{4}})$, $1 \leq i \leq P_T^n$. Define for each $j = 1, 2, \ldots, N$, and each $m = 1, 2, \ldots, M$,

$A_j^{i,n}$, the number of arrivals from flow j in the time interval $[(i-1)n^{\frac{1}{4}}, in^{\frac{1}{4}})$,

$B_m^{i,n}$, the number of (full) time-slots that the channel is in state m in the time interval $[(i-1)n^{\frac{1}{4}}, in^{\frac{1}{4}})$.

Let us denote

$$\begin{aligned} E_j^n(T,\nu) &= \bigcup_{1\le i\le P_T^n} \left\{ \left| \frac{A_j^{i,n}}{n^{\frac{1}{4}}} - \lambda_j \right| > \nu \right\} , \\ G_m^n(T,\nu) &= \bigcup_{1\le i\le P_T^n} \left\{ \left| \frac{B_m^{i,n}}{n^{\frac{1}{4}}} - \pi^m \right| > \nu \right\} . \end{aligned}$$

Also, let $\mathcal{Q}_+$ denote the set of strictly positive rational numbers.

Lemma 4.3 *The following properties hold:*

$$\Pr\left(\bigcup_{\nu,T\in\mathcal{Q}_+} \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} E_j^n(T,\nu) \right) = 0 \quad \forall j \in N , \tag{24}$$

$$\Pr\left(\bigcup_{\nu,T\in\mathcal{Q}_+} \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} G_m^n(T,\nu) \right) = 0 \quad \forall m \in M . \tag{25}$$

Equivalently, with probability 1*, for all rational* $T > 0$ *and* $\nu > 0$*, there exists finite* k *such that for all* $n > k$*,*

$$\max_{j\in N,\ 1\le i\le P_T^n} \left| \frac{A_j^{i,n}}{n^{\frac{1}{4}}} - \lambda_j \right| < \nu , \tag{26}$$

$$\max_{m\in M,\ 1\le i\le P_T^n} \left| \frac{B_m^{i,n}}{n^{\frac{1}{4}}} - \pi^m \right| < \nu . \tag{27}$$

Proof Fix any rational $T > 0$ and $\nu > 0$. Fix any $j \in N$. According to Assumption 2.1 on the arrival process, the following large deviations estimate holds. There exists $a = a(\nu, j) > 0$ and a $k_1 = k_1(\nu, j)$ such that for all $n \ge k_1$, uniformly on $1 \le i \le P_T^n$,

$$\Pr\left(\left| \frac{A_j^{i,n}}{n^{\frac{1}{4}}} - \lambda_j \right| > \nu \right) < \exp(-n^{\frac{1}{4}}a) \tag{28}$$

Hence, for all $n \ge k_1$,

$$\Pr\left(E_j^n(T,\nu)\right) < P_T^n \exp(-n^{\frac{1}{4}}a) ,$$

and therefore

$$\sum_n \Pr\left(E_j^n(T,\nu)\right) < \infty .$$

By Borel-Cantelli lemma, it follows that

$$\Pr\left(\bigcap_{k=1}^{\infty}\bigcup_{n=k}^{\infty} E_j^n(T,\nu)\right) = 0 \ .$$

This implies property (24), since the first union in (24) is over a countable set.

The proof of (25) is essentially same: a large deviation estimate analogous to (28) follows from the fact that the channel state process $\mathbf{m}$ is a finite irreducible Markov chain (see [8]). ∎

4.3 Proof of the Main Result

We now in position to prove Lemma 4.1 and therefore our main result, Theorem 2.1.

