CDMA Data QoS Scheduling on the Forward Link with Variable Channel Conditions

Matthew Andrews Krishnan Kumaran Kavita Ramanan
Alexander Stolyar Rajiv Vijayakumar Phil Whiting

Bell Laboratories, Lucent Technologies
600-700 Mountain Avenue
Murray Hill, NJ 07974-0636
{dmandrews,kumaran,kavita,stolyar,rvijayak,pawhiting}@lucent.com
April 2, 2000

Abstract

We consider the problem of scheduling CDMA data users on the forward link. The goal is to meet their QoS requirements defined in terms of probabilistic packet delay bounds. The constraint is the limit on the total forward link transmit power. Each user's channel condition is characterized by the forward link power required to achieve a unit data rate.

This paper extends the work reported in [1], in which several simplifying assumptions were made, including the assumption that channel conditions are constant in time. In this work, we study a more realistic scenario, in which transmission rates can only be chosen from a discrete finite set, rate scheduling can only be done at discrete scheduling intervals, and, most importantly, the users' channel conditions may vary in time.

We show that if the discrete rate set and discrete scheduling intervals are the only additional constraints, a straightforward adjustment of the Largest Weighted Delay First (LWDF) scheduling scheme [14, 1] provides QoS performance that is very good.

In the case of varying channel conditions a *Modified LWDF* (M-LWDF) scheme with properly chosen parameters provides good QoS and is easily implemented. Moreover, we show how M-LWDF can be used to achieve alternative QoS defined in terms of a predefined minimum long-term throughput for each user.

We also derive theoretical results showing that the M-LWDF rule is optimal in the sense of providing maximal possible throughput; if there is any scheduling discipline at all which can handle the offered traffic for all users, then the M-LWDF will do so too.

1 Introduction

This work is a follow up and extension of work [1]. As in that paper, we consider the forward link of a CDMA cell that supports N data users. Each user i has its own probabilistic *Quality of Service (QoS)* requirement of the form

$$\Pr\{W_i > T_i\} \leq \delta_i \ , \tag{1}$$

where W_i is the steady-state packet delay for user i, and parameters T_i and δ_i are the delay threshold and the maximum probability of exceeding it, respectively.

Depending on the amount of interference (both background noise and other-cell interference) different users require different amounts of power for the same data rate of transmission. We assume that each mobile i is characterized by the *weight* c_i, the transmission power requirement per unit data rate. Generally speaking, c_i depends on time t. We assume that the total Base Station (BS) transmit power is limited and normalized to 1. Therefore the following linear constraint on the mobile data transmission rates $\mu_i(t)$ must hold at all times:

$$\sum_{i=1}^{N} c_i(t)\mu_i(t) \leq 1 \ . \tag{2}$$

We assume that rate scheduling decisions (i.e., μ_i reassignments) can only be made at the boundaries of *scheduling intervals* of constant length a. The goal is to find a good *rate scheduling rule* (in other words a *queueing discipline*) which will try to satisfy the QoS requirement (1) for all (or for as many as possible) users, while satisfying the constraint (2).

The model analyzed in [1] was mostly conceptual; the main goal was to study tradeoffs between different types of scheduling rules, and the capacity implications of users having different weights c_i. For this reason, the following simplifying assumptions were made in [1]:

1) *Continuous rate set.* Transmission rates μ_i may be set to any non-negative value as long as (2) holds.

2) *Continuous time scheduling.* The scheduling interval was very short, 1 ms, which (for the parameter setting used) was essentially equivalent to the ability to change transmission rates continuously in time.

3) *Constant channel conditions.* The weights c_i were different for different mobiles, but constant in time.

It has been shown in [14] that the Largest-Weighted-Delay-First (LWDF) scheduling rule is *asymptotically optimal* when T_i's are large and δ_i's are small. The simulations in [1] show that the LWDF indeed appears to be nearly optimal. This rule clearly showed the best QoS performance among all other rules simulated in [1].

The LWDF scheduling rule is as follows:

In the scheduling interval starting at time t, serve at the maximal possible rate a single user j such that

$$a_j W_j(t) = \max_i a_i W_i(t) \ , \tag{3}$$

where $W_i(t)$ is the user i flow delay at time t (i.e., the maximum waiting time of any packet in the queue), and $a_i > 0,\ i = 1, \ldots, N$, is a fixed set of constants.

The choice of the constants a_i which makes LWDF nearly optimal is

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

The goal of the present paper is to find good rate scheduling rules for a more realistic scenario in which the above assumptions 1)-3) do *not* hold.

In Sections 2 and 3 we show that the LWDF rule provides good performance, even when the assumptions 1) and 2) are removed. In the case of LWDF with *discrete rate set* and *discrete time scheduling*, the delay characteristics of users' flows are affected very little, unless the scheduling interval is large enough to be of the order of a packet transmission time.

In Section 4 we drop the assumption 3) and consider the case of *varying channel conditions*. In this case, if the scheduling interval is short enough to follow time variations of $c_i(t)$ (as is the case in Qualcomm's High Data Rate proposal [3]), a scheduling algorithm can take advantage of channel variations by giving some priority to users with (temporarily) better channels. Since channel conditions of different users vary in time in an asynchronous manner, the QoS of all users can be improved, as compared to scheduling schemes which do not take channel conditions into account. (A scheduling rule providing *proportional fairness* in the achieved long-term throughput of different users is proposed and analyzed in [15].) We show that if QoS is defined by (1), then the *Modified LWDF* (M-LWDF) rule provides very good QoS performance. This rule is as follows:

In the scheduling interval starting at time t, serve at the maximal possible rate a single user j such that

$$\frac{\gamma_j W_j(t)}{c_j(t)} = \max_i \frac{\gamma_i W_i(t)}{c_i(t)}$$

where $\gamma_i > 0,\ i = 1, \ldots, N$, is an arbitrary fixed set of constants.

The choice of constants γ_i which results in good QoS performance of M-LWDF is

$$\gamma_i = a_i \overline{c_i} \ ,$$

where $\overline{c_i}$ is the measured short-term average (or median) of $c_i(t)$.

We also show in Section 4.2 that the M-LWDF discipline can be used to provide Quality of Service defined not by (1), but in terms of ensuring certain *minimum* long-term throughput for each user.

In Sections 5 and 6 we derive theoretical results showing that the M-LWDF rule is optimal in the sense of user flows' *stability*. We show that the M-LWDF, with *arbitrary* choice of constants γ_i, ensures that the system is able to handle the offered traffic of all users, if this is feasible at all with any other rule.

1.1 Simulation Framework

In all simulations in this paper, we consider a cell containing $N = 16$ users uniformly distributed throughout the cell. The c_i values of the 16 users are

listed in Figure 1. Those were generated randomly and independently according to the distribution of a mobile Signal-to-Noise-Ratio derived in [16].

The traffic for each mobile is generated by an ON-OFF source, with ON and OFF periods independent and exponentially distributed with means 93 msec and 907 msec, respectively. When the source is ON the (Poisson) flow of packets is generated at the rate of 9 packets/sec. The packet sizes are independent and exponentially distributed. We adjust the load of the system by changing the mean packet size.

The QoS parameters were always set as follows. The deadline T_i was equal to 3 sec for the 8 "close" (smaller c_i) and 7 sec for the 8 "far" (greater c_i) users. The deadline violation probability $\delta_i = 0.1$ for all users. In the simulations we record the packet delay distribution of the "closest" (smallest c_i) and the furthest, (greatest c_i) user, as they represent two extremes of the delay distribution.

2 Discrete Rate Set

Let $\mathcal{R}$ denote the finite set of rates which the base station can use for a transmission to a mobile. For convenience, let us adopt the convention that $0 \in \mathcal{R}$. Then the maximum rate μ_i at which we can transmit to user i is given by $\mu_i = \max\{\mu \in \mathcal{R} : c_i\mu \leq 1\}$. If we transmit to a single user i, we typically incur a waste in power of $1 - \mu_i c_i > 0$. This excess power can be used to transmit to other mobiles. This leads us to the following greedy scheduling algorithm, which is a straightforward adjustment of the Largest Weighted Delay First scheduling discipline, which proved to provide very good QoS performance in case of the continuous rate set (see [1]) and has some nice asymptotic optimality properties [14].

The transmission rates $\mu_i(t)$ at time t are assigned as follows. Renumber the mobiles so that $i < j \Rightarrow a_i W_i(t) \geq a_j W_j(t)$, where $W_j(t)$ is the user i traffic delay at time t (i.e., the current waiting time of the head-of-the-line packet). For $p \geq 0$ and $c > 0$, define the function $R(p;c) = \max\{\mu \in \mathcal{R} : c\mu \leq p\}$. Then the rate $\mu_i(t)$ assigned to mobile i is computed iteratively by

$$\mu_1(t) = R(1;c_1)\ , \tag{4}$$

$$\mu_i(t) = R\left(1 - \sum_{j=1}^{i-1} \mu_j(t)c_j; c_i\right)\ , \quad i = 2,\ldots,N, \tag{5}$$

where N is the number of mobiles.

The rate set $\mathcal{R}$ used for our simulations was $\{0\} \cup \{9.6 \times 2^j : j = 0,\ldots,6\}$, where all rates are in kbps.

The simulation results comparing the performance of the system with the LWDF discipline adjusted for the discrete rate set system to the performance of the system with "pure" LWDF (i.e., one with a continuous rate set of $[0,\infty)$) are shown in Figures 3 and 4. In those simulations the scheduling interval is 1 ms, as in [1]. The mean packet size is set to 15500 bits, which corresponds to

an average arrival rate of 13 kbps, an arrival rate during the ON period of 139.5 kbps, and a system load of 91%. As might be expected, the use of a discrete rate set pushes the delay profile to the right and increases the average delay. The probability of violating the deadline increased by about 0.02 for the closest user and 0.03 for the furthest user. In a busy period the transmitter was able to utilize (on average) 96.5% of the available power.

The main conclusion we can draw from the simulations is that (at least for our parameter setting) the impact of the discrete rate set restriction is small.

3 Discrete Time Scheduling

In the simulations of the previous section (as well as in [1]), the scheduling interval was 1 ms. This means the transmission rates were recomputed every 1 ms. This means that even for the maximum possible transmission rate (within our rate set) 614.4 kbps, the ratio of scheduling interval duration to the mean packet transmission time was about 0.04. Thus, so far we were modeling a situation when transmission rates could be changed essentially continuously in time. In many practical situations however it is impossible or highly undesirable to have a very short scheduling interval. For example, in CDMA2000 systems the scheduling interval must be a multiple of 20 ms (one frame duration).

In this section we investigate the effects of the scheduling interval being 20 ms or a multiple of 20 ms.

We will assume that

1. delays can only be computed to within multiples of 20 ms,

2. transmissions may start only at the boundaries of scheduling intervals.

With the above assumptions, it is never necessary to assign to a mobile i a rate higher than the lowest rate in $\mathcal{R}$ that would cause its buffer to be emptied during the scheduling interval. Therefore, the following further adjustment of the LWDF scheduling discipline is natural.

For $p \geq 0$, $c > 0$, and $Q > 0$, define the function $R^*(p; c, Q) = R(p; c) \wedge \min\{\mu \in \mathcal{R} : \mu a \geq Q\}$, where a is the scheduling interval duration, the min is equal to ∞ if the condition is infeasible, and $\wedge$ denotes the minimum of two numbers. The transmission rates $\mu_i(t)$ for the scheduling interval $[t, t + a)$ are assigned as follows. Renumber the mobiles so that $i < j \Rightarrow a_i W_i(t) \geq a_j W_j(t)$, and let $Q_i(t)$ be the queue length (number of bits in the buffer) for user i at time t. Then

$$\mu_1(t) = R^*(1; c_1, Q_1(t)) \ , \tag{6}$$

$$\mu_i(t) = R^* \left(1 - \sum_{j=1}^{i-1} \mu_j(t) c_j ; c_i, Q_i(t) \right) \ . \tag{7}$$

The results of running this algorithm with scheduling intervals 20 ms, 40 ms, and 100 ms, are presented in Figures 5 through 10. The delay distributions are compared to those of (the discrete rate adjusted) LWDF discipline with 1 ms scheduling interval. We see that the increase in probability of deadline violation is quite small for 20 and 40 ms scheduling interval, and becomes significant only when the scheduling interval becomes of the order of 100 ms.

4 Varying Channel Conditions

4.1 Introductory Discussion

Let us consider the even more realistic case in which the channel conditions vary with time. In other words, suppose that for each user i, $c_i = (c_i(t), t \geq 0)$ is a random process. Of primary interest to us will be the impact of fast fading. So, we are interested in c_i having the form

$$c_i(t) = \bar{c}_i \xi_i(t) \ , \tag{8}$$

where $\bar{c}_i$ is a constant determined by shadow fading effects, and ξ_i is a fast (say Raleigh) fading process.

The channel time variations can be exploited to achieve better QoS for all users. Very roughly speaking, if user i (temporarily!) has a good channel (i.e., low $c_i(t)$), then this provides an opportunity to give this user a higher data rate and to efficiently use the BS transmission power. Thus, as channel conditions change, it seems sensible and natural to dynamically reallocate users' data rates to give some priority to users which at the time have a good channel.

The first issue that arises immediately is one of the scheduling interval length. The scheduling interval must be short enough for the rate adjustments to follow fast fading. This is feasible, and in fact a 1.67 ms frame size is used in the HDR proposal [3]. In both simulations and analysis, we will assume that c_i values stay constant over a scheduling interval, and that, for the upcoming scheduling interval, these values are known to the base station.