Theorem 4.2 *With probability* 1*, a limiting set of functions, as in Lemma 4.2, satisfies the following additional condition. At every regular point* $t > 0$*,*

$$\max_i a_i q_i(t) > 0 \quad \textit{implies} \quad (\max_i a_i q_i(t))' \leq -\epsilon^* \ ,$$

where

$$\epsilon^* \doteq a_1(y_1 - \lambda_1) > 0$$

is defined by (3).
Consequently, there exists $T > 0$ *such that, with probability* 1*, a limiting set of functions is such that* $\sum_i q_i(t) = 0$ *for any* $t \geq T$*.*

Proof. The subset of outcomes (i.e., elements of the underlying probability space) for which the statements of both Lemma 4.2 and Lemma 4.3 hold, has probability 1. Consider this subset. Suppose statement of the theorem does not hold. Then there exists an outcome within the specified subset, such that a subsequence of scaled processes converges to a "fluid limit" (i.e., a set of limiting functions as in Lemma 4.2) satisfying the following property. For some fixed regular point $t > 0$ and a constant $\eta_1 < \epsilon^*$, we have $h(t) \doteq \max_i \tilde{q}_i(t) \equiv \max_i a_i q_i(t) > 0$ and $h'(t) > -\eta_1$. Let us prove that this assumption leads to a contradiction.

If the assumption holds, then there exist constants $\delta > 0$, $\xi > 0$, and $\eta_2 \in (\eta_1, \epsilon^*)$, such that

$$h(s) > \xi \ , \quad \forall s \in [t, t+\delta] \ ,$$

and

$$\frac{h(t+\delta) - h(t)}{\delta} \quad > \quad -\eta_2 \quad .$$

For each n, let us now divide the interval $[t, t+\delta]$ into $\sqrt{n}$ intervals, each of length $\frac{\delta}{\sqrt{n}}$. (Since $\sqrt{n}$ may not be an integer, we should divide into, say, $\lceil\sqrt{n}\rceil$ intervals. To avoid trivial complications and heavy notation, we assume that $\sqrt{n}$ *is* integer. It will be clear that we do not lose the correctness

of the argument.) Note that in the "unscaled time" (i.e. on the time scale of the original process S), each subinterval is of length $\delta\sqrt{n}$.

Let us denote

$$h^{(n)}(t) \doteq \max_i \tilde{q}_i^{(n)}(t) \ ,$$

and fix any constant $\eta \in (\eta_2, \epsilon^*)$. From the Dirichlet principle, for all sufficiently large n, in at least one of the subintervals (of length $\frac{\delta}{\sqrt{n}}$), the average rate of change of $h^{(n)}(.)$ is greater than or equal to $(-\eta)$. We pick such a subinterval $[s^{(n)}, s^{(n)} + \delta/\sqrt{n}]$ for each n. Let us choose a further subsequence of the sequence of indices $\{n\}$ (which we will still denote by $\{n\}$), such that $s^{(n)} \to s$, for some fixed $s \in [t, t+\delta]$. Obviously, the right end-point $s^{(n)} + \delta/\sqrt{n}$ of the subinterval also converges to s. Let us recall the notation

$$\widetilde{q}_i^{(n)}(t) \equiv \frac{1}{n}\widetilde{Q}_i^{(n)}(nt) \equiv \frac{1}{n}[a_i Q_i^{(n)}(nt) - b_i\sqrt{\bar{Q}^{(n)}(nt)}\ ] \ ,$$

where

$$\bar{Q}^{(n)}(t) \equiv \frac{1}{N}\sum_j a_j Q_j^{(n)}(t) \ .$$

From the subsequence $\{n\}$, we choose a further subsequence such that the order of values of $\widetilde{q}_i^{(n)}(s^{(n)}), i \in N$, remains the same. For example, without loss of generality let us assume that

$$\widetilde{q}_1^{(n)}(s^{(n)}) \geq \ldots \geq \widetilde{q}_N^{(n)}(s^{(n)}) \ .$$

Finally, for each $i \in N$, consider the following processes:

$${}_\diamond q_i^{(n)}(x) \doteq \sqrt{n}[\widetilde{q}_i^{(n)}(s^{(n)} + x/\sqrt{n}) - \widetilde{q}_1^{(n)}(s^{(n)})], \ x \in [0, \delta] \ ,$$

and choose a subsequence such that for each i,

$${}_\diamond q_i^{(n)}(0) \to {}_\diamond q_i(0) \ ,$$

where ${}_\diamond q_1(0) = 0$ (by our construction), and each other ${}_\diamond q_i(0)$ is either finite non-positive or $-\infty$. Let us consider only the case when all ${}_\diamond q_i(0)$ are finite. (If not, it is easy to observe that the flows with ${}_\diamond q_i(0) = -\infty$ receive no service at all in the (unscaled) time interval $[ns^{(n)}, ns^{(n)} + \sqrt{n}\delta]$. So, the same argument, restricted to the remaining subset of flows applies.)

We notice that a process ${}_\diamond q_i^{(n)}(\cdot)$ is obtained from the process $Q_i^{(n)}(\cdot)$ by the time "speedup" of $\sqrt{n}$ and the "space" scaling by the factor $1/\sqrt{n}$ (in addition to the centering). Thus, it is very natural that as $n \to \infty$, the sequence of processes $\{{}_\diamond q_i^{(n)}(\cdot), \ i \in N\}$, all defined in the interval $[0, \delta]$, should converge (over a subsequence of $\{n\}$ to another - "local" - fluid limit. The Proposition 1 formulated below formalizes this observation.

Let us define the following functions, all defined $x \in [0, \delta]$, and associated with the (unscaled) time interval $[ns^{(n)}, ns^{(n)} + \sqrt{n}\delta]$:
$\sqrt{n}{}_\diamond f_i^{(n)}(x)$ is the number of type i customers arrived in the (unscaled) interval $[ns^{(n)}, ns^{(n)} + \sqrt{n}x]$;
$\sqrt{n}{}_\diamond \widehat{f}_i^{(n)}(x)$ is the number of type i customers served in the (unscaled) interval $[ns^{(n)}, ns^{(n)} + \sqrt{n}x]$;
$\sqrt{n}{}_\diamond g_i^{(n)}(x)$ is the number of (complete) time slots in the (unscaled) interval $[ns^{(n)}, ns^{(n)} + \sqrt{n}x]$, in

which the channel was in state m;
$\sqrt{n}\,{}_\diamond\widehat{g}_{mi}^{(n)}(x)$ is the number of (complete) time slots in the (unscaled) interval $[ns^{(n)}, ns^{(n)} + \sqrt{n}x]$, in which the channel was in state m and the service was allocated to queue i.

Proposition 1. There exists a further subsequence of $\{n\}$ such that the following additional properties hold for each $i \in N$ and $m \in M$:

$$({}_\diamond f_i^{(n)}(x),\ 0 \le x \le \delta) \to (\lambda_i x,\ 0 \le x \le \delta) \quad u.o.c. \tag{29}$$

$$({}_\diamond g_i^{(n)}(x),\ 0 \le x \le \delta) \to (\pi^m x,\ 0 \le x \le \delta) \quad u.o.c. \tag{30}$$

$$({}_\diamond \widehat{g}_{mi}^{(n)}(x),\ 0 \le x \le \delta) \to ({}_\diamond \widehat{g}_{mi}(x),\ 0 \le x \le \delta) \quad u.o.c. \tag{31}$$

$$({}_\diamond \widehat{f}_i^{(n)}(x),\ 0 \le x \le \delta) \to ({}_\diamond \widehat{f}_i(x),\ 0 \le x \le \delta) = (\sum_m \mu_i^m\, {}_\diamond \widehat{g}_{mi}(x),\ 0 \le x \le \delta) \quad u.o.c. \tag{32}$$

$$({}_\diamond q_i^{(n)}(x),\ 0 \le x \le \delta) \to ({}_\diamond q_i(x),\ 0 \le x \le \delta) = ({}_\diamond q_i(0) + a_i(\lambda_i x - {}_\diamond \widehat{f}_i(x)),\ 0 \le x \le \delta) \quad u.o.c.\ , \tag{33}$$

where all the functions ${}_\diamond\widehat{g}_{mi}(\cdot)$ and ${}_\diamond\widehat{f}_i(\cdot)$ are non-decreasing Lipschitz continuous with value 0 at $x = 0$, and all functions ${}_\diamond q_i(\cdot)$ are Lipschitz continuous, and in addition

$$\sum_{i=1}^N {}_\diamond \widehat{g}_{mi}(x) = \pi^m x\ .$$

The proof of Proposition 1 is completely analogous to the proof of Lemma 4.2. The convergence properties (29) and (30) trivially follow from the fact that we consider an element of probability space for which the properties (26) and (27) hold.