Recall that our goal is to find algorithms that try to satisfy the QoS of different users. In the case of time varying channels, this problem is much more complicated. Even the issue of *stability* is quite non-trivial in this case. (Informally, by stability we mean the existence of a stationary regime for the queue length process, i.e., the property that queues do not have an inherent tendency to grow to infinity. See Section 5 for precise definitions.) Indeed, in the case of constant c_i's and (for simplicity) a continuous rate set, the queues are stable if and only if

$$\rho \doteq \sum_i \lambda_i c_i < 1$$

where ρ is the system *nominal load*, i.e. average BS power requirement (recall that maximum BS power is 1), λ_i is the mean data rate (in bps) for user i. The situation is not that simple if the c_i's are varying with time. For a system which makes scheduling decisions oblivious of the c_i values (such as the LWDF scheme

previously discussed, which in the case of a continuous rate set always chooses for service the queue with maximal $a_i W_i(t)$), the average transmission rate to mobile i is (roughly) $\mathbf{E}(1/c_i)$. Therefore the nominal load in this case is

$$\rho \doteq \sum_i \lambda_i / \mathbf{E}(1/c_i) \ . \tag{9}$$

The c_i-oblivious schemes are expected to make queues stable if $\rho < 1$. (This statement is certainly correct for LWDF but may need modification for other disciplines.) It is not hard to see that schemes which use the information on the current values of c_i may be stable (and moreover able to meet QoS requirements) for nominal loads of well over 1 (or, 100%). In the sequel, the nominal load is always as defined in (9).

In Sections 5 and 6 we formulate and prove analytic results regarding stability. The main result Theorem 3 basically says that in a system with continuous rate set (or a system where only one queue may be served in one scheduling interval), the following simple *Modified LWDF* (M-LWDF) scheme makes the system stable if it can be stable at all (with any scheduling rule):

In the scheduling interval starting at time t, serve at the maximal possible rate a single user j such that

$$\frac{\gamma_j W_j(t)}{c_j(t)} = \max_i \frac{\gamma_i W_i(t)}{c_i(t)} \tag{10}$$

where $\gamma_i > 0,\ i = 1, \ldots, N$, is an arbitrary fixed set of constants.

Remark. The assumption that "only one user may be served at a time" is not very restrictive. For example, in the HDR proposal [3] this is a "built-in" constraint.

The M-LWDF rule ensures system stability for any set of γ_i. The choice of γ_i's is a "degree of freedom" we can use to try to satisfy QoS (1) for different users.

In what follows we always consider index rules of the form "Serve user $j \in \arg\max_i I_i$" where I_i is some function. Therefore, we will label the rules by the function I_i. For example, the "$\gamma_i W_i / c_i$" rule is a shorthand for the M-LWDF rule (10).

4.2 The M-LWDF and Alternative Notion of QoS

The M-LWDF scheduling scheme can be used to achieve QoS not only of the form of a probabilistic delay bound (1), but also for other forms of QoS as well.

For example, in case of variable channel conditions, even the problem of keeping the long-term throughput R_i for each flow i above the desired minimum level r_i is non-trivial. Namely, suppose the QoS requirements have the form

$$R_i > r_i \ . \tag{11}$$

However, it is easy to observe that the problem (11) can be viewed as the problem of achieving stability of a system in which the actual random flow i is

replaced by a virtual (token) flow of constant input rate r_i. And, as mentioned above, we prove in this paper that the M-LWDF discipline solves the latter problem if it is feasible at all.

Thus, a practical scheme which will achieve QoS (11) (if it is feasible at all) can be as follows. There is a token bucket associated with each flow i, in which tokens arrive at constant rate r_i. (Token flow and token bucket are virtual of course, implemented as counters in software.) In each scheduling interval, a decision which flow to serve is made according to the M-LWDF rule (10), with $W_i(t)$ however being *not* the actual delay of flow i, but the delay of a longest waiting token in the token bucket i. (Note that since tokens arrive at a constant rate, $W_i(t)$ is equal to [Number of tokens in the bucket i]$/r_i$.) After the service of the (actual) queues in the scheduling interval is complete, the number of tokens in each bucket is reduced by the actual amount of data served from each queue. (Of course, reasonable adjustments need to be made for "special situations", like having too small amount of data in the actual queue which would be chosen by M-LWDF according to the token bucket contents.)

Moreover, the greater the M-LWDF parameter γ_i (relatively to γ_j for other flows), the "tighter" the minimum rate assurance for flow i. This means that the desired minimum rate is provided on a finer time scale, i.e., as the average rate over shorter time intervals.

4.3 Simulation Parameters

For ease of comparison with the simulations in the previous section, we use a 20 ms scheduling interval in this section too; it should be clear however that the absolute value of the scheduling interval we pick is not essential, and moreover that the "variable c_i" schemes we study are intended for shorter scheduling intervals.

For the same reason, in the simulations of this section, we always consider the discrete rate set adjustment of each scheduling rule. This means that when we simulate an "I_i"-rule, the rates in a scheduling interval starting at t are assigned recursively according to (6) and (7), with users being ordered by decreasing value of I_i, and c_i being replaced by $c_i(t)$ for all i.

We model the fading process by a Markov chain. For our simulations, we used the simple three state Markov chain shown in Figure 2.

This model assumes that each mobile i has some median fading level $\overline{c_i}$ and that c_i is either $\overline{c_i}$ or $\overline{c_i} \pm 3$dB. The $\overline{c_i}$ values were chosen to be the same as the c_i values listed in Figure 1.

To model the actual fading process more closely, we made the mobiles spend the most time in the median fading level ($c_i = \overline{c_i}$), with

The traffic pattern of each user is the same as in previous sections except for the mean packet length. (The mean packet length for the constant c_i case was taken to be 15500 bits corresponding to a nominal load of about 91%.) For the varying c_i case, the mean packet length was taken to be 16000 bits corresponding to a nominal load of about 98%.

User Number	c_i(W/bps)
0	2.508×10^{-6}
1	2.518×10^{-6}
2	2.518×10^{-6}
3	2.598×10^{-6}
4	2.771×10^{-6}
5	2.924×10^{-6}
6	3.623×10^{-6}
7	4.142×10^{-6}
8	4.307×10^{-6}
9	4.533×10^{-6}
10	5.229×10^{-6}
11	6.482×10^{-6}
12	6.635×10^{-6}
13	7.257×10^{-6}
14	7.395×10^{-6}
15	7.470×10^{-6}

Figure 1: The c_i values

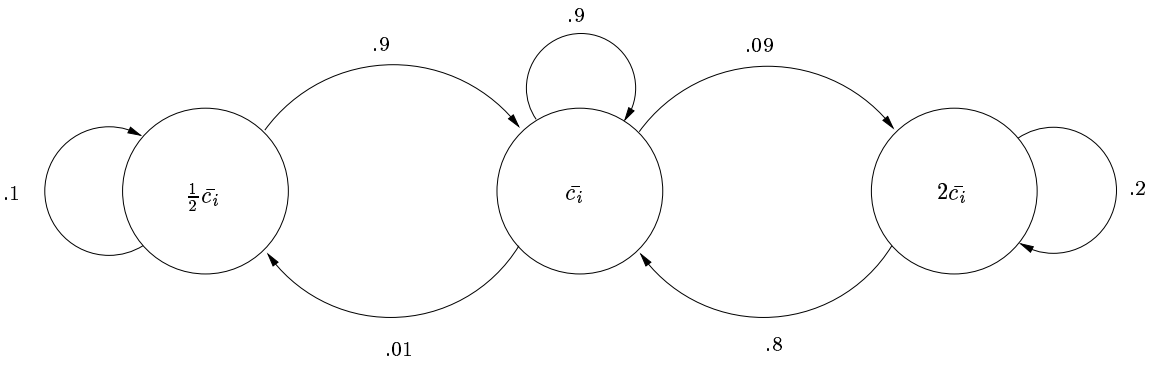

Figure 2: Markov Model for Fading

4.4 Performance of the M-LWDF Scheme

Consider the M-LWDF scheme, "$\gamma_i W_i/c_i$". We will try different reasonable choices of parameter γ_i.

First consider the case $\gamma_i = a_i$, where $a_i = -\log(\delta_i)/T_i$, as defined earlier. Figures 11 and 12 show the results of simulating this scheme, in comparison to the "$a_i W_i$"-rule (i.e., pure LWDF). We observe that, not surprisingly, this rule favors mobiles with lower average c_i. This is of course due to the fact that we use the actual c_i value, rather than the relative value of c_i over $\bar{c}_i$.

We can expect an improvement if we replace the c_i term in the denominator by $c_i/\overline{c_i}$. The results of using this scheduling policy are shown in Figures 13 and 14. We see that delays for the closest user increase and delays for the furthest user decrease. This brings the QoS of different users closer to each other, which is what we are trying to achieve. Indeed, in our particular example, the furthest user just meets its QoS (the probability of delay exceeding 7 sec is about 0.1) and the QoS for the closest user is somewhat better than desired (the probability of delay exceeding 3 sec is about 0.06). We believe that this version of the M-LWDF, namely "$a_i \overline{c_i} W_i/c_i$", is a good scheduling rule which can be used in practice with $\bar{c}_i$ being a measured short-term average or median of $c_i(t)$. However, we notice that (again, not surprisingly) the system is still biased in favor of the users with lower average c_i.

Figures 15 and 16 compare $a_i W_i \overline{c_i}/c_i$ with the following "exponent" rule

$$\gamma_i/c_i \times \exp\left(\frac{a_i W_i - \overline{aW}}{\sqrt{(\overline{aW})} + b}\right), \tag{12}$$

where $\overline{aW} = (1/N)\sum_i a_i W_i$, and $b > 0$ and $\gamma_i > 0$, $i = 1, \cdots, N$ are fixed constants. This rule is *not* a M-LWDF. Its idea is to keep the values of $a_i W_i$ close to each other (which is the main idea of LWDF), but only up to the differences of the order of $\sqrt{(\overline{aW})}$. If some $a_i W_i$ is greater than $\overline{aW}$ by the value of the order $\sqrt{(\overline{aW})}$ or more, this user i will get some priority. If the deviation of $a_i W_i$ is of the order less than $\sqrt{(\overline{aW})}$, then the scheduling decision depends mostly on the current channel conditions, i.e., the weights $c_i(t)$. (Note that the term $-\overline{aW}$ in the numerator of the exponent can be dropped without changing the rule. We put it there to make the above rationale behind the rule clearer.)

Figures 15 and 16 show the performance of the rule with parameters $\gamma_i = \overline{c_i}$ and $b = 2.5$. We are quite surprised to notice that the new scheme reduces delays for *both* the closest and the furthest user. The schemes of type (12) need and deserve more study. This is a subject for future research.

5 Varying Channel Conditions: Stability Results

5.1 Formal Model

Without loss of generality, we assume that the scheduling decisions are made at the boundaries of unit length time intervals, "time slots", i.e., at times $t = 0, 1, 2, \ldots$. By convention, by time slot t we mean the interval $[t, t+1)$. We also assume that traffic is measured in discrete units, "customers". (Typically, customer means one bit of data).

We assume that there is a finite set of (aggregate) channel states

$$M = \{1, \ldots, M\},$$

and the channel state is constant within each time slot. Associated with each state $m \in M$ is a fixed vector of data rates $(\mu_1^m, \ldots, \mu_N^m)$, 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 of individual users.

However, the *service in any time slot may be split* according to a (generally speaking random) stochastic vector $\sigma = (\sigma_1, \ldots, \sigma_N)$, $\sigma_i \geq 0, \forall i$, $\sum_i \sigma_i = 1$. 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.

The random channel state process $\mathbf{m}$ is assumed to be an irreducible [1] discrete time Markov chain with the (finite) state space M. The (unique) stationary distribution of this Markov chain we denote by $\pi = (\pi^1, \ldots, \pi^M)$.

Denote by $A_i(t)$ the number of type i customers arrived in time slot t. To avoid purely technical complications, let us assume that each input process A_i is an ergodic (discrete time) Markov chain, and the input processes are mutually independent. (This condition can be relaxed as follows. The aggregate arrival process $A = \{(A_1(t), \ldots, A_N(t)), t = 0, 1, 2, \ldots\}$ can be described by a finite number of regenerative processes with finite mean regeneration cycles.) Let us denote by λ_i, $i = 1, \ldots, N$, the mean arrival rate for flow i, i.e., the mean number of type i customers arriving in one time slot.

[1] In this paper we refer to ergodicity, aperiodicity and irreducibility of Markov chains. For completeness, we define these concepts here. (See Feller [8].) Let $p_{kj}^{(n)}$ be the probability of transitioning from state k to state j in n steps. A Markov chain is said to be,

- *irreducible* if every state is reachable from every other state with positive probability.
- *aperiodic* if for all k, j the greatest common divisor of the set $\{n : p_{kj}^{(n)} > 0\}$ is 1.
- *ergodic* if $\lim_{n\to\infty} p_{kj}^{(n)} = u_j$, where u_j is the reciprocal of the mean recurrence time of state j.

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.)

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 case of aperiodicity and irreducibility) is equivalent to the existence of a stationary distribution.

5.2 Necessary and Sufficient Stability Condition. Static Service Split Rule

Suppose a stochastic matrix $\phi = (\phi_{mi}, m \in M, i = 1, \dots, N)$ 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, 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 $v = (v_1, \dots, v_N) = v(\phi)$, 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.

Theorem 1 *A scheduling rule H under which the system is stable exists if and only if there exists a stochastic matrix ϕ such that*

$$\lambda_i < v_i(\phi),\ \forall i\ . \tag{13}$$

Proof. Sufficiency of condition (13) is obvious: the SSS rule with the matrix ϕ makes system stable. To prove necessity, consider a rule H under which system

is stable, i.e., Markov chain S is ergodic. Let us denote by $\psi_{mi},\ m \in M,\ i = 1, \ldots, N$, and $\psi_{m0},\ m \in M$, the following stationary expected values:

$$\psi_{mi} = E[\sigma_i(t) I\{m(t) = m, \sum_k Q_k(t) > 0\}],$$

and

$$\psi_{m0} = EI\{m(t) = m, \sum_k Q_k(t) = 0\} \ .$$

(Here $I(\cdot, \cdot)$ denotes the indicator function.) Then the ergodicity of S implies:

$$\psi_{mi} > 0 \ \text{ for all } m \in M \text{ and } i = 0, 1, \ldots, N,$$

$$\sum_{i=0}^{N} \psi_{mi} = \pi^m \ \text{ for all } m \in M \ ,$$

$$\lambda_i \leq \sum_{m \in M} \psi_{mi} \mu_i^m, \ \ i = 1, \ldots, N,$$

If we set

$$\phi_{mi} = \psi_{mi} / \sum_{j=1}^{N} \psi_{mj},\ i = 1, \ldots, N,\ m \in M,$$

we obtain for each i

$$\lambda_i \leq \sum_{m \in M} (\sum_{j=1}^{N} \psi_{mj}) \phi_{mi} \mu_i^m < \sum_{m \in M} \pi^m \phi_{mi} \mu_i^m = v_i(\phi) \ .$$

■

In addition to Theorem 1 (which is quite standard), specifics of our model allow us to characterize the structure of "good" SSS rules. We do that in Theorem 2 below.

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.

Theorem 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 for any m and i,*

$$\phi_{mi}^* > 0 \ \textit{implies} \ i \in \arg\max_j \alpha_j \mu_j^m \ . \tag{14}$$

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 say what to do in case of a tie (although this can be said too), but if for example the number M of channel states is large compared to the number of flows N (equal to the number of constants α_i), then it is natural to expect that typically ties happen rarely, and therefore even an arbitrary tie breaking rule should result in an SSS rule which is a good approximation of a maximal SSS rule.

Proof. Consider the following linear program:

$$\max_{\Lambda, \{\phi_{mi}\}} \Lambda$$

subject to

$$\sum_{m=1}^{M} \pi^m \mu_i^m \phi_{mi} \quad \geq \quad \Lambda v_i^* \tag{15}$$

$$\sum_{i=1}^{N} \phi_{mi} = 1, \qquad 0 \leq \phi_{mi} \leq 1, \ \forall m, i \tag{16}$$

From the definition of v^* we know that $\Lambda = 1$ and $\phi = \phi^*$ solve this linear program, with constraints (15) satisfied as equalities. Then, by Kuhn-Tucker theorem (see for example [9]), there exists a set of non-negative Lagrange multipliers $\alpha_0, \alpha_1, \ldots, \alpha_N$ such that $\Lambda = 1$ and $\phi = \phi^*$ also solve the following linear program (with the same value of the maximum):

$$\max_{\Lambda, \{\phi_{mi}\}} \alpha_0 \Lambda + \sum_{i=1}^{N} \alpha_i \Big(\sum_{m=1}^{M} \pi^m \mu_i^m \phi_{mi} - \Lambda v_i^* \Big) \tag{17}$$

subject to

$$\sum_{i=1}^{N} \phi_{mi} = 1, \qquad 0 \leq \phi_{mi} \leq 1, \ \forall m, i \ . \tag{18}$$

It is easy to verify that all α_i must be strictly positive, and $\alpha_0 = 1$. Then rewriting (17) as

$$\max_{\Lambda, \{\phi_{mi}\}} \Lambda - \Lambda \sum_{i=1}^{N} \alpha_i v_i^* + \sum_{m=1}^{M} \pi^m \sum_{i=1}^{N} \alpha_i \mu_i^m \phi_{mi}$$

we see that the condition (14) must hold, because otherwise the maximum would not be achieved by $\phi = \phi^*$.

■

5.3 Modified LWDF Discipline. Main Result

The following natural question arises. Is there a scheduling rule which (unlike SSS) does not use a priori information about input rates λ_i and the stationary distribution π of the channel state, and yet ensures system stability as long as the necessary and sufficient stability condition (13) is satisfied. Theorem 3 below shows that the answer is **yes**.

Let us call the value

$$W_i(t) \equiv U_{i1}(t)$$

(with $W_i(t) = 0$ if $Q_i(t) = 0$ by convention) the *delay* of flow i at time t.

Let a set of positive constants $\gamma_1, \ldots, \gamma_N$, and a positive constant $\beta > 0$ be fixed. Let us call Modified Largest-Weighted-Delay-First (M-LWDF) a scheduling rule which chooses for service in time slot t a single queue

$$i \in i(S(t)) = \arg\max_j \gamma_j \mu_j^{m(t)} (W_j(t))^\beta.$$

(The previous definition of M-LWDF in (10) is equivalent to the above definition since the meaning of $\mu_j^{m(t)}$ is the inverse of $c_i(t)$.)

An analogous rule, which chooses a single queue

$$i \in i(S(t)) = \arg\max_j \gamma_j \mu_j^{m(t)} (Q_j(t))^\beta.$$

we will call Modified Largest-Weighted-(Unfinished)-Work-First (M-LWWF). (Formally speaking, we need a tie breaking rule which may be, for example, $i = \max\{j \ : \ j \in i(S(t))\}$.)

Theorem 3 *Let an arbitrary set of positive constants $\gamma_1, \ldots, \gamma_N$, and $\beta > 0$ be fixed. Then either of the two scheduling rules, M-LWDF or M-LWWF, make the system stable if it is feasible at all, i.e., if the necessary and sufficient stability condition (13) is satisfied.*

Remark 1. Our Theorems 1 and 3 are closely related to the results by Kahale and Wright [10]. Although our "variable channel" setting may seem quite different, the key idea of the proof of Theorem 3 - the use of a power law Lyapunov function - is similar to the use of quadratic Lyapunov function in [10]. The technique we employ to prove Theorem 3, *fluid limit*, is different though. In particular, use of this technique makes it very intuitive that M-LWDF stability "follows" from the M-LWWF stability. We note that [10] contains the proof of the result analogous to M-LWWF stability, but only a statement of the result analogous to M-LWDF stability. (For recent generalizations of the Kahale-Wright results see [2]. The stability results in [2] are also for the M-LWWF-like scheduling disciplines based on the queue lengths, not for the "delay based" disciplines like M-LWDF.)

Remark 2. It will be clear from the proof of Theorem 3 that it is actually valid for a more general "mixed" M-LWDF/M-LWWF rule:

$$i \in i(S(t)) = \arg\max_j \gamma_j \mu_j^{m(t)} (V_j(t))^\beta,$$

where for each j, V_j may be set to either W_j or Q_j.

6 Proof of Theorem 3

To simplify notation, the proof will be for the case $\beta = 1$. (The M-LWDF discipline with $\beta = 1$ is the one we actually evaluate in our simulations in the previous sections.) The generalization of the proof for arbitrary $\beta > 0$ is trivial: the quadratic Lyapunov function in (45) needs to be replaced by the power law function

$$L(y) = \frac{1}{1+\beta} \sum_{1}^{N} \gamma_i y_i^{1+\beta} \ ;$$

in the formulations of Lemmas 2 and 6, $q_i(t)$, $q_j(t)$, $w_i(t)$, $w_j(t)$, need to be replaced by $q_i(t)^\beta$, $q_j(t)^\beta$, $w_i(t)^\beta$, $w_i(t)^\beta$, respectively; corresponding minor adjustments need to be made throughout the proofs.

6.1 Preliminaries

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

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

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 [11].

Theorem 4 *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*

$$\limsup_{n\to\infty} E[\frac{1}{n}\|S^{(n)}(nT)\|] \leq 1 - \epsilon \ . \tag{19}$$

Then S is ergodic.

It was shown by Rybko and Stolyar [12] that an ergodicity condition of the type (19) naturally leads to a fluid-limit approach to the stability problem of queueing systems. This approach was further developed by Dai [5], Chen [4], Stolyar [13], and Dai and Meyn [6]. As the form of (19) 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. It is sufficient however to show that for some $\epsilon > 0$, $\|s(T)\| \leq 1 - \epsilon$, which is what we are going to do in this paper (In most cases of interest, including ours (with a little bit of extra work), a still weaker condition is sufficient: it is enough to verify that $\inf_{t\geq 0} \|s(t)\| < 1$, as shown in [13].) In

our setting we need to define what the scaling $\frac{1}{n}S^{(n)}(nt)$ means. 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 [12] 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. Let us also denote

$$U_i^{(n)}(t) \equiv t - W_i^{(n)}(t), \quad t \geq 0, \quad i = 1, 2, \ldots, N.$$

Then the following relations obviously hold:

$$Q_i^{(n)}(t) \equiv F_i^{(n)}(t) - \hat{F}_i^{(n)}(t), \quad t \geq 0, \quad i = 1, 2, \ldots, N, \tag{20}$$

$$U_i^{(n)}(t) \leq t, \ t \geq 0,$$

$$U_i^{(n)}(t) = \inf\{s \leq t \ : \ F_i^{(n)}(s) > \hat{F}_i^{(n)}(t)\}, \ t \geq 0. \tag{21}$$

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

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

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

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

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

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

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

$$W^{(n)} = (W_i^{(n)}(t), \quad t \geq 0, \quad i = 1, 2, \ldots, N).$$

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

Let us also adopt the convention

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

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

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

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

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

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

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

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

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

and the scaling is defined as

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

From (20) we get:

$$q_i^{(n)}(t) \equiv f_i^{(n)}(t) - \hat{f}_i^{(n)}(t), \quad t \geq 0, \quad i = 1, 2, \ldots, N. \tag{22}$$

The following lemma establishes convergence to a fluid process and is a variant of Theorem 4.1 in [5].

Lemma 1 *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* $m \in M$,

$$(f_i^{(k)}(t), t \geq -1) \Rightarrow (f_i(t), t \geq -1) \tag{23}$$

$$(f_i^{(k)}(t), t \geq 0) \rightarrow (f_i(t), t \geq 0) \quad u.o.c. \tag{24}$$

$$(\hat{f}_i^{(k)}(t), t \geq 0) \rightarrow (\hat{f}_i(t), t \geq 0) \quad u.o.c. \tag{25}$$

$$(q_i^{(k)}(t), t \geq 0) \rightarrow (q_i(t), t \geq 0) \quad u.o.c. \tag{26}$$

$$(g_m^{(k)}(t), t \geq 0) \rightarrow (g_m(t), t \geq 0) \quad u.o.c. \tag{27}$$

$$(\hat{g}_{mi}^{(k)}(t), t \geq 0) \rightarrow (\hat{g}_{mi}(t), t \geq 0) \quad u.o.c. \tag{28}$$

$$(u_i^{(k)}(t), t \geq 0) \Rightarrow (u_i(t), t \geq 0) \tag{29}$$

$$(w_i^{(k)}(t), t \geq 0) \Rightarrow (w_i(t), t \geq 0) \tag{30}$$

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 $[0, \infty)$, functions u_i are non-decreasing RCLL in $[0, \infty)$, functions w_i are non-negative RCLL 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

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

also satisfies the following properties:

$$\sum_{i=1}^{N} f_i(0) \leq 1, \tag{31}$$

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

$$f_i(t) - f_i(0) = \lambda_i t, \quad t \geq 0, \tag{32}$$

$$\hat{f}_i(0) = 0, \tag{33}$$

$$\hat{f}_i(t) \leq f_i(t), \quad t \geq 0 \ , \tag{34}$$

$$g_m(t) = \pi^m t, \quad t \geq 0, \tag{35}$$

$$q_i(t) = f_i(t) - \hat{f}_i(t), \quad t \geq 0 \ , \tag{36}$$

$$\hat{g}_{mi}(0) = 0, \tag{37}$$

$$\sum_{i=1}^{N} \hat{g}_{mi}(t) = g_m(t), \tag{38}$$

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

$$\hat{f}_i(t_2) - \hat{f}_i(t_1) \leq \sum_{m \in M} \mu_i^m (\hat{g}_{mi}(t_2) - \hat{g}_{mi}(t_1)) \ , \tag{39}$$

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

$$\hat{f}_i(t_2) - \hat{f}_i(t_1) = \sum_{m \in M} \mu_i^m (\hat{g}_{mi}(t_2) - \hat{g}_{mi}(t_1)) \ , \tag{40}$$

$$u_i(t) = t - w_i(t), \tag{41}$$

for any fixed $t_1 > 0$ *the conditions* $u_i(t_1) > 0$ *and* $\hat{f}_i(t_1) > f_i(0)$ *are equivalent and if they hold, then in the interval* $[t_1, \infty)$

$$\lambda_i w_i(t) = q_i(t), \tag{42}$$

which in particular implies that w_i *and* u_i *are Lipschitz continuous in* $[t_1, \infty)$.

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 (24), (31), and (32) 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 $f_i^{(k)}$ and $u_i^{(k)}$ are non-decreasing, we can always choose a further subsequence such that (23) and (29) hold. Then (30) follows from (29).

The properties (27) and (35) 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 (25) and (28) take place, and (39) holds.

The relations (33), (34), (37), (38), and (41), follow from the corresponding relations which trivially hold for the prelimit functions (with index (n)). The convergence (26) and identity (36) trivially follow from identity (22).

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 (40).

The property (42) easily follows from the fact that in the interval $[0,\infty)$ the scaled input flow function $f_i^{(k)}(\cdot)$ converges u.o.c. to the strictly increasing linear function $f_i(0)+\lambda_i t$. We leave details to the reader.

■

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*.

6.2 Proof of Theorem 3 for the M-LWWF discipline

Lemma 2 *Consider a system with the M-LWWF discipline. With probability* 1, *a limiting set of functions* x, *as defined in Lemma 1, satisfies the following additional property. If*

$$\gamma_i\mu_i^m q_i(t) < \max_j \gamma_j\mu_j^m q_j(t) \tag{43}$$

for some regular point $t\geq 0$, *for some* i *and* m, *then*

$$\hat{g}'_{mi}(t)=0. \tag{44}$$

Proof. Let us pick a j at which the maximum in the inequality (43) is attained. Similarly to the proof of property (40) (in Lemma 1), we can fix a small positive $\delta_1>0$, such that for all sufficiently large k, we must have

$$\max_{\zeta\in[(t-\delta_1)k,(t+\delta_1)k]} \gamma_i\mu_i^m Q_i^{(k)}(\zeta) < \min_{\zeta\in[(t-\delta_1)k,(t+\delta_1)k]} \gamma_j\mu_j^m Q_j^{(k)}(\zeta)\ .$$

(If $t=0$ then the time interval should be $[0,\delta_1 k]$.) This means that in the interval $[(t-\delta_1)k+1,(t+\delta_1)k-1]$, queue i can not be served in any time slot when the channel is in state m, because it would contradict the M-LWWF scheduling rule. Thus, for all sufficiently large k we must have:

$$\hat{g}_i^{(k)}(t+\delta_1/2)-\hat{g}_i^{(k)}(t-\delta_1/2)=0\ ,$$

which implies $\hat{g}_i(t+\delta_1/2)-\hat{g}_i(t-\delta_1/2)=0$ and we are done. ■

Let us introduce a quadratic Lyapunov function

$$L(y)=\frac{1}{2}\sum_1^N \gamma_i y_i^2\ , \tag{45}$$

for a vector $y=(y_1,\ldots,y_N)$.