Let us denote

$${}_\diamond h^{(n)}(x) = \max_i {}_\diamond q_i^{(n)}(x),\ x \in [0, \delta]\ ,$$

and

$${}_\diamond h(x) = \max_i {}_\diamond q_i(x),\ x \in [0, \delta]\ .$$

It follows from Proposition 1 that ${}_\diamond h(\cdot)$ is a Lipschitz continuous function, and from our construction that

$${}_\diamond h(\delta) - {}_\diamond h(0) \ge -\eta\delta\ .$$

A point $x \in (0, \delta)$ we will call *regular* if the derivatives of all the functions ${}_\diamond h(\cdot)$, ${}_\diamond q_i(\cdot)$, ${}_\diamond \widehat{g}_{mi}(\cdot)$, and ${}_\diamond \widehat{f}_i(\cdot)$, exist in this point. Almost all points (with respect to Lebesgue measure) of the interval $(0, \delta)$ are regular.

We will now show that in each regular point $x \in (0,\delta)$, ${}_\diamond h'(x) \leq -\epsilon^*$. This will imply that ${}_\diamond h(\delta) - {}_\diamond h(0) \leq -\epsilon^*\delta$, which is the desired contradiction.

For each n consider the (unscaled) time interval $[ns^{(n)}, ns^{(n)} + \sqrt{n}\delta]$, and consider how the coefficient of $\mu_i(t)$ (in the EXP-Q rule) behaves in this interval. Obviously, multiplying the coefficients of all $\mu_i(t)$ by the same positive function of time (not necessarily a constant), does not change the EXP-Q scheduling rule. Therefore, the following functions $\tilde{\gamma}_i^{(n)}(\cdot)$ can be regarded as the coefficients of $\mu_i(t)$. For every n, $i \in N$, and $x \in [0,\delta]$, we have

$$\tilde{\gamma}_i^{(n)}(x) \doteq \gamma_i \exp\left(\frac{a_i Q_i^{(n)}(ns^{(n)} + \sqrt{n}x) - n\tilde{q}_1^{(n)}(s^{(n)})}{\beta + \sqrt{\overline{Q}^{(n)}(ns + \sqrt{n}x)}}\right)$$

$$= \gamma_i \exp\left(\frac{a_i Q_i^{(n)}(ns^{(n)} + \sqrt{n}x) - b_i\sqrt{\overline{Q}^{(n)}(ns^{(n)} + \sqrt{n}x)} - n\tilde{q}_1^{(n)}(s^{(n)}) + b_i\sqrt{\overline{Q}^{(n)}(ns^{(n)} + \sqrt{n}x)}}{\sqrt{n}\left(\frac{\beta}{\sqrt{n}} + \sqrt{\frac{\overline{Q}^{(n)}(ns+\sqrt{n}x)}{n}}\right)}\right)$$

$$= \gamma_i \exp\left(\frac{{}_\diamond q_i^{(n)}(x)}{\frac{\beta}{\sqrt{n}} + \sqrt{\frac{\overline{Q}^{(n)}(ns^{(n)}+\sqrt{n}x)}{n}}}\right) \exp\left(\frac{b_i\sqrt{\frac{\overline{Q}^{(n)}(ns^{(n)}+\sqrt{n}x)}{n}}}{\frac{\beta}{\sqrt{n}} + \sqrt{\frac{\overline{Q}^{(n)}(ns^{(n)}+\sqrt{n}x)}{n}}}\right) .$$

Note that as $n \to \infty$ the convergence

$$\frac{\overline{Q}^{(n)}(ns^{(n)} + \sqrt{n}x)}{n} \longrightarrow \overline{q}(s)$$

is uniform on $x \in [0,\delta]$, where

$$\overline{q}(s) \doteq \frac{1}{N}\sum_i a_i q_i(s) > 0 .$$

Therefore, we have

$$(\tilde{\gamma}_i^{(n)}(x),\ 0 \leq x \leq \delta) \to (\alpha_i \exp({}_\diamond q_i(x)/\sqrt{\overline{q}(s)}),\ 0 \leq x \leq \delta) \quad u.o.c. . \tag{34}$$