Lemma 3 *Consider a system with the M-LWWF discipline. For any* $\delta_1>0$, *there exists* $\delta_2>0$ *such that the following holds. With probability* 1, *a limiting*

set of functions x, as defined in Lemma 1, satisfies the following additional properties:

$L(q(t)), t \geq 0$, is an absolutely continuous function,

$$L(q(0)) \leq \frac{1}{2}\sum_{1}^{N}\gamma_i \ , \tag{46}$$

and at any regular point t,

$$L(q(t)) \geq \delta_1 \ \textit{implies} \ \frac{d}{dt}L(q(t)) \leq -\delta_2 \ . \tag{47}$$

Proof. For any regular $t \geq 0$ such that $L(q(t)) > 0$, the derivative of $L(q(t))$ can be written as follows:

$$\frac{d}{dt}L(q(t)) = \sum_{i=1}^{N}\gamma_i q_i(t)(\lambda_i - \hat{f}_i'(t)) = \tag{48}$$

$$= \sum_{i=1}^{N}\gamma_i q_i(t)(\lambda_i - v_i(\phi)) + K(\phi, q(t)) - K(\hat{\phi}(t), q(t)) \tag{49}$$

where for a stochastic $M \times N$ matrix ξ and a non-negative N-dimensional vector y we use the notation

$$K(\xi, y) \equiv \sum_i \gamma_i y_i \sum_m \pi^m \xi_{mi}\mu_i^m \equiv \sum_m \pi^m \sum_i \xi_{mi}\gamma_i\mu_i^m y_i \ ,$$

$$\hat{\phi}_{mi}(t) \equiv \hat{g}_{mi}'(t)/\pi^m \ ,$$

and we used the fact (following from the property (40)) that

$$\hat{f}_i'(t) = \sum_m \mu_i^m \hat{g}_{mi}'(t) \text{ if } q_i(t) > 0 \ .$$

Let us choose $\delta_3 > 0$ such that $L(y) \geq \delta_1$ implies $\max_i y_i \geq \delta_3$. Then the first sum in (49) is bounded as follows:

$$\sum_{i=1}^{N}\gamma_i q_i(t)(\lambda_i - v_i(\phi)) \leq -(\min_i \gamma_i)\delta_3 \min_i(v_i(\phi) - \lambda_i) \equiv -\delta_2$$

It remains to show that

$$K(\hat{\phi}(t), q(t)) \geq K(\phi, q(t)) \ . \tag{50}$$

It is easy to see that for any non-negative vector y, a stochastic matrix ξ maximizes $K(\xi, y)$ if and only if the following condition holds for every i and m: If $\gamma_i\mu_i^m y_i < \max_j \gamma_j\mu_j^m y_j$, then

$$\xi_{mi} = 0 \ . \tag{51}$$

But, the property (44) shows that the property (51) is satisfied for $y = q(t)$ and $\xi = \hat{\phi}(t)$. This proves (50) and the lemma. ∎

Lemma 4 *Consider a system with the M-LWWF discipline. For any $\delta > 0$, there exists $T > 0$ such that with probability 1, a limiting set of functions x, as defined in Lemma 1, satisfies the following additional property:*

$$L(q(t)) \leq \delta, \quad t \geq T. \tag{52}$$

Proof follows from Lemma 3.

Proof of Theorem 3 for M-LWWF. According to Lemmas 1- 4, for any fixed $\epsilon_1 > 0$ we can always choose a large enough integer $T > 0$ such that for any sequence of processes $X^{(n)}$, there exists a subsequence $X^{(k)}$, $\{k\} \subset \{n\}$ such that with probability 1 the convergence to a limiting set of functions x takes place, and moreover

$$\sum_i q_i(T) \leq \epsilon_1,$$

implying (T is large!)

$$\hat{f}_i(T) = f_i(T) - q_i(T) > f_i(0), \ \forall i, \tag{53}$$

implying

$$w_i(T) = q_i(T)/\lambda_i, \ \forall i, \tag{54}$$

implying (recall ϵ_1 is small)

$$\sum_i q_i(T) + \sum_i w_i(T) \leq (1 + 1/(\min_i \lambda_i))\epsilon_1 \equiv 1 - \epsilon < 1.$$

Therefore, with probability 1,

$$\limsup_{n\to\infty} \frac{1}{n} \|S^{(n)}(nT)\| \leq 1 - \epsilon \ , \tag{55}$$

which along with the uniform integrability of the sequence $\frac{1}{n}\|S^{(n)}(nT)\|$, $n = 1, 2, \ldots$, verifies condition (19), and therefore proves stability. ∎

The following supplemental statement about the M-LWWF discipline will play an important role in the stability proof for the M-LWDF discipline.

Consider a *generalized* system with a given discipline H. The generalization is to assume that some time slots are unavailable for service of any queue. In each available for service time slot, the scheduling rule is H. In a generalized system let $G_m^{(n)}(t)$ denote the number of *available for service* time slots (by time t) when the channel is in state m.

Lemma 5 *Let positive constants K_0 and K_1 be fixed. Consider a sequence (indexed by k) of sample paths $X^{(n)}$ of the generalized system with M-LWWF, such that all properties described in Lemmas 1 and 2 hold with the following modifications:*

property (31) is replaced by

$$\sum_{i=1}^{N} f_i(0) \le K_0 < \infty, \tag{56}$$

property (35) is replaced by

$$g_m(t) = \pi^m t - h_m(t), \quad t \ge 0, \tag{57}$$

where each function h_m is non-decreasing Lipschitz continuous, $h_m(0) = 0$, and

$$\sum_m \lim_{t\to\infty} h_m(t) \le K_1 \ .$$

Then the function $L(q(t))$ has the upper bound $C < \infty$ which depends only on K_0 and K_1:

$$L(q(t)) \le C, \quad t \ge 0 \ . \tag{58}$$

Proof. We will use the notation $\bar{L}(t) \equiv L(q(t))$. Let us choose $\delta > 0$ small enough, so that the condition $\sum_m h'_m(t) \le \delta$ is satisfied at a regular point t. This implies $g'_m(t) \ge \pi^m - \delta$ for each m, which guarantees that $(d/dt)\bar{L}(t) < 0$. (The existence of such a δ is easily obtained using the argument and estimates used in the proof of Lemma 3.)

Let us denote by Λ the Lebesgue measure, and by $\mathcal{L}$ the σ-algebra of Lebesgue measurable subsets of $[0,\infty)$. Consider a subset

$$B \equiv \{t \in [0,\infty) \ : \ t \text{ is regular}, \sum_m h'_m(t) > \delta\} \ .$$

Clearly, $B \in \mathcal{L}$ and

$$\Lambda(B) \le K_1/\delta \ .$$

Define the measure ν on $\mathcal{L}$ as follows:

$$\nu(A) \equiv \Lambda(A \cap B) \ .$$

Notice that $\nu([0,\infty)) = \Lambda(B)$.

Let us also make a simple observation:

$$\bar{L}'(t) = \sum_i \gamma_i q_i q'_i(t) \le (\max \lambda_i) \sum_i \gamma_i q_i \le (\max \lambda_i) \sum_i \gamma_i (1 + q_i^2)$$

implying the existence of positive c_1 and c_2 such that

$$\bar{L}'(t) \le c_1 + c_2 \bar{L}(t) \ . \tag{59}$$

So, we see that the derivative $\bar{L}'(t)$ is bounded above as in (59) at regular points $t \in B$, and is negative at regular points $t \in [0,\infty)\backslash B$. We can write

$$\bar{L}(t) \le \bar{L}(0) + \int_{[0,t]\cap B} \bar{L}'(y)\Lambda(dy) = \bar{L}(0) + \int_0^t \bar{L}'(y)\nu(dy) \le$$

$$\le \bar{L}(0) + c_1\nu([0,t]) + c_2\int_0^t \bar{L}(y)\nu(dy) \le \bar{L}(0) + c_1\nu([0,\infty)) + c_2\int_0^t \bar{L}(y)\nu(dy)$$

Applying Gronwall's inequality ([7], p. 498), we get

$$\bar{L}(t) \le [\bar{L}(0) + c_1\nu([0,\infty))]\exp\{c_2\nu([0,\infty))\}$$

and finally

$$\bar{L}(t) \le [K_0 + c_1K_1/\delta]\exp\{c_2K_1/\delta\},\ t \ge 0,$$

which proves the Lemma. ∎

6.3 Proof of Theorem 3 for the M-LWDF discipline

The following lemma describes the key property of the M-LWDF discipline which is analogous to the M-LWWF property described in Lemma 2.

Lemma 6 *Consider a system with the M-LWDF discipline. With probability 1, a limiting set of functions x, as defined in Lemma 1, satisfies the following additional property. If in some interval $[t_1,t_2]$, $0 \le t_1 < t_2 < \infty$, for some fixed m, and fixed i and j,*

$$\sup_{t_1\le t\le t_2} \gamma_i\mu_i^m w_i(t) < \inf_{t_1\le t\le t_2} \gamma_j\mu_j^m w_j(t)\ , \tag{60}$$

then

$$\hat{g}_{mi}(t_2) - \hat{g}_{mi}(t_1) = 0. \tag{61}$$

Proof is analogous to the proof of Lemma 2. (The only additional difficulty is the fact that the functions $w_i(\cdot)$ may not be continuous.) Let us fix positive constants α and δ such that

$$\sup_{t_1\le t\le t_2} \gamma_i\mu_i^m w_i(t) < \alpha - \delta < \alpha + \delta < \inf_{t_1\le t\le t_2} \gamma_j\mu_j^m w_j(t)\ . \tag{62}$$

Then for all $t \in [t_1,t_2]$ we have

$$u_i(t) > t - (\alpha-\delta)/(\gamma_i\mu_i^m)$$

and

$$u_j(t) < t - (\alpha+\delta)/(\gamma_j\mu_j^m)\ .$$

Since for each i, $u_i(\cdot)$ and all $u_i^{(k)}(\cdot)$ are non-decreasing, and we have convergence $\lim u_i^{(k)}(t) \to u_i(t)$ for every t where u_i is continuous, we see that for all sufficiently large k, for all $t \in [t_1, t_2]$,

$$u_i^{(k)}(t) > t - \alpha/(\gamma_i \mu_i^m)$$

and

$$u_j^{(k)}(t) < t - \alpha/(\gamma_j \mu_j^m) \ .$$

From the latter two inequalities we see that

$$\gamma_i \mu_i^m w_i^{(k)}(t) < \alpha < \gamma_j \mu_j^m w_j^{(k)}(t), \quad t \in [t_1, t_2].$$

Just as in the proof of Lemma 2, we observe that the latter property implies that for all large k,

$$\hat{g}_{mi}^{(k)}(t_2 - 1/k) - \hat{g}_{mi}^{(k)}(t_1 + 1/k) = 0,$$

because in the unscaled system with index k, queue i may not be served in any time slot in the interval $[kt_1 + 1, kt_2 - 1]$ when the channel is in state m. (Otherwise, we would get a violation of the M-LWDF scheduling rule.) Taking the limit $k \to \infty$ completes the proof. ∎

Lemma 7 *Consider a system with the M-LWDF discipline. There exists $T_N > 0$ such that with probability 1, a limiting set of functions x, as defined in Lemma 1, satisfies the following additional property:*

$$\hat{f}_i(T_N) > f_i(0), \quad i = 1, \ldots, N \ .$$

Proof. Let us fix an arbitrary $\epsilon_2 > 0$. So we have

$$f_i(\epsilon_2) = f_i(0) + \lambda_i \epsilon_2 > f_i(0), \quad \forall i,$$

and

$$\sum_i q_i(\epsilon_2) \leq 1 + (\sum_i \lambda_i)\epsilon_2 \equiv K_1 \ .$$

We will show the existence of T_N such that

$$\hat{f}_i(T_N) \geq f_i(\epsilon_2), \quad i = 1, \ldots, N \ . \tag{63}$$

The proof of (63) is by induction.

Induction Base. There exists $T_1 > 0$ such that for at least one i,

$$\hat{f}_i(T_1) \geq f_i(\epsilon_2).$$

Let us set $T_1 = \epsilon_2 + C_2$ with

$$C_2 = \frac{NK_1}{\min_{m,i} \mu_i^m}$$

If for at least one i, there exists $t \in [\epsilon_2, T_1]$ such that $q_i(t) = f_i(t) - \hat{f}_i(t) = 0$, then we are done. If not, we observe that for at least one i

$$\sum_m (\hat{g}_{mi}(T_1) - \hat{g}_{mi}(\epsilon_2)) \geq C_2/N \ .$$

Then, since for that i, $q_i(t) > 0, \ t \in [\epsilon_2, T_1]$, from (40) we get:

$$\hat{f}_i(T_1) - \hat{f}_i(\epsilon_2) = \sum_m \mu_i^m (\hat{g}_{mi}(T_1) - \hat{g}_{mi}(\epsilon_2)) \geq$$

$$\geq (\min_{m,i} \mu_i^m) \sum_m (\hat{g}_{mi}(T_1) - \hat{g}_{mi}(\epsilon_2)) \geq$$

$$\geq (\min_{m,i} \mu_i^m) C_2/N = K_1 \geq q_i(\epsilon_2) \ .$$

This means

$$\hat{f}_i(T_1) \geq \hat{f}_i(\epsilon_2) + q_i(\epsilon_2) = f_i(\epsilon_2) \ ,$$

which proves the induction base.

Induction Step. Suppose there exists $T_l > 0$, $1 \leq l < N$, such that for at least one subset $N_l \subset \{1, \ldots, N\}$ of cardinality l, we have

$$\hat{f}_j(T_l) \geq f_j(\epsilon_2) \tag{64}$$

for all $j \in N_l$. Then there exists $T_{l+1} \geq T_l$ such that (64) holds for all j within at least one subset N_{l+1} of cardinality $l+1$.

We will prove the induction step for $l = 1$. (The generalization for the arbitrary l is straightforward.) So, we need to prove the existence of $T_2 \geq T_1$ such that for at least two different i and k, (64) holds (with $j = i, k$), with T_1 being the constants from the induction base statement.

Let us fix i for which

$$\hat{f}_i(t) \geq f_i(\epsilon_2), \quad t \geq T_1,$$

according to the induction base.

Suppose

$$\hat{f}_k(T_1) < f_k(\epsilon_2), \text{ for all } k \neq i. \tag{65}$$

We can observe that

$$q_i(T_1) \leq 1 + \lambda_i T_1 \equiv K_0$$

and

$$\sum_{k \neq i} (f_k(\epsilon_2) - \hat{f}_k(T_1)) \leq K_1 \ ,$$

where K_1 is already defined above.

Now, let us view the scaled system after time T_1 (i.e., each unscaled system with index n is considered after time nT_1) as a generalized system with the

single input flow of type i, and with time slots allocated to any other flow being unavailable to flow i. Since the simple linear relation $\lambda_i w_i(t) = q_i(t)$ holds for flow i for all $t \geq T_1$, the generalized system with the M-LWDF discipline satisfies all the properties of the generalized system with the M-LWWF discipline with each γ_i replaced by γ_i/λ_i. Let C be the constant defined in Lemma 5, which depends only on the constants K_0 and K_1 defined above in this proof (and of course on the set of system parameters $\lambda_j,\ \forall j,\ \mu_j^m,\ \forall j, m$, and the stationary distribution $\pi^m,\ \forall m$). Note that $L(q(t)) = (1/2)(\gamma_i/\lambda_i)q_i^2(t) \leq C$ (recall, we are considering a generalized system with flow i only) implies

$$q_i(t) \leq 1 + \sqrt{\frac{2C\lambda_i}{\gamma_i}} \equiv C_1 \ .$$

Let us choose $T_2' > T_1$ large enough so that for all $j \neq i$ and all $m \in M$ we have

$$\gamma_i \mu_i^m C_1/\lambda_i < \gamma_j \mu_j^m (T_2' - \epsilon_2) \tag{66}$$

Finally, let us choose $T_2 > T_2'$ large enough so that

$$\lambda_i (T_2 - T_2') > C_1 \ .$$

We claim that for at least one $k \neq i$ we must have

$$\hat{f}_k(T_2) \geq f_k(\epsilon_2) \ . \tag{67}$$

Suppose not, i.e., for all $k \neq i$, $\hat{f}_k(T_2) < f_k(\epsilon_2)$. Then, by Lemma 5, $L(q(t)) \leq C$ for $t \in [T_1, T_2]$, and therefore

$$q_i(t) \leq C_1,\ t \in [T_1, T_2] \ . \tag{68}$$

Our choice of T_2' in (66) guarantees that for at least one $j \neq i$ and all m,

$$\sup_{T_2' \leq t \leq T_2} \gamma_i \mu_i^m w_i(t) < \inf_{T_2' \leq t \leq T_2} \gamma_j \mu_j^m w_j(t) \ .$$

This (according to Lemma 6) implies that for all m,

$$\hat{g}_{mi}(T_2) - \hat{g}_{mi}(T_2') = 0,$$

and therefore

$$\hat{f}_i(T_2) - \hat{f}_i(T_2') = 0,$$

implying in turn that

$$q_i(T_2) = q_i(T_2') + \lambda_i (T_2 - T_2') > C_1,$$

which is a contradiction to (68). This proves our claim (67). Our choice of T_2 depended on i but since there is only a finite number of possible values of i, we can choose T_2 so that (67) holds for some $k \neq i$ no matter what i is.

Thus we proved the existence of T_2 such that we have (67) for some $k \neq i$, assuming condition (65). But the opposite of the condition (65) implies (67) for some $k \neq i$ trivially. The proof of the induction step is complete. ∎

Proof of Theorem 3 for M-LWDF. We proved the existence of $T_N > 0$ such that for any sequence of processes $X^{(n)}$, there exists a subsequence $X^{(k)}$, $\{k\} \subset \{n\}$ such that with probability 1 the convergence to a limiting set of functions x takes place, and moreover x is such that the strict linear relation

$$\lambda_i w_i(t) = q_i(t), \ t \geq T_N,$$

exists for all i. The latter fact, along with Lemma 6, means that, with probability 1 in the interval $[T_N, \infty)$ the set x also satisfies all the properties described in Lemmas 2-4 if only in their formulations we replace γ_i by γ_i/λ_i, replace (46) by condition

$$L(q(T_N)) \leq \frac{1}{2} \sum_1^N \gamma_i (1 + \lambda_i T_N)^2 \ ,$$

and move the time origin to T_N. Therefore, for any $\epsilon_1 > 0$ there exists $T \geq T_N$ such that, with probability 1, x satisfies the condition

$$\sum_i q_i(T) \leq \epsilon_1 \ .$$

The rest is exactly as in the proof of the Theorem for M-LWWF. The only difference is that we get (54) directly from the property (42) and Lemma 7, and not from (53). ∎

7 Conclusions

The main conclusions of this work are as follows.

- In case of **constant channel conditions**, a simple adjustment of the LWDF scheduling discipline provides good Quality of Service for the users even when there are additional "discrete rate set" and "discrete time" scheduling constraints. Very naturally, the condition for this is that the scheduling interval length is small compared to a typical packet transmission time.
- In case of **variable channel conditions**, we have shown that the M-LWDF rule is optimal in the sense that it can handle all the offered traffic if this is feasible at all. Moreover, with the appropriate choice of parameters, which we specify in the paper, the M-LWDF provides very good Quality of Service. The M-LWDF rule can also be used to satisfy different QoS requirements, namely, the desired minimum long-term throughput for each user.

An important subject of future research is finding good scheduling rules which are less dependent on the "proper" parameter setting. The "exponent" rule we considered in this paper is a step in this direction.

Acknowledgment We would like to thank Sem Borst for numerous useful discussions.

References

[1] M.Andrews, K.Kumaran, K.Ramanan, A.Stolyar, P.Whiting. Data Rate Scheduling Algorithms and Capacity Estimates for the CDMA Forward Link. *Bell Labs Technical Memorandum* BL0112120-990922-32TM, 1999.

[2] M.Armony, N.Bambos. Queueing Networks with Interacting Service Resources. *Preprint*, 2000.

[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. Preprint, 1999.

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

[5] 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.

[6] 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.

[7] S. N. Ethier and T. G. Kurtz. *Markov Process: Characterization and Convergence.* John Wiley and Sons, New York, 1986.

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

[9] P.E.Gill and W.Murray. *Numerical Methods for Constrained Optimization.* Academic Press, London, 1974.

[10] N.Kahale and P.E.Wright. Dynamic Global Packet Routing in Wireless Networks. *Proceedings of the INFOCOM'97*, 1997, pp. 1414-1421.

[11] 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.

[12] 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.

[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. L. Stolyar and K. Ramanan. Largest Weighted Delay First Scheduling: Large Deviations and Optimality. *Bell Labs Technical Memorandum* BL0112120-990823-25TM, 1999. To appear in *Annals of Applied Probability*, 2000.

[15] D. Tse. Forward Link Multiuser Diversity Through Proportional Fair Scheduling. *Presentation at Bell Labs*, 8/12/1999.

[16] Viterbi A.J. (1995) CDMA. Principles of Spread Spectrum Communication, *Addison-Wesley*.