Now, consider any regular point $x \in (0,\delta)$. Consider the subset of flows $I^{**} \subseteq N$ for which ${}_\diamond q_i(x)$ is maximal, i.e. ${}_\diamond q_i(x) = {}_\diamond h(x)$. Also, let us denote

$$M^{**} = \{m \in M \mid \phi^*_{mi} > 0 \text{ for at least one } i \in I^{**}\} .$$

For every $m \in M^{**}$, let us pick an element $i(m) \in I^{**}$ for which $\phi^*_{mi} > 0$. Observe that the value of $\alpha_{i(m)}\mu^m_{i(m)}$ will be the same regardless of which of those elements we pick. From the form of the EXP-Q scheduling rule and the uniform convergence (34) we can make the following observation. There exists a small $\epsilon_1 > 0$ such that for any $z \in (x, x+\epsilon_1)$ and any $\epsilon_2 > 0$, we have the following estimate for all sufficiently large n,

$$\sum_{i\in I^{**}} \alpha_i({}_\diamond\widehat{f}_i^{(n)}(z) - {}_\diamond\widehat{f}_i^{(n)}(x)) =$$

$$\sum_{i\in I^{**}} \alpha_i \sum_m \mu_i^m({}_\diamond g_{mi}^{(n)}(z) - {}_\diamond g_{mi}^{(n)}(x)) \geq$$

$$\sum_{m\in M^{**}}\sum_{i\in I^{**}}\alpha_i\mu_i^m({}_\diamond g_{mi}^{(n)}(z)-{}_\diamond g_{mi}^{(n)}(x))\geq$$

$$\sum_{m\in M^{**}}\pi^m(1-\epsilon_2)(z-x)\alpha_{i(m)}\mu_{i(m)}^m\geq$$

$$\sum_{m\in M^{**}}\pi^m(1-\epsilon_2)(z-x)\sum_{i\in I^{**}}\phi_{mi}^*\alpha_i\mu_i^m=$$

$$(1-\epsilon_2)(z-x)\sum_{i\in I^{**}}\alpha_i v_i(\phi^*)\ .$$

Since $\epsilon_2>0$ is arbitrary, this implies that for any $z\in(x,x+\epsilon_1)$,

$$\sum_{i\in I^{**}}\alpha_i({}_\diamond\widehat{f}_i(z)-{}_\diamond\widehat{f}_i(x))\geq(z-x)\sum_{i\in I^{**}}\alpha_i v_i(\phi^*)\ ,$$

and therefore

$$\sum_{i\in I^{**}}\alpha_{i\diamond}\widehat{f}_i'(x)\geq\sum_{i\in I^{**}}\alpha_i v_i(\phi^*)\ .$$

Then, ${}_\diamond\widehat{f}_i'(x)\geq v_i(\phi^*)$ for at least one $i\in I^{**}$, and therefore ${}_\diamond q_i'(x)\leq-\epsilon^*$ holds for this i. Point x is regular. By definition, in any regular point the derivatives ${}_\diamond h'(\cdot)$ and ${}_\diamond q_i'(\cdot)$ for all $i\in I^{**}$ are all equal. Thus

$${}_\diamond h'(x)\leq-\epsilon^*\ , \tag{35}$$

and we are done. ∎

Proof of Lemma 4.1.
The results of this section imply the following property.
There exists $T<\infty$ such that, with probability 1, any subsequence of processes $X^{(n)}$, has a further subsequence (still denoted $X^{(n)}$), such that

$$\sum_i q_i^{(n)}(T)\to 0\ .$$

This in particular means that w.p.1, any sequence of processes $X^{(n)}$ is such that

$$\lim_{n\to\infty}\frac{1}{n}\|S^{(n)}(nT)\|=0\ ,$$

which, along with the (easily verified) uniform integrability of the family $\{\|S^{(n)}(nT)\|/n\}$, implies