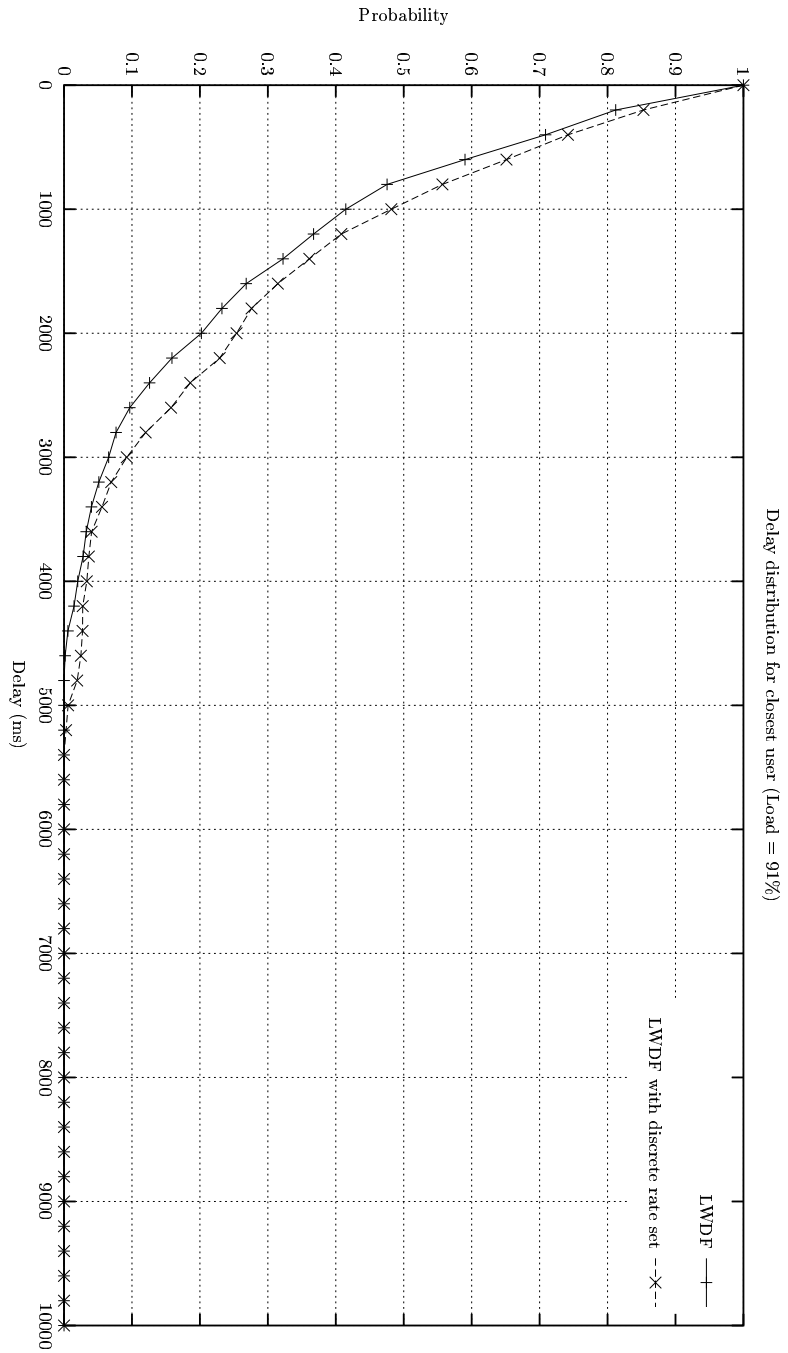

Figure 3: LWDF: Discrete Vs. Continuous Rate Set

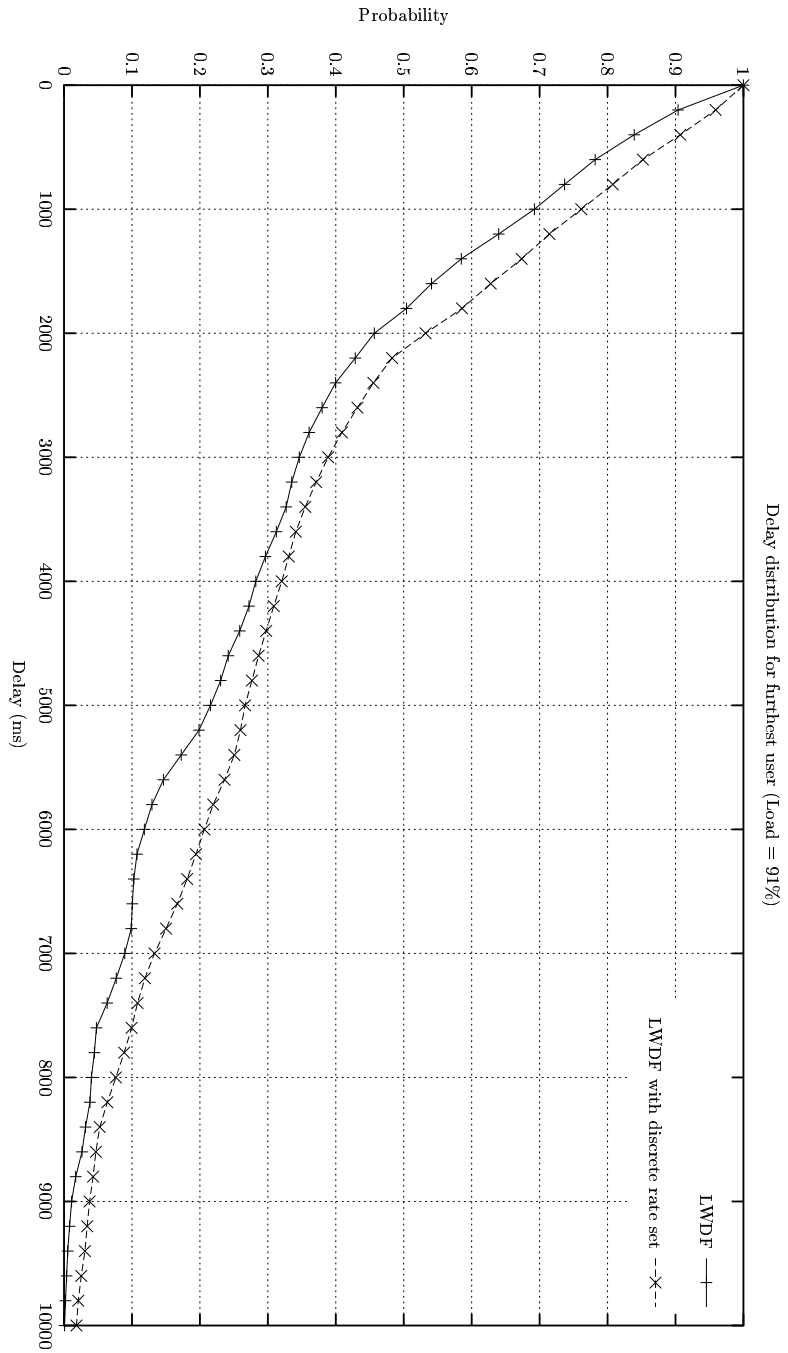

Figure 4: LWDF: Discrete Vs. Continuous Rate Set

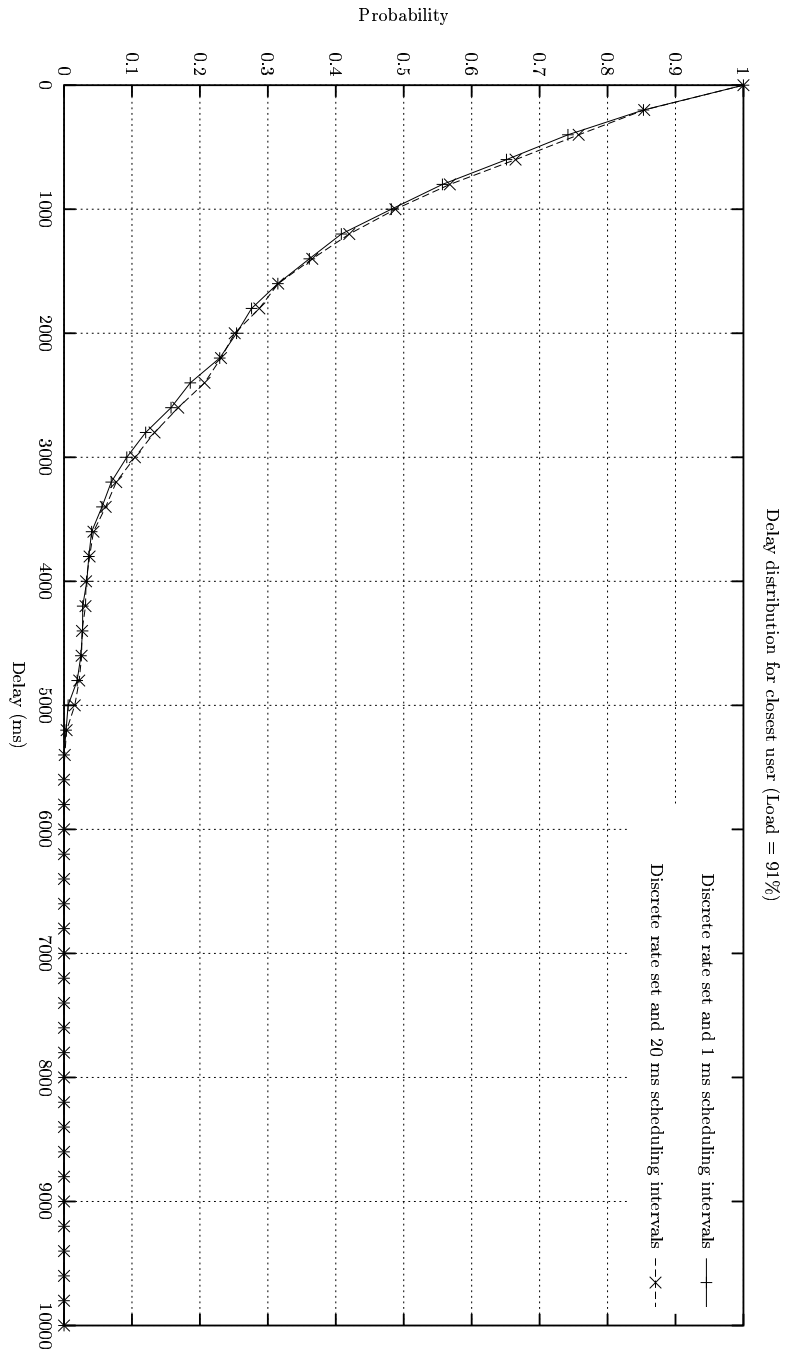

Figure 5: LWDF: Discrete Vs. Continuous Time Scheduling

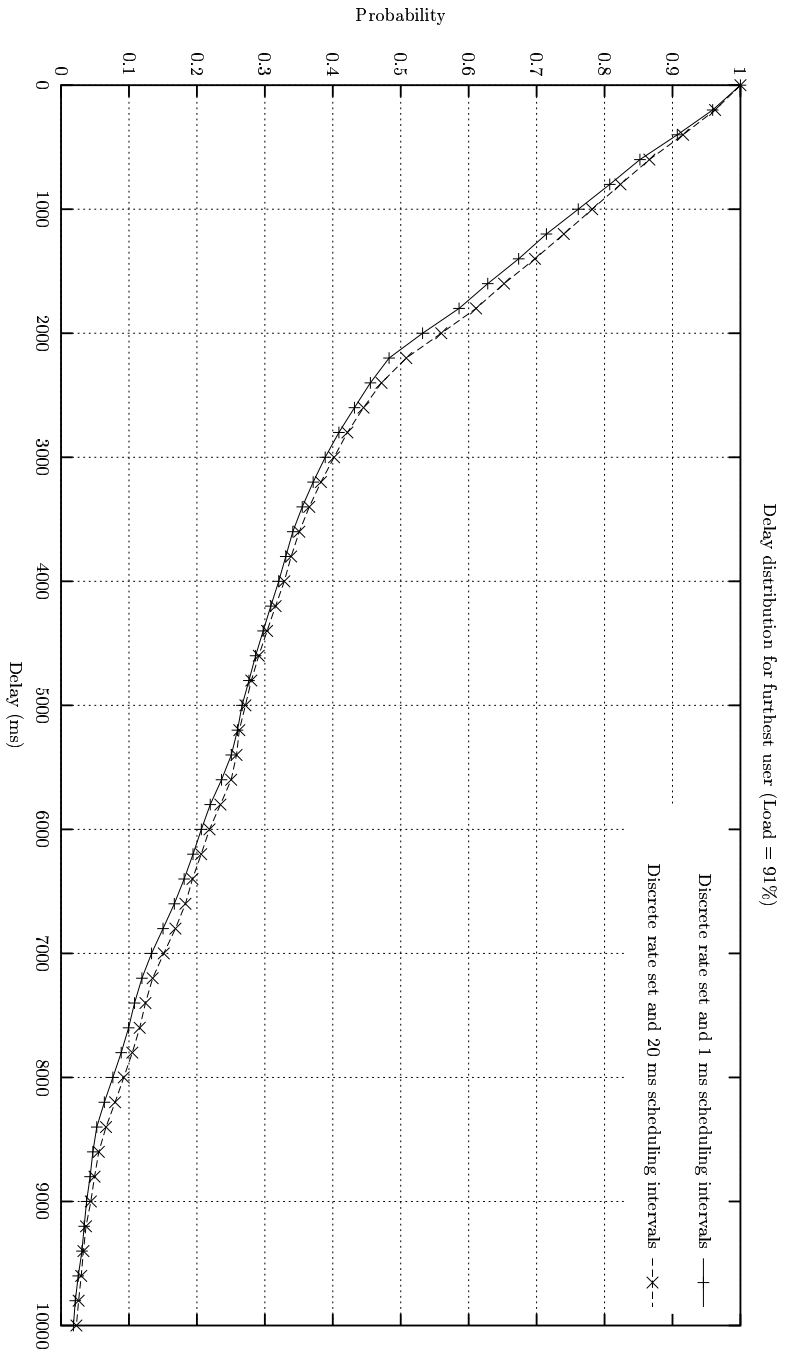

Figure 6: LWDF: Discrete Vs. Continuous Time Scheduling

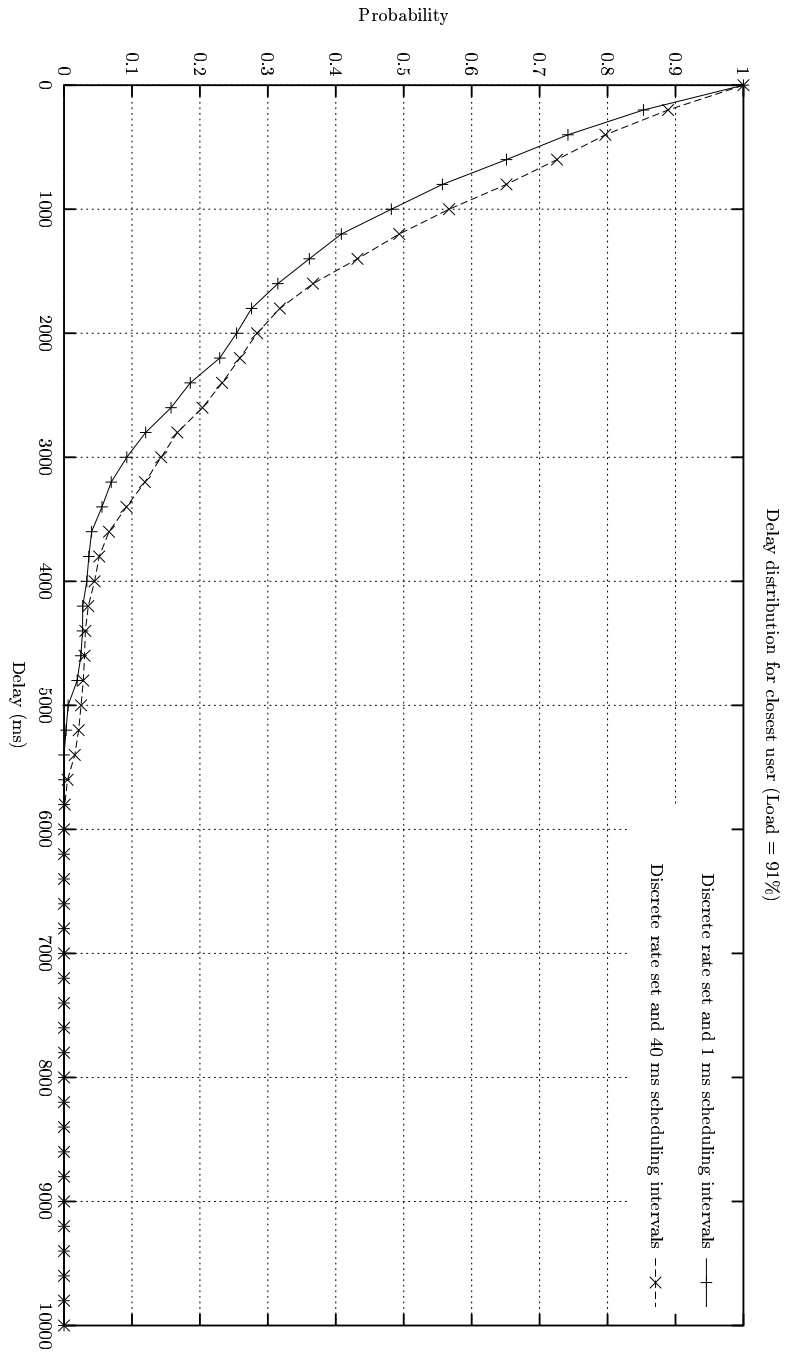

Figure 7: LWDF: Discrete Vs. Continuous Time Scheduling

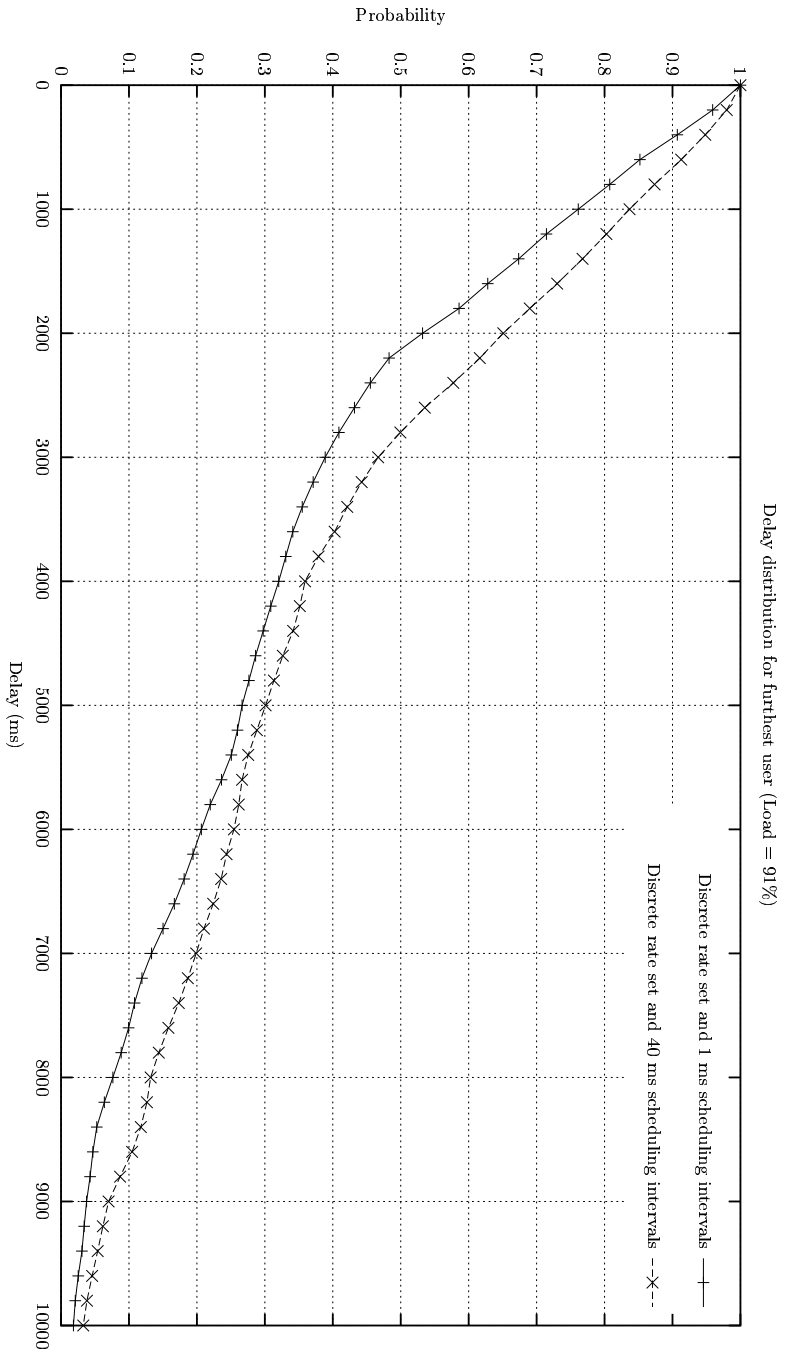

Figure 8: LWDF: Discrete Vs. Continuous Time Scheduling

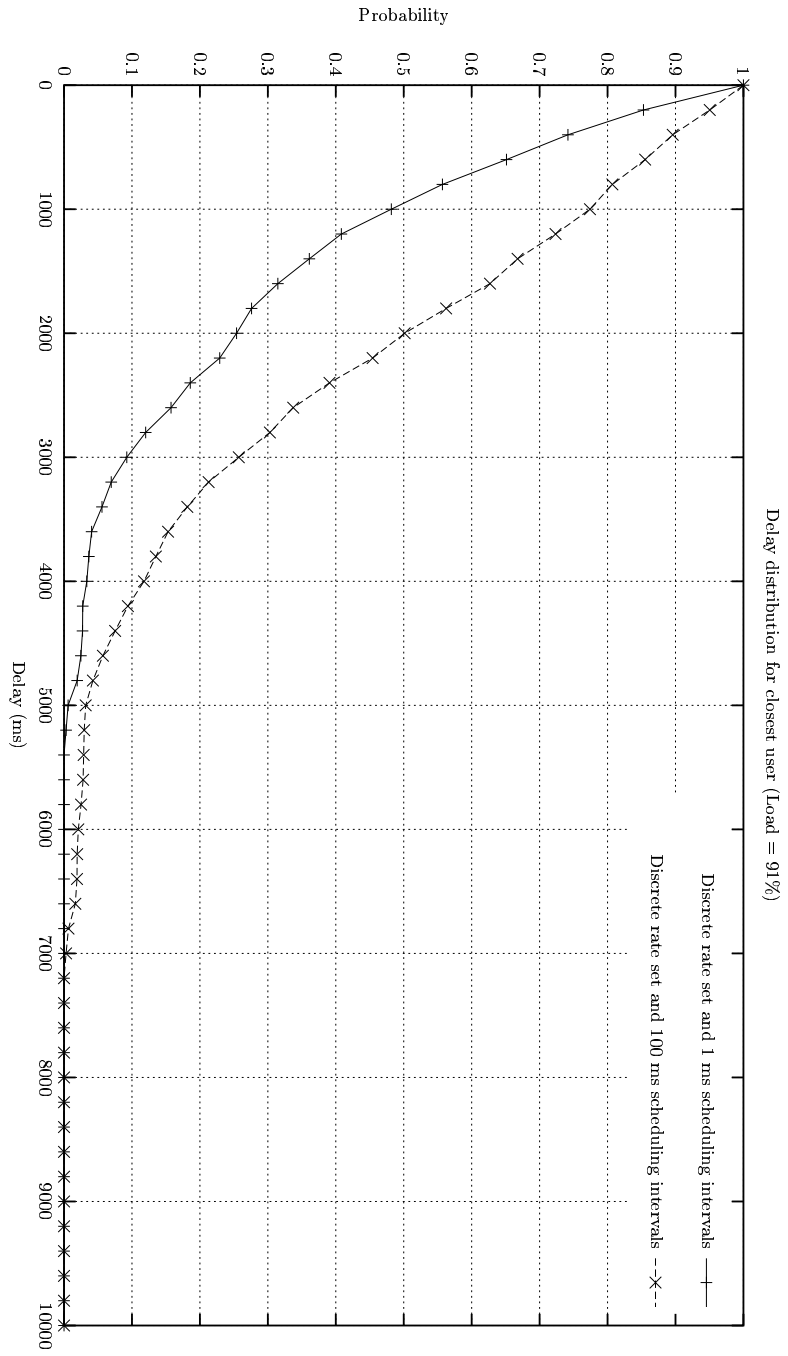

Figure 9: LWDF: Discrete Vs. Continuous Time Scheduling

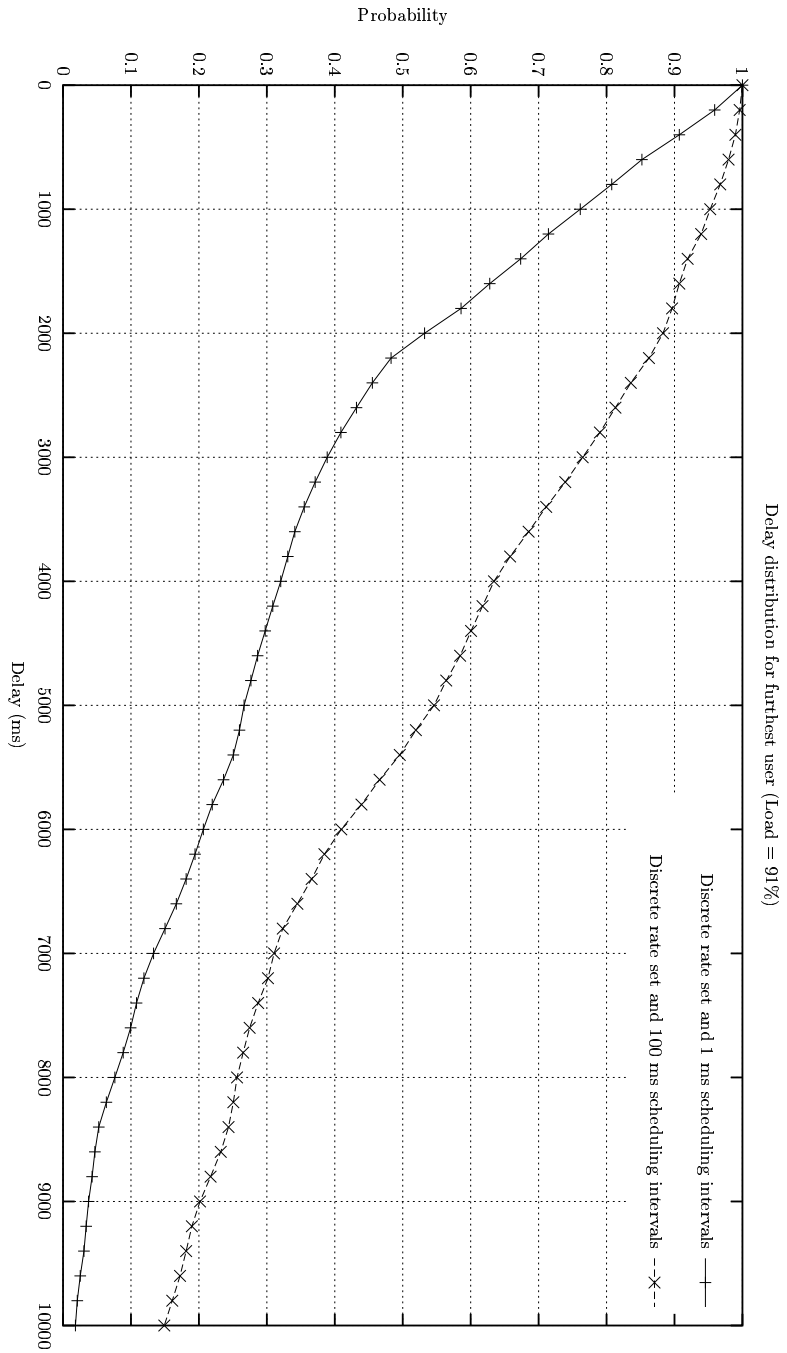

Figure 10: LWDF: Discrete Vs. Continuous Time Scheduling

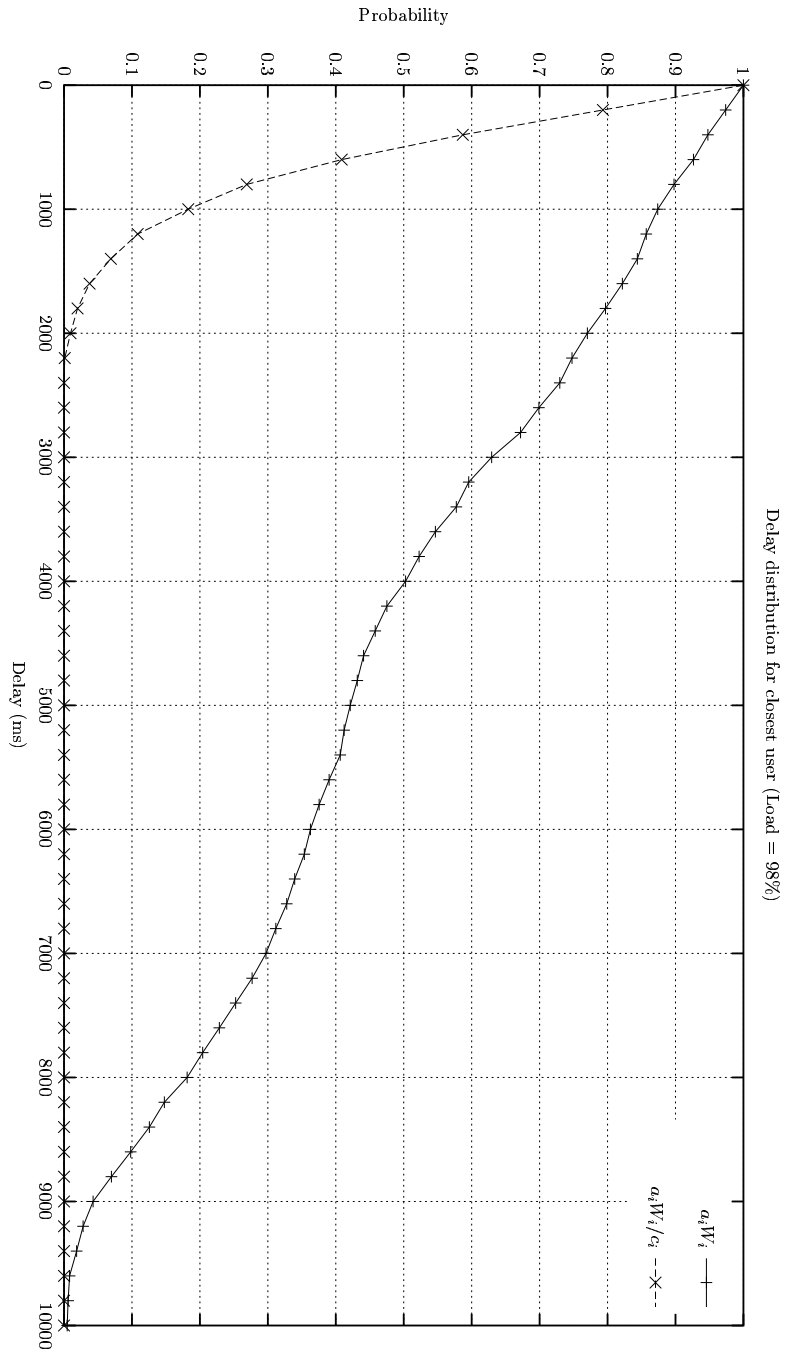

Figure 11: Time Varying Channels: M-LWDF Vs LWDF

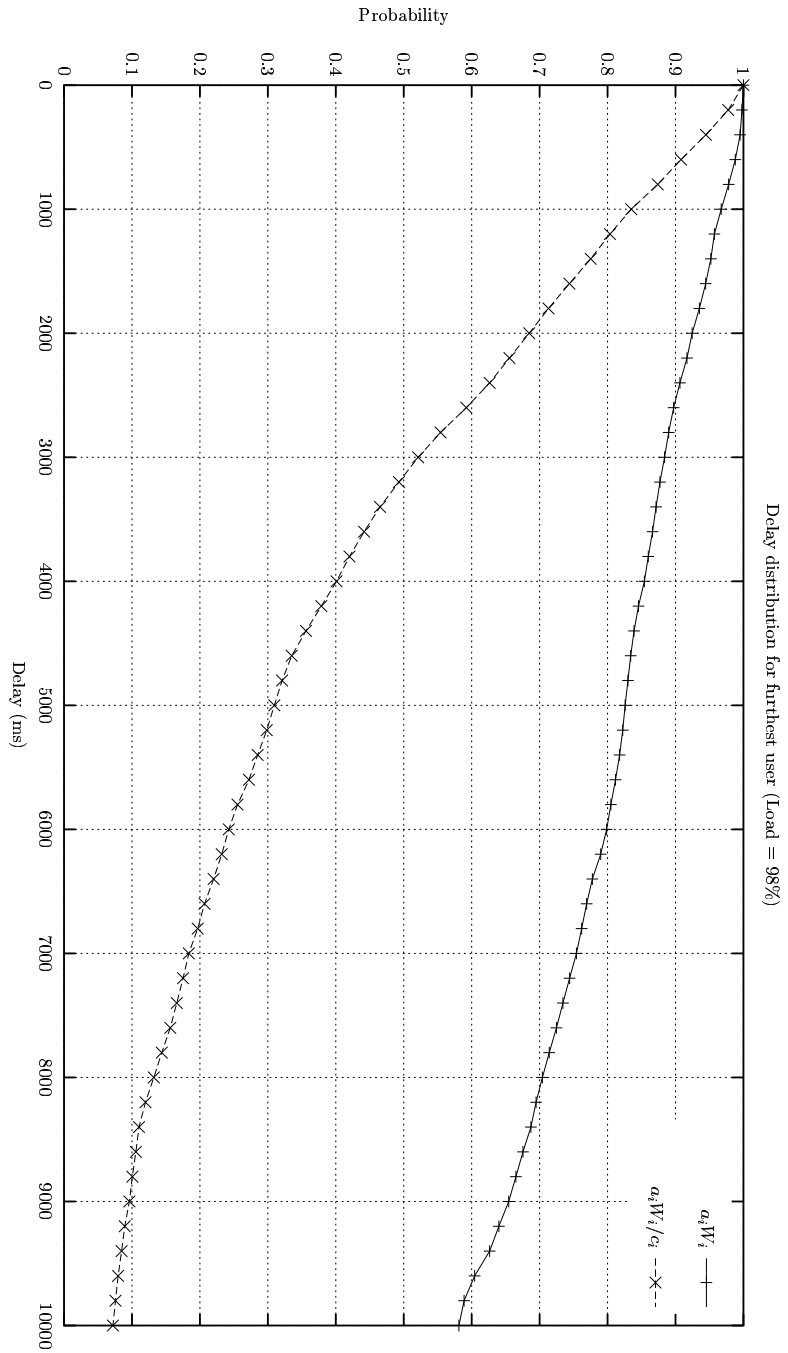