$$\lim_{n\to\infty}E[\frac{1}{n}\|S^{(n)}(nT)\|]=0\ .$$

This verifies the condition (4), and we are done. ∎

4.4 Proof for the EXP-W Rule: A Sketch

The proof for the EXP-W rule requires an additional preliminary step, which goes after Lemma 4.2. This step is analogous to the one in [2] for the M-LWDF rule. (But its proof for the EXP-W is much simpler than that for the M-LWDF.) Namely, it needs to be shown that with probability 1 each limiting set of functions, described in Lemma 4.2, is such that by a finite time $T_1 > 0$ all the work present at time 0 has been served, i.e.,

$$t - w_i(t) > 0, \quad \forall t \geq T_1, \quad \forall i \ ,$$

where w_i is the "$\Rightarrow$" limit of the (scaled) delay process $(1/n)W_i^{(n)}(nt)$.

After this step, the rest of the proof for EXP-W essentially repeats the proof for EXP-Q, because for $t \geq T_1$ the linear relation (Little's law) $q_i(t) = \lambda_i w_i(t)$ exists for both the "conventional" and "local" fluid limits.

5 Conclusions

In this paper, we have studied the scheduling problem in a system with multiple flows served by a single channel with capacity varying in time randomly and asynchronously with respect to different flows. This problem arises, for example, in wireless communications.

We have proved that the EXP scheduling rule is throughput optimal. The rule also shows good performance in "practical" situations (see [12]). This motivates an important subject of the future work: a more detailed study of the EXP rule properties, beyond throughput optimality.

References

[1] M. Andrews, K. Kumaran, K. Ramanan, A. L. Stolyar, R. Vijayakumar, P. Whiting "CDMA Data QoS Scheduling on the Forward Link with Variable Channel Conditions," Bell Laboratories Technical Report, April, 2000.

[2] M. Andrews, K. Kumaran, K. Ramanan, A. L. Stolyar, R. Vijayakumar, P. Whiting, "Scheduling in a Queueing System with Asynchronously Varying Service Rates," 2000. (Submitted.)

[3] W. C. Jakes, "*Microwave Mobile Communications*," Wiley-Interscience, 1974.

[4] 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.

[5] H. Chen, "Fluid Approximations and Stability of Multiclass Queueing Networks: Work-conserving Disciplines," *Annals of Applied Probability*, Vol. 5, (1995), pp. 637-665.

[6] J. G. Dai, "On the Positive Harris Recurrence for Open Multiclass Queueing Networks: A Unified Approach Via Fluid Limit Models," *Annals of Applied Probability*, Vol. 5, (1995), pp. 49-77.

[7] J. G. Dai and S. P. Meyn, "Stability and Convergence of Moments for Open Multiclass Queueing Networks Via Fluid Limit Models," *IEEE Transactions on Automatic Control*, Vol. 40, (1995), pp. 1889-1904.

[8] A. Dembo and O. Zeitouni, "*Large Deviations, Techniques and Applications*," Springer-Verlag, 1998.

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

[10] V.A. Malyshev and M.V. Menshikov, "Ergodicity, Continuity, and Analyticity of Countable Markov Chains," *Transactions of Moscow Mathematical Society*, Vol. 39, (1979), pp. 3-48.

[11] A.N. Rybko and A.L. Stolyar, "Ergodicity of Stochastic Processes Describing the Operation of Open Queueing Networks," *Problems of Information Transmission*, Vol. 28, (1992), pp. 199-220.

[12] S. Shakkottai and A. L. Stolyar, "Scheduling Algorithms for a Mixture of Real-Time and Non-Real-Time Data in HDR," *Proceedings of the International Teletraffic Congress - ITC-17, Salvador, Brazil, September 2001*, Elsevier, 2001, pp. 793-804.

[13] A.L. Stolyar, "On the Stability of Multiclass Queueing Networks: A Relaxed Sufficient Condition via Limiting Fluid Processes," *Markov Processes and Related Fields*, 1(4), 1995, pp.491-512.

[14] A. J. Viterbi, "*CDMA. Principles of Spread Spectrum Communication*," Addison-Wesley, 1995.