Figure 12: Time Varying Channels: M-LWDF Vs LWDF

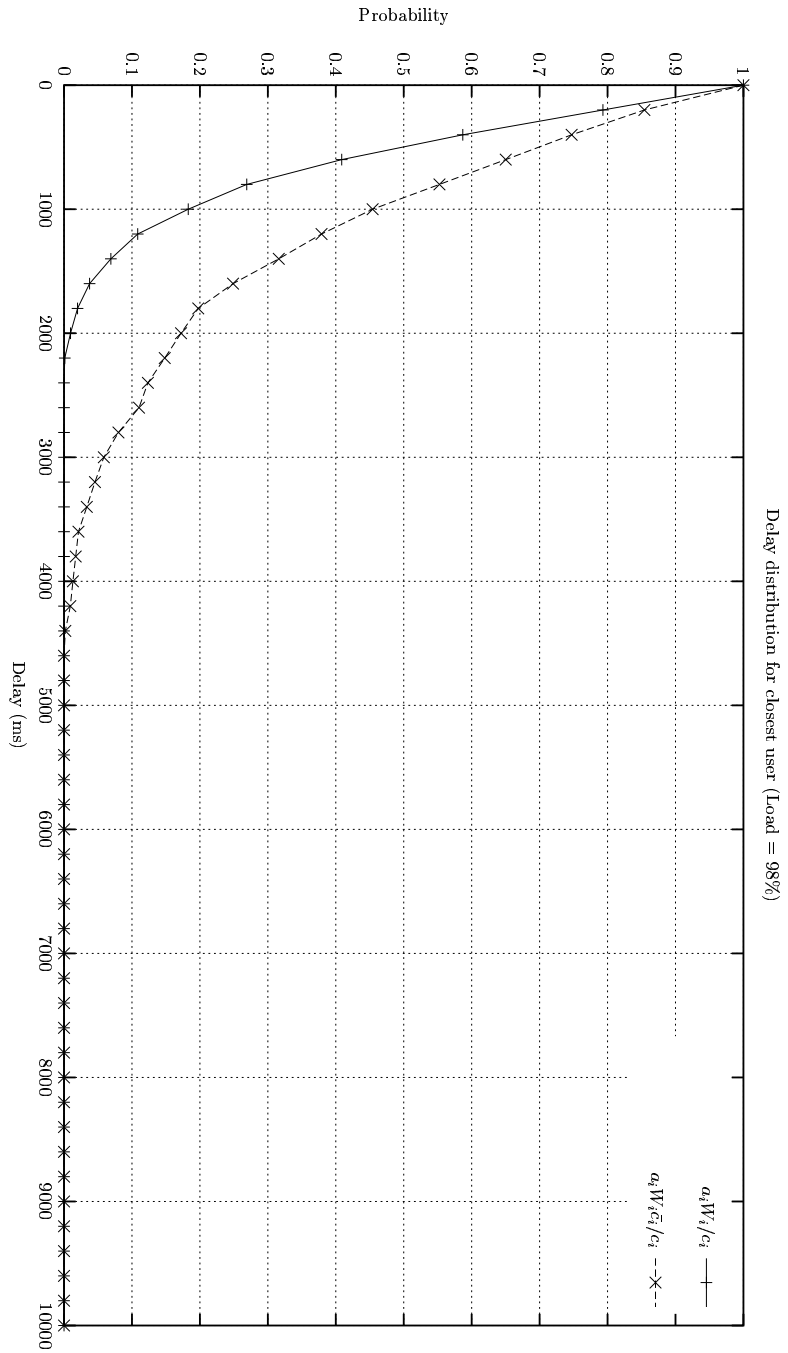

Figure 13: Time Varying Channels: M-LWDF with Different Parameter Settings

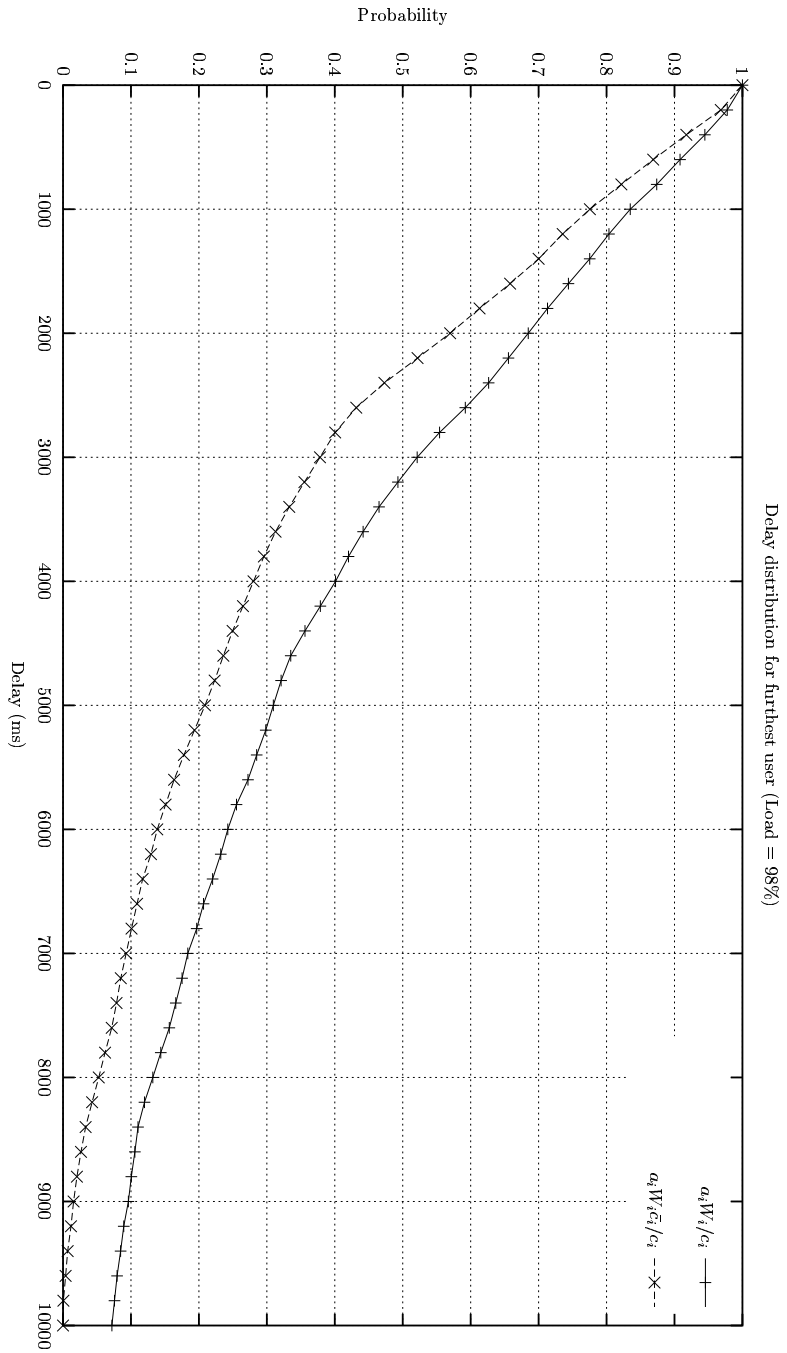

Figure 14: Time Varying Channels: M-LWDF with Different Parameter Settings

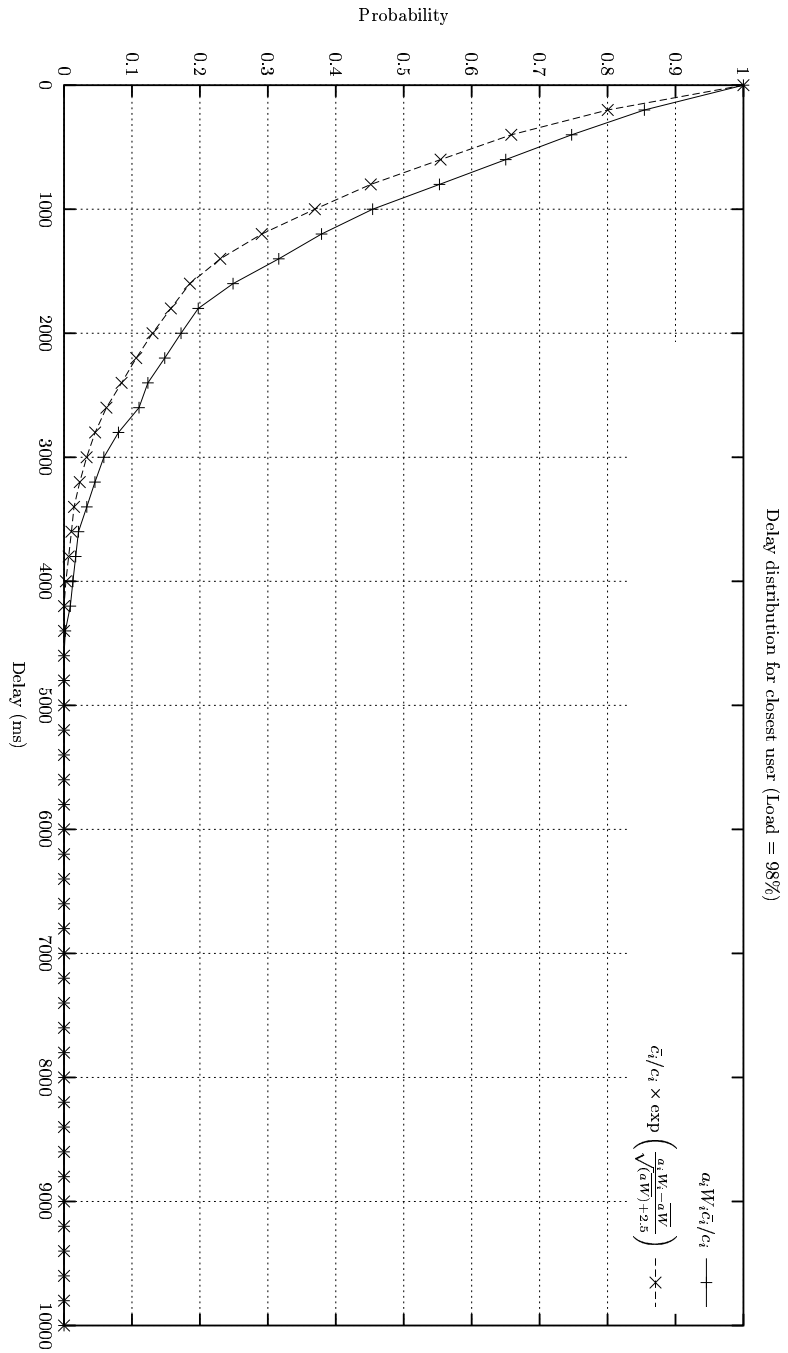

Figure 15: Time Varying Channels: M-LWDF Vs. "Exponent Rule"

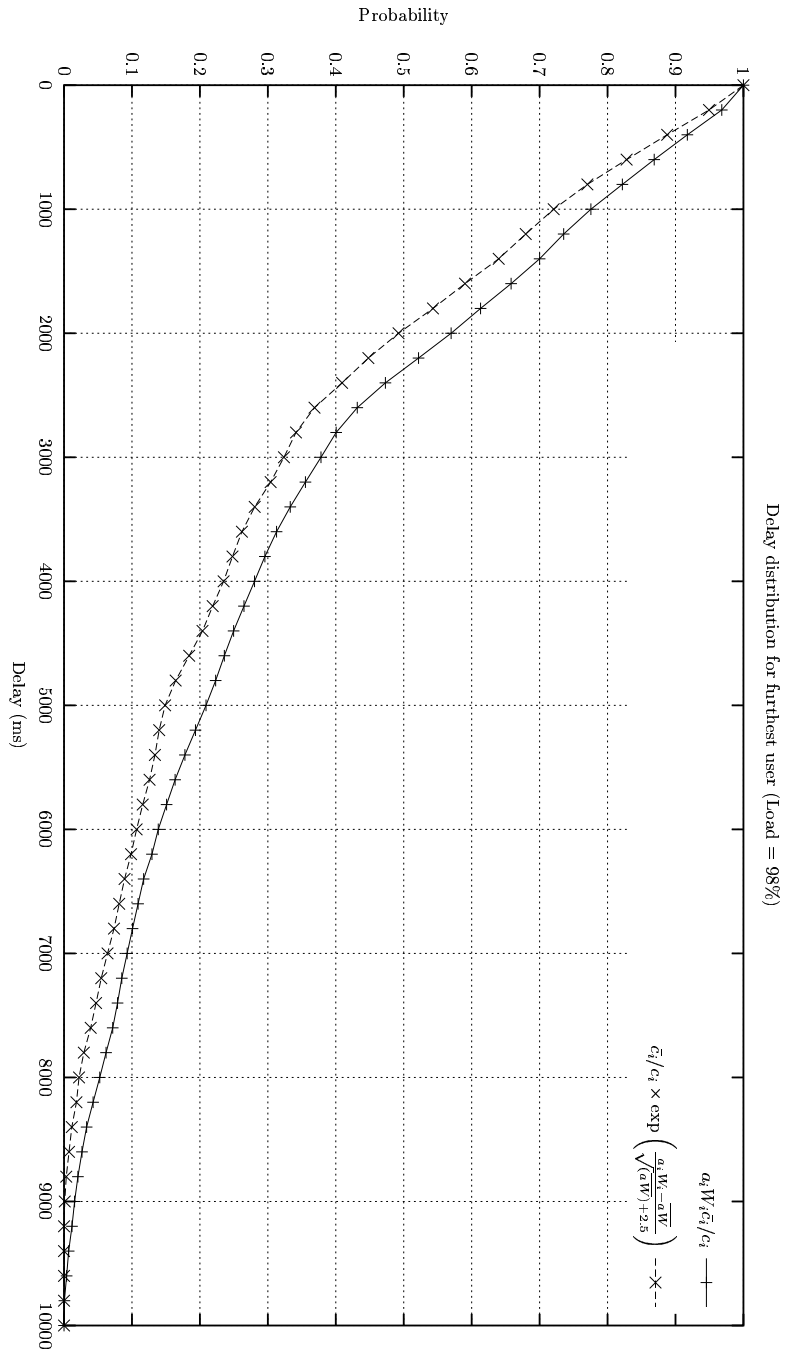

Figure 16: Time Varying Channels: M-LWDF Vs. "Exponent Rule"